Tight Bounds for ℓ_1 Oblivious Subspace Embeddings

RUOSONG WANG and DAVID P. WOODRUFF, Carnegie Mellon University

An ℓ_p oblivious subspace embedding is a distribution over $r \times n$ matrices Π such that for any fixed $n \times d$ matrix A,

$$\Pr_{\Pi}[\text{for all } x,\ \|Ax\|_p \leq \|\Pi Ax\|_p \leq \kappa \|Ax\|_p] \geq 9/10,$$

where r is the *dimension* of the embedding, κ is the *distortion* of the embedding, and for an n-dimensional vector y, $\|y\|_p = (\sum_{i=1}^n |y_i|^p)^{1/p}$ is the ℓ_p-norm. Another important property is the *sparsity* of Π, that is, the maximum number of non-zero entries per column, as this determines the running time of computing ΠA. While for $p = 2$ there are nearly optimal tradeoffs in terms of the dimension, distortion, and sparsity, for the important case of $1 \leq p < 2$, much less was known. In this article, we obtain nearly optimal tradeoffs for ℓ_1 oblivious subspace embeddings, as well as new tradeoffs for $1 < p < 2$. Our main results are as follows:

(1) We show for every $1 \leq p < 2$, any oblivious subspace embedding with dimension r has distortion

$$\kappa = \Omega\left(\frac{1}{\left(\frac{1}{d}\right)^{1/p} \log^{2/p} r + \left(\frac{r}{n}\right)^{1/p-1/2}}\right).$$

When $r = \text{poly}(d) \ll n$ in applications, this gives a $\kappa = \Omega(d^{1/p} \log^{-2/p} d)$ lower bound, and shows the oblivious subspace embedding of Sohler and Woodruff (STOC, 2011) for $p = 1$ is optimal up to $\text{poly}(\log(d))$ factors.

(2) We give sparse oblivious subspace embeddings for every $1 \leq p < 2$. Importantly, for $p = 1$, we achieve $r = O(d \log d)$, $\kappa = O(d \log d)$ and $s = O(\log d)$ non-zero entries per column. The best previous construction with $s \leq \text{poly}(\log d)$ is due to Woodruff and Zhang (COLT, 2013), giving $\kappa = \Omega(d^2 \text{poly}(\log d))$ or $\kappa = \Omega(d^{3/2}\sqrt{\log n} \cdot \text{poly}(\log d))$ and $r \geq d \cdot \text{poly}(\log d)$; in contrast our $r = O(d \log d)$ and $\kappa = O(d \log d)$ are optimal up to $\text{poly}(\log(d))$ factors even for dense matrices.

We also give (1) ℓ_p oblivious subspace embeddings with an expected $1 + \varepsilon$ number of non-zero entries per column for arbitrarily small $\varepsilon > 0$, and (2) the first oblivious subspace embeddings for $1 \leq p < 2$ with $O(1)$-distortion and dimension independent of n. Oblivious subspace embeddings are crucial for distributed and streaming environments, as well as entrywise ℓ_p low-rank approximation. Our results give improved algorithms for these applications.

CCS Concepts: • **Theory of computation → Random projections and metric embeddings;**

Additional Key Words and Phrases: Subspace embedding, ℓ_p norm, linear regression

The authors were supported in part by Office of Naval Research (ONR) Grant No. N00014-18-1-2562. Part of this work was done while the authors were visiting the Simons Institute for the Theory of Computing.

Authors' address: R. Wang and D. P. Woodruff, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania, 15217; emails: ruosongw@andrew.cmu.edu, dwoodruf@cs.cmu.edu.

1549-6325/2022/01-ART8 $15.00
https://doi.org/10.1145/3477537

ACM Reference format:
Ruosong Wang and David P. Woodruff. 2022. Tight Bounds for ℓ_1 Oblivious Subspace Embeddings. *ACM Trans. Algorithms* 18, 1, Article 8 (January 2022), 32 pages.
https://doi.org/10.1145/3477537

1 INTRODUCTION

An ℓ_p *oblivious subspace embedding* with distortion κ is a distribution over $r \times n$ matrices Π such that for any given $A \in \mathbb{R}^{n\times d}$, with constant probability, $\|Ax\|_p \leq \|\Pi Ax\|_p \leq \kappa\|Ax\|_p$ simultaneously for all $x \in \mathbb{R}^d$. The goal is to minimize r, κ and the time to calculate ΠA.

Oblivious subspace embeddings have proven to be an essential ingredient for approximately solving numerical linear algebra problems, such as regression and low-rank approximation. Sárlos [29] first used ℓ_2 oblivious subspace embeddings to solve ℓ_2-regression and Frobenius-norm low-rank approximation. To see the connection, suppose one wishes to solve the ℓ_2-regression problem $\text{argmin}_x \|Ax - b\|_2$ in the overconstrained setting, i.e., $A \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^n$ where $n \gg d$. Sárlos showed that to solve this problem approximately, it suffices to solve a much smaller instance $\text{argmin}_x \|\Pi Ax - \Pi b\|_2$, provided Π is an ℓ_2 oblivious subspace embedding for the matrix formed by concatenating columns of $A \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^n$. Sárlos further showed that using the Fast Johnson-Lindenstrauss Transform in Reference [1] as the ℓ_2 oblivious subspace embedding with $\kappa = 1 + \varepsilon$, one can get a $(1 + \varepsilon)$-approximate solution to the ℓ_2-regression problem in $O(nd \log d) + \text{poly}(d/\varepsilon)$ time, which is a substantial improvement over the standard approach based on the normal equation, which runs in $O(nd^2)$ time. The advatange of the Fast Johnson-Lindenstrauss Transform is that for any $A \in \mathbb{R}^{n\times d}$, for any matrix Π in the support of the distribution, ΠA can be computed in $O(nd \log n)$ time.

Subsequent to the work of Sárlos, the "sketch and solve" approach became an important way to solve numerical linear algebra problems. We refer interested readers to the monograph of Woodruff [34] for recent developments.

The bottleneck of Sárlos's approach is the step to calculate ΠA, which requires $\Omega(nd)$ time due to the structure of the Fast Johnson-Lindenstrauss Transform. Although this is already nearly optimal for dense matrices, when A is large and sparse, one may wish to solve the problem faster than $O(nd)$ time by exploiting the sparsity of A. Clarkson and Woodruff [10] showed that there exist ℓ_2 oblivious subspace embeddings with $r = \text{poly}(d/\varepsilon)$ rows, $s = 1$ non-zero entries per column, and $\kappa = 1 + \varepsilon$. The property that $s = 1$ is significant, since it implies calculating ΠA requires only $O(\text{nnz}(A))$ time, where $\text{nnz}(A)$ is the number of non-zero entries of A. In fact, the oblivious subspace embedding they used is the CountSketch matrix from the data stream literature [6]. By using the CountSketch embedding in Reference [10], one can reduce an ℓ_2-regression instance of size $n \times d$ into a smaller instance of size $\text{poly}(d/\varepsilon) \times d$ in $O(\text{nnz}(A))$ time. The original proof in Reference [10] used a technique based on splitting coordinates by leverage scores. The number of rows can be further reduced to $r = O((d/\varepsilon)^2)$ using the same construction and a finer analysis based on second moment method, shown independently in References [24, 25].

One may wonder if it is possible to further reduce the number of rows in the CountSketch embedding, since this affects the size of the smaller instance to solve. In Reference [26], Nelson and Nguyễn showed that any ℓ_2 oblivious subspace embedding with constant distortion and $s = 1$ non-zero entries per column requires $\Omega(d^2)$ rows. Although this rules out the possibility of further reducing the number of rows in the CountSketch embedding, this lower bound can be circumvented by considering embeddings with $s > 1$ non-zero entries in each column. This idea is implemented by the same authors in Reference [25], obtaining a result showing that for any

$B > 2$, for r about $B \cdot d \log^8 d/\varepsilon^2$ and s about $\log_B^3 d/\varepsilon$, one can achieve an ℓ_2 oblivious subspace embedding with $\kappa = 1 + \varepsilon$. The bound on r and s was further improved in Reference [11] (see also Reference [4]), where Cohen showed that for any $B > 2$, it suffices to have $r = O(B \cdot d \log d/\varepsilon^2)$ and $s = O(\log_B d/\varepsilon)$. Cohen's result matches the lower bound in Reference [27] up to a multiplicative $\log d$ factor in the number of rows.

Another line of research focused on the case when $p \neq 2$, as the corresponding regression and low-rank approximation problems are often considered to be more robust, or less sensitive to outliers. Moreover, the $p = 1$ error measure for regression yields the maximum likelihood estimator under Laplacian noise models. When $p = 1$, using Cauchy random variables, Sohler and Woodruff [30] showed there exist ℓ_1 oblivious subspace embeddings with $r = O(d \log d)$ rows and $\kappa = O(d \log d)$. This approach was generalized by using p-stable random variables in work of Meng and Mahoney [24] to ℓ_p-norms when $1 < p < 2$, where they showed there exist ℓ_p oblivious subspace embeddings with $r = O(d \log d)$ rows and $\kappa = O(d \log d)$.[1] Unlike the case when $p = 2$, due to the large distortion incurred in such upper bounds, one cannot directly get a $(1 + \varepsilon)$-approximate solution to the ℓ_p-regression problem by solving $\operatorname{argmin}_x \|\Pi A x - \Pi b\|_p$. A natural question then, is whether it is possible to obtain $(1 + \varepsilon)$-distortion with ℓ_p oblivious subspace embeddings; prior to our work there were no lower bounds ruling out the existence of ℓ_p oblivious subspace embeddings with $r = \text{poly}(d/\varepsilon)$ and $\kappa = 1 + \varepsilon$.

Although it was unknown if better oblivious subspace embeddings exist for $p \neq 2$ prior to our work, ℓ_p oblivious subspace embeddings still played a crucial role in solving ℓ_p-regression problems in earlier work, since they provide a way to *precondition* the matrix A, which enables one to further apply non-oblivious (sampling-based) subspace embeddings. We refer interested readers to Chapter 3 of Reference [34] and references therein for further details. Recent developments in entrywise ℓ_p low-rank approximation [31] also used ℓ_p oblivious subspace embeddings as an important ingredient. Furthermore, such ℓ_1 oblivious subspace embeddings are the only known way to achieve single-pass streaming algorithms for ℓ_1-regression (see, e.g., Section 5 of Reference [30], where it is shown how to implement the preconditioning and sampling in parallel in a single pass), a model that has received considerable interest for linear algebra problems (see, e.g., Reference [9]). We note that recent algorithms for ℓ_p-regression based on Lewis weights sampling require at least $\Omega(\log \log n)$ passes in the streaming model.

Due to these applications, speeding up the computation of ΠA for ℓ_p oblivious subspace embeddings is an important goal. In Reference [8], Clarkson et al. combined the idea of Cauchy random variables and Fast Johnson-Lindenstrauss Transforms to obtain a more structured family of subspace embeddings, which enables one to calculate ΠA in $O(nd \log n)$ time. Meng and Mahoney [24] showed that when $1 \leq p < 2$, there exist ℓ_p oblivious subspace embeddings with $r = \widetilde{O}(d^5)$ rows and $s = 1$ non-zero entries per column, where the distortion $\kappa = \widetilde{O}(d^3)$.[2] The structure of the embedding by Meng and Mahoney is very similar to the CountSketch embedding by Clarkson and Woodruff [10]. In fact, to prove the distortion bound, Meng and Mahoney also used techniques of splitting coordinates based on leverage scores.

Inspired by the technique by Andoni [2], which used exponential random variables to estimate the ℓ_p-norm of a data steam, Woodruff and Zhang [35] improved the embedding given in Reference [24]. They showed there exist ℓ_1 oblivious subspace embeddings with $r = \widetilde{O}(d)$ rows and $s = \text{polylog}(d)$ non-zero entries per column, where the distortion $\kappa = \min\{\widetilde{O}(d^2), \widetilde{O}(d^{1.5})\sqrt{\log n}\}$. Note that to achieve such a small polylogarithmic sparsity, the distortion κ given by the analysis

[1] In Reference [24] the authors incorrectly claimed that the distortion of their subspace embedding is $\kappa = O((d \log d)^{1/p})$. See Section 1.3 for more details.

[2] Again, in Reference [24] the authors incorrectly claimed that the distortion is $\kappa = \widetilde{O}(d^{3/p})$.

Table 1. Summary of Previous and New Upper Bounds When $p = 1$

Reference	r	κ	s
Sohler and Woodruff [30]	$O(d \log d)$	$O(d \log d)$	Dense
Clarkson et al. [8]	$O(d \log d)$	$O(d^4 \log^4 d)$ $O(d(d + \log n)^{1+\eta} \log d)$	Dense
Meng and Mahoney [24]	$O(d^5 \log^5 d)$	$O(d^3 \log^3 d)$	1
Woodruff and Zhang [35]	$\widetilde{O}(d)$	$\min\{\widetilde{O}(d^2), \widetilde{O}(d^{1.5})\sqrt{\log n}\}$	$\text{poly}(\log d)$
Theorem 4.1	$O(d^2)$	$O(d)$	2
Theorem 4.2	$O(B \cdot d \log d)$	$O(d \log_B d)$	$O(\log_B d)$
Theorem 6.1	$O(d^2)$	$O(d)$	$1 + \varepsilon$ in expectation
Theorem 6.6	$\widetilde{O}(d^4)$	$\widetilde{O}(d)$	1

Here, $\eta > 0$ is an arbitrarily small constant. $B > 0$ is required to be sufficiently large and provides a tradeoff between r, κ, and s.

in Reference [35] had to either increase to $\widetilde{O}(d^2)$ or to $\widetilde{O}(d^{1.5})\sqrt{\log n}$, the latter also depending on n. See Table 1 for a summary of previous and new upper bounds when $p = 1$.

The above works leave many gaps in our understanding on the tradeoffs between dimension, distortion, and sparsity for ℓ_p oblivious subspace embeddings. For instance, it is natural to ask what the optimal distortion bound for ℓ_p oblivious subspace embeddings is when $1 \le p < 2$, provided that the number of rows $r = \text{poly}(d)$. Results in References [24, 30] showed that $\kappa = O(d \log d)$ is achievable. Is this optimal? Also, it is unknown whether there exist *sparse* ℓ_1 oblivious subspace embeddings with dimension $\widetilde{O}(d)$ and distortion $\kappa = \widetilde{O}(d)$. In this article, we resolve these questions.

1.1 Our Results

Distortion Lower Bound. We first show a distortion lower bound for ℓ_p oblivious subspace embeddings, when $1 \le p < 2$.

Theorem 1.1. *For $1 \le p < 2$, if a distribution over $r \times n$ matrices Π is an ℓ_p oblivious subspace embedding, then the distortion*

$$\kappa = \Omega\left(\frac{1}{\left(\frac{1}{d}\right)^{1/p} \cdot \log^{2/p} r + \left(\frac{r}{n}\right)^{1/p-1/2}}\right).$$

When $1 \le p < 2$ and $r = \text{poly}(d)$, the denominator of the lower bound is dominated by the $(\frac{1}{d})^{1/p} \cdot \log^{2/p} r$ term, provided n is large enough. In that case, our lower bound is $\Omega(d^{1/p} \log^{-2/p} d)$. It was shown in Reference [30] that there exist ℓ_1 oblivious subspace embeddings with $r = O(d \log d)$ rows and distortion $\kappa = O(d \log d)$. Our lower bound matches this result up to an $O(\log^3 d)$ factor. Thus, our lower bound is nearly optimal for $r = \text{poly}(d)$ when $p = 1$ (which is the main regime of interest in the above applications).

The dependence on $(r/n)^{1/p-1/2}$ reflects the fact that

- When the number of rows $r = n$, one can get a trivial ℓ_p oblivious subspace embedding with $\kappa = 1$, i.e., the identity matrix I;
- As $p \to 2$, there exist ℓ_2 oblivious subspace embeddings [4, 10, 11, 24, 25, 29] with $\kappa = 1 + \varepsilon$ and $r = \text{poly}(d/\varepsilon)$, where ε can be an arbitrarily small constant.

It is possible that the $\log^{2/p} r$ factor (in the $(1/d)^{1/p} \cdot \log^{2/p} r$ term) could be somewhat improved. However, we show that some dependence on r is, in fact, necessary.

Theorem 1.2 (Informal version of Theorem 3.5). *For $1 \le p < 2$, there exists an ℓ_p oblivious subspace embedding over $\exp(\exp(O(d))) \times n$ matrices Π, where the distortion κ is a constant.*

Even though Theorem 1.2 has a doubly exponential dependence on d in the number of rows, it is the first ℓ_p oblivious subspace embedding with constant distortion, when $1 \le p < 2$ and r does not depend on n. This new embedding suggests that it is impossible to get a lower bound of

$$\kappa = \Omega\left(\frac{1}{\left(\frac{1}{d}\right)^{1/p} + \left(\frac{r}{n}\right)^{1/p-1/2}}\right),$$

i.e., the $(\frac{1}{d})^{1/p}$ term should have some dependence on r.

New ℓ_p oblivious subspace embeddings. We next show there exist *sparse* ℓ_1 oblivious subspace embeddings with nearly optimal distortion, and *sparse* ℓ_p oblivious subspace embeddings with $O(d \log d)$ distortion when $1 < p < 2$.

Theorem 1.3 (Summary of Theorem 4.1, 4.2, 5.1, and 5.2.). *For $1 \le p < 2$, there exist ℓ_p oblivious subspace embeddings over $r \times n$ matrices Π with s non-zero entries per column and distortion κ, where*

(1) *When $p = 1$,*
   (a) *$r = O(d^2)$, $s = 2$ and $\kappa = O(d)$; or*
   (b) *For sufficiently large B, $r = O(B \cdot d \log d)$, $s = O(\log_B d)$ and $\kappa = O(d \log_B d)$.*
(2) *When $1 < p < 2$, $\kappa = O(d \log d)$,*
   (a) *$r = O(d^2)$, $s = 2$; or*
   (b) *For sufficiently large B, $r = O(B \cdot d \log d)$, $s = O(\log_B d)$.*

Notably, the distortion of our embeddings is never worse than the dense constructions in References [24, 30]. Also, when $p = 1$, if we set $r = O(d^2)$ (Case 1(a)) or $r = O(d^{1+\eta})$ for any constant $\eta > 0$ (Case 1(b)), then the distortion can be further improved to $O(d)$. This is the first known ℓ_1 oblivious subspace embedding with $r = \text{poly}(d)$ rows and distortion $\kappa = o(d \log d)$. We remark that by using the dense construction in Reference [30], it is also possible to reduce the distortion to $O(d)$ by increasing the number of rows.

Similar to the OSNAP embedding in Reference [25], our results in Case 1(b) and Case 2(b) provide a tradeoff between the number of rows and the number of non-zero entries in each column.

Sparser ℓ_p oblivious subspace embeddings. Finally, we show that the sparsity of Case 1(a) and Case 2(a) in Theorem 1.3 can be further reduced by using two different approaches.

The first approach is based on random sampling, which leads to the following theorem.

Theorem 1.4 (Summary of Theorem 6.1 and 6.2). *For $1 \le p < 2$ and any constant $0 < \varepsilon < 1$, there exists an ℓ_p oblivious subspace embedding over $O(d^2) \times n$ matrices Π where each column of Π has at most two non-zero entries and $1+\varepsilon$ non-zero entries in expectation, and the distortion $\kappa = O(d)$ (when $p = 1$) or $\kappa = O(d \log d)$ (when $1 < p < 2$).*

The second approach is based on the construction in Reference [24] and a truncation argument, which leads to the following theorem.

Theorem 1.5 (Summary of Theorem 6.5 and 6.6). *For $1 \le p < 2$, there exists an ℓ_p oblivious subspace embedding over $\widetilde{O}(d^4) \times n$ matrices Π where each column of Π has a single non-zero entry and distortion $\kappa = \widetilde{O}(d)$.*

It has been shown in Reference [26] that for any distribution over $r \times n$ matrices Π with $s = 1$ non-zero entries per column, if for any fixed matrix $A \in \mathbb{R}^{n \times d}$, $\text{rank}(\Pi A) = \text{rank}(A)$ with constant

probability, then Π should have $r = \Omega(d^2)$ rows. Since oblivious subspace embeddings with finite distortion always preserve the rank, this lower bound can also be applied. We show also that this lower bound holds even if the columns of Π have $1+\varepsilon$ non-zero entries in expectation for a constant $0 < \varepsilon < 1$, thereby showing Theorem 1.4 is optimal.

1.2 Comparison with Previous Work

To compare our results with previous work, it is crucial to realize the difference between *oblivious* embeddings and *non-oblivious* embeddings. An oblivious subspace embedding Π is a universal distribution over $\mathbb{R}^{r\times n}$, which does not depend the given matrix $A \in \mathbb{R}^{n\times d}$. A non-oblivious subspace embedding, however, is a distribution that possibly depends on the given matrix A. Most known non-oblivious subspace embeddings involve importance sampling according to the *leverage scores* or *Lewis weights* of the rows, and so are inherently non-oblivious. We refer the interested reader to Reference [20] for an excellent survey on leverage scores and References [12, 13, 17] for recent developments on non-oblivious subspace embeddings.

Previous impossibility results for dimension reduction in ℓ_1 [5, 7, 16] are established by creating a set of $O(n)$ points in $\mathbb{R}^n$ and showing that any (non-oblivious) embedding on them incurs a large distortion. In this article, we focus on embedding a d-dimensional subspace of $\mathbb{R}^n$ into $\mathbb{R}^{\mathrm{poly}(d)}$ using oblivious embeddings. We stress that $O(n)$ points in a d-dimensional subspace have a very different structure from $O(n)$ arbitrary points in $\mathbb{R}^n$. Previous results [13] showed that any d-dimensional subspace in $\mathbb{R}^n$ can be embedded into $\mathbb{R}^{O(d(\log d)\varepsilon^{-2})}$ with $(1+\varepsilon)$ distortion in ℓ_1 using non-oblivious linear embeddings, where $\varepsilon > 0$ is an arbitrarily small constant. Here the subspace structure is critically used, since Charikar and Sahai [7] showed that there exist $O(n)$ points such that any linear embedding $\mathbb{R}^n \to \mathbb{R}^d$ must incur a distortion of $\Omega(\sqrt{n/d})$, even for non-oblivious linear embeddings.

Our hardness result in Theorem 1.1 establishes a separation between oblivious and non-oblivious subspace embeddings in ℓ_p when $1 \le p < 2$. This result suggests that to construct a subspace embedding with $(1+\varepsilon)$ distortion, it is essential to use non-oblivious subspace embeddings.

Although our main focus in this article is to understand oblivious subspace embeddings, we remark that our technique for proving the hardness result in Theorem 1.1 can also be applied to embed any d points in $\mathbb{R}^n$ into $\mathbb{R}^{\mathrm{poly}(d)}$ in ℓ_p using oblivious linear embeddings, when $1 \le p < 2$. In particular, it is possible to reproduce the result of Reference [7] using our techniques, although in a weaker setting where the embeddings are oblivious.

1.3 Errors in Prior Work and the Conference Version

In the conference version of this article [32], we incorrectly claimed that the distortion of the embeddings in Theorems 1.3 and 1.4 is $\kappa = O((d\log d)^{1/p})$, and $\kappa = \widetilde{O}(d^{1/p})$ for the embedding in Theorem 1.5. The source of the error is Lemma 2.22 (Lemma 2.16 in the conference version), in which it was claimed that by the existence of an Auerbach basis, the existence of a certain well-conditioned basis follows. However, it is unclear whether the existence of an Auerbach basis implies the existence of such a well-conditioned basis. The claim that we made was actually already made in previous work. It first appeared in the technical report version [23] of Reference [24]. In the proof of Theorem 6 in Section A.7 of Reference [23], the authors claimed that the existence of an Auerbach basis implied the existence of a $(d^{1/p}, 1, p)$-conditioned basis. The authors of Reference [24] confirmed that this is an error in their work in Reference [22]. Besides propagating to our work, that error also propagates to Reference [35], in which a similar claim was made.

After fixing this, the best existing ℓ_p oblivious subspace embeddings with $r = \text{poly}(d)$ rows have distortion $\kappa = O(d \log d)$ when $1 < p < 2$. Thus, while we obtain tight bounds up to polylog(d) factors for the important case of $p = 1$, when $1 < p < 2$, our distortion lower bound in Theorem 1.1 is not necessarily tight. We leave it as an open problem to resolve the case $1 < p < 2$.

1.4 Applications of Our Subspace Embeddings

Using the sparse ℓ_p oblivious subspace embeddings in Theorem 1.3, we obtain improvements to many related problems. We list a few examples in this section.

ℓ_p-regression in the streaming model. Using dense Cauchy embeddings and a sampling data structure from Reference [3], a single-pass streaming algorithm for ℓ_1-regression $\text{argmin}_x \|Ax-b\|_1$ was designed in Reference [30]. To get a $(1+\varepsilon)$-approximate solution to the regression problem, the algorithm uses $\text{poly}(d\varepsilon^{-1}\log n)$ bits of space, where $A \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^n$. The total running time of the algorithm, however, is $O(\text{nnz}(A)\cdot d + \text{poly}(d\varepsilon^{-1}\log n))$.

By replacing the dense Cauchy embedding with our new oblivious subspace embeddings in Theorem 1.3, the total running time can be further improved to $\widetilde{O}(\text{nnz}(A)) + \text{poly}(d\varepsilon^{-1}\log n)$ while the space complexity remains unchanged. We note that using earlier sparse Cauchy embeddings [24] would also give such a running time, but with a significantly worse $\text{poly}(d\varepsilon^{-1}\log n)$ factor. The same approach can also be applied to design input-sparsity time algorithms for ℓ_p-regression in the streaming model when $1 < p < 2$.

Entrywise ℓ_p low-rank approximation. Given a matrix $A \in \mathbb{R}^{n\times d}$ and approximation factor α, the goal of the ℓ_1-low-rank approximation problem is to output a matrix $\widehat{A}$ for which

$$\|A - \widehat{A}\|_1 \le \alpha \cdot \min_{\text{rank-}k \text{ matrices } A'} \|A - A'\|_1,$$

where $\|\cdot\|_1$ is the entrywise ℓ_1-norm.

In Reference [31], the authors devised an algorithm that runs in $T = O(\text{nnz}(A) + (n+d)\cdot\text{poly}(k))$ time to solve this problem, with $\alpha = \text{poly}(k)\cdot\log d$. The exact expression of the poly(k) factor in the approximation factor α and the running time T, depends on the number of rows r and the distortion κ of the ℓ_1 oblivious subspace embedding used. Both poly(k) factors can be directly improved by replacing the sparse Cauchy embedding [24], which is originally used in Reference [31], with our new oblivious subspace embeddings in Theorem 1.3. This improvement also propagates to other problems considered in Reference [31] such as ℓ_p-low-rank approximation, entrywise ℓ_p-norm CUR decomposition and ℓ_p-low-rank approximation in distributed and streaming models.

Quantile Regression. Given a matrix $A \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^n$, the goal of *quantile regression* is to solve

$$\text{argmin}_x \rho_\tau(b - Ax),$$

where $\rho_\tau(b-Ax) = \sum_{i=1}^n \rho_\tau((b-Ax)_i)$ and for any $z \in \mathbb{R}$,

$$\rho_\tau(z) = \begin{cases} \tau z & z \ge 0 \\ (\tau-1)z & z < 0 \end{cases}.$$

Here τ is a parameter in $(0,1)$.

An efficient algorithm to calculate a $(1+\varepsilon)$-approximate solution to quantile regression was proposed in Reference [36]. Using their approach, one can reduce a quantile regression instance of size $n \times d$ to a smaller instance of size $O(\text{poly}(d)\varepsilon^{-2}\log(1/\varepsilon)) \times d$ in $O(\text{nnz}(A)) + \text{poly}(d)$ time. By replacing the sparse Cauchy embedding, which is used in the conditioning step of their algorithm,

with our new oblivious subspace embeddings in Theorem 1.3, the poly(d) term in the running time can be directly improved.

1.5 Our Techniques

Distortion lower bound. We use the case when $p = 1$ to illustrate our main idea for proving our distortion lower bounds. We start with Yao's minimax principle, which enables us to deal only with deterministic embeddings. Here our goal is to construct a distribution over matrices $A \in \mathbb{R}^{n\times d}$ such that for any $\Pi \in \mathbb{R}^{r\times n}$, if

$$\|Ax\|_1 \le \|\Pi Ax\|_1 \le \kappa\|Ax\|_1 \tag{1}$$

holds simultaneously for all $x \in \mathbb{R}^d$ with constant probability, then $\kappa = \widetilde{\Omega}(d)$.

Roughly speaking, our proof is based on the crucial observation that, the histogram of the ℓ_1-norm of columns in the deterministic embedding Π/κ should look like that of a discretized standard Cauchy distribution. That is, there are at most 2^i columns in Π/κ with ℓ_1-norm larger than $\Theta((n/d)2^{-i})$. This is because if we choose a matrix $A \in \mathbb{R}^{n\times d}$ such that each column contains $(n/d)2^{-i}$ non-zero entries at random positions and all these $(n/d)2^{-i}$ non-zero entries are i.i.d. sampled from the standard Gaussian distribution $\mathcal{N}(0,1)$, then for each column in A, the ℓ_1-norm of that column is $\Theta((n/d)2^{-i})$ with constant probability. However, if the embedding Π/κ contains more than 2^i columns with ℓ_1-norm larger than $\Theta((n/d)2^{-i})$, then with constant probability, there exists some $i \in [n]$ and $j \in [d]$ such that $A_{i,j} \sim \mathcal{N}(0,1)$ and the ith column of Π/κ has ℓ_1-norm larger than $\Theta((n/d)2^{-i})$. In that case, it can be shown that after projection by Π/κ, the jth column of A has ℓ_1-norm larger than $\Theta((n/d)2^{-i})$, which violates the condition in Equation (1).

To prove $\kappa = \widetilde{\Omega}(d)$, let $c \in \mathbb{R}^n$ be a vector whose entries are all i.i.d. sampled from $\mathcal{N}(0,1)$. With constant probability $\|c\|_1 = \Omega(n)$. However, we are able to show that the constraint we put on the histogram of the ℓ_1-norm of columns in Π/κ implies that $\|\Pi c/\kappa\|_1 = \widetilde{O}(n/d)$ and hence $\kappa = \widetilde{\Omega}(d)$. The formal analysis in Section 3.1 shows that $\kappa = \Omega(d\log^{-2} r)$ when $n \gg r$.

To show that the dependence on r in the lower bound is necessary, we construct an ℓ_1 oblivious subspace embedding with $\exp(\exp(O(d)))$ rows and constant distortion. The construction itself is the same as the dense construction in Reference [30]. Unlike previous approaches [24, 30, 35], we do not use the existence of an Auerbach basis to prove the dilation bound. Our analysis is based on tighter tail bounds for sums of absolute values of independent standard Cauchy (and also p-stable) random variables in Lemmas 2.12 and 2.14. Let $\{X_i\}$ be $R = \exp(\exp(O(d)))$ independent standard Cauchy random variables. Based on the tighter tail bounds, it can be shown that with probability $1 - \exp(-\Omega(d))$,

$$\sum_{i=1}^{R} |X_i| = \Theta(R\log R),$$

which enables us to now apply a standard net argument to prove the constant distortion bound. The formal analysis is given in Section 3.2.

New ℓ_p oblivious subspace embeddings. For ease of notation, here we focus on $p = 1$. Before getting into our results, we first review the construction in Reference [24] and its analysis. The sparse Cauchy embedding in Reference [24] has $\widetilde{O}(d^5)$ rows. In each column, there is a single non-zero entry that is sampled from the standard Cauchy distribution. The $\widetilde{O}(d)$ dilation bound follows the standard approach [30] of using the existence of an Auerbach basis and upper tail bounds for dependent standard Cauchy random variables. The contraction bound is based on the technique of splitting coordinates, which was first proposed in Reference [10] to analyze the CountSketch embedding. A coordinate is heavy if its ℓ_1 leverage score is larger than $1/d$ and light otherwise. For any vector $y = Ax$, if light coordinates contribute more to the ℓ_1-norm of y, then standard

concentration bounds and Cauchy lower tail bounds imply a constant distortion. If heavy coordinates contribute more to the ℓ_1-norm, since there will be at most $O(d^2)$ heavy coordinates and the embedding has $\Omega(d^4)$ rows, then all the heavy coordinates will be perfectly hashed. An $\Omega(d^{-2})$ contraction bound follows by setting up a global event saying that the absolute values of all of the $O(d^2)$ standard Cauchy random variables associated with the heavy coordinates are at least $\Omega(d^{-2})$, which holds with constant probability.

Although the dilation bound seems to be tight, the contraction bound can be improved. Indeed, the ℓ_1-norm of columns in the embedding of Reference [24] almost follows the histogram predicted by our lower bound argument, except for the lower tail part. As predicted by our lower bound argument, for an embedding Π, which has the optimal $\kappa = \widetilde{O}(d)$ distortion, the ℓ_1-norm of each column in Π should be larger than a constant. However, the standard Cauchy distribution is heavy-tailed in both directions.[3] This leads to the idea of truncation, which is formalized in Section 6.2. The rough idea is that we make sure the absolute values of the standard Cauchy random variables are never smaller than a constant and thus the contraction bound can be improved to be a constant. It is shown in Corollary 6.4 that standard Cauchy random variables are still "approximately 1-stable" after truncation, which enables one to use Cauchy tail inequalities to analyze the dilation bound. However, even though the distortion bound of this new embedding is nearly optimal, the number of rows is $\widetilde{O}(d^4)$, which seems difficult to improve.

Our alternate approach is still based on the technique of splitting coordinates. Unlike the approach in Reference [24], which is based on splitting coordinates according to the ℓ_1 leverage scores, in this new approach, for any vector $y = Ax$, a coordinate i is heavy if $|y_i| \geq \frac{1}{d^2}\|y\|_1$ and light otherwise. When light coordinates contribute more to the ℓ_1-norm of y, we show that the sparse Cauchy embedding in Reference [24] with only $O(d \log d)$ rows is already sufficient to deal with such vectors. This is due to a tighter analysis based on negative association theory [15], which also greatly simplifies the proof. When heavy coordinates contribute more to the ℓ_1-norm of y, the idea is to use known ℓ_2 oblivious subspace embeddings. The key observation is that when heavy coordinates contribute more to the ℓ_1-norm, we have $\|y\|_2 \geq \Omega(\frac{1}{d})\|y\|_1$ and thus any ℓ_2 oblivious subspace embedding with constant distortion will also be an ℓ_1 oblivious subspace embedding with $O(d)$ distortion. See Section 5 for a formal analysis and Section 5 for how to generalize this idea to ℓ_p-norms when $1 < p < 2$.

Our final embedding consists of two parts. The ℓ_2 oblivious subspace embedding part could be the CountSketch embedding or the OSNAP embedding, which also provides a tradeoff between the number of non-zero entries per column and number of rows. For the sparse Cauchy part, although it would be sufficient to prove the $O(d \log d)$ distortion bound as long as this part has $O(d \log d)$ rows, an analysis based on a tighter Cauchy lower tail bound in Lemma 2.14 shows that it is possible to further reduce the dilation to $O(d)$ by increasing the number of rows in this part.

Using this approach, the sparsest embedding we can construct has $O(d^2)$ rows and two non-zero entries per column. We further show how to construct even sparser embeddings using random sampling. Since we only use the sparse Cauchy part to deal with vectors in which light coordinates contribute most of the ℓ_1-norm, even if we zero out each coordinate with probability $1 - \varepsilon$ for a small constant ε, the resulting vector will still have a sufficiently ℓ_1-norm, with large enough probability. Thus, if we zero out each standard Cauchy random variable in the sparse Cauchy part with probability $1 - \varepsilon$, then the resulting embedding will still have the same distortion bound, up to a constant factor. By doing so, there will be $1 + \varepsilon$ non-zero entries in expectation in each column of the new embedding. This idea is formalized in Section 6.1.

[3] This is also observed in Reference [35], but the authors use exponential random variables there to remedy this issue instead of the idea of truncation that we use here.

1.6 Followup Work

Building upon our Theorem 1.2, followup work in Reference [18] improved our ℓ_1 oblivious subspace embedding from $r = \exp(\exp(O(d)))$ to $r = \exp(\text{poly}(d))$, and generalized it from a constant distortion to a $1+\varepsilon$ distortion, obtaining overall dimension $r = \exp(\text{poly}(d/\varepsilon))$. Our Theorem 1.2 is still useful though, as a suitably generalized version of it in terms of ε (see Reference [18] for details) is composed with another oblivious subspace embedding in the work of Reference [18], which is important for making their subspace embedding have a dimension r independent of n.

Additional followup work includes Reference [19], which uses our framework for proving Theorem 1.3, and obtains a more general reduction in terms of any embedding for the ℓ_2-norm; we refer the reader to Reference [19] for further details.

2 PRELIMINARIES

Throughout this article, we use $[n]$ to denote the set $\{1, 2, \ldots, n\}$. We use $\|\cdot\|_p$ to denote the ℓ_p-norm of a vector or the entry-wise ℓ_p-norm of a matrix. The following lemma is a direct application of Hölder's inequality.

Lemma 2.1. *For any $x \in \mathbb{R}^n$ and $1 \le p \le q \le 2$, we have*

$$\|x\|_q \le \|x\|_p \le n^{1/p-1/q}\|x\|_q.$$

For $u \in \mathbb{R}^n$ and $1 \le a \le b \le n$, let $u_{a:b}$ denote the vector with ith coordinate equal to u_i when $a \le i \le b$, and zero otherwise. For a matrix $S \in \mathbb{R}^{n\times m}$, we use $S_{i,*}$ to denote the ith row of S, and $S_{*,j}$ to denote the jth column of S.

For two vectors $u, v \in \mathbb{R}^n$, we use $\langle u, v\rangle$ to denote the inner product of u and v.

Definition 2.2. For $p \in [1, 2]$, a distribution over $r \times n$ matrices Π is an ℓ_p oblivious subspace embedding, if for any fixed $A \in \mathbb{R}^{n\times d}$,

$$\Pr_{\Pi}\left[\|Ax\|_p \le \|\Pi Ax\|_p \le \kappa\|Ax\|_p, \forall x \in \mathbb{R}^d\right] \ge 0.99.$$

Here κ is the *distortion* of Π.

Throughout the article, we use $X \simeq Y$ to mean that X and Y have the same distribution. We use $X \ge Y$ to denote stochastic dominance, i.e., $X \ge Y$ iff for any $t \in \mathbb{R}$, $\Pr[X \ge t] \ge \Pr[Y \ge t]$.

2.1 Stable Distribution

Definition 2.3 (p-stable Distribution). A distribution $\mathcal{D}$ is p-stable if for any n real numbers $a_1, a_2, \ldots, a_n$, we have

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

Here X_i are i.i.d. drawn from $\mathcal{D}$ and $X \sim \mathcal{D}$.

p-stable distributions exist for any $0 < p \le 2$ (see, e.g., Reference [28]). We let $\mathcal{D}_p$ denote the p-stable distribution. It is also well known that the standard Cauchy distribution is 1-stable and the standard Gaussian distribution $\mathcal{N}(0, 1)$ is 2-stable.

We use the following lemma due to Nolan [28].

Lemma 2.4 (Theorem 1.2 in Reference [28]). *For $1 \le p < 2$, let $X_p \sim \mathcal{D}_p$.*

$$\lim_{t\to\infty} \Pr[X_p > t]/t^{-p} = c_p,$$

where $c_p > 0$ is a constant that depends only on p.

The following lemma is established in Reference [24] by using Lemma 2.4.

Lemma 2.5 (Lemma 8 in Reference [24]). *For $1 \le p < 2$, let $X_p \sim \mathcal{D}_p$. There exists a constant α_p such that*

$$\alpha_p |C| \ge |X_p|^p,$$

where C is a standard Cauchy random variable and α_p is a constant that depends only on p.

2.2 Tail Inequalities

We use the following standard form of the Chernoff bound and Bernstein's inequality.

Lemma 2.6 (Chernoff Bound). *Suppose $X_1, X_2, \ldots, X_n$ are independent random variables taking values in $[0,1]$. Let $X = \sum_{i=1}^n X_i$.*

For any $\delta > 0$, we have

$$\Pr[X > (1+\delta)\,\mathrm{E}[X]] \le \exp(-\delta^2\,\mathrm{E}[X]/3),$$

$$\Pr[X < (1-\delta)\,\mathrm{E}[X]] \le \exp(-\delta^2\,\mathrm{E}[X]/2).$$

For $t > 2e\,\mathrm{E}[X]$, we have

$$\Pr[X > t] \le 2^{-t}.$$

Lemma 2.7 (Bernstein's Inequality). *Suppose $X_1, X_2, \ldots, X_n$ are independent random variables taking values in $[0,b]$. Let $X = \sum_{i=1}^n X_i$ and $\mathrm{Var}[X] = \sum_{i=1}^n \mathrm{Var}[X_i]$ be the variance of X. For any $t > 0$, we have*

$$\Pr[X > \mathrm{E}[X] + t] \le \exp\left(-\frac{t^2}{2\,\mathrm{Var}[X] + 2bt/3}\right).$$

The following Bernstein-type lower tail inequality is due to Maurer [21].

Lemma 2.8 ([21]). *Suppose $X_1, X_2, \ldots, X_n$ are independent positive random variables that satisfy $\mathrm{E}[X_i^2] < \infty$. Let $X = \sum_{i=1}^n X_i$. For any $t > 0$, we have*

$$\Pr[X \le \mathrm{E}[X] - t] \le \exp\left(-\frac{t^2}{2\sum_{i=1}^n \mathrm{E}[X_i^2]}\right).$$

We use the following tail inequality of a Gaussian random vector, whose proof can be found in Appendix A.

Lemma 2.9. *Let $(a_1, a_2, \ldots, a_n)$ be a fixed vector. For $i \in [n]$, let $\{X_i\}$ be n possibly dependent standard Gaussian random variables. For any $1 \le p \le 2$, we have*

$$\Pr\left[\left(\sum_{i=1}^n |a_i X_i|^p\right)^{1/p} \in \left[C_p^{-1}\|a\|_p, C_p\|a\|_p\right]\right] \ge 0.99.$$

Here $C_p > 1$ is an absolute constant that depends only on p.

The following upper tail inequality for dependent standard Cauchy random variables is established in Reference [24].

Lemma 2.10 (Lemma 3 in Reference [8]). *For $i \in [n]$, let $\{X_i\}$ be n possibly dependent standard Cauchy random variables and $\gamma_i > 0$ with $\gamma = \sum_{i\in[n]} \gamma_i$. For any $t \ge 1$ and $n \ge 3$,*

$$\Pr\left[\sum_{i\in[n]} \gamma_i |X_i| > \gamma t\right] \le \frac{2\log(nt)}{t}.$$

The following corollary is a direct implication of Lemmas 2.10 and 2.5.

Corollary 2.11. *For $i \in [n]$, let $\{X_i\}$ be n possibly dependent p-stable random variables and $\gamma_i > 0$ with $\gamma = \sum_{i\in[n]} \gamma_i$. For any $t \geq 1$ and $n \geq 3$,*

$$\Pr\left[\sum_{i\in[n]} \gamma_i |X_i|^p > \alpha_p \gamma t\right] \leq \frac{2\log(nt)}{t},$$

where α_p is the constant in Lemma 2.5.

For the sum of absolute values of *independent* standard Cauchy random variables, it is possible to prove an upper tail inequality stronger than that in Lemma 2.10. The proof of the lemma can be found in Appendix A.

Lemma 2.12. *For $i \in [n]$, let $\{X_i\}$ be n independent standard Cauchy random variables. There exists a constant U_1, such that for any $n \geq 3$,*

$$\Pr\left[\sum_{i=1}^{n} |X_i| \leq U_1 n \log n\right] \geq 1 - \frac{\log\log n}{\log n}.$$

The following corollary is a direct implication of Lemma 2.12 and Lemma 2.5.

Corollary 2.13. *Suppose $1 \leq p < 2$. For $i \in [n]$, let $\{X_i\}$ be n independent p-stable random variables. There exists a constant U_p that depends only on p, such that for any $n \geq 3$,*

$$\Pr\left[\sum_{i=1}^{n} |X_i|^p \leq U_p n \log n\right] \geq 1 - \frac{\log\log n}{\log n}.$$

We use the following lower tail inequality for the sum of absolute values of independent p-stable random variables, whose proof can be found in Appendix A.

Lemma 2.14. *Suppose $1 \leq p < 2$. For $i \in [n]$, let $\{X_i\}$ be n independent p-stable random variables. There exists a constant L_p that depends only on p, such that for sufficiently large n and T,*

$$\Pr\left[\sum_{i=1}^{n} |X_i|^p \geq L_p n \log\left(\frac{n}{\log T}\right)\right] \geq 1 - \frac{1}{T}.$$

2.3 ε-nets

We use the following standard ε-net construction in the analysis of our subspace embeddings.

Definition 2.15. For any $1 \leq p \leq 2$, for a given $A \in \mathbb{R}^{n\times d}$, let $B = \{Ax \mid x \in \mathbb{R}^d, \|Ax\|_p = 1\}$. We say $\mathcal{N} \subseteq B$ is an ε-net of B if for any $y \in B$, there exists a $\hat{y} \in \mathcal{N}$ such that $\|y - \hat{y}\|_p \leq \varepsilon$.

Lemma 2.16 (see, e.g., Reference [33, p. 74]). *For a given $A \in \mathbb{R}^{n\times d}$, there exists an ε-net $\mathcal{N} \subseteq B = \{Ax \mid x \in \mathbb{R}^d, \|Ax\|_p = 1\}$ with size $|\mathcal{N}| \leq (1 + 1/\varepsilon)^d$.*

2.4 Known ℓ_2 Oblivious Subspace Embeddings

In References [4, 10, 11, 24, 25], a series of results on sparse ℓ_2 oblivious subspace embedding are obtained.

Lemma 2.17 (CountSketch [10, 24, 25]). *There exists an ℓ_2 oblivious subspace embedding over $O(d^2) \times n$ matrices Π, where each column of Π has a single non-zero entry and the distortion $\kappa = 2$.*

Lemma 2.18 (OSNAP [11, 25]). *For any $B > 2$, there exists an ℓ_2 oblivious subspace embedding over $O(B \cdot d\log d) \times n$ matrices Π, where each column of Π has at most $O(\log_B d)$ non-zero entries and the distortion $\kappa = 2$.*

For completeness, we include the construction for CountSketch and OSNAP here. In the CountSketch embedding, each column is chosen to have $s = 1$ non-zero entries chosen in a uniformly random location and the non-zero value is uniformly chosen in $\{-1, 1\}$. In the OSNAP embedding, each column is chosen to have $s = O(\log_B d)$ non-zero entries in random locations, each equal to $\pm s^{-1/2}$ uniformly at random. All other entries in both embeddings are set to zero.

We need a few additional properties of the CountSketch embedding and the OSNAP embedding. The following lemma is a direct calculation of the operator ℓ_1-norm of the matrices stated above.

Lemma 2.19. *For any $y \in \mathbb{R}^n$,*

- $\|\Pi y\|_1 \le \|y\|_1$ *if* Π *is sampled from the CountSketch embedding;*
- $\|\Pi y\|_1 \le O(\log_B^{1/2} d)\|y\|_1$ *if* Π *is sampled from the OSNAP embedding.*

The following lemma deals with the ℓ_p-norm of a vector and its ℓ_p-norm after projection using CountSketch or OSNAP. Its proof can be found in Appendix A.

Lemma 2.20. *For any $y \in \mathbb{R}^n$ and sufficiently large ω, with probability $1 - \exp(\Omega(\omega d \log d))$,*

- $\|\Pi(y_{1:d^2})\|_p \le (\omega d \log d)^{1-1/p}\|y\|_p$ *if* Π *is sampled from the CountSketch embedding;*
- $\|\Pi(y_{1:d^2})\|_p \le (O(\log_B d))^{1/p-1/2}(\omega d \log d)^{1-1/p}\|y\|_p$ *if* Π *is sampled from the OSNAP embedding.*

We remark that since one can permute entries of y arbitrarily, Lemma 2.20 gives a bound for any subset of d^2 entries of y.

2.5 Well-conditioned Bases

We recall the definition and some existential results on well-coditioned matrices with respect to ℓ_p-norms.

Definition 2.21 ((α, β, p)-well-conditioning [14]). For a given matrix $U \in \mathbb{R}^{n\times d}$ and $p \in [1, 2]$, let $\|\cdot\|_q$ be the dual norm of $\|\cdot\|_p$, i.e., $1/p + 1/q = 1$. We say U is (α, β, p)-well-conditioned if (i) $\|U\|_p \le \alpha$ and (ii) $\|x\|_q \le \beta\|Ux\|_p$ for any $x \in \mathbb{R}^d$.

Lemma 2.22. *For any full rank matrix $A \in \mathbb{R}^{n\times d}$ and $p \in [1, 2]$, there exists a basis matrix $U \in \mathbb{R}^{n\times d}$ of A such that U is $(d, 1, p)$-well-conditioned.*

Proof. By Auerbach's Lemma (see, e.g., Reference [33, p. 75]), there exists a set of basis vectors $u_1, u_2, \ldots, u_d$ of the column space of A, and a set of basis vectors $v_1, v_2, \ldots, v_d$, such that

- $\|u_i\|_p = 1$;
- $\|v_i\|_q = 1$;
- $\langle u_i, v_j\rangle = \begin{cases} 1 & i = j \\ 0 & i \ne j \end{cases}$.

Here, $\|\cdot\|_q$ is the dual norm of $\|\cdot\|_p$.

We let U be the matrix whose first column is $u_1 \cdot d^{1/q}$, second column is $u_2 \cdot d^{1/q}, \ldots$, and the last column is $u_d \cdot d^{1/q}$. Clearly,

$$\|U\|_p^p = \sum_{i=1}^{d} \|u_i\|_p^p \cdot d^{p/q} = d^p,$$

which implies $\alpha = d$.

For any $x \in \mathbb{R}^d$, since the ith column of U is $u_i \cdot d^{1/q}$, by Hölder's inequality,

$$|x_i| = |\langle v_i, u_i x_i\rangle| \le \|Ux\|_p\|v_i\|_q \cdot d^{-1/q} = \|Ux\|_p d^{-1/q}.$$

Thus, by Lemma 2.1,

$$\|x\|_q \le d^{1/q} \cdot \|x\|_\infty \le d^{1/q} \cdot \|Ux\|_p d^{-1/q} = \|Ux\|_p,$$

which implies $\beta = 1$. □

3 HARDNESS RESULT

3.1 The Lower Bound

The goal of this section is to prove Theorem 1.1. We restate it here for convenience.

Theorems 1.1 (Restated). *For $1 \le p < 2$, if a distribution over $r \times n$ matrices Π is an ℓ_p oblivious subspace embedding, then the distortion*

$$\kappa = \Omega\left(\frac{1}{\left(\frac{1}{d}\right)^{1/p} \cdot \log^{2/p} r + \left(\frac{r}{n}\right)^{1/p-1/2}}\right).$$

By Yao's minimax principle [37], it suffices to show that there exists a hard distribution $\mathcal{A}$ over $\mathbb{R}^{n\times d}$ such that for any $\Pi \in \mathbb{R}^{r\times n}$, if

$$\Pr_{A\sim\mathcal{A}}\left[\|Ax\|_p \le \|\Pi Ax\|_p \le \kappa\|Ax\|_p, \forall x \in \mathbb{R}^d\right] \ge 0.99, \tag{2}$$

then

$$\kappa = \Omega\left(\frac{1}{\left(\frac{1}{d}\right)^{1/p} \cdot \log^{2/p} r + \left(\frac{r}{n}\right)^{1/p-1/2}}\right).$$

The columns in our construction of $\mathcal{A}$ consist of three parts:

- The first column is a vector where all the n entries are i.i.d. standard Gaussian random variables. We call this column the D-column.
- For the next $d/4$ columns, each column has $4n/d$ non-zero entries, where all these non-zero entries are i.i.d. standard Gaussian random variables. The indices of the $4n/d$ non-zero entries of the ith column are $(4n/d)\cdot(i-1)+1, (4n/d)\cdot(i-1)+2, \ldots, (4n/d)\cdot i$. We call each such column an M-column.
- We divide the next $d/2$ columns into $\log(n/d)$ *blocks*, where each block contains $\frac{d}{2\log(n/d)}$ columns. For $0 \le i < \log(n/d)$, columns in the ith block contain 2^{i+1} non-zero entries and all of these non-zero entries are i.i.d. standard Gaussian random variables. For the $\frac{d}{2\log(n/d)}$ columns in the ith block, the indices of the $\frac{d}{2\log(n/d)}\cdot 2^{i+1} = \frac{d}{\log(n/d)}2^i$ non-zero entries are sampled from $\{1, 2, \ldots, n\}$ without replacement (which implies the sets of indices of non-zero entries are disjoint for two different columns in the same block). We call each such column an S-column.

All entries in other columns are zero. This finishes our construction of $\mathcal{A}$.
The following lemma is a direct implication of Lemma 2.9 and our construction.

Lemma 3.1. *For each column c in $\mathcal{A}$, with probability at least 0.99, the following holds:*

(1) *If c is an S-column in the ith block, then $\|c\|_p \le C_p 2^{(i+1)/p}$.*
(2) *If c is an M-column, then $\|c\|_p \le C_p(4n/d)^{1/p}$.*
(3) *If c is a D-column, then $\|c\|_p \ge C_p^{-1} n^{1/p}$.*

Here C_p is the constant in Lemma 2.9.

Lemma 3.2. *For any matrix $\Pi \in \mathbb{R}^{r\times n}$ that satisfies the inequality in Equation (2), the ℓ_p-norm of each column of Π is at most $C_p^2\kappa(4n/d)^{1/p}$, where C_p is the constant in Lemma 2.9.*

Proof. Suppose for contradiction that there exists an $i \in [n]$ for which the ith column of Π has ℓ_p-norm larger than $C_p^2\kappa(4n/d)^{1/p}$. Consider the vector M_j, which is the jth M-column, whose ith entry is a standard Gaussian random variable, i.e., $(4n/d)\cdot(j-1)+1 \le i \le (4n/d)\cdot j$. We first show that with probability at least 0.99, $\|\Pi\mathsf{M}_j\|_p > C_p\kappa(4n/d)^{1/p}$. According to the 2-stability of the standard Gaussian distribution, for any $k \in [r]$,

$$(\Pi\mathsf{M}_j)_k \sim \left(\sum_{l=(4n/d)\cdot(j-1)+1}^{(4n/d)\cdot j} \Pi_{k,l}^2\right)^{1/2} \mathcal{N}(0,1).$$

Since

$$\left(\sum_{l=(4n/d)\cdot(j-1)+1}^{(4n/d)\cdot j} \Pi_{k,l}^2\right)^{1/2} \ge \Pi_{k,i},$$

according to Lemma 2.9, with probability at least 0.99,

$$\|\Pi\mathsf{M}_j\|_p \ge C_p^{-1}\|\Pi_{*,i}\|_p > C_p\kappa(4n/d)^{1/p}.$$

According to Lemma 3.1, with probability at least 0.99,

$$\|\mathsf{M}_j\|_p \le C_p(4n/d)^{1/p},$$

which implies the condition in Equation (2) is violated. □

Lemma 3.3. *For any matrix $\Pi \in \mathbb{R}^{r\times n}$ that satisfies the condition in Equation (2), for any $0 \le i < \log(n/d)$, the number of columns in Π with ℓ_p-norm larger than $C_p^2\kappa 2^{(i+1)/p}$ is at most $\frac{n\log(n/d)}{d}2^{-i}$, where C_p is the constant in Lemma 2.9.*

Proof. Suppose for contradiction that for some $0 \le i < \log(n/d)$, the number of columns in Π with ℓ_p-norm larger than $C_p^2\kappa 2^{(i+1)/p}$ is larger than $\frac{n\log(n/d)}{d}2^{-i}$. Let $\pi^1, \pi^2, \ldots, \pi^{d\log^{-1}(n/d)/2}$ be the $d\log^{-1}(n/d)/2$ S-column in the ith block. With probability at least $1-(1-\frac{\log(n/d)}{d}2^{-i})^{\frac{d}{\log(n/d)}2^i} \ge 1-1/e$, there exists a $j \in [d\log^{-1}(n/d)/2]$ and $l \in [n]$ such that (i) $\|\Pi_{*,l}\|_p \ge C_p^2\kappa 2^{(i+1)/p}$ and (ii) π_l^j is a standard Gaussian random variable. According to Lemma 3.1, with probability at least 0.99, $\|\pi^j\|_p \le C_p 2^{(i+1)/p}$. Now, we show that with probability at least 0.99, $\|\Pi\pi^j\|_p \ge C_p^{-1}\|\Pi_{*,l}\|_p > C_p\kappa 2^{(i+1)/p}$. Suppose $P \subseteq [n]$ is the set of indices at which π^j contains a standard Gaussian random variable. We know that $l \in P$. Thus, due to the 2-stability of the standard Gaussian distribution, for any $k \in [r]$,

$$(\Pi\pi^j)_k \sim \left(\sum_{m\in P}\Pi_{k,m}^2\right)^{1/2}\mathcal{N}(0,1).$$

Since

$$\left(\sum_{m\in P}\Pi_{k,m}^2\right)^{1/2} \ge \Pi_{k,l},$$

according to Lemma 2.9, with probability at least 0.99, $\|\Pi\pi^j\|_p \ge C_p^{-1}\|\Pi_{*,l}\|_p > C_p\kappa 2^{(i+1)/p} \ge \kappa\|\pi^j\|_p$, which implies the condition in Equation (2) is violated. □

Lemma 3.4. *For any matrix $\Pi \in \mathbb{R}^{r\times n}$ that satisfies the condition in Equation (2), we have*

$$\left(\sum_{i=1}^{r} \|\Pi_{i,*}\|_2^p\right)^{1/p} = O\left(\kappa(n/d)^{1/p}\log^{2/p}(n/d) + \kappa r^{\frac{1}{p}-\frac{1}{2}}\sqrt{n}\right).$$

Proof. We partition the columns of Π into two parts. We let Π^L be the submatrix of Π formed by columns with ℓ_p-norm at most $2^{1/p}C_p^2\kappa$ and Π^H be the submatrix formed by columns with ℓ_p-norm larger than $2^{1/p}C_p^2\kappa$. For Π^H, by Lemmas 3.2 and 3.3, we have

$$\begin{aligned}
&\left(\sum_{i=1}^{r} \|\Pi^H_{i,*}\|_2^p\right)^{1/p} \le \left(\sum_{i=1}^{r} \|\Pi^H_{i,*}\|_p^p\right)^{1/p} = \|\Pi^H\|_p = \left(\sum_i \|\Pi^H_{*,i}\|_p^p\right)^{1/p}\\
\le&\left(\sum_{i=0}^{\log(n/d)-1} (C_p^2\kappa)^p 2^{i+2}\frac{n\log(n/d)}{2^i d} + (C_p^2\kappa)^p(4n/d)2\log(n/d)\right)^{1/p}\\
=&\,O(\kappa(n/d)^{1/p}\log^{2/p}(n/d)).
\end{aligned}$$

For Π^L, since all the columns have ℓ_p-norm at most $2^{1/p}C_p^2\kappa$, we have

$$\begin{aligned}
&\left(\sum_{i=1}^{r} \|\Pi^L_{i,*}\|_2^p\right)^{1/p}\\
\le& r^{\frac{1}{p}-\frac{1}{2}}\left(\sum_{i=1}^{r} \|\Pi^L_{i,*}\|_2^2\right)^{1/2} = r^{\frac{1}{p}-\frac{1}{2}}\|\Pi^L\|_2 = r^{\frac{1}{p}-\frac{1}{2}}\left(\sum_i \|\Pi^L_{*,i}\|_2^2\right)^{1/2}\\
\le& r^{\frac{1}{p}-\frac{1}{2}}\left(\sum_i \|\Pi^L_{*,i}\|_p^2\right)^{1/2} = O\left(\kappa r^{\frac{1}{p}-\frac{1}{2}}\sqrt{n}\right),
\end{aligned}$$

where the first inequality follows from Lemma 2.1 and the last equality follows from the fact that Π^L has at most n columns.

Notice that for any $1 \le i \le r$, $\|\Pi_{i,*}\|_2 \le \|\Pi^H_{i,*}\|_2 + \|\Pi^L_{i,*}\|_2$, which implies

$$\left(\sum_{i=1}^{r} \|\Pi_{i,*}\|_2^p\right)^{1/p} \le \left(\sum_{i=1}^{r} \|\Pi^H_{i,*}\|_2^p\right)^{1/p} + \left(\sum_{i=1}^{r} \|\Pi^L_{i,*}\|_2^p\right)^{1/p} = O\left(\kappa(n/d)^{1/p}\log^{2/p}(n/d) + \kappa r^{\frac{1}{p}-\frac{1}{2}}\sqrt{n}\right).$$

□

Now consider the vector D, which is the D-column in $\mathcal{A}$. According to Lemma 3.1, with probability at least 0.99, $\|\mathrm{D}\|_p = \Omega(n^{1/p})$. Due to the 2-stability of the standard Gaussian distribution,

$$(\Pi\mathrm{D})_i \sim \|\Pi_{i,*}\|_2\mathcal{N}(0,1).$$

According to Lemma 2.9, with probability at least 0.99,

$$\|\Pi\mathrm{D}\|_p = O\left(\sum_{i=1}^{r} \|\Pi_{i,*}\|_2^p\right)^{1/p} = O\left(\kappa(n/d)^{1/p}\log^{2/p}(n/d) + \kappa r^{\frac{1}{p}-\frac{1}{2}}\sqrt{n}\right).$$

According to the condition in Equation (2), we have

$$\Omega(n^{1/p}) = \|\mathrm{D}\|_p \le \|\Pi\mathrm{D}\|_p = O\left(\kappa(n/d)^{1/p}\log^{2/p}(n/d) + \kappa r^{\frac{1}{p}-\frac{1}{2}}\sqrt{n}\right),$$

which implies

$$\kappa = \Omega\left(\frac{1}{\left(\frac{1}{d}\right)^{1/p} \cdot \log^{2/p}(n/d) + \left(\frac{r}{n}\right)^{1/p-1/2}}\right).$$

Now, we show that the lower bound can be further improved to

$$\kappa = \Omega\left(\frac{1}{\left(\frac{1}{d}\right)^{1/p} \cdot \log^{2/p} r + \left(\frac{r}{n}\right)^{1/p-1/2}}\right). \tag{3}$$

We first note that r should be at least d; otherwise, if we take a full-rank matrix $A \in \mathbb{R}^{n\times d}$, then $\operatorname{rank}(\Pi A) < d = \operatorname{rank}(A)$, which means we can find a non-zero vector $y = Ax$ in the column space of A and $\Pi y = 0$, which implies the distortion κ is not finite.

When $n \le rd^{2/(2-p)}$, $\log^{2/p}(n/d) = O(\log^{2/p} r)$, which means the lower bound in Equation (3) holds. When $n > rd^{2/(2-p)}$, we repeat the argument above but only consider the first $rd^{2/(2-p)}$ columns of Π. By doing so, we get a lower bound of

$$\kappa = \Omega\left(\frac{1}{\left(\frac{1}{d}\right)^{1/p} \cdot \log^{2/p}(rd^{2/(2-p)-1}) + \left(\frac{1}{d^{2/(2-p)}}\right)^{1/p-1/2}}\right) = \Omega(d^{1/p}/\log^{2/p}(r)),$$

which is always stronger than the lower bound of

$$\kappa = \Omega\left(\frac{1}{\left(\frac{1}{d}\right)^{1/p} \cdot \log^{2/p} r + \left(\frac{r}{n}\right)^{1/p-1/2}}\right).$$

3.2 Necessity of Dependence on r

The goal of this section is to prove Theorem 3.5.

Theorem 3.5. *Let $r = \exp(4 \cdot 10^4 \cdot (24(U_p L_p^{-1})^{1/p})^{2d})$, where U_p and L_p are the constants in Corollary 2.13 and Lemma 2.14, respectively. For $1 \le p < 2$, there exists an ℓ_p oblivious subspace embedding over $r \times n$ matrices Π, where the distortion κ is a constant that depends only on p.*

Our construction for the embedding in Theorem 3.5 is actually the same as the dense p-stable embedding in Reference [30] (for $p = 1$) and Theorem 6 in Reference [24] (for $1 < p < 2$), whose entries are i.i.d. sampled from the scaled p-stable distribution $(r \log r)^{-1/p}\mathcal{D}_p$.

For any given matrix $A \in \mathbb{R}^{n\times d}$ and any $x \in \mathbb{R}^d$, we show that

$$\Pr_{\Pi}\left[\left(L_p/2\right)^{1/p} \|Ax\|_p \le \|\Pi Ax\|_p \le U_p^{1/p}\|Ax\|_p\right] \ge 1 - 10^{-2}\left(24\left(U_p L_p^{-1}\right)^{1/p}\right)^{-d}.$$

According to the definition of the p-stable distribution in Definition 2.3, for any $i \in [r]$,

$$(\Pi Ax)_i \sim (r \log r)^{-1/p}\|Ax\|_p \mathcal{D}_p.$$

Since the entries in Π are independent, the entries in the vector ΠAx are also independent. Thus, according to Corollary 2.13, with probability at least $1 - \frac{\log\log r}{\log r} \ge 1 - 200^{-1}(24(U_p L_p^{-1})^{1/p})^{-d}$, we have

$$\|\Pi Ax\|_p^p \le U_p(r \log r)^{-1}\|Ax\|_p^p \cdot r \log r = U_p\|Ax\|_p^p,$$

which implies

$$\|\Pi Ax\|_p \le U_p^{1/p}\|Ax\|_p.$$

However, according to Lemma 2.14, by setting $T = 200(24(U_pL_p^{-1})^{1/p})^d$, with probability at least $1 - 1/T = 1 - 200^{-1}(24(U_pL_p^{-1})^{1/p})^{-d}$, we have

$$\|\Pi Ax\|_p^p \geq L_p(r\log r)^{-1}\|Ax\|_p^p \cdot r\log\frac{r}{\log T} \geq L_p/2 \cdot \|Ax\|_p^p,$$

which implies

$$\|\Pi Ax\|_p \geq \left(L_p/2\right)^{1/p}\|Ax\|_p.$$

It follows by a union bound that for any $x \in \mathbb{R}^d$,

$$\Pr_{\Pi}\left[(L_p/2)^{1/p}\|Ax\|_p \leq \|\Pi Ax\|_p \leq U_p^{1/p}\|Ax\|_p\right] \geq 1 - 10^{-2}\left(24\left(U_pL_p^{-1}\right)^{1/p}\right)^{-d}.$$

We build an ε-net $\mathcal{N} \subseteq B = \{Ax \mid x \in \mathbb{R}^d, \|Ax\|_p = 1\}$ by setting $1/\varepsilon = 8(U_pL_p^{-1})^{1/p}$. According to Lemma 2.16, $|\mathcal{N}| \leq (1 + 1/\varepsilon)^d \leq (3/\varepsilon)^d = (24(U_pL_p^{-1})^{1/p})^d$. Again by a union bound, with probability at least 0.99, we have for any $y \in \mathcal{N}$,

$$(L_p/2)^{1/p}\|y\|_p \leq \|\Pi y\|_p \leq U_p^{1/p}\|y\|_p.$$

Condition on the event stated above. Now, we show that for any $x \in \mathbb{R}^d$,

$$(L_p/4)^{1/p}\|Ax\|_p \leq \|\Pi Ax\|_p \leq 2U_p^{1/p}\|Ax\|_p.$$

For any $x \in \mathbb{R}^d$, let $y = Ax$. By homogeneity, we can assume $\|y\|_p = 1$. We claim that y can be written as

$$y = y^0 + y^1 + y^2 + \ldots,$$

where for any $i \geq 0$, we have (i) $\frac{y^i}{\|y^i\|_p} \in \mathcal{N}$ and (ii) $\|y^i\|_p \leq \varepsilon^i$.

According to the definition of an ε-net, there exists a vector $y^0 \in \mathcal{N}$ for which $\|y - y^0\|_p \leq \varepsilon$ and $\|y^0\|_p = 1$. If $y = y^0$, then we stop. Otherwise, we consider the vector $\frac{y-y^0}{\|y-y^0\|_p}$. Again, we can find a vector $\hat{y}^1 \in \mathcal{N}$ such that $\|\frac{y-y^0}{\|y-y^0\|_p} - \hat{y}^1\|_p \leq \varepsilon$ and $\|\hat{y}^1\|_p = 1$. Here, we set $y^1 = \|y - y^0\|_p \cdot \hat{y}^1$ and continue this process inductively.

It follows that

$$\begin{aligned}\|\Pi y\|_p &\geq \|\Pi y^0\| - \sum_{i>0}\|\Pi y^i\| \\ &\geq (L_p/2)^{1/p} - \sum_{i>0} U_p^{1/p}\varepsilon^i \\ &\geq (L_p/2)^{1/p} - 2U_p^{1/p}\varepsilon \geq (L_p/4)^{1/p}\end{aligned}$$

and

$$\|\Pi y\|_p \leq \sum_{i\geq 0}\|\Pi y^i\| \leq \sum_{i\geq 0} U_p^{1/p}\varepsilon^i \leq 2U_p^{1/p}.$$

Thus, Π is a valid ℓ_p oblivious subspace embedding with $\kappa \leq 2(4U_pL_p^{-1})^{1/p}$, which is a constant that depends only on p.

4 NEW SUBSPACE EMBEDDINGS FOR ℓ_1

In this section, we present new *sparse* ℓ_1 oblivious subspace embeddings with nearly optimal distortion.

THEOREM 4.1. *For any given $A \in \mathbb{R}^{n\times d}$, let U be a $(d,1,1)$-well-conditioned basis of A. There exists an ℓ_1 oblivious subspace embedding over $O(d^2) \times n$ matrices Π where each column of Π has two non-zero entries and with probability 0.99, for any $x \in \mathbb{R}^d$,*

$$\Omega(\log d)\|Ux\|_1 \le \|\Pi Ux\|_1 \le O(d\log d)\|Ux\|_1.$$

THEOREM 4.2. *For any given $A \in \mathbb{R}^{n\times d}$ and sufficiently large B, let U be a $(d,1,1)$-well-conditioned basis of A. There exists an ℓ_1 oblivious subspace embedding over $O(B\cdot d\log d)\times n$ matrices Π where each column of Π has $O(\log_B d)$ non-zero entries and with probability 0.99, for any $x \in \mathbb{R}^d$,*

$$\Omega(\log B)\|Ux\|_1 \le \|\Pi Ux\|_1 \le O(d\log d)\|Ux\|_1.$$

Our embedding for Theorems 4.1 and 4.2 can be written as $\Pi = (\Pi_1, \Pi_2)^T$. For Theorem 4.1, Π_1 is sampled from the CountSketch embedding in Lemma 2.17, scaled by a $d\log d$ factor. For Theorem 4.2, Π_1 is sampled from the OSNAP embedding in Lemma 2.18 with $O(B\cdot d\log d)$ rows, $O(\log_B d)$ non-zero entries per column, and scaled by a $d\log B$ factor. Suppose Π_1 has R_1 rows. Let $R_2 = \min\{R_1, d^{1.1}\}$. Π_2 can be written as $\Phi D : \mathbb{R}^n \to \mathbb{R}^{R_2}$ as follows:

- $h : [n] \to [R_2]$ is a random map so that for each $i \in [n]$ and $t \in [R_2]$, $h(i) = t$ with probability $1/R_2$.
- Φ is an $R_2 \times n$ binary matrix with $\Phi_{h(i),i} = 1$ and all remaining entries 0.
- D is an $n \times n$ random diagonal matrix where the diagonal entries are i.i.d. sampled from the standard Cauchy distribution.

It is immediate to see that the number of rows in Π_2 is at most that in Π_1. Furthermore, Π_2 has a single non-zero entry per column.

In the remainder of this section, we prove the dilation bound in Section 4.1, and the contraction bound in Section 4.2. In the analysis, we will define three events $\mathcal{E}_1$, $\mathcal{E}_2$ and $\mathcal{E}_3$, which we will condition on later in the analysis. We will prove that each of these events holds with probability at least 0.999. By a union bound, all of these events hold with probability at least 0.997. Thus, these conditions will not affect our overall failure probability by more than 0.003.

4.1 No Overestimation

Let $\mathcal{E}_1$ be the event that $\|\Pi_2 U\| \le O(d\log d)$. We first prove that $\mathcal{E}_1$ holds with probability at least 0.999.

LEMMA 4.3. *$\mathcal{E}_1$ holds with probability at least 0.999.*

PROOF.

$$\|\Pi_2 U\|_1 = \sum_{i=1}^{R_2}\sum_{j=1}^{d}|(\Pi_2 U)_{i,j}| = \sum_{i=1}^{R_2}\sum_{j=1}^{d}\left|\sum_{k|h(k)=i} D_{k,k}U_{k,j}\right| \simeq \sum_{i=1}^{R_2}\sum_{j=1}^{d}\left(\sum_{k|h(k)=i}|U_{k,j}|\right)|\hat{X}_{i,j}|.$$

Here $\{\hat{X}_{i,j}\}$ are dependent standard Cauchy random variables. Since U is a $(d,1,1)$-well-conditioned basis of A, we have

$$\sum_{i=1}^{R_2}\sum_{j=1}^{d}\left(\sum_{k|h(k)=i}|U_{k,j}|\right) \le d.$$

By Lemma 2.10, we have

$$\Pr\left[\|\Pi_2 U\|_1 > td\right] \leq \frac{2\log(R_2 td)}{t}.$$

Taking $t = \omega \log d$ where ω is a sufficiently large constant, we have

$$\Pr\left[\|\Pi_2 U\|_1 \leq td\right] \geq 0.999. \qquad \square$$

Lemma 4.4. *Conditioned on $\mathcal{E}_1$, for any $x \in \mathbb{R}^d$, we have*

$$\|\Pi_2 Ux\|_1 \leq O(d \log d)\|Ux\|_1.$$

Proof.

$$\|\Pi_2 Ux\|_1 \leq \|\Pi_2 U\|_1 \|x\|_\infty \leq \|\Pi_2 U\|_1 \|Ux\|_1 \leq O(d \log d)\|Ux\|_1.$$

The first inequality follows from Hölder's inequality, and the second inequality follows from the definition of a $(d, 1, 1)$-well-conditioned basis. $\square$

Since Π_1 is the CountSketch embedding scaled by a $d \log d$ factor, or the OSNAP embedding scaled by a $d \log B$ factor, the following lemma is a direct implication of Lemma 2.19.

Lemma 4.5. *For any $x \in \mathbb{R}^d$, we have*

$$\|\Pi_1 Ux\|_1 \leq O(d \log d)\|Ux\|_1.$$

Combining Lemma 4.4 and Lemma 4.5, we can bound the overall dilation of our embedding.

Lemma 4.6. *Conditioned on $\mathcal{E}_1$, for any $x \in \mathbb{R}^d$, we have*

$$\|\Pi Ux\|_1 \leq O(d \log d)\|Ux\|_1.$$

Proof.

$$\|\Pi Ux\|_1 = \|\Pi_1 Ux\|_1 + \|\Pi_2 Ux\|_1 = O(d \log d)\|Ux\|_1. \qquad \square$$

4.2 No Underestimation

We let $\mathcal{E}_2$ be the event that for any $x \in \mathbb{R}^d$,

$$d \log d\|Ux\|_2 \leq \|\Pi_1 Ux\|_2 \leq 2d \log d\|Ux\|_2 \qquad \text{(for Theorem 4.1)}$$

or

$$d \log B\|Ux\|_2 \leq \|\Pi_1 Ux\|_2 \leq 2d \log B\|Ux\|_2 \qquad \text{(for Theorem 4.2)}.$$

Since Π_1 is sampled from an ℓ_2 oblivious subspace embedding with $\kappa = 2$ and scaled by a factor of $d \log d$ (for Theorem 4.1) or $d \log B$ (for Theorem 4.2), $\mathcal{E}_2$ holds with probability at least 0.999.[4]

Without loss of generality, we assume $|x_1| \geq |x_2| \geq |x_3| \geq \cdots \geq |x_n|$. Of course, this order is unknown and is not used by our embedding.

We first show that for any $y = Ux$, if we can find a "heavy" part inside y, then the scaled ℓ_2 oblivious subspace embedding Π_1 also works well for ℓ_1. Formally, we have the following lemma.

Lemma 4.7. *Conditioned on $\mathcal{E}_2$, for any $x \in \mathbb{R}^d$, if $\|(Ux)_{1:d^2}\|_1 \geq 0.5\|Ux\|_1$, then*

- $\|\Pi_1 Ux\|_1 \geq \Omega(\log d)\|Ux\|_1$ *for Theorem 4.1;*
- $\|\Pi_1 Ux\|_1 \geq \Omega(\log B)\|Ux\|_1$ *for Theorem 4.2.*

[4]Notice that by Definition 2.2, $\mathcal{E}_2$ only holds with probability 0.99. However, for the CountSketch embedding in Lemma 2.17 and the OSNAP embedding in Lemma 2.18, we can boost the failure probability to an arbitrarily small constant, by increasing the dimension by a constant factor. By doing so, we can now assume $\mathcal{E}_2$ holds with probability 0.999.

Proof. Notice that

$$\|Ux\|_2 \geq \|(Ux)_{1:d^2}\|_2 \geq \frac{1}{d}\|(Ux)_{1:d^2}\|_1 \geq \frac{1}{2d}\|Ux\|_1,$$

where the second inequality follows from Lemma 2.1. Thus, for Theorem 4.1, $\|\Pi_1 Ux\|_1 \geq \|\Pi_1 Ux\|_2 \geq \Omega(\log d)\|Ux\|_1$, since Π_1 is sampled from an ℓ_2 oblivious subspace embedding and scaled by a factor of $d \log d$. For Theorem 4.2, $\|\Pi_1 Ux\|_1 \geq \|\Pi_1 Ux\|_2 \geq \Omega(\log B)\|Ux\|_1$, since Π_1 is sampled from an ℓ_2 oblivious subspace embedding and scaled by a factor of $d \log B$. □

Now, we analyze those vectors Ux that do *not* contain a "heavy" part. We show that they can be handled by the Π_2 part of our embedding.

Lemma 4.8. *For any $x \in \mathbb{R}^d$, if $\|(Ux)_{d^2+1:n}\|_1 \geq 0.5\|Ux\|_1$, then with probability at least $1 - \exp(-32d \log d)$, we have*

- $\|\Pi_2 Ux\|_1 \geq \Omega(\log d)\|Ux\|_1$ *for Theorem 4.1;*
- $\|\Pi_2 Ux\|_1 \geq \Omega(\log B)\|Ux\|_1$ *for Theorem 4.2.*

Proof. Let $y = Ux$. By homogeneity, we assume $\|y\|_1 = 1$. According to the given condition, we have $\|y_{d^2+1:n}\|_1 \geq 0.5$. Notice that $\|y_{d^2+1:n}\|_\infty \leq 1/d^2$, since, otherwise, $\|y_{1:d^2}\|_1 > d^2 \cdot 1/d^2 = 1$.

For $i \in [R_2]$, let $B_i = \sum_{d^2 < j \leq n} B_{i,j}$, where

$$B_{i,j} = \begin{cases} |y_j| & \text{if } h(j) = i \\ 0 & \text{otherwise} \end{cases}.$$

It follows that $\sum_{i=1}^{R_2} B_i = \|y_{d^2+1:n}\|_1 \geq 0.5$.

Since $\|y_{d^2+1:n}\|_\infty \leq 1/d^2$ and $1/2 \leq \|y_{d^2+1:n}\|_1 \leq 1$, for any $i \in [R_2]$ and $j > d^2$, we have

$$B_{i,j} \leq \frac{1}{d^2}$$

and

$$\frac{1}{2R_2} \leq \mathrm{E}[B_i] \leq \frac{1}{R_2}.$$

Furthermore, by Hölder's inequality, we have

$$\mathrm{Var}[B_i] = \sum_{j=1}^{n} \mathrm{Var}[B_{i,j}] \leq \frac{1}{R_2} \sum_{j=d^2+1}^{n} y_j^2 \leq \frac{1}{R_2}\|y_{d^2+1:n}\|_\infty \cdot \|y_{d^2+1:n}\|_1 \leq \frac{1}{R_2 d^2}.$$

Thus, by Bernstein's inequality in Lemma 2.7, we have

$$\Pr[B_i \geq 1/R_2 + t] \leq \exp\left(-\frac{t^2}{\frac{2}{R_2 d^2} + \frac{2t}{3d^2}}\right). \tag{4}$$

Let $t = d^{0.2}/R_2$. Since $R_2 \leq d^{1.1}$, by Equation (4), we have

$$\Pr[B_i > (d^{0.2} + 1)/R_2] \leq \exp(-3d^{2.2}/4R_2) \leq \exp(-3d^{1.1}/4).$$

By a union bound, with probability at least

$$1 - \exp(-3d^{1.1}/4) \cdot R_2 \geq 1 - \exp(-32d \log d)/4,$$

simultaneously for all $i \in [R_2]$, we have $B_i \leq (d^{0.2} + 1)/R_2$.

Let $t = 1/R_2$. Since $R_2 \leq d^{1.1}$, by Equation (4), we have

$$\Pr[B_i > 2/R_2] \leq \exp(-3d^2/8R_2) \leq \exp(-3d^{0.9}/8).$$

According to Reference [15], B_i are negatively associated, which implies for any $I \subseteq [R_2]$, we have

$$\Pr[B_i \geq t_i, i \in I] \leq \prod_{i \in I} \Pr[B_i \geq t_i].$$

Thus, the probability that the number of B_i which satisfy $B_i > 2/R_2$ is larger than $d^{0.2}$, is at most

$$\binom{R_2}{d^{0.2}} \exp((-3d^{0.9}/8) \cdot d^{0.2}) < \exp(-32d \log d)/4.$$

It follows that with probability at least $1 - \exp(-32d \log d)/2$, for any $i \in [R_2]$, we have $B_i \leq (d^{0.2} + 1)/R_2$, and the number of B_i which satisfy $B_i > 2/R_2$ is at most $d^{0.2}$. In the rest of the proof, we condition on this event.

Since $R_2 \geq d \log d$,

$$\sum_{i \in [R_2] | B_i \leq 2/R_2} B_i \geq 0.5 - d^{0.2} \cdot (d^{0.2} + 1)/R_2 \geq 1/4.$$

Thus, the number of B_i which satisfy $B_i \geq \frac{1}{8R_2}$ is at least $R_2/16$, since, otherwise,

$$\sum_{i \in [R_2] | B_i \leq 2/R_2} B_i < R_2/16 \cdot 2/R_2 + R_2 \cdot \frac{1}{8R_2} = 1/4.$$

Now consider $\Pi_2 y$. According to the 1-stability of the standard Cauchy distribution,

$$|(\Pi_2 y)_i| \simeq \left(\sum_{j \in [n] | h(j) = i} |y_j| \right) \cdot |X_i|,$$

where $\{X_i\}$ are independent standard Cauchy random variables. Notice that conditioned on the event stated above, the number of B_i which satisfy $B_i \geq \frac{1}{8R_2}$ is at least $R_2/16$. Furthermore, for any $i \in [R_2]$, $\sum_{j \in [n] | h(j) = i} |y_j| \geq \sum_{d^2 < j \leq n | h(j) = i} |y_j| = B_i$. Thus,

$$\sum_{i=1}^{R_2} |(\Pi_2 y)_i| \geq \sum_{i=1}^{R_2/16} \frac{1}{8R_2} |\overline{X}_i|,$$

where $\{\overline{X}_i\}$ are independent standard Cauchy random variables.

According to Lemma 2.14, by setting $T = 2 \exp(32d \log d)$, with probability at least $1 - 1/T$, we have

$$\sum_{i=1}^{R_2} |(\Pi_2 y)_i| \geq L_1 \cdot \frac{R_2}{16} \cdot \log(R_2/(16 \log T)) \cdot \frac{1}{8R_2} = \Omega(\log(R_2/\log T)).$$

Thus, for Theorem 4.1, we have $\|\Pi_2 y\|_1 \geq \Omega(\log d)$, since $R_2 = d^{1.1}$ and $\log T = O(d \log d)$. For Theorem 4.2, when $R_1 \leq d^{1.1}$, we have $\|\Pi_2 y\|_1 \geq \Omega(\log B)$, since $R_2 = R_1 = O(B \cdot d \log d)$ and $\log T = O(d \log d)$. When $R_1 > d^{1.1}$, we have $\|\Pi_2 y\|_1 \geq \Omega(\log d) = \Omega(\log B)$, since $R_2 = d^{1.1}$ and $\log T = O(d \log d)$. □

Set $\varepsilon = 1/d^2$ and create an ε-net $\mathcal{N} \subseteq B = \{Ux \mid x \in \mathbb{R}^d \text{ and } \|Ux\|_1 = 1\}$. According to Lemma 2.16, $|\mathcal{N}| \leq (1 + d^2)^d \leq (3d^2)^d$. Let $\mathcal{E}_3$ be the event that for all $y \in \mathcal{N}$, if $\|y_{d^2+1:n}\|_1 \geq 0.5$, then $\|\Pi_2 y\|_1 \geq \Omega(\log d)\|y\|_1$ (for Theorem 4.1) or $\|\Pi_2 y\|_1 \geq \Omega(\log B)\|y\|_1$ (for Theorem 4.2).

Now, we show that $\mathcal{E}_3$ holds with constant probability.

Lemma 4.9. *$\mathcal{E}_3$ holds with probability at least* 0.999.

Proof. According to Lemma 4.8, by using a union bound, we have

$$\Pr[\mathcal{E}_3 \text{ holds}] \geq 1 - |\mathcal{N}| \exp(-32d \log d) > 0.999. \qquad \square$$

We are now ready to prove the contraction bound.

Lemma 4.10. *Conditioned on $\mathcal{E}_1$, $\mathcal{E}_2$ and $\mathcal{E}_3$, for all $x \in \mathbb{R}^d$, we have*

- $\|\Pi_2 Ux\|_1 \geq \Omega(\log d)\|Ux\|_1$ *for Theorem 4.1;*
- $\|\Pi_2 Ux\|_1 \geq \Omega(\log B)\|Ux\|_1$ *for Theorem 4.2.*

Proof. By homogeneity, we can assume $\|Ux\|_1 = 1$. According to Lemma 4.7, conditioned on $\mathcal{E}_2$ and $\mathcal{E}_3$, for all $y = Ux \in \mathcal{N}$, we have $\|\Pi_2 Ux\|_1 \geq \Omega(\log d)\|Ux\|_1$ (for Theorem 4.1) or $\|\Pi_2 Ux\|_1 \geq \Omega(\log B)\|Ux\|_1$ (for Theorem 4.2). For any given $y = Ux$ where $\|y\|_1 = 1$, there exists some $\hat{y} \in \mathcal{N}$ for which $\|y - \hat{y}\|_1 \leq \varepsilon = 1/d^2$. Thus, conditioned on $\mathcal{E}_1$, notice that both $\hat{y}$ and $y - \hat{y}$ are in the column space of U, so according to Lemma 4.6, we have

$$\|\Pi y\|_1 \geq \|\Pi\hat{y}\|_1 - \|\Pi(y - \hat{y})\|_1 \geq \Omega(\log d) - (1/d^2) \cdot O(d \log d) = \Omega(\log d) \qquad \text{(for Theorem 4.1)}$$

or

$$\|\Pi y\|_1 \geq \|\Pi\hat{y}\|_1 - \|\Pi(y - \hat{y})\|_1 \geq \Omega(\log B) - (1/d^2) \cdot O(d \log d) = \Omega(\log B) \qquad \text{(for Theorem 4.2).}$$

$\square$

5 NEW SUBSPACE EMBEDDINGS FOR ℓ_p

In this section, we show how to generalize the constructions in Section 4 to ℓ_p-norms, when $1 < p < 2$.

Theorem 5.1. *Suppose $1 \leq p < 2$. For any given $A \in \mathbb{R}^{n \times d}$, let U be a $(d, 1, p)$-well-conditioned basis of A. There exists an ℓ_p oblivious subspace embedding over $O(d^2) \times n$ matrices Π where each column of a matrix drawn from Π has two non-zero entries and with probability 0.99, for any $x \in \mathbb{R}^d$,*

$$\Omega(1)\|Ux\|_p \leq \|\Pi Ux\|_p \leq O\left(d \log d\right)\|Ux\|_p.$$

Theorem 5.2. *Suppose $1 \leq p < 2$. For any given $A \in \mathbb{R}^{n \times d}$ and sufficiently large B, let U be a $(d, 1, p)$-well-conditioned basis of A. There exists an ℓ_p oblivious subspace embedding over $O(B \cdot d \log d) \times n$ matrices Π where each column of a matrix drawn from Π has $O(\log_B d)$ non-zero entries and with probability 0.99, for any $x \in \mathbb{R}^d$,*

$$\Omega(1)\|Ux\|_p \leq \|\Pi Ux\|_p \leq O\left(d \log d\right)\|Ux\|_p.$$

Our embeddings for Theorems 5.1 and 5.2 can be written as $\Pi = (\Pi_1, \Pi_2)^T$. Similar to the constructions in Section 4, for Theorem 5.1, Π_1 is sampled from the CountSketch embedding in Lemma 2.17, scaled by a $d^{2/p-1}$ factor. For Theorem 5.2, Π_1 is sampled from the OSNAP embedding in Lemma 2.18 with $O(B \cdot d \log d)$ rows and $O(\log_B d)$ non-zero entries per column and also scaled by a $d^{2/p-1}$ factor. The construction for Π_2 is almost the same as that for Theorems 4.1 and 4.2, except for replacing the standard Cauchy random variables in the diagonal entries of D with p-stable random variables. Most parts of the proof for the distortion bound resemble that for Theorems 4.1 and 4.2. We will omit similar proofs.

The following lemma can be proved in the same way as Lemmas 4.3 and 4.4, except for replacing the upper tail inequality for standard Cauchy random variables in Lemma 2.10 with that for p-stable random variables in Corollary 2.11, and replacing the properties of a $(d, 1, 1)$-well-conditioned basis with those of a $(d, 1, p)$-well-conditioned basis.

Lemma 5.3. *Let $\mathcal{E}_1$ be the event that $\|\Pi_2 U\|_p \le O(d\log d)$. $\mathcal{E}_1$ holds with probability at least 0.999. Furthermore, conditioned on $\mathcal{E}_1$, for any $x \in \mathbb{R}^d$, we have*

$$\|\Pi_2 Ux\|_p \le O\left(d\log d\right)\|Ux\|_p.$$

Let $\mathcal{E}_2$ be the event that for any $x \in \mathbb{R}^d$,

$$d^{2/p-1}\|Ux\|_2 \le \|\Pi_1 Ux\|_2 \le 2d^{2/p-1}\|Ux\|_2.$$

Since Π_1 is sampled from an ℓ_2 oblivious subspace embedding with $\kappa = 2$, and scaled by a factor of $d^{2/p-1}$, $\mathcal{E}_2$ holds with probability at least 0.999.[5]

Without loss of generality, we assume $|x_1| \ge |x_2| \ge |x_3| \ge \cdots \ge |x_n|$. Of course, this order is unknown and is not used by our embeddings.

Lemma 5.4. *Conditioned on $\mathcal{E}_2$, for any $x \in \mathbb{R}^d$, we have*

$$\|\Pi_1(Ux)_{d^2+1:n}\|_p \le O\left(d\log d\right)\|Ux\|_p.$$

Proof. By homogeneity, we can assume $\|Ux\|_p = 1$. Notice that $\|(Ux)_{d^2+1:n}\|_\infty \le d^{-2/p}$, since otherwise $\|Ux\|_p \ge \|(Ux)_{1:d^2}\|_p > 1$. By Hölder's inequality,

$$\|(Ux)_{d^2+1:n}\|_2 = \left(\sum_{i=d^2+1}^{n}(Ux)_i^2\right)^{1/2} \le \left(\sum_{i=d^2+1}^{n}|(Ux)_i|^p \cdot \max_{d^2+1\le i\le n}|Ux_i|^{2-p}\right)^{1/2} \le d^{1-2/p}.$$

Thus,

$$\begin{aligned}\|\Pi_1(Ux)_{d^2+1:n}\|_p &\le R_1^{1/p-1/2}\|\Pi_1(Ux)_{d^2+1:n}\|_2\\ &\le O(d^{2/p-1})\cdot 2d^{2/p-1}\|(Ux)_{d^2+1:n}\|_2 = O(d^{2/p-1}) = O\left(d\log d\right).\end{aligned}$$

Here the first inequality follows from Lemma 2.1 and the fact that $\Pi_1 Ux$ has R_1 rows, the second inequality holds, since $R_1 \le O(d^2)$ and $\mathcal{E}_2$ holds. □

Lemma 5.5. *Conditioned on $\mathcal{E}_2$, for any $x \in \mathbb{R}^d$, if $\|(Ux)_{1:d^2}\|_p \ge 0.5\|Ux\|_p$, then $\|\Pi_1 Ux\|_p \ge \Omega(1)\|Ux\|_p$.*

Proof. Notice that

$$\|Ux\|_2 \ge \|(Ux)_{1:d^2}\|_2 \ge d^{1-2/p}\|(Ux)_{1:d^2}\|_p \ge d^{1-2/p}/2\cdot\|Ux\|_p,$$

where the second inequality follows from Lemma 2.1 and the third inequality follows from the condition that $\|(Ux)_{1:d^2}\|_p \ge 0.5\|Ux\|_p$. Thus, $\|\Pi_1 Ux\|_p \ge \|\Pi_1 Ux\|_2 \ge \Omega(1)\|Ux\|_p$, since Π_1 is sampled from an ℓ_2 oblivious subspace embedding and scaled by a factor of $d^{2/p-1}$. □

The proof of the following lemma is almost identical to that of Lemma 4.8. We omit the proof here.

Lemma 5.6. *For any $x \in \mathbb{R}^d$, if $\|(Ux)_{d^2+1:n}\|_p \ge 0.5\|Ux\|_p$, then with probability at least $1-\exp(-32d\log d)$, we have $\|\Pi_2 Ux\|_p \ge \Omega(1)\|Ux\|_p$.*

Set $\varepsilon = 1/d^2$ and create an ε-net $\mathcal{N} \subseteq B = \{Ux \mid x \in \mathbb{R}^d \text{ and } \|Ux\|_p = 1\}$. According to Lemma 2.16, $|\mathcal{N}| \le (1+d^2)^d \le (3d^2)^d$. Let $\mathcal{E}_3$ be the event that for all $y \in \mathcal{N}$,

(1) if $\|y_{d^2+1:n}\|_p \ge 0.5\|y\|_p$, then $\|\Pi_2 y\|_p \ge \Omega(1)\|y\|_p$;
(2) $\|\Pi_1(y_{1:d^2})\|_p \le O(d\log d)\|y\|_p$.

[5]Again, we can assume $\mathcal{E}_2$ holds with probability 0.999, by increasing the dimension of the CountSketch embedding and the OSNAP embedding by a constant factor.

Lemma 5.7. *$\mathcal{E}_3$ holds with probability at least* 0.999.

Proof. Notice that Π_1 is sampled from the CountSketch embedding or the OSNAP embedding and scaled by a $d^{2/p-1}$ factor. According to Lemma 2.20, by setting ω sufficiently large, with probability $1-\exp(-32d\log d)$ we have $\|\Pi_1(y_{1:d^2})\|_p \le d^{2/p-1}(O(\log_B d))^{1/p-1/2}(\omega d\log d)^{1-1/p}\|y\|_p = O(d\log d)\|y\|_p$. Combining this with Lemma 5.6 and a union bound, we have

$$\Pr[\mathcal{E}_3 \text{ holds}] \ge 1-2|\mathcal{N}|\exp(-32d\log d) > 0.999. \qquad \square$$

Lemma 5.8. *Conditioned on $\mathcal{E}_1$, $\mathcal{E}_2$, and $\mathcal{E}_3$, for all $x\in\mathbb{R}^d$, we have*

$$\Omega(1)\|Ux\|_p \le \|\Pi Ux\|_p \le O\left(d\log d\right)\|Ux\|_p.$$

Proof. For any $x\in\mathbb{R}^d$, let $y = Ux$. By homogeneity, we can assume $\|y\|_p = 1$. As in the proof of Theorem 3.5, y can be written as

$$y = y^0 + y^1 + y^2 + \dots,$$

where for any $i\ge 0$ we have (i) $\frac{y^i}{\|y^i\|_p}\in\mathcal{N}$ and (ii) $\|y^i\|_p \le \varepsilon^i$.

It follows by Lemmas 5.3, 5.4, and 5.5 that

$$\|\Pi y\|_p \ge \|\Pi y^0\|_p - \sum_{i>0}\|\Pi y^i\|_p \ge \Omega(1) - \sum_{i>0} O\left(d\log d\right)\varepsilon^i \ge \Omega(1) - O\left(d\log d\right)\cdot 2\varepsilon \ge \Omega(1)$$

and

$$\|\Pi y\|_p \le \sum_{i\ge 0}\|\Pi y^i\|_p \le \sum_{i\ge 0} O\left(d\log d\right)\varepsilon^i \le O\left(d\log d\right). \qquad \square$$

6 SUBSPACE EMBEDDINGS WITH IMPROVED SPARSITY

In this section, we present two approaches to constructing *sparser* ℓ_p oblivious subspace embeddings for $1\le p<2$. In Section 6.1, we present our first approach based on random sampling, which yields an ℓ_p oblivious subspace embedding where each column of the embedding has at most two non-zero entries and $1+\varepsilon$ non-zero entries in expectation, where the number of rows $r = O(d^2)$. In Section 6.2, we present another approach based on the construction in Reference [24] and a truncation argument, which yields an ℓ_p oblivious subspace embedding where each column of the embedding has a single non-zero entry, at the cost of increasing the number of rows r to $\widetilde{O}(d^4)$.

6.1 Improved Sparsity Based on Random Sampling

In this section, we show how to further improve the sparsity in the constructions of Theorems 4.1 and 5.1.

Theorem 6.1. *For any given $A\in\mathbb{R}^{n\times d}$, let U be a $(d,1,1)$-well-conditioned basis of A. For any constant $0<\varepsilon<1$, there exists an ℓ_1 oblivious subspace embedding over $O(d^2)\times n$ matrices Π where each column of Π has at most two non-zero entries and $1+\varepsilon$ non-zero entries in expectation, such that with probability* 0.99, *for any $x\in\mathbb{R}^d$,*

$$\Omega(\log d)\|Ux\|_1 \le \|\Pi Ux\|_1 \le O(d\log d)\|Ux\|_1.$$

Our embedding for Theorem 6.1 is almost identical to that for Theorem 4.1 except for the Π_2 part. Recall that the Π_2 part of the construction for Theorem 4.1 can be written as ΦD, where $\Phi_{h(i),i}=1$ and all remaining entries are 0. In the new construction for Π_2, $\Phi_{h(i),i}$ are i.i.d. samples from the Bernoulli distribution $\mathrm{Ber}(\varepsilon)$. I.e., $\Phi_{h(i),i}=1$ with probability ε and 0 otherwise. All other parts of the construction are the same as in Theorem 4.1.

We note that the proof for Theorem 4.1 can still go through for the new construction. The only difference occurs when proving Lemma 4.8. In fact, the Π_2 part of the new construction for

Theorem 6.1 can be viewed as the following equivalent two-step procedure. For any given vector $y = Ux$, we first zero out each coordinate of y with probability $1-\varepsilon$, which results in a new vector $\overline{y}$, and then apply the Π_2 part of the embedding in Theorem 4.1 on the new vector $\overline{y}$.

Now, we show that with probability at least $1-\exp(-\Omega(d^2\varepsilon))$, we have

$$\|\overline{y}_{d^2+1:n}\|_1 \geq \Omega(\varepsilon)\|y_{d^2+1:n}\|_1.$$

Notice that $\mathrm{E}[|\overline{y}_i|] = \varepsilon|y_i|$ and $\mathrm{E}[\overline{y}_i^2] = \varepsilon y_i^2$, which implies

$$\mathrm{E}[\|\overline{y}_{d^2+1:n}\|_1] = \varepsilon \cdot \|y_{d^2+1:n}\|_1$$

and

$$\sum_{i>d^2} \mathrm{E}[\overline{y}_i^2] = \sum_{i>d^2} \varepsilon y_i^2 \leq \varepsilon\|y_{d^2+1:n}\|_1 \cdot \|y_{d^2+1:n}\|_\infty$$
$$= \varepsilon d^{-2}\|y_{d^2+1:n}\|_1^2.$$

Thus, by Maurer's inequality in Lemma 2.8, with probability at least $1-\exp(-\Omega(d^2\varepsilon))$, we have

$$\|\overline{y}_{d^2+1:n}\|_1 \geq \Omega(\varepsilon)\|y_{d^2+1:n}\|_1.$$

The rest of the proof is identical to the original proof for Lemma 4.8. Similarly, the same argument can also be applied to Theorem 5.1.

Theorem 6.2. *Suppose $1 \leq p < 2$. For any given $A \in \mathbb{R}^{n\times d}$, let U be a $(d, 1, p)$-well-conditioned basis for A. For any constant $0 < \varepsilon < 1$, there exists an ℓ_p oblivious subspace embedding over $O(d^2)\times n$ matrices Π where each column of Π has at most two non-zero entries and $1+\varepsilon$ non-zero entries in expectation, such that with probability 0.99, for any $x \in \mathbb{R}^d$,*

$$\Omega(1)\|Ux\|_p \leq \|\Pi Ux\|_p \leq O(d\log d)\,\|Ux\|_p.$$

The number of rows in Theorems 6.1 and 6.2 cannot be further reduced. It is shown in Reference [26] (Theorem 16) that for any distribution over $r\times n$ matrices Π such that any matrix in its support has at most one non-zero entry per column, if $\mathrm{rank}(\Pi A) = \mathrm{rank}(A)$ holds with constant probability, then $r = \Omega(d^2)$. Now, we sketch how to generalize this lower bound to distributions over $r \times n$ matrices for which each column has at most $1+\varepsilon$ non-zero entries in expectation, for any constant $0 < \varepsilon < 1$. Notice that such a lower bound already implies the number of rows of Theorems 6.1 and 6.2 are optimal up to constant factors, since any oblivious subspace embedding preserves the rank with constant probability.

For each column in the matrix Π, by Markov's inequality, with probability at least $1-\frac{1+\varepsilon}{2}$, there will be at most one non-zero entry in that column. By the Chernoff bound in Lemma 2.6, with probability $1-\exp(-\Omega(n))$, the number of columns in Π with at most one non-zero entry is $\Omega(n)$. Furthermore, the balls and bins analysis in the proof of Theorem 16 in Reference [26] can be applied to distributions over $r \times n$ matrices such that for any matrix in the support of the distribution, the number of columns with at most one non-zero entry is $\Omega(n)$. Indeed, with constant probability the rank will drop if the embedding matrix has $o(d^2)$ rows. This establishes the desired lower bound of $r = \Omega(d^2)$.

6.2 Improving Sparsity Based on Truncation

In this section, we show how to use a truncation argument to improve the construction in Reference [24].

Before formally stating the construction, we first define the truncation operation. For a given parameter $\alpha > 0$, for any $x \in \mathbb{R}$, define

$$\mathsf{trunc}_\alpha(x) = \begin{cases} \alpha & \text{if } x \in [0, \alpha] \\ -\alpha & \text{if } x \in [-\alpha, 0) \\ x & \text{otherwise} \end{cases}.$$

Here, we note some elementary properties of the truncation operation.

Lemma 6.3. *For a given parameter $\alpha > 0$, for any $x \in \mathbb{R}$, we have*

- $|\mathsf{trunc}_\alpha(x)| \geq \alpha$.
- $\mathsf{trunc}_\alpha(x) - \alpha \leq x \leq \mathsf{trunc}_\alpha(x) + \alpha$.

When applying the truncation operation to standard Cauchy random variables, the following properties are direct implications of Lemma 6.3 and the 1-stability of standard Cauchy random variables.

Corollary 6.4. *For $i \in [n]$, let $\{X_i\}$ be n independent standard Cauchy random variables. The following holds.*

- $|\mathsf{trunc}_\alpha(X_i)| \geq \alpha$.
- *For any $a = (a_1, a_2, \ldots, a_n) \in \mathbb{R}^n$,*

$$\|a\|_1 \cdot \hat{X} - \|a\|_1 \cdot \alpha \leq \sum_{i=1}^{n} a_i \cdot \mathsf{trunc}_\alpha(X_i) \leq \|a\|_1 \cdot \hat{X} + \|a\|_1 \cdot \alpha,$$

where $\hat{X}$ is a standard Cauchy random variable.

Now, we are ready to state the main result of this section.

Theorem 6.5. *There exists an ℓ_1 oblivious subspace embedding over $\widetilde{O}(d^4) \times n$ matrices Π where each column of Π has a single non-zero entry. The distortion $\kappa = \widetilde{O}(d)$.*

Our embedding for Theorem 6.5 is almost identical to the embedding for Theorem 2 in Reference [24] and the Π_2 part of the embedding for Theorems 4.1 and 4.2, except for replacing standard Cauchy random variables with truncated standard Cauchy random variables. Let $R = \widetilde{O}(d^4)$ be the number of rows of Π. Here Π can be written as $\Phi D : \mathbb{R}^n \to \mathbb{R}^R$, defined as follows:

- $h : [n] \to [R]$ is a random map so that for each $i \in [n]$ and $t \in [R]$, $h(i) = t$ with probability $1/R$.
- Φ is an $R \times n$ binary matrix with $\Phi_{h(i),i} = 1$ and all remaining entries 0.
- D is an $n \times n$ random diagonal matrix where $D_{i,i} = \mathsf{trunc}_\alpha(X_i)$. Here $\{X_i\}$ are i.i.d. samples from the standard Cauchy distribution and $\alpha < 1/4$ is a positive constant.

Now, we sketch how to modify the proof of Theorem 2 in Reference [24] to prove the distortion bound of our new embedding.

In the proof of Theorem 2 in Reference [24], the authors define five events: $\mathcal{E}_U$, $\mathcal{E}_L$, $\mathcal{E}_H$, $\mathcal{E}_C$ and $\mathcal{E}_{\hat{L}}$. Notice that for our new embedding, the event $\mathcal{E}_C$ is no longer needed, since by Corollary 6.4, the absolute values of standard Cauchy random variables are never smaller than α after truncation, where α is a small constant. We also change the number of rows of Π to $O(d^4 \log^5 d)$, and the definition of the event $\mathcal{E}_{\hat{L}}$ is changed to $\|\Pi U^{\hat{L}}\|_1 \leq O(1/(d \log^2 d))$ correspondingly.

Lemma 16 and Lemma 22 in the proof for Theorem 2 in Reference [24] show that $\mathcal{E}_U$ and $\mathcal{E}_{\hat{L}}$ hold with constant probability. The proofs for these two lemmas almost remain unchanged, except

for replacing the 1-stability of standard Cauchy random variables with the upper bound part of the "approximate 1-stability" of truncated standard Cauchy random variables in Corollary 6.4.

Lemma 13 is changed to the following: Given $\mathcal{E}_L$, for any fixed $y \in Y^L$, we have

$$\Pr\left[\|\Pi y\|_1 \leq \left(\frac{1}{4} - \alpha\right)\|y\|_1\right] \leq \exp(-\Omega(d \log d)).$$

The proof of the new version of Lemma 13 is also similar to the original proof, except for replacing the 1-stability of standard Cauchy random variables with the lower bound part of the "approximate 1-stability" of truncated standard Cauchy random variables in Corollary 6.4. This also explains why we need α to be a constant smaller than 1/4. Similarly, the constant 1/8 in Lemma 14 also needs to be modified to reflect the changes in Lemma 13.

Finally, since the absolute values of standard Cauchy random variables are never smaller than α after truncation, Lemma 15 is changed to the following: Given $\mathcal{E}_H$ and $\mathcal{E}_{\hat{L}}$, for any $y \in Y^H$, we have $\|\Pi y\|_1 \geq \Omega(\alpha)\|y\|_1$. This finishes our modification to the proof of Theorem 2 in Reference [24].

By applying the truncation argument to p-stable random variables, a similar result can be obtained for ℓ_p oblivious subspace embeddings.

Theorem 6.6. *For $1 \leq p < 2$, there exists an ℓ_p oblivious subspace embedding over $\widetilde{O}(d^4) \times n$ matrices Π where each column of Π has a single non-zero entry. The distortion $\kappa = \widetilde{O}(d)$.*

APPENDIX

A MISSING PROOFS IN SECTION 2

A.1 Proof of Lemma 2.9

Proof.

$$\mathrm{E}\left[\sum_{i=1}^{n} |a_i X_i|^p\right] = \sum_{i=1}^{n} |a_i|^p \,\mathrm{E}\left[|X_i|^p\right] = A_p \sum_{i=1}^{n} |a_i|^p,$$

where $A_p = \mathrm{E}[|X_i|^p]$ is a constant that depends only on p.

Thus, by Markov's inequality, with probability at least 0.995,

$$\left(\sum_{i=1}^{n} |a_i X_i|^p\right)^{1/p} \leq (200 A_p)^{1/p} \|a\|_p.$$

There exists a constant B_p that depends only p, such that

$$\Pr[|X_i|^p < B_p] \leq \frac{1}{400}.$$

We let Y_i be an indicator variable such that

$$Y_i = \begin{cases} 1 & \text{if } |X_i|^p < B_p \\ 0 & \text{otherwise} \end{cases}.$$

We know that $\mathrm{E}[Y_i] \leq \frac{1}{400}$, which also implies $\mathrm{E}[\sum_{i=1}^{n} |a_i|^p \cdot Y_i|] \leq \frac{1}{400}\|a\|_p^p$. Thus, by Markov's inequality, with probability at least 0.995, we have

$$\sum_{i=1}^{n} |a_i|^p \cdot Y_i \leq \frac{1}{2}\|a\|_p^p.$$

Notice that

$$\sum_{i=1}^{n} |a_i X_i|^p \geq B_p \sum_{i=1}^{n} |a_i|^p (1 - Y_i).$$

Thus, with probability at least 0.995,

$$\left(\sum_{i=1}^{n} |a_i X_i|^p\right)^{1/p} \geq \left(\frac{B_p}{2}\right)^{1/p} \|a\|_p.$$

Thus, the lemma holds by taking $C_p = \max\{(200A_p)^{1/p}, (\frac{2}{B_p})^{1/p}\}$ and using a union bound. □

A.2 Proof of Lemma 2.12

Proof. Let $\mathcal{E}_i$ be the event that $|X_i| \leq \frac{n \log n}{\log\log n}$. According to the cumulative density function of the standard Cauchy distribution, we have

$$\Pr[\mathcal{E}_i] = 1 - \frac{2}{\pi} \arctan\left(n \log n / \log\log n\right) \geq 1 - \frac{2 \log\log n}{\pi n \log n}.$$

Let $\mathcal{E} = \bigcap_{1 \leq i}^{n} \mathcal{E}_i$. By a union bound, $\mathcal{E}$ holds with probability at least $1 - \frac{2 \log\log n}{\pi \log n}$. Next, we calculate $\mathrm{E}[|X_i| \mid \mathcal{E}]$. Since the X_i are independent, by using the probability density function of the standard Cauchy distribution,

$$\mathrm{E}[|X_i| \mid \mathcal{E}] = \mathrm{E}[|X_i| \mid \mathcal{E}_i] = \frac{1}{\Pr[\mathcal{E}_i]} \frac{1}{\pi} \log\left(1 + (n \log n / \log\log n)^2\right) = O(\log n).$$

Notice that conditioned on $\mathcal{E}$, $|X_i|$ are still independent. Furthermore, conditioned on $\mathcal{E}$, for any $i \in [n]$, $|X_i| \in [0, n \log n / \log\log n]$. Thus, for sufficiently large U_1, by applying the Chernoff bound in Lemma 2.6 on $|X_i| \log\log n (n \log n)^{-1}$,

$$\Pr\left[\sum_{i=1}^{n} |X_i| > U_1 n \log n \mid \mathcal{E}\right] \leq 2^{-\frac{U_1 n \log n}{n \log n / \log\log n}} = 2^{-U_1 \log\log n}.$$

Thus, for sufficiently large U_1,

$$\Pr\left[\sum_{i=1}^{n} |X_i| \leq U_1 n \log n\right] \geq \Pr\left[\sum_{i=1}^{n} |X_i| \leq U_1 n \log n \mid \mathcal{E}\right] \cdot \Pr[\mathcal{E}]$$
$$\geq \left(1 - 2^{-U_1 \log\log n}\right) \cdot \left(1 - \frac{2 \log\log n}{\pi \log n}\right) \geq 1 - \frac{\log\log n}{\log n}.$$ □

A.3 Proof of Lemma 2.14

Proof. According to Lemma 2.4, there exists a constant $t_p \geq 1$ that depends only on p, such that for any $t \geq t_p$,

$$\Pr[X_i > t] \geq \frac{c_p}{2} t^{-p}.$$

Thus, for $t \geq t_p^p$,

$$\Pr[|X_i|^p > t] = \Pr[|X_i| > t^{1/p}] = 2 \Pr[X_i > t^{1/p}] \geq c_p t^{-1}.$$

For $i \geq 0$ and $j \in [n]$, we let N_j^i denote the indicator variable such that

$$N_j^i = \begin{cases} 1 & \text{if } |X_j|^p > 2^i t_p^p \\ 0 & \text{otherwise} \end{cases},$$

and $N^i = \sum_{j=1}^{n} N_j^i$. We have that $\mathrm{E}[N_j^i] \geq 2^{-i} c_p t_p^{-p}$ and thus $\mathrm{E}[N^i] \geq n \cdot 2^{-i} c_p t_p^{-p}$. According to the Chernoff bound in Lemma 2.6, we have $\Pr[N^i \geq n 2^{-i-1} c_p t_p^{-p}] \geq 1 - \exp(-n \cdot 2^{-i-3} c_p t_p^{-p})$. Let

l_{max} be the largest i such that

$$\exp\left(-n2^{-i-3}c_p t_p^{-p}\right) \leq \frac{1}{2T}. \tag{5}$$

By a union bound, with probability at least

$$1 - \sum_{i=0}^{l_{max}} \exp\left(-n2^{-i-3}c_p t_p^{-p}\right) \geq 1 - 1/T,$$

simultaneously for all $0 \leq i \leq l_{max}$, $N^i \geq n2^{-i-1}c_p t_p^{-p}$, which implies

$$\sum_{i=1}^{n} |X_i|^p \geq \sum_{i=0}^{l_{max}} 2^i t_p^p \cdot N^i/2 \geq c_p/4 \cdot l_{max} \cdot n.$$

Solving Equality (5) and substituting the value of l_{max}, for sufficiently large T and n, with probability at least $1 - 1/T$,

$$\sum_{i=1}^{n} |X_i|^p \geq L_p n \log\left(\frac{n}{\log T}\right),$$

where L_p is a constant that depends only on p. □

A.4 Proof of Lemma 2.20

Proof. Suppose Π has R rows and s non-zero entries per column. For the CountSketch embedding, we have $R = O(d^2)$ and $s = 1$, while for the OSNAP embedding, we have $R = O(B \cdot d \log d)$ and $s = O(\log_B d)$. In either case, we have $R \leq O(d^2)$.

For $i \in [R]$, define $B_i = \{j \mid j \leq d^2 \text{ and } \Pi_{i,j} \neq 0\}$. According to the Chernoff bound in Lemma 2.6, with probability at least $1 - \exp(-\Omega(\omega d \log d))$, $|B_i| \leq \omega d \log d$. It follows by a union bound that with probability at least $1 - \exp(-\Omega(\omega d \log d)) \cdot R = 1 - \exp(-\Omega(\omega d \log d))$, simultaneously for all $i \in [R]$, we have $|B_i| \leq \omega d \log d$. We condition on this event in the rest of the proof.

Notice that

$$|(\Pi\,(y_{1:d^2}))_i|^p \leq \left(s^{-1/2} \sum_{j \in B_i} |y_j|\right)^p$$

$$\leq \left(s^{-1/2}(\omega d \log d)^{1-1/p} \left(\sum_{j \in B_i} |y_j|^p\right)^{1/p}\right)^p = s^{-p/2}(\omega d \log d)^{p-1} \left(\sum_{j \in B_i} |y_j|^p\right).$$

Here the second inequality follows from Lemma 2.1 and $|B_i| \leq \omega d \log d$. For each $j \in [d^2]$, the number of $i \in [R]$ for which $j \in B_i$ is exactly s, which implies

$$\sum_{i=1}^{R} |(\Pi\,(y_{1:d^2}))_i|^p \leq s^{-p/2}(\omega d \log d)^{p-1} \sum_{i=1}^{R} \sum_{j \in B_i} |y_j|^p = s^{1-p/2}(\omega d \log d)^{p-1} \sum_{j=1}^{d^2} |y_j|^p.$$

Thus,

$$\|\Pi\,(y_{1:d^2})\,\|_p \leq s^{1/p-1/2}(\omega d \log d)^{1-1/p} \|y_{1:d^2}\|_p \leq s^{1/p-1/2}(\omega d \log d)^{1-1/p} \|y\|_p.$$ □

ACKNOWLEDGMENTS

The authors thank Lin F. Yang, Hanrui Zhang, and Peilin Zhong for helpful discussions, to the anonymous reviewers for helpful comments, and to Xiangrui Meng for discussing the details of References [23, 24].

REFERENCES

[1] N. Ailon and B. Chazelle. 2006. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *Proceedings of the STOC*. 557–563.

[2] A. Andoni. 2017. High frequency moments via max-stability. In *Proceedings of the ICASSP*. 6364–6368.

[3] A. Andoni, K. Do Ba, P. Indyk, and D. P. Woodruff. 2009. Efficient sketches for earth-mover distance, with applications. In *Proceedings of the FOCS*. 324–330.

[4] J. Bourgain, S. Dirksen, and J. Nelson. 2015. Toward a unified theory of sparse dimensionality reduction in euclidean space. *Geom. Funct. Anal.* 25, 4 (2015), 1009–1088.

[5] B. Brinkman and M. Charikar. 2005. On the impossibility of dimension reduction in ℓ_1. *J. ACM* 52, 5 (2005), 766–788.

[6] M. Charikar, K. Chen, and M. Farach-Colton. 2004. Finding frequent items in data streams. *Theor. Comput. Sci.* 312, 1 (2004), 3–15.

[7] M. Charikar and A. Sahai. 2002. Dimension reduction in the ℓ_1 norm. In *Proceedings of the FOCS*. 551–560.

[8] K. L. Clarkson, P. Drineas, M. Magdon-Ismail, M. W. Mahoney, X. Meng, and D. P. Woodruff. 2016. The fast cauchy transform and faster robust linear regression. *SIAM J. Comput.* 45, 3 (2016), 763–810.

[9] K. L. Clarkson and D. P. Woodruff. 2009. Numerical linear algebra in the streaming model. In *Proceedings of the STOC*. 205–214.

[10] K. L. Clarkson and D. P. Woodruff. 2017. Low-rank approximation and regression in input sparsity time. *J. ACM* 63, 6 (2017), 54.

[11] M. B. Cohen. 2016. Nearly tight oblivious subspace embeddings by trace inequalities. In *Proceedings of the SODA*. 278–287.

[12] M. B. Cohen, Y. T. Lee, C. Musco, C. Musco, R. Peng, and A. Sidford. 2015. Uniform sampling for matrix approximation. In *Proceedings of the ITCS*. 181–190.

[13] M. B. Cohen and R. Peng. 2015. ℓ_p row sampling by lewis weights. In *Proceedings of the STOC*. 183–192.

[14] A. Dasgupta, P. Drineas, B. Harb, R. Kumar, and M. W. Mahoney. 2009. Sampling algorithms and coresets for ℓ_p regression. *SIAM J. Comput.* 38, 5 (2009), 2060–2078.

[15] D. Dubhashi and D. Ranjan. 1996. Balls and bins: A study in negative dependence. *BRICS Report Series* 3, 25 (1996).

[16] J. R. Lee and A. Naor. 2004. Embedding the diamond graph in L_p and dimension reduction in L_1. *Geom. Funct. Anal.* 14, 4 (2004), 745–747.

[17] M. Li, G. L. Miller, and R. Peng. 2013. Iterative row sampling. In *Proceedings of the FOCS*. 127–136.

[18] Y. Li, D. P. Woodruff, and T. Yasuda. 2021. Exponentially improved dimensionality reduction for ℓ_1: Subspace embeddings and independence testing. Retrieved from https://arXiv:2104.12946.

[19] M. Magdon-Ismail and A. Gittens. 2019. Fast fixed dimension L2-subspace embeddings of arbitrary accuracy, with application to L1 and L2 tasks. Retrieved from https://arXiv:1909.12580.

[20] M. W. Mahoney. 2011. Randomized algorithms for matrices and data. *Found. Trends Mach. Learn.* 3, 2 (2011), 123–224.

[21] A. Maurer. 2003. A bound on the deviation probability for sums of non-negative random variables. *J. Inequal. Pure Appl. Math.* 4, 1 (2003), 15.

[22] X. Meng. 2019. Private Communication.

[23] X. Meng and M. W. Mahoney. 2012. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. Retrieved from https://arXiv:1210.3135.

[24] X. Meng and M. W. Mahoney. 2013. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Proceedings of the STOC*. 91–100.

[25] J. Nelson and H. L. Nguyễn. 2013. OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *Proceedings of the FOCS*. 117–126.

[26] J. Nelson and H. L. Nguyễn. 2013. Sparsity lower bounds for dimensionality reducing maps. In *Proceedings of the STOC*. 101–110.

[27] J. Nelson and H. L. Nguyễn. 2014. Lower bounds for oblivious subspace embeddings. In *Proceedings of the ICALP*. 883–894.

[28] J. P. Nolan. 2020. *Univariate Stable Distributions*. Springer.

[29] T. Sárlos. 2006. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the FOCS*. 143–152.

[30] C. Sohler and D. P. Woodruff. 2011. Subspace embeddings for the L1-norm with applications. In *Proceedings of the STOC*. 755–764.

[31] Z. Song, D. P. Woodruff, and P. Zhong. 2017. Low-rank approximation with entrywise ℓ_1-norm error. In *Proceedings of the STOC*. 688–701.

[32] R. Wang and D. P. Woodruff. 2019. Tight bounds for ℓ_p oblivious subspace embeddings. In *Proceedings of the SODA*. 1825–1843.

[33] P. Wojtaszczyk. 1991. *Banach Spaces for Analysts*. Cambridge University Press.

[34] D. P. Woodruff. 2014. Sketching as a tool for numerical linear algebra. *Found. Trends Theor. Comput. Sci.* 10, 1–2 (2014), 1–157.

[35] D. P. Woodruff and Q. Zhang. 2013. Subspace embeddings and ℓ_p-regression using exponential random variables. In *Proceedings of the COLT*. 546–567.

[36] J. Yang, X. Meng, and M. W. Mahoney. 2014. Quantile regression for large-scale applications. *SIAM J. Sci. Comput.* 36, 5 (2014), S78–S110.

[37] A. C. Yao. 1977. Probabilistic computations: Toward a unified measure of complexity. In *Proceedings of the FOCS*. 222–227.

Received March 2019; revised May 2021; accepted July 2021