SIAM J. COMPUT.
Vol. 50, No. 4, pp. 1287–1335

TIGHT BOUNDS FOR THE SUBSPACE SKETCH PROBLEM WITH APPLICATIONS*

YI LI†, RUOSONG WANG‡, AND DAVID P. WOODRUFF‡

Abstract. In the *subspace sketch problem* one is given an $n \times d$ matrix A with $O(\log(nd))$ bit entries, and would like to compress it in an arbitrary way to build a small space data structure Q_p, so that for any given $x \in \mathbb{R}^d$, with probability at least 2/3, one has $Q_p(x) = (1 \pm \varepsilon)\|Ax\|_p$, where $p \geq 0$ and the randomness is over the construction of Q_p. The central question is, *how many bits are necessary to store Q_p?* This problem has applications to the communication of approximating the number of nonzeros in a matrix product, the size of coresets in projective clustering, the memory of streaming algorithms for regression in the row-update model, and embedding subspaces of L_p in functional analysis. A major open question is the dependence on the approximation factor ε. We show if $p \geq 0$ is *not a positive even integer* and $d = \Omega(\log(1/\varepsilon))$, then $\Omega(\varepsilon^{-2}d)$ bits are necessary. On the other hand, if p is a positive even integer, then there is an upper bound of $O(d^p \log(nd))$ bits independent of ε. Our results are optimal up to logarithmic factors. As corollaries of our main lower bound, we obtain new lower bounds for a wide range of applications, including the above, which in many cases are optimal.

Key words. subspace sketch, lower bound, sketching, M-estimator, projective clustering

AMS subject classifications. 68Q17, 68Q87, 46B07, 46B85

DOI. 10.1137/20M1311831

1. Introduction.

The explosive growth of available data has necessitated new models for processing such data. A particularly powerful tool for analyzing such data is *sketching*, which has found applications to communication complexity, data stream algorithms, functional analysis, machine learning, numerical linear algebra, sparse recovery, and many other areas. Here one is given a large object, such as a graph, a matrix, or a vector, and one seeks to compress it while still preserving useful information about the object. One of the main goals of a sketch is to use as little memory as possible in order to compute functions of interest. Typically, to obtain nontrivial space bounds, such sketches need to be both randomized and approximate. By now there are nearly optimal bounds on the memory required of sketching many fundamental problems, such as graph sparsification, norms of vectors, and problems in linear algebra such as low-rank approximation and regression. We refer the reader to the surveys [32, 44] as well as the compilation of lecture notes [3].

In this paper we consider the *subspace sketch problem.*

*Received by the editors March 16, 2020; accepted for publication (in revised form) April 19, 2021; published electronically August 5, 2021. A preliminary version of this paper was presented at the *Thirty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, 2020, Salt Lake City, Utah, SIAM, Philadelphia, 2020, pp. 1655–1674.

https://doi.org/10.1137/20M1311831

Funding: The first author was supported in part by Singapore Ministry of Education (AcRF) Tier 2 grant MOE2018-T2-1-013. The second and third authors were supported in part by an Office of Naval Research (ONR) grant N00014-18-1-2562 as well as the Simons Institute for the Theory of Computing where part of this work was done.

†Divison of Mathematical Sciences, Nanyang Technological University, Singapore, 637371 (yili@ntu.edu.sg).

‡Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15213 USA (ruosongw@andrew.cmu.edu, dwoodruf@cs.cmu.edu).

DEFINITION 1.1. *Given an $n\times d$ matrix A with entries specified by $O(\log(nd))$ bits, an accuracy parameter $\varepsilon > 0$, and a function $\Phi : \mathbb{R}^n \to \mathbb{R}^{\geq 0}$, design a data structure Q_Φ so that for any $x \in \mathbb{R}^d$, with probability at least 0.9, $Q_\Phi(x) = (1 \pm \varepsilon)\Phi(Ax)$.*

The subspace sketch problem captures many important problems as special cases. We will show how to use this problem to bound the communication of approximating statistics of a matrix product, the size of coresets in projective clustering, the memory of streaming algorithms for regression in the row-update model, and the embedding dimension in functional analysis. We will describe these applications in more detail below.

The goal in this work is to determine the memory, i.e., the size of Q_Φ, required for solving the subspace sketch problem for different functions Φ. We first consider the classical ℓ_p-norms $\Phi(x) = \sum_{i=1}^n |x_i|^p$, in which case the problem is referred to as the ℓ_p subspace sketch problem.[1] We later extend our techniques to their robust counterparts $\Phi(x) = \sum_{i=1}^n \phi(x_i)$, where $\phi(t) = |t|^p$ if $|t| \leq \tau$ and $\phi(t) = \tau^p$ otherwise. Here Φ is a so-called M-estimator and known as the Tukey loss p-norm. It is less sensitive to "outliers" since it truncates large coordinate values at τ. We let Q_p denote Q_Φ when $\Phi(x) = \sum_i |x_i|^p$, and use $Q_{p,\tau}$ when Φ is the Tukey loss p-norm.

It is known that for $p \in (0, 2]$ and $r = O(\varepsilon^{-2})$, if one chooses a matrix $S \in \mathbb{R}^{r\times n}$ of independent and identically distributed (i.i.d.) p-stable random variables, then for any fixed $y \in \mathbb{R}^n$, from the sketch $S \cdot y$ one can output a number z for which $(1 - \varepsilon)\|y\|_p \leq z \leq (1 + \varepsilon)\|y\|_p$ with probability at least 0.9 [20]. We say z is a $(1 \pm \varepsilon)$-approximation of $\|y\|_p$. For $p = 1$, the output is just $\text{med}(Sy)$, where $\text{med}(\cdot)$ denotes the median of the absolute values of the coordinates in a vector. A sketch S with $r = O(\varepsilon^{-2} \log n)$ rows is also known for $p = 0$ [24]. For $p > 2$, there is a distribution on $S \in \mathbb{R}^{r\times n}$ with $r = O(\varepsilon^{-2} n^{1-2/p} \log n)$ for which one can output a $(1 \pm \varepsilon)$-approximation of $\|y\|_p$ given Sy with probability at least 0.9 [18]. By appropriately discretizing the entries, one can solve the ℓ_p subspace sketch problem by storing SA for an appropriate sketching matrix S, and estimating $\|Ax\|_p$ using SAx. In this way, one obtains a sketch of size $\widetilde{O}(\varepsilon^{-2}d)$ bits[2] for $p \in [0, 2]$, and a sketch of size $\widetilde{O}(\varepsilon^{-2} n^{1-2/p} \cdot d)$ bits for $p > 2$. Note, however, that this is only one particular approach, based on choosing a random matrix S, and better approaches may be possible. Indeed, note that for $p = 2$, one can simply store $A^T A$ and output $Q_2(x) = x^T A^T Ax$. This is exact (i.e., holds for $\varepsilon = 0$) and only uses $O(d^2 \log(nd))$ bits of space, which is significantly smaller than $\widetilde{O}(\varepsilon^{-2}d)$ for small enough ε. We note that the ε^{-2} term may be extremely prohibitive in applications. For example, if one wants high accuracy such as $\varepsilon = 0.1\%$, the ε^{-2} factor is a severe drawback of existing algorithms.

A natural question is what makes it possible for $p = 2$ to obtain $\widetilde{O}(d^2)$ bits of space, and whether it is also possible to achieve $\widetilde{O}(d^2)$ space for $p = 1$. One thing that makes this possible for $p = 2$ is the singular value decomposition (SVD), namely, that $A = U\Sigma V^T$ for matrices $U \in \mathbb{R}^{n\times d}$ and $V \in \mathbb{R}^{d\times d}$ with orthonormal columns, and Σ is a nonnegative diagonal matrix. Then $\|Ax\|_2^2 = \|\Sigma V^T x\|_2^2$ since U has orthonormal columns. Consequently, once Σ and V are obtained, one can discard A and recover $\|Ax\|_2^2$ from the d inner products $\langle \Sigma_{1,1} v_1, x\rangle, \ldots, \langle \Sigma_{d,d} v_d, x\rangle$, where the v_i's are the

[1] Note we are technically considering the pth power of the ℓ_p-norms, but for the purposes of $(1+\varepsilon)$-approximation, they are the same for constant p. Also, when $p < 1$, ℓ_p is not a norm, though it is still a well-defined quantity. Finally, ℓ_0 denotes the number of nonzero entries of x.

[2] Throughout we use $\widetilde{O}, \widetilde{\Omega}$, and $\widetilde{\Theta}$ to hide factors that are polynomial in $\log(nd/\varepsilon)$. We note that our lower bounds are actually independent of n.

TABLE 1.1

Summary of results. The lower bound column suppresses $\widetilde{\Omega}$-notation. Except for the last row, all ℓ_p subspace sketch problems in this table refer to the for-each version as defined in Definition 1.1. *For the subspace embedding problem, the lower bound is on the dimension of the target space. For the last two rows, the lower bound is on the length of the sketch. In all problems it is assumed that n is sufficiently large.*

Problem	Lower bound	Theorem(s)	Notes
ℓ_p subspace sketch	$\varepsilon^{-2}d$ bits	Theorem 1.2	$d=\Omega(\log(1/\varepsilon))$, $p\in[0,\infty)\setminus 2\mathbb{Z}^+$
	$d^{p/2}$ bits	Theorem 1.4	$p\geq 2$
M-estimator sketch	$\varepsilon^{-2}d$ bits	Theorem 1.3	$d=\Omega(\log(1/\varepsilon))$
Projective clustering	$\varepsilon^{-2}kj$ bits	Theorem 1.5	$j=\Omega(\log(k/\varepsilon))$
Streaming coreset for linear regression	$\varepsilon^{-2}d$ bits	Corollary 1.6	$d=\Omega(\log(1/\varepsilon))$
Subspace embedding	ε^{-2}	Corollary 1.7	$d=\Omega(\log(1/\varepsilon))$, $p\in[1,\infty)\setminus 2\mathbb{Z}^+$
	$d^{\max\{p/2,1\}}$	Paragraph below Theorem 1.8	$p\geq 1$
ℓ_p subspace sketch via sampling matrices	$\varepsilon^{-2}d$	Corollary 1.9	$d=\Omega(\log(1/\varepsilon))$, $p\in[1,\infty)\setminus 2\mathbb{Z}^+$
ℓ_p subspace sketch (for-all) via oblivious sketches	$\varepsilon^{-2}d$	Theorem 1.10	$p\in[1,2]$

rows of V^T. Thus one can "compress" A to d "directions" $\Sigma_{i,i}v_i$. A natural question is whether for $p=1$ it is also possible to find $O(d)$ directions $v_1,\ldots,v_{O(d)}$ such that for any x, $\|Ax\|_1$ can be well-approximated from some function of $O(d)$ inner products $\langle v_1,x\rangle,\ldots,\langle v_{O(d)},x\rangle$. Here we need the function to be of low space, which, together with $v_1,\ldots,v_{O(d)}$, forms a "compressed" version of A for calculating $\|Ax\|_1$. Indeed, this would be the analogue of the SVD for $p=1$, for which little is known.

The central question of our work is, *how much memory is needed to solve the subspace sketch problem as a function of Φ?*

1.1. Our contributions. A summary of the results is given in Table 1.1. Up to polylogarithmic factors, we resolve the above question for ℓ_p-norms and Tukey loss p-norms for any $p\in[0,2)$. For $p\geq 2$ we also obtain a surprising separation for even integers p from other values of p.

Our main theorem is the following. We denote by $\mathbb{Z}^+$ the set of positive integers.

THEOREM 1.2 (informal version of Corollaries 3.13 and 3.14). *Let $p\in[0,\infty)\setminus 2\mathbb{Z}^+$ be a constant. For any $d=\Omega(\log(1/\varepsilon))$ and $n=\widetilde{\Omega}(\varepsilon^{-2}d)$, we have that $\widetilde{\Omega}(\varepsilon^{-2}d)$ bits are necessary to solve the ℓ_p subspace sketch problem.*

When $p\in 2\mathbb{Z}^+$, there is an upper bound of $O(d^p\log(nd))$ bits, independent of ε (see Remark 3.15). This gives a surprising separation between positive even integers and other values of p; in particular for positive even integers p it is possible to obtain $\varepsilon=0$ with at most $O(d^p\log(nd))$ bits of space, whereas for other values of p the space becomes arbitrarily large as $\varepsilon\to 0$. This also shows it is not possible, for $p=1$ for example, to find $O(d)$ representative directions for $\varepsilon=0$ analogously to the SVD for $p=2$. Note that the lower bound in Theorem 1.2 is much stronger than this, showing that there is no data structure whatsoever which uses fewer than $\widetilde{\Omega}(\varepsilon^{-2}\cdot d)$ bits, and so as ε gets smaller, the space complexity becomes arbitrarily large.

In addition to the ℓ_p-norm, in the subspace sketch problem we also consider a more general entry-decomposable Φ, that is, $\Phi(v)=\sum_i\phi(v_i)$ for $v\in\mathbb{R}^n$ and some

$\phi : \mathbb{R} \to \mathbb{R}^{\geq 0}$. We show the same $\widetilde{\Omega}(\varepsilon^{-2}d)$ lower bounds for a number of M-estimators ϕ.

THEOREM 1.3 (informal version of Corollaries 8.2 and 8.4). *The subspace sketch problem requires $\widetilde{\Omega}(\varepsilon^{-2}d)$ bits for the following functions ϕ when $d = \Omega(\log(1/\varepsilon))$ and $n = \widetilde{\Omega}(\varepsilon^{-2}d)$:*

- *(L_1-L_2 estimator) $\phi(t) = 2(\sqrt{1 + t^2/2} - 1)$;*
- *(Huber estimator) $\phi(t) = t^2/(2\tau) \cdot \mathbf{1}_{\{|t| \leq \tau\}} + (|t| - \tau/2) \cdot \mathbf{1}_{\{|t| > \tau\}}$;*
- *(Fair estimator) $\phi(t) = \tau^2(|t|/\tau - \ln(1 + |t|/\tau))$;*
- *(Cauchy estimator) $\phi(t) = (\tau^2/2)\ln(1 + (t/\tau)^2)$;*
- *(Tukey loss p-norm) $\phi(t) = |t|^p \cdot \mathbf{1}_{\{|t| \leq \tau\}} + \tau^p \cdot \mathbf{1}_{\{|t| > \tau\}}$.*

We also consider the mollified version of the Tukey loss functions ($0 < p < 2$), for which the lower bound of $\widetilde{\Omega}(\varepsilon^{-2}d)$ bits still holds. Furthermore, this lower bound is tight up to logarithmic factors, since we design a new algorithm which approximates $\Phi(x)$ using $\widetilde{O}(\varepsilon^{-2})$ bits, which implies an upper bound of $\widetilde{O}(\varepsilon^{-2}d)$ for the subspace sketch problem. See section 10 for details.

While Theorem 1.2 gives a tight lower bound for $p \in [0, 2)$, matching the simple sketching upper bounds described earlier, and also gives a separation from the $O(d^p \log(nd))$ bit bound for even integers $p \geq 2$, one may ask what exactly the space required is for even integers $p \geq 2$ and arbitrarily small ε. For $p = 2$, the $O(d^2 \log(nd))$ upper bound is tight up to logarithmic factors since the previous work [5, Theorem 2.2] implies an $\widetilde{\Omega}(d^2)$ lower bound once $\varepsilon = O(1/\sqrt{d})$. For $p > 2$, we show the following: for a constant $\varepsilon \in (0, 1)$, there is an upper bound of $\widetilde{O}(d^{p/2})$ bits (see Remark 4.4), which is nearly tight in light of the following lower bound, which holds for constant ε.

THEOREM 1.4 (informal version of Theorem 4.8). *Let $p \geq 2$ and $\varepsilon \in (0, 1)$ be constants. Suppose that $n = \widetilde{\Omega}(d^{p/2})$, then $\widetilde{\Omega}(d^{p/2})$ bits are necessary to solve the ℓ_p subspace sketch problem.*

Note that Theorem 1.4 holds even if p is not an even integer, and shows that a lower bound of $d^{\Omega(p)}$ holds for every $p \geq 2$.

We next turn to concrete applications of Theorems 1.2 and 1.3.

Statistics of a matrix product. In [45], an algorithm was given for estimating $\|A \cdot B\|_p$ for integer matrices A and B with $O(\log n)$ bit integer entries (see Algorithm 1 in [45] for the general algorithm). When $p = 0$, this estimates the number of nonzero entries of $A \cdot B$, which may be useful since there are faster algorithms for matrix product when the output is sparse; see [34] and the references therein. More generally, norms of the product $A \cdot B$ can be used to determine how correlated the rows of A are with the columns of B. The bit complexity of this problem was studied in [42, 45]. In [42] a lower bound of $\Omega(\varepsilon^{-2}n)$ bits was shown for estimating $\|AB\|_0$ for $n \times n$ matrices A, B up to a $(1 + \varepsilon)$ factor, assuming $n \geq 1/\varepsilon^2$ (this lower bound holds already for binary matrices A and B). This lower bound implies an ℓ_0-subspace sketch lower bound of $\Omega(\varepsilon^{-2}d)$ *assuming that* $d \geq 1/\varepsilon^2$. Our lower bound in Theorem 1.2 considerably strengthens this result by showing the same lower bound (up to polylog(d/ε) factors) for a much smaller value of $d = \Omega(\log(1/\varepsilon))$. For any $p \in [0, 2]$, there is a matching upper bound up to polylogarithmic factors (such an upper bound is given implicitly in the description of Algorithm 1 of [45], where the ε there is instantiated with $\sqrt{\varepsilon}$, and also follows from the random sketching matrices S discussed above).

Projective clustering. In the task of projective clustering, we are given a set $X \subset \mathbb{R}^d$ of n points, a positive integer k, and a nonnegative integer $j \leq d$. A center $\mathcal{C}$ is a k-tuple $(V_1, V_2, \ldots, V_k)$, where each V_i is a j-dimensional affine subspace in $\mathbb{R}^d$. Given a function $\phi : \mathbb{R} \to \mathbb{R}^{\geq 0}$, the objective is to find a center $\mathcal{C}$ that minimizes the projective cost, defined to be

$$\operatorname{cost}(X, \mathcal{C}) = \sum_{x \in X} \phi(\operatorname{dist}(x, \mathcal{C})),$$

where $\operatorname{dist}(x, \mathcal{C}) = \min_i \operatorname{dist}(x, V_i)$, the Euclidean distance from a point p to its nearest subspace V_i in $\mathcal{C} = (V_1, V_2, \ldots, V_k)$. The coreset problem for projective clustering asks to design a data structure Q_ϕ such that for any center $\mathcal{C}$, with probability at least 0.9, $Q_\phi(\mathcal{C}) = (1 \pm \varepsilon) \operatorname{cost}(X, \mathcal{C})$. Note that in this and other computational geometry problems, the dimension d may be small (e.g., $d = \log(1/\varepsilon)$), though one may want a high accuracy solution. Although possibly far from optimal, surprisingly our lower bound below is the first nontrivial lower bound on the size of coresets for projective clustering.

THEOREM 1.5 (informal version of Corollary 9.4). *Suppose that $\phi(t) = |t|^p$ for $p \in [0, \infty) \setminus 2\mathbb{Z}^+$ or ϕ is one of the functions in Theorem 1.3. For $k \geq 1$ and $j = \Omega(\log(k/\varepsilon))$, any coreset for projective clustering requires $\widetilde{\Omega}(\varepsilon^{-2}kj)$ bits.*

Linear regression. In the linear regression problem, there is an $n \times d$ data matrix A and a vector $b \in \mathbb{R}^n$. The goal is to find a vector $x \in \mathbb{R}^d$ so as to minimize $\Phi(Ax - b)$, where $\Phi(v) = \sum_i \phi(v_i)$ for $v \in \mathbb{R}^n$ and some $\phi : \mathbb{R} \to \mathbb{R}^{\geq 0}$. Here we consider streaming coresets for linear regression in the row-update model. In the row-update model, the streaming coreset is updated online during one pass over the n rows of $\begin{pmatrix} A & b \end{pmatrix}$, and outputs a $(1 \pm \varepsilon)$-approximation to the optimal value $\min_x \Phi(Ax - b)$ at the end. By a simple reduction, our lower bound for the subspace sketch problem implies lower bounds on the size of streaming coresets for linear regression in the row-update model. To see this, we note that by taking sufficiently large λ,

$$\min_y (\Phi(Ay) + \lambda \Phi(x - y)) = \Phi(Ax).$$

Thus, a streaming coreset for linear regression can solve the subspace sketch problem, which we formalize in the following corollary.

COROLLARY 1.6. *Suppose that $\phi(t) = |t|^p$ for $p \in [0, \infty) \setminus 2\mathbb{Z}^+$ or ϕ is one of the functions in Theorem 1.3. Any streaming coreset for linear regression in the row-update model requires $\widetilde{\Omega}(\varepsilon^{-2}d)$ bits when $d = \Omega(\log(1/\varepsilon))$.*

Subspace embeddings. Let $p \geq 1$ and n be sufficiently large. Given $A \in \mathbb{R}^{n \times d}$, the ℓ_p subspace embedding problem asks to find a linear map $T : \mathbb{R}^n \to \mathbb{R}^r$ such that for all $x \in \mathbb{R}^d$,

$$(1 - \varepsilon)\|Ax\|_p \leq \|TAx\|_p \leq (1 + \varepsilon)\|Ax\|_p. \tag{1.1}$$

The smallest r which admits a T for every A is denoted by $N_p(d, \varepsilon)$, which is of interest in functional analysis. When T is allowed to be random, we require (1.1) to hold with probability at least 0.9. This problem can be seen as a special case of the "for-all" version of the subspace sketch problem in Definition 1.1. In the for-all version of the subspace sketch problem, the data structure Q_p is required to, with probability at least 0.9, satisfy $Q_p(x) = (1 \pm \varepsilon)\|Ax\|_p$ simultaneously for all $x \in \mathbb{R}^d$. In this case, the same lower bound of $\widetilde{\Omega}(\varepsilon^{-2}d)$ bits holds for $p \in [1, \infty) \setminus 2\mathbb{Z}$.

Since the data structure can store T if it exists, we can turn our bit lower bound into a dimension lower bound on $N_p(d,\varepsilon)$. Doing so will incur a loss of an $\widetilde{O}(d)$ factor (Theorem 5.1). We give an $\widetilde{\Omega}(\varepsilon^{-2})$ lower bound, which is the first such lower bound giving a dependence on ε for general p.

COROLLARY 1.7. *Suppose that $p \in [1,\infty)\setminus 2\mathbb{Z}$ and $d = \Omega(\log(1/\varepsilon))$. It holds that $N_p(d,\varepsilon) = \widetilde{\Omega}(\varepsilon^{-2})$.*

The dependence on ε in this lower bound is tight, up to polylog$(1/\varepsilon)$ factors, for all values of $p \in [1,\infty)\setminus 2\mathbb{Z}$ [37]. When $p \in 2\mathbb{Z}$, no lower bound with a dependence on ε should exist, since a d-dimensional subspace of ℓ_p^n always embeds into ℓ_p^r isometrically with $r = \binom{d+p-1}{p} - 1$ [25]. See more discussion below in section 1.2 on functional analysis. We also prove a bit complexity lower bound for the aforementioned for-all version of the subspace sketch problem. We refer the reader to section 4.2 for details.

THEOREM 1.8. *Let $p \geq 1$ be a constant. Suppose that $\varepsilon > 0$ is a constant. The for-all version of the subspace sketch problem requires $\Omega(d^{\max\{p/2,1\}+1})$ bits.*

This lower bound immediately implies a dimension lower bound of $N_p(d,\varepsilon) = \widetilde{\Omega}(d^{\max\{p/2,1\}})$ for the subspace embedding problem for constant ε, recovering existing lower bounds (up to logarithmic factors), which are known to be tight.

Sampling by Lewis weights. While it is immediate that $N_p(d,\varepsilon) \geq d$, our lower bound above thus far has not precluded the possibility that $N_p(d,\varepsilon) = \widetilde{O}(d + 1/\varepsilon^2)$. However, the next corollary, which lower bounds the target dimension for sampling-based embeddings, indicates this is impossible to achieve using a prevailing existing technique.

COROLLARY 1.9. *Let $p \geq 1$ and $p \notin 2\mathbb{Z}$. Suppose that $Q_p(x) = \|TAx\|_p^p$ solves the ℓ_p subspace sketch problem for some $T \in \mathbb{R}^{r\times n}$ for which each row of T contains exactly one nonzero element. Then $r = \widetilde{\Omega}(\varepsilon^{-2}d)$, provided that $d = \Omega(\log(1/\varepsilon))$ and $n = \widetilde{\Omega}(\varepsilon^{-2}d)$.*

The same lower bound holds for the for-all version of the ℓ_p subspace sketch problem. As a consequence, since the upper bounds of $N_p(d,\varepsilon)$ in (1.2) for $1 \leq p < 2$ are based on subsampling with the "change of density" technique (also known as sampling by Lewis weights [15]), they are, within the framework of this classical technique, best possible up to polylog(d/ε) factors.

Oblivious sketches. For the for-all version of the ℓ_p subspace sketch problem, we note that there exist general sketches such as the Cauchy sketch [14] which are beyond the reach of the corollary above. Note that the Cauchy sketch is an oblivious sketch, which means the distribution is independent of A. We also prove a dimension lower bound of $\widetilde{\Omega}(\varepsilon^{-2}\cdot d)$ on the target dimension for oblivious sketches (see section 7), which is tight up to logarithmic factors since the Cauchy sketch has a target dimension of $O(\varepsilon^{-2}d\log(d/\varepsilon))$.

THEOREM 1.10 (informal version of Theorem 7.3). *Let $p \in [1,2)$ be a constant. Any oblivious sketch that solves the for-all version of the ℓ_p subspace sketch problem has a target dimension of $\widetilde{\Omega}(\varepsilon^{-2}d)$.*

Therefore, it is natural to ask in general whether $N_p(d,\varepsilon) = \widetilde{\Omega}(d/\varepsilon^2)$. A proof using the framework of this paper would require an $\widetilde{\Omega}(d^2/\varepsilon^2)$ lower bound for the for-all version of the ℓ_p subspace sketch problem. We conjecture it is true; however, our current methods, giving almost-tight lower bounds (in the for-each sense), do not extend to give this result and so we leave it as a main open problem.

1.2. Connection with Banach space theory. In the language of functional analysis, the ℓ_p subspace embedding problem is a classical problem in the theory of L_p spaces with a rich history. For two Banach spaces X and Y, we say X K-embeds into Y if there exists an injective homomorphism $T : X \to Y$ satisfying $\|x\|_X \le \|Tx\|_Y \le K\|x\|_X$ for all $x \in X$. Such a T is called an *isomorphic embedding*. A classical problem in the theory of Banach spaces is to consider the isomorphic embedding of finite-dimensional subspaces of $L_p = L^p(0,1)$ into $\ell_p^n = (\mathbb{R}^n, \|\cdot\|_p)$, where $p \ge 1$ is a constant. Specifically, the problem asks what is the minimum value of n, denoted by $N_p(d,\varepsilon)$, for which all d-dimensional subspaces of L_p $(1+\varepsilon)$-embed into ℓ_p^n. A comprehensive survey of this problem can be found in [22].

The case of $p = 2$ is immediate, in which case one can take $n = d$ and $\varepsilon = 0$, obtaining an isometric embedding, and thus we assume $p \ne 2$. We remark that, when p is an even integer, it is also possible to attain an isometric embedding into ℓ_p^n with $n = \binom{d+p-1}{p} - 1$ [25]. In general, the best[3] known upper bounds on $N_p(d,\varepsilon)$ are as follows:

$$N_p(d,\varepsilon) \le \begin{cases} C\varepsilon^{-2} d \log d, & p = 1, \\ C\varepsilon^{-2} d (\log \varepsilon^{-2} d)(\log\log \varepsilon^{-2} d + \log(1/\varepsilon))^2, & p \in (1,2), \\ C_p \varepsilon^{-2} d^{p/2} \log^2 d \log(d/\varepsilon), & p \in (2,\infty) \setminus 2\mathbb{Z}, \\ C\varepsilon^{-2}(10d/p)^{p/2}, & p \in 2\mathbb{Z}^+, \end{cases} \tag{1.2}$$

where $C > 0$ is an absolute constant and $C_p > 0$ is a constant that depends only on p. The cases of $p = 1$ and $p \in (1,2)$ are due to Talagrand [40, 41]. The case of noneven integers $p > 2$ is taken from [26, Theorem 15.13], based on the earlier work of Bourgain, Lindenstrauss, and Milman [9]. The case of even integers p is due to Schechtman [38].

The upper bounds in (1.2) are established by subsampling with a technique called change of density [22]. First observe that it suffices to consider embeddings from ℓ_p^N to ℓ_p^n since any d-dimensional subspace of L_p $(1+\varepsilon)$-embeds into ℓ_p^N for some large N. Now suppose that E is a d-dimensional subspace of ℓ_p^N. One can show that randomly subsampling coordinates induces a low-distortion isomorphism between E and E restricted onto the sampled coordinates, provided that each element of E is "spread out" among the coordinates, which is achieved by first applying the technique of change of density to E.

Regarding lower bounds, a quick lower bound follows from the tightness of Dvoretzky's theorem for ℓ_p spaces (see, e.g., [30, p. 21]), which states that if ℓ_2^d 2-embeds into ℓ_p^n, then $n \ge cd$ for $1 \le p < 2$ and $n \ge (cd/p)^{p/2}$ for $p \ge 2$, where $c > 0$ is an absolute constant. Since ℓ_2^d embeds into L_p isometrically for all $p \ge 1$ [21, p. 16], identical lower bounds for $N_p(d,\varepsilon)$ follow. Hence the upper bounds in (1.2) are, in terms of d, tight for $p \in 2\mathbb{Z}$ and tight up to logarithmic factors for other values of p. However, the right dependence on ε is a long-standing open problem and little is known. See [22, p. 845] for a discussion on this topic. It is known that $N_1(d,\varepsilon) \ge c(d)\varepsilon^{-2(d-1)/(d+2)}$ [9], whose proof critically relies upon the fact that the unit ball of a finite-dimensional space of ℓ_1 is the polar of a zonotope (a linear image of cube $[-1,1]^d$) and the ℓ_1-norm for vectors in the subspace thus admits a nice representation [7]. However, a lower bound for general p is unknown. Our Corollary 1.7 shows that $n \ge c\varepsilon^{-2}/\operatorname{poly}(\log(1/\varepsilon))$ for all $p \ge 1$ and $p \notin 2\mathbb{Z}$, which is the first lower bound on the dependence of ε for general

[3] A few upper bounds for the case $p \in (2,\infty) \setminus 2\mathbb{Z}$ are known, none of which dominates the rest. Here we choose the one having the best dependence on both d and ε, up to polylog(d/ε) factors.

p, and is optimal up to logarithmic factors. We would like to stress that except for the very special case of ℓ_1, no lower bound on the dependence on ε whatsoever was known for $p \notin 2\mathbb{Z}$. We consider this to be significant evidence of the generality and novelty of our techniques. Moreover, even our lower bound for $p = 1$ is considerably wider in scope, as discussed more below.

1.3. Comparison with prior work.

1.3.1. Comparison with previous results in functional analysis. As discussed, the mentioned lower bounds on $N_p(d, \varepsilon)$ come from the tightness of Dvoretzky's theorem, which shows the impossibility of embedding ℓ_2^d into a Banach space with low distortion. Here the hardness comes from the geometry of the target space. In contrast, we emphasize that the hardness in our ℓ_p subspace sketch problem comes from the source space, since the target space is unconstrained and the output function $Q_p(\cdot)$ does not necessarily correspond to an embedding. The lower bound via tightness of Dvoretzky's theorem cannot show that ℓ_p^d does not $(1+\varepsilon)$-embed into ℓ_q^n for $d = \Theta(\log(1/\varepsilon))$ and $n = O(1/\varepsilon^{1.99})$, where $q \notin 2\mathbb{Z}$.

When the target space is not ℓ_p, lower bounds via functional analysis are more difficult to obtain since they require understanding the geometry of the target space. Since our data structure problem has no constraints on $Q_p(\cdot)$, the target space does not even need to be normed. In theoretical computer science and machine learning applications, the usual "sketch and solve" paradigm typically just requires the target space to admit an efficient algorithm for the optimization problem at hand.[4] Our lower bounds are thus much wider in scope than those in geometric functional analysis.

1.3.2. Comparison with previous results for graph sparsifiers. Recently, the bit complexity of cut sparsifiers was studied in [5, 12]. Given an undirected graph $G = (V, E)$, $|V| = d$, a function $f : 2^V \to \mathbb{R}$ is a $(1+\varepsilon)$-*cut sketch*, if for any vertex set $S \subseteq V$,

$$(1-\varepsilon)C(S, V \setminus S) \le f(S) \le (1+\varepsilon)C(S, V \setminus S),$$

where $C(S, V \setminus S)$ denotes the capacity of the cut between S and $V \setminus S$. The main result of these works is that any $(1+\varepsilon)$-cut sketch requires $\Omega(\varepsilon^{-2} d \log d)$ bits to store. Note that a cut sketch can be constructed using a for-all version of the ℓ_p subspace sketch for any p, by just taking the matrix A to be the edge-vertex incidence matrix of the graph G and querying all vectors $x \in \{0,1\}^d$. Thus, one may naturally ask if the lower bounds in [5, 12] imply any lower bounds for the subspace sketch problem.

We note that both works [5, 12] have explicit constraints on the value of ε. In [5], in order to prove the $\Omega(\varepsilon^{-2}d)$ lower bound, it is required that $\varepsilon = \Omega(1/\sqrt{d})$. In [12] the lower bound of $\Omega(d \log d/\varepsilon^2)$ requires $\varepsilon = \omega(1/d^{1/4})$. Thus, the strongest lower bound that can be proved using such an approach is $\widetilde{\Omega}(d^2)$. This is natural, since one can always store the entire adjacency matrix of the graph in $\widetilde{O}(d^2)$ bits. Our lower bound, in contrast, becomes arbitrarily large as $\varepsilon \to 0$.

1.4. Follow-up work.

After the publication of the preliminary version of this paper in a conference proceedings, Andoni et al. applied our technique to obtain a nearly tight lower bound for point-querying the objective function of a bias-regularized support vector machine problem [4].

[4] For example, consider the space $\mathbb{R}^n$ endowed with a premetric $d(x, y) = \sum_i f(x_i - y_i)$, where $f(x) = \tau x \mathbf{1}_{\{x \ge 0\}} + (\tau - 1)x\mathbf{1}_{\{x \le 0\}}$ ($\tau \in (0,1)$), which is not even symmetric when $\tau \ne \frac{1}{2}$. See [46] for an embedding into this space.

1.5. Our techniques. We use the case of $p = 1$ to illustrate our ideas behind the $\widetilde{\Omega}(\varepsilon^{-2})$ lower bound for the ℓ_p subspace sketch problem, when $d = \Theta(\log(1/\varepsilon))$. We then extend this to an $\widetilde{\Omega}(\varepsilon^{-2}d)$ lower bound for general d via a simple padding argument. We first show how to prove a weaker $\widetilde{\Omega}(\varepsilon^{-1})$ lower bound for the for-all version of the problem, and then show how to strengthen the argument to obtain both a stronger $\widetilde{\Omega}(\varepsilon^{-2})$ lower bound and in the weaker original version of the problem (the "for-each" model, where we only need to be correct on a fixed query x with constant probability).

Note that the condition that $d = \Theta(\log(1/\varepsilon))$ is crucial for our proof. As shown in section 11, when $d = 2$, there is actually an $\widetilde{O}(\varepsilon^{-1})$ upper bound, and thus our $\widetilde{\Omega}(\varepsilon^{-2})$ lower bound does not hold universally for all values of d. It is thus crucial that we look at a larger value of d, and we show that $d = \Theta(\log(1/\varepsilon))$ suffices.

To prove our bit lower bounds for the ℓ_1 subspace sketch problem, we shall encode random bits in the matrix A such that having a $(1+\varepsilon)$-approximation to $\|Ax\|_1$ will allow us to recover, in the for-each case, some specific random bit, and in the for-all case, all the random bits using different choices of x. A standard information-theoretic argument then implies that the lower bound for the subspace sketch problem is proportional to the number of random bits we can recover.

Warmup: An $\widetilde{\Omega}(\varepsilon^{-1})$ lower bound for the for-all version. In our hard instance, we let $d = \Theta(\log(1/\varepsilon))$ be such that $n = 2^d = \widetilde{\Theta}(1/\varepsilon)$. Form a matrix $A \in \mathbb{R}^{n\times d}$ by including all vectors $i \in \{-1,1\}^d$ as its rows and then scaling the ith row by a nonnegative scalar $r_i \le \mathrm{poly}(d)$. We can think of r as a vector in $\mathbb{R}^n$ with $\|r\|_\infty \le \mathrm{poly}(d)$. Now, we query $Q_1(i)$ for all vectors $i \in \{-1,1\}^d$. For an appropriate choice of $d = \Theta(\log(1/\varepsilon))$, for all $i \in \{-1,1\}^d$, we have

$$\|Ai\|_1 = \sum_{j\in\{-1,1\}^d} r_j \cdot |\langle i,j\rangle| \le 2^d \cdot \mathrm{poly}(d) < \frac{1}{\varepsilon}. \tag{1.3}$$

Since $Q_1(i)$ is a $(1\pm\varepsilon)$-approximation to $\|Ai\|_1$, and $\|Ai\|_1$ is always an integer, we can recover the exact value of $\|Ai\|_1$ using $Q_1(i)$ for all $i \in \{-1,1\}^d$.

Now we define a matrix $M \in \mathbb{R}^{n\times n}$, where $M_{i,j} = |\langle i,j\rangle|$, where i,j are interpreted as vectors in $\{-1,1\}^d$. A simple yet crucial observation is that $\|Ai\|_1$ is exactly the ith coordinate of Mr. Notice that this critically relies on the assumption that r has nonnegative coordinates. Thus, the problem can be equivalently viewed as designing a vector $r \in \mathbb{R}^n$ with $\|r\|_\infty \le \mathrm{poly}(d)$ and recovering r from the vector Mr. At this point, a natural idea is to show that the matrix M has a sufficiently large rank, say, $\mathrm{rank}(M) = \widetilde{\Omega}(\varepsilon^{-1})$, and carefully design r to show an $\Omega(\mathrm{rank}(M)) = \widetilde{\Omega}(\varepsilon^{-1})$ lower bound.

Fourier analysis on the hypercube shows that the eigenvectors of M are the rows of the normalized Hadamard matrix, while the eigenvalues of M are the Fourier coefficients associated with the function $g(s) = |d - 2w_H(s)|$, where $w_H(s)$ is the Hamming weight of a vector $s \in \mathbb{F}_2^d$. Considering all vectors of Hamming weight $d/2$ in $\mathbb{F}_2^d$ and their associated Fourier coefficients, we arrive at the conclusion that there are at least $\binom{d}{d/2}$ eigenvalues of M with absolute value

$$\left| \sum_{\substack{0\le i\le d \\ i \text{ is even}}} (-1)^{i/2} \binom{d/2}{i/2} |d-2i| \right|,$$

which can be shown to be at least $\Omega(2^{d/2}/\mathrm{poly}(d))$. The formal argument is given in

section 3.1. Hence $\operatorname{rank}(M) \geq \binom{d}{d/2} = \Omega(2^d/\operatorname{poly}(d))$. Without loss of generality we assume the $\operatorname{rank}(M) \times \operatorname{rank}(M)$ upper-left block of A is nonsingular.

Now an $\widetilde{\Omega}(1/\varepsilon)$ lower bound follows readily. Set r so that

$$r_i = \begin{cases} s_i, & i \leq \operatorname{rank}(M), \\ 0, & i > \operatorname{rank}(M), \end{cases}$$

where $\{s_i\}_{i=1}^{\operatorname{rank}(M)}$ is a set of i.i.d. Bernoulli random variables. Since the exact value of Mr is known and the $\operatorname{rank}(M) \times \operatorname{rank}(M)$ upper-left block of A is nonsingular, one can recover the values of $\{s_i\}_{i=1}^{\operatorname{rank}(M)}$ by solving a linear system, which implies an $\Omega(\operatorname{rank}(M)) = \widetilde{\Omega}(\varepsilon^{-1})$ lower bound.

Before proceeding, let us first review why our argument fails for $p = 2$. For the ℓ_p-norm, the Fourier coefficients associated with the vectors of Hamming weight $d/2$ on the Boolean cube are

$$\left| \sum_{\substack{0 \leq i \leq d \\ i \text{ is even}}} (-1)^{i/2} \binom{d/2}{i/2} |d - 2i|^p \right| = \Theta\left(\frac{2^{d/2}}{\sqrt{d}} \left| \sin \frac{p\pi}{2} \right| \right).$$

Therefore this sum vanishes if and only if p is an even integer, in which case $\operatorname{rank}(A)$ will no longer be $\Omega(2^d/\operatorname{poly}(d))$ and the lower bound argument will fail.

An $\widetilde{\Omega}(\varepsilon^{-2})$ lower bound for the for-each version. To strengthen this to an $\widetilde{\Omega}(\varepsilon^{-2})$ lower bound, it is tempting to increase d so that $n = 2^d = \widetilde{\Omega}(\varepsilon^{-2})$. In this case, however, we can no longer recover the exact value of Mr, since each entry of Mr now has magnitude $\widetilde{\Theta}(\varepsilon^{-2})$ and the function $Q_1(\cdot)$ only gives a $(1 \pm \varepsilon)$-approximation. We still obtain a noisy version of Mr, but with a $\widetilde{\Theta}(1/\varepsilon)$ additive error on each entry. One peculiarity of the model here is that if some entries of r are negative, then $\|Ai\|_1 = (M|r|)_i$ (cf. (1.3)), where $|r|$ denotes the vector formed by taking the absolute value of each coordinate of r, i.e., $\|Ai\|_1$ depends only on the absolute values of entries of r, which suggests that the constraint that each entry of Mr has magnitude $\widetilde{\Theta}(1/\varepsilon^2)$ with an additive error of $\widetilde{\Theta}(1/\varepsilon)$ is somehow intrinsic.

To illustrate our idea for overcoming the issue of large additive error, for the time being let us forget the actual form of M previously defined in the argument for our $\widetilde{\Omega}(\varepsilon^{-1})$ lower bound and consider instead a general $M \in \mathbb{R}^{n \times n}$ with orthogonal rows, each row having ℓ_2 norm $\Omega(2^{d/2}/\operatorname{poly}(d))$. For now we also allow r to contain negative entries such that $\|r\|_\infty \leq \operatorname{poly}(d)$, and pretend that the noisy version of Mr has an $\widetilde{\Theta}(1/\varepsilon)$ additive error on each entry. Now, let

$$r = \sum_{i=1}^{n} s_i \cdot \frac{M_i}{\|M_i\|_2},$$

where $\{s_i\}_{i=1}^n$ is a set of i.i.d. Rademacher random variables. By a standard concentration inequality, $\|r\|_\infty \leq \operatorname{poly}(d)$ holds with high probability (recall that $n = 2^d$). Consider the vector Mr. Due to the orthogonality of the rows of M, the ith coordinate of Mr will be

$$\langle M_i, r \rangle = s_i \cdot \|M_i\|_2.$$

Provided that $\|M_i\|_2$ is larger than the additive error $\widetilde{\Theta}(1/\varepsilon)$, we can still recover s_i by just looking at the sign of $\langle M_i, r \rangle$. Thus, for an appropriate choice of d such that $2^{d/2}/\operatorname{poly}(d) = \widetilde{\Omega}(1/\varepsilon)$, we can obtain an $\Omega(2^d) = \widetilde{\Omega}(1/\varepsilon^2)$ lower bound.

Now we return to the original M with $M_{i,j} = |\langle i, j\rangle|$, whose rows are not necessarily orthogonal. The previous argument still goes through so long as we can identify a subset $\mathcal{R} \subseteq [n] = [2^d]$ of size $|\mathcal{R}| \geq \Omega(2^d/\operatorname{poly}(d))$ such that the rows $\{M_i\}_{i\in\mathcal{R}}$ are nearly orthogonal, meaning that the ℓ_2 norm of the orthogonal projection of M_i onto the subspace spanned by other rows $\{M_j\}_{j\in\mathcal{R}\setminus\{i\}}$ is much smaller than $\|M_i\|_2$.

To achieve this goal, we study the spectrum of M and, as far as we are aware, this is the first such study of spectral properties of this matrix. The Fourier argument mentioned above implies that at least $\Omega(2^d/\operatorname{poly}(d))$ eigenvalues of A have the same absolute value $\Omega(2^{d/2}/\operatorname{poly}(d))$. If all other eigenvalues of A were zero, then we could identify a set of $|\mathcal{R}| \geq \Omega(2^d/\operatorname{poly}(d))$ nearly orthogonal rows using rows of A each with ℓ_2 norm $\Omega(2^{d/2}/\operatorname{poly}(d))$, using a procedure similar to the standard Gram–Schmidt process. The full details can be found in section 3.2. Although the other eigenvalues of M are not all zero, we can simply ignore the associated eigenvectors since they are orthogonal to the set of nearly orthogonal rows we obtain above.

Last, recall that what we truly obtain is $M|r|$ rather than Mr unless $r \geq 0$. To fix this, note that $\|r\|_\infty \leq \operatorname{poly}(d)$ with high probability, and so we can just shift each coordinate of r by a fixed amount of $\operatorname{poly}(d)$ to ensure that all entries of r are positive. We can still obtain $\langle M_i, r\rangle$ with an additive error $\widetilde{\Theta}(1/\varepsilon)$, since the amount of the shift is fixed and bounded by $\operatorname{poly}(d)$.

Notice that the above argument in fact holds even for the for-each version of the subspace sketch problem. By querying the ith vector on the Boolean cube for some $i \in \mathcal{R}$, we are able to recover the sign of s_i with constant probability. Given this, a standard information-theoretic argument shows that our lower bound holds for the for-each version of the problem.

The formal analysis given in section 3.3 is a careful combination of all the ideas mentioned above.

Applications: M-estimators and projective clustering coresets. Our general strategy for proving lower bounds for M-estimators is to relate one M-estimator, for which we want to prove a lower bound, to another M-estimator for which a lower bound is easy to derive. For the L_1-L_2 estimator, the Huber estimator, and the Fair estimator, when $|t|$ is sufficiently large, $\phi(t) = (1 \pm \varepsilon)|t|$ (up to rescaling of t and the function value), and thus the lower bounds follow from those for the ℓ_1 subspace sketch problem.

For the Cauchy estimator, we relate it to another estimator $\phi_{\text{aux}}(t) = \ln|x| \cdot \mathbf{1}_{\{|x|\geq 1\}}$. In section 8, we show that our Fourier analytic arguments also work for $\phi_{\text{aux}}(t)$. Since for sufficiently large t, the Cauchy estimator satisfies $\phi(t) = (1 \pm \varepsilon)\phi_{\text{aux}}(t)$ (up to rescaling of t and the function value), a lower bound for the Cauchy estimator follows.

To prove lower bounds on coresets for projective clustering, the main observation is that when $k = 1$ and $j = d - 1$, by choosing the query subspace to be the orthogonal complement of a vector z, the projection cost is just $\sum_{x\in X} \phi(\langle x, z\rangle)$, and thus we can invoke our lower bounds for the subspace sketch problem. We use a coding argument to handle general k. In Lemma 9.1, we show there exists a set $S = \{(s_1, t_1), (s_2, t_2), \ldots, (s_k, t_k)\}$, where $s_i, t_i \in \mathbb{R}^{O(\log k)}$, $\langle s_i, t_i\rangle = 0$, and $\langle s_i, t_j\rangle$ is arbitrarily large when $i \neq j$. Now for k copies of the hard instance of the subspace sketch problem, we add s_i as a prefix to all data points in the ith hard instance, and set the query subspace to be the orthogonal complement of a vector z, to which we add t_i as a prefix. Now, the data points in the ith hard instance will always choose the ith center in the optimal solution, since otherwise an arbitrarily large cost will

incur. Thus, we can solve k independent copies of the subspace sketch problem, and the desired lower bound follows.

In the rest of the section, we shall illustrate our techniques for proving lower bounds that depend on p for the ℓ_p subspace sketch problem. These lower bounds hold even when ε is a constant. We again resort to information theory, trying to recover, using Q_p queries, the entire matrix A among a collection $\mathcal{S}$ of matrices. The lower bound is then $\Omega(\log|\mathcal{S}|)$ bits.

An $\widetilde{\Omega}(d^{p/2})$ lower bound for the for-each version. Our approach for proving the $\widetilde{\Omega}(d^{p/2})$ lower bound is based on the following crucial observation: consider a uniformly random matrix $A \in \{-1,1\}^{\Theta(d^{p/2})\times d}$ and a uniformly random vector $x \in \{-1,1\}^d$. Then $\mathbb{E}\,\|Ax\|_p^p = O(d^p)$ with the constant hidden in the O-notation less than 1, whereas for each row $A_i \in \mathbb{R}^d$ of A, interpreted as a column vector, $\|AA_i\|_p^p \geq d^p$. Intuitively, the lower bound comes from the fact that one can recover the whole matrix A by querying all Boolean vectors $x \in \{-1,1\}^d$ using the function $Q_p(\cdot)$, since if x is a row of A, then $\|Ax\|_p^p$ would be slightly larger than its typical value, by adjusting constants.

To implement this idea, one can generate a set of almost orthogonal vectors $S \subseteq \mathbb{R}^d$ and require that all rows of A come from S. A simple probabilistic argument shows that one can construct a set of $|S| = d^p$ vectors such that for any distinct $s,t \in S$, $|\langle s,t\rangle| \leq O(\sqrt{d\log d})$.[5] If we form the matrix A using $n = \widetilde{\Omega}(d^{p/2})$ vectors from S as its rows, then for any vector t that is *not* a row of A,

$$\|At\|_p^p \leq n\cdot(d\log d)^{p/2} \ll d^p$$

for some appropriate choice of n. Thus, by querying $Q_p(s)$ for all vectors $s \in S$, one can recover the whole matrix A, even when ε is a constant. By a standard information-theoretic argument, this leads to a lower bound of $\Omega(\log\binom{d^p}{\widetilde{\Omega}(d^{p/2})}) = \widetilde{\Omega}(d^{p/2})$ bits. Furthermore, one only needs to query $|S| = d^p$ vectors, which means the lower bound in fact holds for the for-each version of the ℓ_p subspace sketch problem, by a standard repetition argument and losing a $\log d$ factor in the lower bound.

An $\Omega(d^{\max\{p/2,1\}+1})$ lower bound for the for-all version. In order to obtain the nearly optimal $\Omega(d^{\max\{p/2,1\}+1})$ lower bound for the for-all version, we must abandon the constraint that all rows of the A matrix come from a set S of $\mathrm{poly}(d)$ vectors. Our plan is still to construct a large set of matrices $\mathcal{S} \subseteq \{+1,-1\}^{\Theta(d^{p/2})\times d}$, and show that for any distinct matrices $S,T \in \mathcal{S}$, it is possible to distinguish them using the function $Q_p(\cdot)$, thus proving an $\Omega(\log|\mathcal{S}|)$ lower bound. The new observation is that, to distinguish two matrices $S,T \in \mathcal{S}$, it suffices to have a *single* row of T, say T_i, such that $\|ST_i\|_p^p \ll d^p$. Again using the probabilistic method, we show the existence of such a set $\mathcal{S}$ with size $\exp\left(\Omega(d^{p/2+1})\right)$, which implies an $\Omega(\log|\mathcal{S}|) = \Omega(d^{p/2+1})$ lower bound.

Our main technical tool is Talagrand's concentration inequality, which shows that for any $p \geq 2$ and vector $x \in \{-1,1\}^d$, for a matrix $A \in \mathbb{R}^{\Theta(d^{p/2})\times d}$ with i.i.d. Rademacher entries, $\|Ax\|_p = \Theta(d)$ with probability $1-\exp(-\Omega(d))$. This implies that for two random matrices $S,T \in \mathbb{R}^{\Theta(d^{p/2})\times d}$ with i.i.d. Rademacher entries, the probability that there exists some row T_i of T such that $\|ST_i\|_p^p \ll d^p$ is at least $1-\exp\left(\Omega(d^{p/2+1})\right)$, since the $\Theta(d^{p/2})$ rows of T are independent. By a probabilistic argument, the existence of the set $\mathcal{S}$ follows. The formal analysis is given in section 4.2.1.

[5] The $O(\sqrt{\log d})$ factor can be removed using more sophisticated constructions based on coding theory (see Lemma 4.1).

The above argument fails to give an $\Omega(d^2)$ lower bound when $p < 2$. However, for any $p < 2$, since ℓ_2^n embeds into ℓ_p^m with $m = O_p(n)$ and a constant distortion, we can directly reduce the case of $p < 2$ to the case of $p = 2$. The formal analysis can be found in section 4.2.2. Combining these two results yields the $\Omega(d^{\max\{p/2,1\}+1})$ lower bound.

2. Preliminaries. For two functions f and g, we write $f(x) \sim g(x)$, $x \to \infty$, if $\lim_{x\to\infty} f(x)/g(x) = 1$.

For a vector $x \in \mathbb{R}^n$, we use $\|x\|_p$ to denote its ℓ_p-norm, i.e., $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$. When $p < 1$, it is not a norm but it is still a well-defined quantity and we call it the ℓ_p-norm for convenience. When $p = 0$, $\|x\|_0$ is defined to be the number of nonzero coordinates of x.

For two vectors $x, y \in \mathbb{R}^n$, we use $\mathrm{proj}_y x \in \mathbb{R}^n$ to denote the orthogonal projection of x onto y. For a matrix $A \in \mathbb{R}^{n\times d}$, we use $A_i \in \mathbb{R}^d$ to denote its ith row, treated as a column vector. We use $\|A\|_2$ to denote its spectral norm, i.e., $\|A\|_2 = \sup_{\|x\|_2=1} \|Ax\|_2$, and $\|A\|_F$ to denote its Frobenius norm, i.e., $\|A\|_F = \left(\sum_{i=1}^n \sum_{j=1}^d A_{ij}^2\right)^{1/2}$.

Suppose that $A \in \mathbb{R}^{m\times n}$ has singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$, where $r = \min\{m, n\}$. It holds that $\sigma_1 = \|A\|_2 \leq \|A\|_F = \left(\sum_{i=1}^r \sigma_i^2\right)^{1/2}$. The condition number of A is defined to be

$$\kappa(A) = \frac{\sup_{\|x\|_2=1} \|Ax\|_2}{\inf_{\|x\|_2=1} \|Ax\|_2}.$$

THEOREM 2.1 (Eckart–Young–Mirsky theorem). *Suppose that $A \in \mathbb{R}^{m\times n}$ has singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$, where $\mathrm{rank}(A) = r \leq \min\{m, n\}$. For any matrix $B \in \mathbb{R}^{m\times n}$ such that $\mathrm{rank}(B) \leq k \leq r$, it holds that*

$$\|A - B\|_F^2 \geq \sum_{i=k+1}^{r} \sigma_i^2.$$

Below we list a handful of concentration inequalities which will be useful in our arguments.

LEMMA 2.2 (Hoeffding's inequality, [8, p. 34]). *Let $s_1, \ldots, s_n$ be i.i.d. Rademacher random variables and $a_1, \ldots, a_n$ be real numbers. Then*

$$\Pr\left\{\sum_i s_i a_i > t\right\} \leq \exp\left(-\frac{t^2}{2\sum_i a_i^2}\right).$$

LEMMA 2.3 (Khintchine's inequality, [8, p. 145]). *Let $a_1, \ldots, a_n$ be real numbers and $s_1, \ldots, s_n$ be i.i.d. Rademacher random variables. There exist absolute constants $A, B > 0$ such that*

$$A\left(\sum_i a_i^2\right)^{1/2} \leq \left(\mathbb{E}\,|\sum_i s_i a_i|^p\right)^{1/p} \leq B\sqrt{p}\left(\sum_i a_i^2\right)^{1/2}.$$

LEMMA 2.4 (Talagrand's inequality, [8, p. 204]). *Let $X = (X_1, \ldots, X_n)$ be a random vector with independent coordinates taking values in $[-1, 1]$. Let $f : [-1, 1]^n \to \mathbb{R}$ be a convex 1-Lipschitz function. It holds for all $t \geq 0$ that*

$$\Pr\{f(X) - \mathbb{E}\, f(X) \geq t\} \leq e^{-t^2/8}.$$

LEMMA 2.5 (Gaussian concentration, [43, p. 105]). *Let $p \geq 1$ be a constant. Consider a random vector $X \sim N(0, I_n)$ and a nonnegative 1-Lipschitz function $f : (\mathbb{R}^n, \|\cdot\|_2) \to \mathbb{R}$, then*

$$\Pr\left\{|f(x) - (\mathbb{E}(f(x))^p)^{1/p}| \geq t\right\} \leq 2e^{-ct^2},$$

where $c = c(p) > 0$ is a constant that depends only on p.

LEMMA 2.6 (extreme singular values, [43, p. 91]). *Let A be an $N \times n$ matrix with i.i.d. Rademacher entries. Let $\sigma_{\min}(A)$ and $\sigma_{\max}(A)$ be the smallest and largest singular values of A. Then for every $t \geq 0$, with probability at least $1 - 2\exp(-ct^2)$, it holds that*

$$\sqrt{N} - C\sqrt{n} - t \leq \sigma_{\min}(A) \leq \sigma_{\max}(A) \leq \sqrt{N} + C\sqrt{n} + t,$$

where $C, c > 0$ are absolute constants.

LEMMA 2.7. *Let $U_1, \ldots, U_k \in \mathbb{R}^n$ be orthonormal vectors and $s_1, \ldots, s_k$ be independent Rademacher random variables. It holds that*

$$\Pr\left\{\left\|\sum_{i=1}^k s_i \cdot U_i\right\|_\infty \leq 3\sqrt{\ln k}\right\} \geq 1 - \frac{1}{k^{1.3}}.$$

Proof. Let $Z = \sum_{i=1}^k s_i U_i$, then

$$Z_j = \sum_{i=1}^k s_i U_{i,j}.$$

Since $\{U_i\}$ is a set of orthonormal vectors, we have that

$$\sum_{i=1}^k U_{i,j}^2 \leq 1.$$

It follows from Hoeffding's inequality (Lemma 2.2) that for each $j \in [k]$,

$$\Pr\{|Z_j| \geq 3\sqrt{\ln k}\} \leq \exp(-2\ln k).$$

The claimed inequality follows by taking a union bound over all $j \in [k]$. ∎

We also need a result concerning uniform approximation of smooth functions by polynomials. Let P_n denote the space of polynomials of degree at most n. For a given function $f \in C[a, b]$, the best degree-n approximation error $E_n(f; [a, b])$ is defined to be

$$E_n(f; [a, b]) = \inf_{p \in P_n} \|f - p\|_\infty,$$

where the $\|\cdot\|_\infty$ norm is taken over $[a, b]$. The following bound on approximation error is a classical result.

LEMMA 2.8 (see [36, p. 23]). *Let $f(x)$ have a kth derivative on $[-1, 1]$. If $n > k$,*

$$E_n(f; [-1, 1]) \leq \frac{6^{k+1}e^k}{(k+1)n^k}\ \omega_k\left(\frac{1}{n-k}\right),$$

where ω_k is the modulus of continuity of $f^{(k)}$, defined as

$$\omega_k(\delta) = \sup_{\substack{x,y \in [-1,1] \\ |x-y| \leq \delta}} \left|f^{(k)}(x) - f^{(k)}(y)\right|.$$

3. An $\widetilde{\Omega}(\varepsilon^{-2})$ lower bound. To prove the space lower bound of the data structure Q_p, we appeal to information theory. We shall encode random bits in A such that if for each x, $Q_p(x)$ approximates $\|Ax\|_p^p$ (or $\|Ax\|_p$ when $p = 0$) up to a $1 \pm \varepsilon$ factor with probability at least 0.9, we can recover from $Q_p(x)$ some random bit (depending on x) with at least constant probability. A standard information-theoretic argument implies a lower bound on the size of Q_p which is proportional to the number of random bits we can recover.

For each $p \geq 0$, we define a family of matrices $M^{(p)} = \{M^{(d,p)}\}_{d=1}^{\infty}$, where $M^{(d,p)}$ is a $2^d \times 2^d$ matrix with entries defined as

$$M_{i,j}^{(d,p)} = |\langle i, j \rangle|^p,$$

where i and j are interpreted as vectors on the Boolean cube $\{-1, 1\}^d$. We assume $0^0 = 0$ throughout the paper.

3.1. Spectrum of matrices M.

LEMMA 3.1. *For any $d \geq 1$, $M^{(d,p)}$ can be rewritten as $H^{(d)}\Lambda^{(d,p)}(H^{(d)})^T$ in its spectral decomposition form, where $\Lambda^{(d,p)}$ is a $2^d \times 2^d$ diagonal matrix, and $H^{(d)}$ is a $2^d \times 2^d$ normalized Hadamard matrix.*

Proof. Let T be the natural isomorphism from the multiplicative group $\{-1, 1\}^d$ to the additive group $\mathbb{F}_2^d$. Then $|\langle i, j \rangle|^p = g(Ti + Tj)$ for some function g defined on $\mathbb{F}_2^d$. It can be computed (see [27, Lemma 5]) that the singular values of a matrix with entries $g(Ti + Tj)$ are the absolute values of the Fourier coefficients of g. In our particular case, the singular values of $M^{(d,p)}$ are

$$|\hat{g}(s)| = \left| \sum_{x \in \mathbb{F}_2^d} (-1)^{\langle s,x \rangle} g(x) \right|, \quad s \in \mathbb{F}_2^d.$$

Furthermore, the proof of that lemma shows that $H^{(d)}$ in the spectral decomposition is given by

$$(H_s^{(d)})_z = \frac{1}{2^{d/2}}(-1)^{\langle s,z \rangle}, \quad s, z \in \mathbb{F}_2^d,$$

which implies that $H^{(d)}$ is a normalized Hadamard matrix. ∎

LEMMA 3.2. *When d is even, there are at least $\binom{d}{d/2}$ entries in $\Lambda^{(d,p)}$ with absolute value*

$$\left| \sum_{\substack{0 \leq i \leq d \\ i \text{ is even}}} (-1)^{i/2} \binom{d/2}{i/2} |d - 2i|^p \right| \triangleq \Lambda_0^{(d,p)} \geq 0.$$

Proof. We shall use the notation in the proof of Lemma 3.1. Consider the Fourier coefficients $\hat{g}(s)$ for $s \in \mathbb{F}_2^d$ with Hamming weight $d/2$, which is the same for all $\binom{d}{d/2}$ such s's. Note that

$$\hat{g}(s) = \sum_{i=0}^{d} \sum_{j=0}^{i} (-1)^j \binom{d/2}{j} \binom{d/2}{i-j} g(i).$$

By comparing the coefficients of x^i on both sides of the identity $(1+x)^{d/2}(1-x)^{d/2} = (1-x^2)^{d/2}$, we see that

$$\sum_{j=0}^{i} (-1)^j \binom{d/2}{j} \binom{d/2}{i-j} = \begin{cases} (-1)^{i/2}\binom{d/2}{i/2}, & i \text{ is even}, \\ i \text{ is odd}. \end{cases}$$

Hence

$$\hat{g}(s) = \sum_{\text{even } i} (-1)^{i/2} \binom{d/2}{i/2} g(i).$$

Finally, observe that, for $x, y \in \{+1, -1\}^d$, we have $(d - \langle x, y\rangle)/2 = d_H(x, y) = w_H(Tx + Ty)$, where $d_H(x, y)$ denotes the Hamming distance between x and y and $w_H(s)$ denotes the Hamming weight of $s \in \mathbb{F}_2^d$. Hence $g(i) = |d - 2i|^p$ and the conclusion follows. ∎

Let $N^{(d)}$ be the multiplicity of the singular value $\Lambda_0^{(d,p)}$ of $M^{(d,p)}$. We know from the preceding lemma that $N^{(d)} \geq \binom{d}{d/2}$. By permuting the columns of $H^{(d)}$, we may assume the absolute value of the first $N^{(d)}$ diagonal entries of $\Lambda^{(d,p)}$ are all equal to $\Lambda_0^{(d,p)}$, i.e.,

$$\left|\Lambda_1^{(d,p)}\right| = \left|\Lambda_2^{(d,p)}\right| = \cdots = \left|\Lambda_{N^{(d)}}^{(d,p)}\right| = \Lambda_0^{(d,p)}.$$

The following lemma is critical in lower bounding $\Lambda_0^{(d,p)}$. We found the result in a post on `math.stackexchange.com` [1] but could not find it in any published literature and so we reproduce the proof in full from [1], with small corrections regarding convergence of integrals.

LEMMA 3.3. *It holds for all complex p satisfying $0 < \operatorname{Re} p < 2n$ that*

$$\sum_{k=1}^{n} (-1)^{k+1} \binom{2n}{n+k} k^p = 2^{2n-p} \frac{\Gamma(p+1)}{\pi} \left(\sin \frac{\pi p}{2}\right) \int_0^\infty \frac{\sin^{2n} t}{t^{p+1}} dt. \tag{3.1}$$

Proof. By the binomial theorem,

$$(z-1)^{2n} = \sum_{k=0}^{2n} \binom{2n}{k} (-z)^k = \sum_{k=-n}^{n} \binom{2n}{k+n} (-z)^{k+n}.$$

Splitting the sum at $k = -1$ and $k = 1$, we have

$$\sum_{k=1}^{n} (-1)^k \binom{2n}{n+k} \left(z^k + z^{-k}\right) = (-1)^n (z-1)^{2n} z^{-n} - \binom{2n}{n}.$$

Plugging in $z = \exp(2it)$ yields

$$(2 \sin t)^{2n} = 2 \sum_{k=1}^{n} (-1)^k \binom{2n}{n+k} \cos(2kt) + \binom{2n}{n}. \tag{3.2}$$

Plug (3.2) into the integral on the right-hand side of (3.1) and introduce a regularizer $\exp(-st)$ $(s > 0)$ under the integral sign:

$$\int_0^\infty e^{-st} \frac{(2 \sin t)^{2n}}{t^{p+1}} dt = 2 \sum_{k=1}^{n} (-1)^k \binom{2n}{n+k} \int_0^\infty \frac{e^{-st} \cos(2kt)}{t^{p+1}} dt + \binom{2n}{n} \Gamma(-p) s^p,$$
$$-1 < \operatorname{Re} p < 0.$$

One can compute that

$$\begin{aligned}\int_0^\infty \frac{e^{-st}\cos(2kt)}{t^{p+1}}dt &= \frac{1}{2}\int_0^\infty \frac{e^{-st}(e^{i2kt}+e^{-i2kt})}{t^{p+1}}dt \\ &= \frac{(s-2ki)^p+(s+2ki)^p}{2}\Gamma(-p) \\ &= (4k^2+s^2)^{\frac{p}{2}}\cos\left(p\arctan\frac{2k}{s}\right)\Gamma(-p), \quad -1<\operatorname{Re}p<0.\end{aligned}$$

It follows that

$$\begin{aligned}&\int_0^\infty e^{-st}\frac{(2\sin t)^{2n}}{t^{p+1}}dt \\ &\quad = 2\sum_{k=1}^n(-1)^k\binom{2n}{n+k}(4k^2+s^2)^{\frac{p}{2}}\cos\left(p\arctan\frac{2k}{s}\right)\Gamma(-p)+\binom{2n}{n}\Gamma(-p)s^p, \\ &\qquad\qquad -1<\operatorname{Re}p<0.\end{aligned}$$

It is easy to verify that the integral on the left-hand side is analytic whenever the integral converges. Analytic continuation permits p to be extended to $\{p : -1 < \operatorname{Re}p < 2n\} \setminus \mathbb{Z}$. Now, for p such that $0 < \operatorname{Re}p < 2n$ and $p \notin \mathbb{Z}$, let $s \to 0^+$ on both sides. It is also easy to verify (for example, by Lebesgue's dominated convergence theorem) that we can take the limit $s \to 0^+$ under the integral sign, hence

$$\int_0^\infty \frac{(2\sin t)^{2n}}{t^{p+1}}dt = 2\sum_{k=1}^n(-1)^k\binom{2n}{n+k}(2k)^p\cos\left(\frac{\pi p}{2}\right)\Gamma(-p), \quad 0<\operatorname{Re}p<2n, p\notin\mathbb{Z}. \tag{3.3}$$

Invoking the reflection identity (see, e.g., [6, p. 9])

$$\Gamma(-p)\Gamma(1+p) = -\frac{\pi}{\sin(p\pi)}, \quad p\notin\mathbb{Z}, \tag{3.4}$$

we obtain that

$$\sum_{k=1}^n(-1)^k\binom{2n}{n+k}k^p = -\frac{2^{2n-p}\Gamma(p+1)}{\pi}\left(\sin\frac{p\pi}{2}\right)\int_0^\infty\frac{\sin^{2n}t}{t^{p+1}}dt, \quad 0<\operatorname{Re}p<2n, p\notin\mathbb{Z}.$$

Finally, analytic continuation extends p to the integers in $(0, 2n)$. ∎

As an immediate corollary of Lemma 3.3, we have the following.

COROLLARY 3.4. *Suppose that $d \in 8\mathbb{Z}$. There exists an absolute constant $c > 0$ such that*

$$\Lambda_0^{(d,p)} \geq c\frac{2^{d/2}}{\sqrt{d}}\left|\sin\frac{p\pi}{2}\right|.$$

Proof. Letting $2n = d/2$ and $k = n - i/2$, the summation in Lemma 3.2 becomes

$$2^{2p+1}\left|\sum_{k=1}^n(-1)^k\binom{2n}{n+k}k^p\right| = \frac{2^{d/2}2^{p+1}\Gamma(p+1)}{\pi}\left|\sin\frac{p\pi}{2}\right|\int_0^\infty\frac{\sin^d t}{t^{p+1}}dt.$$

Since (see, e.g., [16, p. 511])

$$\int_0^\pi \sin^d x\,dx = \frac{\sqrt{\pi}\Gamma(\frac{d+1}{2})}{\Gamma(\frac{d}{2}+1)} \geq \frac{C}{\sqrt{d}},$$

where $C > 0$ is an absolute constant, we have that

$$\int_0^\infty \frac{\sin^d t}{t^{p+1}}dt \geq \sum_{n=0}^\infty \frac{1}{((n+1)\pi)^{p+1}} \int_{n\pi}^{(n+1)\pi} \sin^d x\,dx \geq \frac{C}{\sqrt{d}} \cdot \frac{\zeta(p+1)}{\pi^{p+1}}.$$

Notice that $h(p) = \Gamma(p+1)\zeta(p+1)/(\pi/2)^{p+1}$ is a positive continuous function on $(0,\infty)$ and $h(p) \to \infty$ as $p \to \infty$ and $p \to 0^+$, it must hold that $\inf_{p>0} h(p) > 0$. The conclusion follows. ∎

3.2. Orthogonalizing rows. Suppose we are given a matrix $\Pi \in \mathbb{R}^{n\times n}$ in its spectral decomposition form $\Pi = H\Sigma H^T$, where

$$\Sigma_{i,i} = \begin{cases} \pm\sigma, & i \leq r, \\ 0, & r < i \leq n, \end{cases}$$

and H is the normalized Hadamard matrix. The goal of this section is to identify a set of orthogonal vectors, using rows of Π.

LEMMA 3.5. *Each row of* Π *has the same* ℓ_2 *norm* $\|\Pi_i\|_2 = \sigma\sqrt{r/n}$.

Proof.

$$\|\Pi_i\|_2 = \|H_i\Sigma H^T\|_2 = \|H_i\Sigma\|_2.$$

The lemma follows since all entries in H have absolute value $1/\sqrt{n}$, and the r nonzero entries on the diagonal of Σ have absolute value σ. ∎

To identify a set of orthogonal vectors using the rows of Π, we run a procedure similar to the standard Gram–Schmidt process.

LEMMA 3.6. *There is a set* $\mathcal{R} \subseteq [n]$ *with size* $|\mathcal{R}| = r/100$ *such that for each* $i \in \mathcal{R}$, Π_i *can be written as*

$$\Pi_i = R_i + P_i, \tag{3.5}$$

where $\{R_i\}_{i\in\mathcal{R}}$ *is a set of orthogonal vectors, and* P_i *is the orthogonal projection of* Π_i *onto the subspace spanned by* $\{R_j\}_{j\in\mathcal{R}\setminus\{i\}}$. *Furthermore, for each* $i \in \mathcal{R}$, $\|R_i\|_2^2 \geq 99/100\|\Pi_i\|_2^2 = 99/100 \cdot \sigma^2 r/n$.

Proof. We show how to construct such a set $\mathcal{R}$. Suppose that we have found a set $\mathcal{R}$ with size strictly less than $r/100$ with $\Pi_i = R_i + P_i$ satisfying the stated constraints. We shall show how to increase the size of $\mathcal{R}$ by one.

Let $\Pi = S + Q$. Here, for each $i \in [n]$ we have $\Pi_i = S_i + Q_i$, where Q_i is the orthogonal projection of Π_i onto the subspace spanned by $\{R_j\}_{j\in\mathcal{R}}$ and $S_i = \Pi_i - Q_i$. Notice that for all $j \in \mathcal{R}$ we have $Q_j = \Pi_j$ and $S_j = 0$. Since $\|\Pi\|_F^2 = r\sigma^2$ and $\operatorname{rank}(Q) \leq |\mathcal{R}|$, by Theorem 2.1 we have

$$\sum_{i\in[n]} \|S_i\|_2^2 = \|S\|_F^2 = \|\Pi - Q\|_F^2 \geq \|\Pi\|_F^2 - |\mathcal{R}| \cdot \sigma^2 > \frac{99}{100}\|\Pi\|_F^2 = \frac{99}{100}\sum_{i\in[n]} \|\Pi_i\|_2^2.$$

Thus, by averaging, there exists $i \notin \mathcal{R}$ such that $\|S_i\|_2^2 > 99/100\|\Pi_i\|_2^2$. We add i into $\mathcal{R}$ and set R_i in (3.5) to be S_i and P_i to be Q_i. It is easy to verify that the stated constraints still hold. We continue this process inductively until $|\mathcal{R}| = r/100$. ∎

LEMMA 3.7. *Let $x \in \mathbb{R}^n$ be a random vector defined as*

$$x = \sum_{i \in \mathcal{R}} s_i \cdot \frac{R_i}{\|R_i\|_2},$$

where $\{s_i\}_{i \in \mathcal{R}}$ is a set of i.i.d. Rademacher random variables. Here the set $\mathcal{R}$ and the orthogonal vectors $\{R_i\}_{i \in \mathcal{R}}$ are as defined in Lemma 3.6. Let $e \in \mathbb{R}^n$ be an arbitrary vector (that could depend on s_i's) satisfying that $\|e\|_\infty \le 0.1\sigma\sqrt{r/n}$. For each $i \in \mathcal{R}$, it holds that

$$\Pr_x \{\operatorname{sign}((\Pi x + e)_i) = \operatorname{sign}(s_i)\} \ge \frac{4}{5}.$$

Proof. For each $i \in \mathcal{R}$, we have

$$\langle \Pi_i, x \rangle = \langle R_i, x \rangle + \langle P_i, x \rangle = s_i \cdot \|R_i\|_2 + \sum_{j \in \mathcal{R} \setminus \{i\}} s_j \cdot \|\operatorname{proj}_{R_j} \Pi_i\|_2.$$

We first analyze the second term:

$$\mathbb{E}\,|\langle P_i, x \rangle| \le \left(\sum_{j \in \mathcal{R} \setminus \{i\}} \|\operatorname{proj}_{R_j} \Pi_i\|_2^2 \right)^{1/2} = \|P_i\|_2 \le \frac{1}{10}\|\Pi_i\|_2.$$

By Markov's inequality, with probability at least $4/5$, we have $|\langle P_i, x \rangle| \le \|\Pi_i\|_2/2$.

Recall that $\|R_i\|_2 \ge 99/100\|\Pi_i\|_2$ (Lemma 3.6) and $\|\Pi_i\|_2 = \sigma\sqrt{r/n}$ (Lemma 3.5). It happens with probability at least $4/5$ that $|e_i| + |\langle P_i, x \rangle| < |\langle R_i, x \rangle|$, in which case we have $\operatorname{sign}((\Pi x + e)_i) = \operatorname{sign}(s_i)$. ∎

3.3. Space lower bound on Q_p. In this section, we describe a reduction from the subspace sketch problem to the INDEX problem, a classical problem in communication complexity. We shall rephrase the problem in the context of a data structure. The INDEX data structure stores an input string $s \in \{-1, 1\}^n$ and supports a query function, which receives an input $i \in [n]$ and outputs $s_i \in \{-1, 1\}$ which is the ith bit of the underlying string. To prove the lower bound for the subspace sketch problem, we need the following lower bound for the distributional INDEX problem.

LEMMA 3.8 (see [31]). *In the* INDEX *problem, suppose that the underlying string s is drawn uniformly from $\{-1, 1\}^n$ and the input i of the query function is drawn uniformly from $[n]$. Any (randomized) data structure for* INDEX *that succeeds with probability at least $2/3$ requires $\Omega(n)$ bits of space, where the randomness is taken over both the randomness in the data structure and the randomness of s and i.*

Throughout the reduction, d is a fixed parameter with value to be determined later. For the matrix $M^{(d,p)}$, we consider its spectrum-truncated version

$$\tilde{M}^{(d,p)} \triangleq H^{(d)} \operatorname{diag}(\Lambda_1^{(d,p)}, \Lambda_2^{(d,p)} \ldots, \Lambda_{N^{(d)}}^{(d,p)}, 0, 0, \ldots, 0)(H^{(d)})^T.$$

LEMMA 3.9. *Each row of $\tilde{M}^{(d,p)}$ is orthogonal to all eigenvectors associated with eigenvalues other than $\Lambda_1^{(d,p)}, \Lambda_2^{(d,p)}, \ldots, \Lambda_{N^{(d)}}^{(d,p)}$.*

Proof. Let $v_1, \ldots, v_{2^d}$ be the columns of $H^{(d)}$. Then $\tilde{M}^{(d,p)} = \sum_{i=1}^{N^{(d)}} \Lambda_i^{(d,p)} v_i v_i^T$. Let w be an eigenvector corresponding to another eigenvalue. Then

$$\tilde{M}^{(d,p)} w = \sum_{i=1}^{N^{(d)}} \Lambda_i^{(d,p)} v_i (v_i^T w) = 0,$$

since v_i and w are orthogonal as they are associated with distinct eigenvalues. ∎

Now we invoke Lemma 3.6 on the matrix $\tilde{M}^{(d,p)}$ and obtain a set $\mathcal{R} \subseteq [2^d]$ and a set of orthogonal vectors $\{R_i\}_{i\in\mathcal{R}}$. We shall encode $|\mathcal{R}|$ random bits in A and show how to recover them.

Let

$$x = \sum_{i\in\mathcal{R}} s_i \cdot \frac{R_i}{\|R_i\|_2}.$$

By Lemma 2.7, with probability $1 - \exp(-\Omega(d))$, it holds that $\|x\|_\infty \le 3\sqrt{d}$. We condition on $\|x\|_\infty \le 3\sqrt{d}$ in the rest the proof, since we can include the alternative case $\|x\|_\infty > 3\sqrt{d}$ in the overall failure probability.

Next we define a vector $y \in \mathbb{R}^{2^d}$ to be $y_i = (x_i + \Delta^{(d)})^{1/p}$, where $\Delta^{(d)} = 5\sqrt{d}$ is a constant that depends only on d. Clearly, it holds for all $i \in [2^d]$ that $2\sqrt{d} \le y_i^p \le 8\sqrt{d}$. Round each entry of y to its nearest integer multiple of $\delta = 1/(p(8\sqrt{d})^{1-1/p}2^d)$, obtaining $\tilde{y}$. A simple calculation using the mean-value theorem shows that for all $i \in [2^d]$,

$$|\tilde{y}_i^p - (x_i + \Delta^{(d)})^p| = |\tilde{y}_i^p - y_i^p| \le p(8\sqrt{d})^{\frac{p-1}{p}}\delta \le 2^{-d}. \tag{3.6}$$

Finally we construct the matrix $A \in \mathbb{R}^{2^d\times d}$ to be used in the ℓ_p subspace sketch problem. The jth row of A is the jth vector of $\{-1,1\}^d$, scaled by $\tilde{y}_j$.

LEMMA 3.10. *The matrix A constructed above for the ℓ_p subspace sketch problem satisfies $\kappa(A) \le C$ for some constant C that depends on p only.*

Proof. Let B be the $2^d \times d$ matrix whose rows are all vectors in $\{-1,1\}^d$. Then,

$$\|Bx\|_2^2 = 2^d\,\mathbb{E}\left|\sum_{i=1}^d s_i x_i\right|^2,$$

where $s_1,\dots,s_d$ is a Rademacher sequence. It follows from Khintchine's inequality that

$$C_1 2^{d/2}\|x\|_2 \le \|Bx\|_2 \le C_2 2^{d/2}\|x\|_2 \tag{3.7}$$

for some constants C_1, C_2. Notice that the rows of A are rescaled rows of B with the scaling factors in $[(2\sqrt{d})^{1/p}, (8\sqrt{d})^{1/p}]$. Hence $\kappa(A) \le C$ for some constant C that depends on p only. ∎

The recovery algorithm is simple. The vector to be used for querying the data structure is the ith vector on the Boolean cube $\{-1,1\}^d$, where $i \in \mathcal{R}$. Given $Q_p(i)$, we guess the sign of s_i to be just the sign of $Q_p(i) - \langle M_i^{(d,p)}, \Delta^{(d)}\cdot\mathbf{1}\rangle$. Next we prove the correctness of the recovery algorithm.

For each $i \in \{-1,1\}^d$, the guarantee of the subspace sketch problem states that, with probability at least 0.9,

$$\|Ai\|_p^p \le Q_p(i) \le (1+\varepsilon)\|Ai\|_p^p. \tag{3.8}$$

We condition on this event in the remaining part of the analysis.

First we notice that

$$\|Ai\|_p^p = \sum_{j=1}^{2^d} |\langle A_j, i\rangle|^p = \sum_{j\in\{-1,1\}^d} |\langle i, \tilde{y}_j \cdot j\rangle|^p = \sum_{j\in\{-1,1\}^d} \tilde{y}_j^p |\langle i, j\rangle|^p. \tag{3.9}$$

Next we give an upper bound on the value of $\|Ai\|_p^p$.

LEMMA 3.11. *For each $i \in \{-1,1\}^d$, the matrix A constructed for the ℓ_p subspace sketch problem satisfies*

$$\|Ai\|_p^p \le 2^d \cdot (8d^{1.5})^p.$$

Proof. Each term in the summation (3.9) is upper bounded by $(8d^{1.5})^p$, which implies the stated lemma. ∎

Combining the preceding lemma with the query guarantee (3.8), the preceding lemma implies it holds that

$$|Q_p(i) - \|Ai\|_p^p| \le \varepsilon \cdot 2^d \cdot (8d^{1.5})^p.$$

On the other hand, by (3.6) and (3.9),

$$\left| \|Ai\|_p^p - \langle M_i^{(d,p)}, (x + \mathbf{1} \cdot \Delta^{(d)}) \rangle \right| \le \sum_{j \in \{-1,1\}^d} |\langle i, j \rangle|^p \cdot |\tilde{y}_i^p - (x_i + \Delta^{(d)})| \le d^p.$$

Thus by the triangle inequality,

$$\left| (Q_p(i) - \langle M_i^{(d,p)}, \Delta^{(d)} \cdot \mathbf{1} \rangle) - (\langle M_i^{(d,p)}, x \rangle) \right| \le \varepsilon \cdot 2^d \cdot (8d^{1.5})^p + d^p.$$

Notice that x is a linear combination of rows of $\tilde{M}^{(d,p)}$. By Lemma 3.9,

$$M^{(d,p)} x = \tilde{M}^{(d,p)} x.$$

By Lemma 3.7, if

$$\varepsilon \cdot 2^d \cdot (8d^{1.5})^p + d^p \le 0.1 \Lambda_0^{(d,p)} \sqrt{N^{(d)}} / 2^{d/2}, \tag{3.10}$$

then with probability $4/5$, $(Q_p(i) - \langle M_i^{(d,p)}, \Delta^{(d)} \cdot \mathbf{1} \rangle)$ has the same sign as $(\tilde{M}^{(d,p)} x)_i$, in which case we recover the correct sign. By Lemma 3.8, the size of Q_p is lower bounded by $\Omega(|\mathcal{R}|)$.

Now for each $\varepsilon > 0$ and $p \in (0, \infty) \setminus 2\mathbb{Z}$, by Lemma 3.2, (3.10) can be satisfied by setting

$$2^{d/2} \ge \frac{\sin(p\pi/2)}{\varepsilon \cdot \operatorname{polylog}(1/\varepsilon)},$$

which implies a space complexity lower bound of

$$\Omega(|\mathcal{R}|) = \Omega(N^{(d)}) = \Omega(2^d/\sqrt{d}) = \Omega\left(\frac{1}{\varepsilon^2 \cdot \operatorname{polylog}(1/\varepsilon)} \right)$$

bits.

Formally, we have proved the following theorem.

THEOREM 3.12. *Let $p \in (0, \infty) \setminus 2\mathbb{Z}$. There exist constants $C \in (0,1]$ and $\varepsilon_0 > 0$ that depend only on p such that the following holds. For any $\varepsilon \in (0, \varepsilon_0)$, $d \ge d_0$, and $n \ge 2^{d_0}$, any data structure for the ℓ_p subspace sketch requires $\Omega(\frac{1}{\varepsilon^2 \cdot \operatorname{polylog}(1/\varepsilon)})$ bits. The lower bound holds even when $\kappa(A) \le K$ for some constant K that only depends on p. Here $d_0 = 2\log_2(C/(\varepsilon \operatorname{polylog}(1/\varepsilon)))$.*

We note that the $\operatorname{polylog}(1/\varepsilon)$ factors in the definition of d_0 and the bit lower bound may not have the same exponent.

Next we strengthen the lower bound to $\widetilde{\Omega}(d/\varepsilon^2)$ bits.

COROLLARY 3.13. *Under the assumptions of C, ε_0, d, in Theorem* 3.12 *and the assumption that $n = \Omega(\frac{d}{\varepsilon^2 \cdot \mathrm{polylog}(1/\varepsilon)})$, any data structure for the ℓ_p subspace sketch problem requires $\Omega(\frac{d}{\varepsilon^2 \cdot \mathrm{polylog}(1/\varepsilon)})$ bits. The* $\mathrm{polylog}(1/\varepsilon)$ *factors may not have the same exponent in the two Ω-notations above.*

Proof. Let $A' \in \mathbb{R}^{n' \times d'}$ be the hard instance matrix for Theorem 3.12, where $d' = 2\log_2(C/(\varepsilon \,\mathrm{polylog}(1/\varepsilon))$ and $n' = 2^{d'}$. We construct a block diagonal matrix A with $b = d/d'$ blocks, each being an independent copy of A', so that A has d columns. The number of rows in A is $bn' = \Omega(\frac{d}{\varepsilon^2 \,\mathrm{polylog}(1/\varepsilon)})$. In this case, the ℓ_p sketch problem on A' requires a data structure of $\widetilde{\Omega}(b/\varepsilon^2) = \Omega(\frac{d}{\varepsilon^2 \cdot \mathrm{polylog}(1/\varepsilon)})$ bits, since we are now solving the INDEX problem with $\widetilde{\Omega}(b \cdot 1/\varepsilon^2)$ random bits. ∎

The corollary above is also true for $p = 0$.

COROLLARY 3.14. *Under the assumptions of C, ε_0, d, and n in Corollary* 3.13, *any data structure for the ℓ_0 subspace sketch problem requires $\Omega(\frac{d}{\varepsilon^2 \cdot \mathrm{polylog}(1/\varepsilon)})$ bits.*

Proof. The matrix $M^{(d,0)}$ is defined as $(M^{(d,0)})_{i,j} = \mathbf{1}_{\{\langle i,j\rangle \neq 0\}}$. Note that each row of $M^{(d,0)}$ has the same number of 1's; let W_d denote this number. Observe that Corollary 3.4 continues to hold because we have by symmetry

$$\sum_{k=1}^{n} (-1)^{k+1} \binom{2n}{n+k} = \frac{1}{2}\binom{2n}{n} \geq c\frac{2^{2n}}{\sqrt{n}}$$

for some absolute constant $c > 0$. Let $y_j = x_i + \Delta^{(d)}$, where x_i and $\Delta^{(d)}$ are as defined before. In the construction of A, replicate $\tilde{y}_j$ times (rounded to an integer multiple of $\delta = 2^{-d}$) the jth vector of $\{-1, 1\}^d$. Our guess of the sign s_i is then the sign of $\delta Q_0(i) - \Delta^{(d)} W_d$. Similarly to the procedure above, we have that

$$\delta \left| Q_0(i) - \|Ai\|_0 \right| \leq \delta\varepsilon \|Ai\|_0 \leq \varepsilon \cdot 8\sqrt{d} \cdot 2^d$$

and

$$\left| \delta\|Ai\|_0 - \langle M_i^{(d,0)}, (x + \mathbf{1} \cdot \Delta^{(d)}) \rangle \right| \leq \sum_j |\hat{y}_j - x_i - \Delta^{(d)}| \mathbf{1}_{\{\langle i,j\rangle \neq 0\}} \leq \delta 2^d = 1.$$

And therefore it suffices to have

$$\varepsilon \cdot 8\sqrt{d} \cdot 2^d + 1 \leq 0.1 \Lambda_0^{(d,p)} \sqrt{N^{(d)}} / 2^{d/2},$$

which holds when $2^{d/2} = 1/(\varepsilon/\,\mathrm{polylog}(1/\varepsilon))$ as before. Therefore the analogue of Theorem 3.12 holds and so does the analogue of Corollary 3.13. ∎

Remark 3.15. The condition that $p \notin 2\mathbb{Z}^+$ is necessary for the lower bound. When $p \in 2\mathbb{Z}^+$, it is possible to achieve $\varepsilon = 0$ with $O(d^p \log(nd))$ words. Recall that a d-dimensional subspace of ℓ_p space can be isometrically embedded into ℓ_p^r with $r = \binom{d+p-1}{p} - 1$ [25]. In general the data structure does not necessarily correspond to a linear map and can be of any form. Indeed, there is a much simpler data structure as follows, based on ideas in [38]. For each $x \in \mathbb{R}^d$, let $y_x \in \mathbb{R}^d$ be defined as $(y_x)_i = ((Ax)_i)^{p/2}$, then $\|y_x\|_2^2 = \|Ax\|_p^p$. Observe that each coordinate $(y_x)_i$ is a polynomial of $d^{p/2}$ terms in $x_1, \ldots, x_d$. Form an $n \times d^{p/2}$ matrix B, where the ith row consists of the coefficients in the polynomial corresponding to $(y_x)_i$. The data structure stores $B^T B$. To answer the query $Q_p(x)$, one first calculates from x a $d^{p/2}$-dimensional

vector x' whose coordinates are all possible monomials of total degree $p/2$. Note that $Bx' = y_x$. Hence one can just answer $Q_p(x) = (x')^T B^T Bx' = \|Bx'\|_2^2 = \|Ax\|_p^p$ without error. This Q_p does not give an isometric embedding but is much simpler than known isometric embeddings, and the space complexity is $O(d^p \log(nd))$ bits.

4. Lower bounds for $p > 2$.

4.1. Lower bounds for the subspace sketch problem for $p > 2$.

In this section, we prove a lower bound on the ℓ_p subspace sketch problem, in the case that ε is a constant and $p \geq 2$. We need the following result from coding theory.

LEMMA 4.1 (see [35]). *For any $p \geq 1$ and $d = 2^k - 1$ for some integer k, there exist a set $S \subset \{-1, 1\}^d$ and a constant C_p depending only on p which satisfy*

(i) $|S| = d^p$;

(ii) *for any $s, t \in S$ such that $s \neq t$, $|\langle s, t\rangle| \leq C_p\sqrt{d}$.*

LEMMA 4.2. *For any $p \geq 1, C \geq 1$, and $d = 2^k - 1$ for an integer k, there exist a set $S \subset \{-1, 1\}^d$ with size $|S| = d^p$, a set $\mathcal{M} \subset \mathbb{R}^{R\times d}$ for some R, and a constant C_p depending only on p which satisfy*

(i) *for any $M_1, M_2 \in \mathcal{M}$ such that $M_1 \neq M_2$, there exists $x \in S$ such that $\|M_2 x\|_p < d/C$ and $\|M_1 x\|_p \geq d$;*

(ii) $|\mathcal{M}| \geq \exp\left(d^{p/2}/(C_p C^p)\right)$.

Proof. Set $R = d^{p/2}/(C_p C^p)$. Then $R \leq d^p/e$. We set $\mathcal{M}$ to be the set of $R \times d$ matrices whose rows are all possible combinations of R distinct vectors in S, where S is the set constructed in Lemma 4.1. Clearly, $|\mathcal{M}| = \binom{d^p}{R} \geq e^R$. Furthermore, consider two different $M_1, M_2 \in M$. There exists an $x \in S$ which is a row of M_1 but not a row of M_2. Thus, $\|M_1 x\|_p \geq d$ and

$$\|M_2 x\|_p \leq C_p\sqrt{d} \cdot R^{1/p} < d/C. \qquad \square$$

THEOREM 4.3. *Solving the ℓ_p subspace sketch problem requires $\widetilde{\Omega}(d^{p/2})$ bits when $0 < \varepsilon < 1$ and $p \geq 2$ are constants and $n = \Omega(d^{p/2})$.*

Proof. We first prove a lower bound for randomized data structures for the ℓ_p subspace sketch problem with failure probability $d^{-p}/100$. Let $\mathcal{M} \subset \mathbb{R}^{R\times d}$ and $S \subset \{-1, 1\}^d$ be as constructed in Lemma 4.2. Choose a matrix M from $\mathcal{M}$ uniformly at random. Since for each $x \in \{-1, 1\}^d$, with probability at least $1 - d^{-p}/100$,

$$\|Mx\|_p^p \leq Q_p(x) \leq (1 + O(\varepsilon))\|Mx\|_p^p, \tag{4.1}$$

by a union bound, with probability at least 0.99, (4.1) holds simultaneously for all $x \in S$. It follows from Lemma 4.2(i) that by querying $\|Mx\|_p$ for all $x \in S$, one can distinguish all different $M \in \mathcal{M}$. A standard information-theoretic argument leads to a lower bound of $\Omega(\log|\mathcal{M}|) = \Omega(d^{p/2})$.

For randomized data structures for the ℓ_p subspace sketch problem with constant failure probability, a standard repetition argument implies that the failure probability can be reduced to $d^{-p}/100$ using $O(\log d)$ independent repetitions. Therefore a lower bound of $\widetilde{\Omega}(d^{p/2})$ bits follows. $\square$

Remark 4.4. The lower bound in Theorem 4.3 is nearly optimal. To obtain an ℓ_p subspace sketch with constant ε and $\widetilde{O}(d^{p/2})$ bits, one can first apply Lewis weights sampling [15] to reduce the size of A to $\widetilde{O}(d^{p/2}) \times d$, and then apply the embedding

in [18] to further reduce the number of rows of A to $\widetilde{O}(d^{(p/2)\cdot(1-2/p)}) = \widetilde{O}(d^{p/2-1})$. Therefore the data structure takes $\widetilde{O}(d^{p/2})$ bits to store.

4.2. Lower bounds for the for-all version. In this section, we prove a lower bound on the for-all version of the ℓ_p subspace sketch problem for the case of $p \geq 2$ and constant ε. In the for-all version of the ℓ_p subspace sketch problem, the data structure Q_p is required to, with probability at least 0.9, satisfy $Q_p(x) = (1 \pm \varepsilon)\|Ax\|_p$ simultaneously for all $x \in \mathbb{R}^d$.

4.2.1. Lower bound for $p \geq 2$. Throughout this section we assume that $p \geq 2$ is a constant.

Let $N = c_p d^{p/2}$ in this section, where $c_p > 0$ is a constant that depends only on p. Denote the unit ball in ℓ_p^n by B_p^n. For each $x \in B_2^n$, we define a function $f_x : \mathbb{R}^{N\times d} \to \mathbb{R}$ by

$$f_x(A) = \|Ax\|_p.$$

LEMMA 4.5. *The function $f_x(\cdot)$ satisfies the following properties:*

(i) $\mathbb{E}[f_x(A)] \leq C c_p^{1/p}\sqrt{p}\sqrt{d}$, *where entries of A are i.i.d. Rademacher random variables and C is an absolute constant;*

(ii) $f_x(\cdot)$ *is 1-Lipschitz with respect to the Frobenius norm;*

(iii) $f_x(\cdot)$ *is a convex function.*

Proof. By Khintchine's inequality, $(\mathbb{E}\,|(Ax)_i|^p)^{1/p} \leq C\sqrt{p}\|x\|_2 = C\sqrt{p}$, where C is an absolute constant. It follows that $\mathbb{E}\,\|Ax\|_p^p \leq N(C\sqrt{p})^p$ and by Jensen's inequality, $\mathbb{E}\,\|Ax\|_p \leq (\mathbb{E}\,\|Ax\|_p^p)^{1/p} \leq N^{1/p}C\sqrt{p} = Cc_p^{1/p}\sqrt{p}\sqrt{d}$, which implies (i). To prove (ii), note that

$$f_x(A-B) = \|Ax - Bx\|_p \leq \|Ax - Bx\|_2 \leq \|A-B\|_2 \leq \|A-B\|_F.$$

(iii) is a simple consequence of the convexity of the ℓ_p norm. ∎

The following lemma is a direct application of Talagrand's concentration inequality (Lemma 2.4) with Lemma 4.5.

LEMMA 4.6. *Let $A \in \mathbb{R}^{N\times d}$ and $x \in \mathbb{R}^d$ have i.i.d. Rademacher random variables. It holds that*

$$\Pr_{A,x}\left\{f_x(A) \geq Cc_p^{1/p}\sqrt{p}d\right\} \leq e^{-cd},$$

where C is an absolute constant and c_p is a constant depending only on p.

Proof. Let $\hat{x} = x/\sqrt{d}$. We have $\|\hat{x}\|_2 = 1$. By Lemmas 4.5 and 2.4, we have

$$\Pr_{A}\left\{f_{\hat{x}}(A) \geq Cc_p^{1/p}\sqrt{p}\sqrt{d}\right\} \leq e^{-cd}.$$

Since $f_x(A) = \sqrt{d}f_{\hat{x}}(A)$, we have

$$\Pr_{A}\left\{f_x(A) \geq Cc_p^{1/p}\sqrt{p}d\right\} \leq e^{-cd},$$

which implies the stated lemma. ∎

LEMMA 4.7. *There exists a set $\mathcal{S} \subseteq \{+1,-1\}^{N\times d}$ such that*

(i) $|\mathcal{S}| \geq \exp(c_1 Nd)$;

(ii) *for any $S,T \in \mathcal{S}$ such that $S \neq T$, there exists $i \in [N]$, such that $\|ST_i\|_p \leq Cc_p^{1/p}\sqrt{p}d$;*

(iii) *when $p > 2$, for any $S \in \mathcal{S}$, $\kappa(S) \leq 2$.*

Proof. We first define a set of bad matrices $\mathsf{Bad} \subseteq \{+1,-1\}^{N\times d}$ to be

$$\mathsf{Bad} = \left\{ A \in \{+1,-1\}^{N\times d} : \Pr_x \left\{ \|Ax\|_p \geq C c_p^{1/p} \sqrt{p} d \right\} \geq 3e^{-cd} \right\},$$

where $x \in \{+1,-1\}^d$ is an i.i.d. Rademacher vector and C_p, c are the same constants in Lemma 4.6. It follows from Lemma 4.6 that

$$\Pr_A\{A \in \mathsf{Bad}\} \leq \frac{1}{3},$$

since otherwise

$$\begin{aligned}\Pr_{A,x}\left\{ f_x(A) \geq C c_p^{1/p}\sqrt{p}d \right\} &\geq \Pr_A\{A \in \mathsf{Bad}\} \cdot \Pr_x\left\{ f_x(A) \geq C c_p^{1/p}\sqrt{p}d \mid A \in \mathsf{Bad} \right\} \\ &> e^{-cd}.\end{aligned}$$

Let the multiset $\mathcal{T} \subseteq \{+1,-1\}^{N\times d}$ of size $|\mathcal{T}| = \exp(c_2 N d)$ consist of independent uniform samples of matrices in $\{+1,-1\}^{N\times d}$. We define three events as follows.

- $\mathcal{E}_1$: There are at least $(1/12)|\mathcal{T}|$ distinct matrices in $\mathcal{T} \setminus \mathsf{Bad}$.
- $\mathcal{E}_2$: For each $S \in \mathcal{T} \setminus \mathsf{Bad}$ and each $T \in \mathcal{T} \setminus \{S\}$, there exists some $i \in [N]$ such that $\|ST_i\|_p \leq Cpd$.
- $\mathcal{E}_3$: There are at least $(23/24)|\mathcal{T}|$ matrices $T \in \mathcal{T}$ such that $\kappa(T) \leq 2$.

We analyze the probability of each event below.

First, notice that $\mathbb{E}\,|\mathcal{T} \cap \mathsf{Bad}| \leq |\mathcal{T}|/3$. Thus, by Markov's inequality we have $\Pr(|\mathcal{T} \cap \mathsf{Bad}| \geq 2|\mathcal{T}|/3) \leq 1/2$. Let X denote the number of distinct matrices in T. Note that $\mathbb{E}\,X = 2^{Nd}(1-(1-2^{-Nd})^{|\mathcal{T}|}) \approx |\mathcal{T}|$ when c_2 is small. It follows from a standard balls-into-bins argument with the bounded difference method (see, e.g., [17, section 6.3]) that $\Pr(X < (3/4)\,\mathbb{E}\,X) \leq 2\exp(-(\mathbb{E}\,X)^2/(8|\mathcal{T}|)) < 0.01$. This implies that $\Pr(\mathcal{E}_1^c) \leq 1/2 + 0.01 = 0.51$.

Next, consider a fixed matrix $S \in \{+1,-1\}^{N\times d} \setminus \mathsf{Bad}$. For a random matrix $T \in \{+1,-1\}^{N\times d}$ whose entries are i.i.d. Rademacher random variables, for each row T_i of T, by the definition of Bad, we have

$$\Pr\left\{ \|ST_i\|_p \geq C c_p^{1/p}\sqrt{p}d \right\} \leq 3e^{-cd}.$$

Since the rows of T are independent,

$$\Pr\left\{ \|ST_i\|_p \geq C c_p^{1/p}\sqrt{p}d,\ \forall i \in [N] \right\} \leq 3^N e^{-cNd} \leq e^{-c'Nd}.$$

Choosing appropriate constants for C and c (and thus c') allows for a union bound over all pairs $S \in \mathcal{T} \setminus \mathsf{Bad}$ and $T \in \mathcal{T} \setminus \{S\}$, and we have $\Pr(\mathcal{E}_2^c) \leq 1/3$.

Last, for the condition number, recall the classical result that for a random matrix T of i.i.d. Rademacher entries, it holds with probability $\geq 1 - \exp(-c_3 d)$ that $s_{\min}(T) \geq \sqrt{N} - c_4\sqrt{d}$ and $s_{\min}(T) \geq \sqrt{N} + c_4\sqrt{d}$, which implies that $\kappa(T) \leq (\sqrt{N} + c_4\sqrt{d})/(\sqrt{N} - c_4\sqrt{d}) \leq 2$ when d is sufficiently large. Letting $\mathcal{T}_1 = \{T \in |\mathcal{T}| : \kappa(T) > 2\}$, we have $\mathbb{E}\,|\mathcal{T}_1| \leq e^{-c_3 d}|\mathcal{T}|$. Thus by a Markov bound, $\Pr\{|\mathcal{T}_1| \geq 6e^{-c_3 d}|\mathcal{T}|\} \leq 1/10$, and thus $\Pr(\mathcal{E}_3^c) \leq 1/10$.

Since $\Pr(\mathcal{E}_1^c) + \Pr(\mathcal{E}_2^c) + \Pr(\mathcal{E}_3^c) < 1$, there exists a set $\mathcal{T}$ for which all $\mathcal{E}_1$, $\mathcal{E}_2$, $\mathcal{E}_3$ hold. Taking $\mathcal{S}$ to be the distinct well-conditioned matrices in $\mathcal{T} \setminus \mathsf{Bad}$, we see that $\mathcal{S}$ satisfies conditions (i)–(iii). ∎

THEOREM 4.8. *The for-all version of the ℓ_p subspace sketch problem requires $\Omega((d/p)^{p/2} \cdot d)$ bits to solve when $p \geq 2$ and $\varepsilon < 1$ are constants and $n = \Omega(d^{p/2})$. The lower bound holds even when $\kappa(A) \leq 2$ if $p > 2$, and all entries in A are in $\{+1, -1\}$.*

Proof. Choose a matrix A uniformly at random from the set $\mathcal{S}$ in Lemma 4.7. Suppose that $Q_p : \mathbb{R}^d \to \mathbb{R}$ satisfies

$$\|Ax\|_p^p \leq Q_p(x) \leq (1 + O(\varepsilon))\|Ax\|_p^p, \quad x \in \mathbb{R}^d.$$

For any row $A_i \in \mathbb{R}^d$ of A, interpreted as a column vector, $\|AA_i\|_p \geq d$, whereas for any $B \in \mathcal{S} \setminus \{A\}$, there exists a row A_i of A such that $\|BA_i\|_p \leq Cc_p^{1/p}\sqrt{p}d < d/3$, provided that c_p is small enough. Thus, by appropriate choice of the constants in Lemma 4.7, we can use Q_p to determine which matrix $A \in \mathcal{S}$ has been chosen. By property (ii) of the set $\mathcal{S}$, it must hold that all elements of $\mathcal{S}$ are distinct from each other. It then follows from a standard information-theoretic argument that the size of the data structure for the ℓ_p sketch problem is lower bounded by $\Omega(\log|\mathcal{S}|) = \Omega(Nd) = \Omega((d/p)^{p/2} \cdot d)$. ∎

4.2.2. Lower bound for $1 \leq p \leq 2$. The lower bound for $1 \leq p < 2$ follows from the lower bound for $p = 2$ by embedding ℓ_p into ℓ_2. It is known that ℓ_2^n K-embeds into ℓ_p^m for some $m \leq cn$, where $c = c(p)$ and $K = K(p)$ are constants that depend only on p. Furthermore, the embedding $T : \ell_2^n \to \ell_p^m$ can be realized using a rescaled matrix of i.i.d. Rademacher entries (with high probability). See [29, section 2.5] for a proof for $p = 1$, which can be generalized easily to a general p. Thus, one can reduce the for-all version of the ℓ_2 subspace sketch problem to the for-all version of the ℓ_p subspace sketch with $1 \leq p \leq 2$. Thus the lower bound of $\Omega(d^2)$ also holds when $1 \leq p \leq 2$.

5. Linear embeddings. In this section, our goal is to show that isomorphic embeddings into low-dimensional spaces induce solutions to the subspace sketch problem. Therefore a lower bound on the subspace sketch problem implies a lower bound on the embedding dimension.

THEOREM 5.1. *Let $p, q \geq 1$, $\varepsilon > 0$, and $A \in \mathbb{R}^{N \times d}$ with full column rank. Let $E \subseteq \ell_p^N$ be the column space of A and suppose that $T : E \to \ell_q^n$ is a $(1 + \varepsilon)$-isomorphic embedding. Then there exists a data structure for the for-all version of the ℓ_p subspace sketch problem on A with approximation ratio $1 \pm 6p\varepsilon$ and $O(nd \log(N^{|1/p-1/2|} dn\kappa(A)/\varepsilon))$ bits.*

Proof. Without loss of generality, we may assume that $\frac{1}{\kappa(A)}\|x\|_2 \leq \|Ax\|_2 \leq \|x\|_2$. Let $B \in \mathbb{R}^{n \times d}$ be such that $B = TA$. Then $\|Ax\|_p^p \leq \|Bx\|_q^p \leq (1 + \varepsilon)^p\|Ax\|_p^p$. Round each entry of B to an integer multiple of $\delta = \varepsilon/(D_1 n^{1/q} d^{1/2} \kappa(A))$, where $D_1 = \max\{1, N^{1/2-1/p}\}$, obtaining $\tilde{B}$. First we claim that the rounding causes only a minor loss,

$$(1 - \varepsilon)^p\|Bx\|_q^p \leq \|\tilde{B}x\|_q^p \leq (1 + \varepsilon)^p\|Bx\|_q^p. \tag{5.1}$$

Indeed, write $B = \tilde{B} + \Delta B$, where each entry of ΔB is bounded by δ. Then

$$\begin{aligned}\|(\Delta B)x\|_q \leq n^{1/q} d^{1/2}\delta\|x\|_2 \leq n^{1/q} d^{1/2}\delta\|x\|_2 &\leq n^{1/q} d^{1/2}\delta \cdot \kappa(A) \cdot \|Ax\|_2 \\ &\leq \varepsilon\|Ax\|_p \\ &\leq \varepsilon\|Bx\|_q.\end{aligned}$$

This proves (5.1) and so

$$(1-\varepsilon)^p\|Ax\|_p^p \le \|\tilde{B}x\|_q^p \le (1+\varepsilon)^{2p}\|Ax\|_p^p,$$

which implies that the matrix $\tilde{B}$ can be used to solve the ℓ_p subspace sketch problem on A with approximation ratio $1 \pm 6p\varepsilon$. Since

$$\|Bx\|_q \le (1+\varepsilon)^p\|Ax\|_p \le (1+\varepsilon)^p D_2\|Ax\|_2 \le (1+\varepsilon)^p D_2\|x\|_2,$$

where $D_2 = \max\{1, N^{1/p-1/2}\}$, each entry of B is at most eD_2. Hence, after rounding, each entry of $\tilde{B}$ can be described in $O(\log(D_2/\delta)) = O(\log(dnD\kappa(A)/\varepsilon))$ bits, where $D = D_1D_2 = N^{|1/2-1/p|}$. The matrix $\tilde{B}$ can be described in $O(nd\log(D_2/\delta))$ bits. The value of δ can be described in $O(\log(1/\delta))$ bits, which is dominated by the complexity for describing $\tilde{B}$. Therefore the size of the data structure is at most $O(nd\log(D_2/\delta)) = O(nd\log(Ddn\kappa(A)/\varepsilon))$ bits. ∎

The dimension lower bound for linear embeddings now follows as a corollary from combining the preceding theorem with Theorem 3.12, where we choose $d = C\log(1/\varepsilon)$ and note that $N = O(1/\varepsilon^2)$ and $\kappa(A) = O(1)$ in our hard instance.

COROLLARY 5.2. *Let $p \in [1,\infty) \setminus 2\mathbb{Z}$ and suppose that $d \ge C\log(1/\varepsilon)$. It holds that*

$$N_p(d,\varepsilon) \ge c_p \cdot 1/(\varepsilon^2 \cdot \operatorname{polylog}(1/\varepsilon)),$$

where $c_p > 0$ is a constant that depends only on p and $C > 0$ is an absolute constant.

Remark 5.3. It is not clear how much the assumption $d \ge C\log\frac{1}{\varepsilon}$ can be weakened. The best known results for $p = 1$ are as follows [25]:

$$N_1(d,\varepsilon) \le \begin{cases} c_2\varepsilon^{-1/2}, & d = 2, \\ c(d)\left(\frac{1}{\varepsilon^2}\log\frac{1}{\varepsilon}\right)^{(d-1)/(d+2)}, & d = 3,4, \\ c(d)\left(\frac{1}{\varepsilon^2}\right)^{(d-1)/(d+2)}, & d \ge 5, \end{cases}$$

which is substantially better than $1/(\varepsilon^2\operatorname{polylog}(1/\varepsilon))$ for constant d. In a similar lower bound [9],

$$N_1(d,\varepsilon) \ge c(d)\varepsilon^{-2(d-1)/(d+2)},$$

where $c(d) \approx e^{-c'd\ln d}$, so the lower bound is nontrivial only when $d = O\left(\log\frac{1}{\varepsilon}/\log\log\frac{1}{\varepsilon}\right)$. Since $N_1(d,\varepsilon)$ is increasing, optimizing d w.r.t. ε yields that

$$N_1(d,\varepsilon) = \Omega\big(\varepsilon^{-2}\exp(-c''\sqrt{\ln(1/\varepsilon)\ln\ln(1/\varepsilon)})\big)$$

for all $d = \Omega\big(\sqrt{\ln(1/\varepsilon)}\big)$. Our result improves the lower bound to $\varepsilon^{-2}/\operatorname{polylog}(1/\varepsilon)$ for larger d and, more importantly, works for general $p \ge 1$ that is not an even integer.

Remark 5.4. In the case of $p > 2$, it is an immediate corollary from Theorem 4.8 that $N_p(d,\varepsilon) = \Omega(d^{p/2}/\log d)$, which recovers the known (and nearly tight) lower bound up to a logarithmic factor.

6. Sampling-based embeddings. Our goal in this section is to prove the following lower bound.

THEOREM 6.1. *Let $p \ge 1$ and $p \notin 2\mathbb{Z}$. Suppose that $Q_p(x) = \|TAx\|_p^p$ solves the for-all version of the ℓ_p subspace sketch problem on A and ε for some $T \in \mathbb{R}^{m\times n}$ such that each row of T contains exactly one nonzero element. Then it must hold that $m \ge c_p d/(\varepsilon^2\operatorname{polylog}(1/\varepsilon))$, provided that $d \ge C_1\log(1/\varepsilon)$ and $n \ge C_2 d/(\varepsilon^2\operatorname{polylog}(1/\varepsilon))$, where $c_p > 0$ is a constant that depends only on p and $C_1, C_2 > 0$ are absolute constants.*

Proof. Let A be the hard instance matrix for the $\widetilde{\Omega}(1/\varepsilon^2)$ lower bound in Corollary 5.2. Recall that A is a block diagonal matrix with $k = \Theta(d/\log(1/\varepsilon))$ diagonal blocks. Each block has dimension $2^s \times s$ with $2^{s/2} = 1/\varepsilon^{1-o(1)}$. Furthermore, each block can be written as DB, where D is a $2^s \times 2^s$ diagonal matrix and B is the $2^s \times s$ matrix whose rows are all vectors in $\{-1, 1\}^s$, and each entry in D has magnitude in $[(2\sqrt{s})^{1/p}, (8\sqrt{s})^{1/p}]$. The matrix B can be described using $O(s2^s)$ bits and the matrix D using $O(2^s \log s)$ bits. Without loss of generality we may assume that each nonzero entry of T is an integer multiple of $\varepsilon^{1/p}$, since the loss of rounding, by the triangle inequality, is at most $\varepsilon\|Ax\|_p^p$. Next, we shall bound the number of bits needed to describe T.

Letting $x = e_j$ be a canonical basis vector,

$$(1+\varepsilon)\|Ax\|_p^p \geq \|TAx\|_p^p = \sum_{i=1}^{m} |t_i A_{i,j}|^p \geq 2\sqrt{s}\sum_{i=1}^{m} |t_i|^p.$$

On the other hand,

$$\|Ax\|_p^p \leq k \cdot 8\sqrt{s}\|Bx\|_p^p \leq 8k\sqrt{s}(2^{s/2}\|Bx\|_2)^p \leq C'k\sqrt{s}2^{sp},$$

where we used (3.7) for the last inequality. It follows from the AM–GM inequality that

$$\begin{aligned}\sum_i \log \frac{|t_i|}{\varepsilon^{1/p}} = \log \prod_i \frac{|t_i|}{\varepsilon^{1/p}} &\leq \frac{m}{p}\log\left(\frac{1}{m}\sum_i \frac{|t_i|^p}{\varepsilon}\right)\\ &\leq C''m\left(s + \log\frac{1}{\varepsilon} + \log\frac{k}{m}\right)\\ &\leq C'''ms,\end{aligned}$$

that is, T can be described using $O(ms)$ bits, provided that $m \geq k$. Therefore TA can be described in $O(s2^s + 2^s \log s + ms) = O((m + 1/\varepsilon^2)\log(1/\varepsilon))$ bits. Combining with the lower bound of $\Omega(d/(\varepsilon^2 \operatorname{poly}\log(1/\varepsilon)))$ bits, we have $m = \Omega(d/(\varepsilon^2 \operatorname{polylog}(1/\varepsilon)))$.

A similar argument shows that when $m < k$, the matrix T can be described in $O(d)$ bits, which leads to a contradiction to the lower bound. Hence it must hold that $m \geq k$ and, as we proved above, $m = \Omega(d/(\varepsilon^2 \operatorname{polylog}(1/\varepsilon)))$. ∎

The lower bound for the (for-each) ℓ_p subspace sketch problem loses further a factor of $\log d$.

COROLLARY 6.2. *Let $p \geq 1$ and $p \notin 2\mathbb{Z}$. Suppose that $Q_p(x) = \|TAx\|_p^p$ solves the (for-each) version of the ℓ_p subspace sketch problem on A and ε for some $T \in \mathbb{R}^{m\times n}$ such that each row of T contains exactly one nonzero element. Then it must hold that $m \geq c_p d/(\varepsilon^2 \cdot \log d \cdot \operatorname{polylog}(1/\varepsilon))$, provided that $d \geq C_1 \log(1/\varepsilon)$ and $n \geq C_2 d/(\varepsilon^2 \operatorname{polylog}(1/\varepsilon))$, where $c_p > 0$ is a constant that depends only on p and $C_1, C_2 > 0$ are absolute constants.*

Proof. Observe that we used the approximation to $\|Ae_i\|_p^p$ for each canonical basis vector e_i in the proof of Theorem 6.1, which holds with a constant probability if we make $O(\log d)$ independent copies of the (randomized) data structure. This incurs a further loss of a $\log d$ factor in the lower bound. ∎

7. Oblivious sketches. An oblivious subspace embedding for d-dimensional subspaces E in ℓ_p^n is a distribution on linear maps $T : \ell_p^n \to \ell_p^m$ such that it holds for

any d-dimensional subspace $E \subseteq \ell_p^n$ that

$$\Pr_T \{(1-\varepsilon)\|x\|_p \le \|Tx\|_p \le (1+\varepsilon)\|x\|_p \ \forall x \in E\} \ge 0.99.$$

More generally, an *oblivious sketch* is a distribution on linear maps $T : \ell_p^n \to \mathbb{R}^m$, accompanied by a recovery algorithm $\mathcal{A}$, such that it holds for any d-dimensional subspace $E \subseteq \ell_p^n$ that

$$\Pr_T \{(1-\varepsilon)\|x\|_p \le \mathcal{A}(Tx) \le (1+\varepsilon)\|x\|_p \ \forall x \in E\} \ge 0.99.$$

It is clear that an oblivious embedding is a special case of an oblivious sketch, where $\mathcal{A}(Tx) = \|Tx\|_p$.

In this section we shall show that when $1 \le p < 2$, any oblivious sketch requires $m = \widetilde{\Omega}(d/\varepsilon^2)$.

Before proving the lower bound, let us prepare some concentration results. We use $\mathbb{S}^{n-1}$ to denote the unit sphere in $(\mathbb{R}^n, \|\cdot\|_2)$. First, observe that the norm function $x \mapsto \|x\|_p$ is a Lipschitz function of Lipschitz constant $\max\{1, n^{1/p-1/2}\}$. Also note that $(\mathbb{E}_{g\sim N(0,I_n)} \|g\|_p^p)^{1/p} = \beta_p n^{1/p}$, where $\beta_p = (\mathbb{E}_{g\sim N(0,1)} |g|^p)^{1/p}$. Standard Gaussian concentration (Lemma 2.5) leads to the following.

LEMMA 7.1. *Let $p \ge 1$ be a constant and $g \sim N(0, I_n)$. It holds with probability at least $1 - \exp(-c\varepsilon^2 n^{\min\{1,2/p\}})$ that $(1-\varepsilon)\beta_p n^{1/p} \le \|g\|_p \le (1+\varepsilon)\beta_p n^{1/p}$, where $c = c(p) > 0$ is a constant that depends only on p.*

Suppose that G is an $n \times d$ Gaussian random matrix of i.i.d. $N(0,1)$ entries. Observe that for a fixed $x \in \mathbb{S}^{d-1}$, $Gx \sim N(0, I_n)$. A typical ε-net argument on $\mathbb{S}^{d-1}$ allows us to conclude the following lemma. We remark that this gives Dvoretzky's theorem for ℓ_p spaces.

LEMMA 7.2. *Let $1 \le p < 2$ be a constant and G be an $n \times d$ Gaussian random matrix. There exist constants $C = C(p) > 0$ and $c = c(p) > 0$ such that whenever $n \ge Cd\log(1/\varepsilon)/\varepsilon^2$, it holds $\Pr\{(1-\varepsilon)\beta_p n^{1/p} \le \|Gx\|_p \le (1+\varepsilon)\beta_p n^{1/p} \ \forall x \in \mathbb{S}^{d-1}\} \ge 1 - 2\exp(-c\varepsilon^2 n)$.*

Now, consider two distributions on $n \times d$ matrices, where $n = \Theta(d\varepsilon^{-2}\log(1/\varepsilon))$. The first distribution $\mathcal{L}_1$ is just the distribution of a Gaussian random matrix G of i.i.d. $N(0,1)$ entries, and the second distribution $\mathcal{L}_2$ is the distribution of $G + \sigma uv^T$, where G is the Gaussian random matrix of i.i.d. $N(0,1)$ entries, $u \sim N(0, I_n)$ and $v \sim N(0, I_d)$ and $\sigma = \alpha\sqrt{\varepsilon/d}$ for some constant α to be determined later, and G, u, and v are independent.

THEOREM 7.3. *Let $1 \le p < 2$ be a constant. Suppose that $S \in \mathbb{R}^{m\times n}$ is an oblivious sketch for d-dimensional subspaces in ℓ_p^n, where $n = \Omega(d\varepsilon^{-2}\log(1/\varepsilon))$. It must hold that $m \ge cd/\varepsilon^2$, where $c = c(p) > 0$ is a constant depending only on p.*

Proof. It follows from the preceding lemma that, if $A \sim \mathcal{L}_1$, we have that $\sup_{x\in\mathbb{S}^{d-1}} \|Ax\|_p \le (1+\varepsilon)\beta_p n^{1/p}$ with probability at least 0.999 with an appropriate choice of constant in the Θ-notation of n. Next we consider the supremum of $\|Ax\|_p$ when $A \sim \mathcal{L}_2$. Observe that

$$\sup_{x\in\mathbb{S}^{d-1}} \left\|(G+\sigma uv^T)x\right\|_p \ge \left\|(G+\sigma uv^T)\frac{v}{\|v\|_2}\right\|_p = \left\|G\frac{v}{\|v\|_2} + \sigma u\|v\|_2\right\|_p.$$

Since $v \sim N(0, I_d)$, the direction $v/\|v\|_2 \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ and the magnitude $\|v\|_2$ are independent, and by rotational invariance of the Gaussian distribution, $Gx \sim N(0, I_d)$

for any $x \in \mathbb{S}^{d-1}$. Hence

$$\left\| G\frac{v}{\|v\|_2} + \sigma u\|v\|_2 \right\|_p \overset{\text{dist}}{=\!=} \|u_1 + \sigma t u_2\|_p \overset{\text{dist}}{=\!=} \sqrt{1+\sigma^2 t^2}\|u\|_p,$$

where t follows the distribution of $\|v\|_2$ and u_1, u_2 are independent $N(0, I_n)$ vectors. Applying the preceding two lemmas, we see that with probability at least 0.998, it holds that $t \geq 0.99\sqrt{d}$ and $\|u\|_p \geq (1-\varepsilon)\beta_p n^{1/p}$. Therefore, when $A \sim \mathcal{L}_2$ with probability at least 0.998, we have $\sup_{x\in\mathbb{S}^{d-1}} \|Ax\|_p \geq \sqrt{1+0.99^2\alpha^2\varepsilon}(1-\varepsilon)\beta_p n^{1/p} \geq (1+4\varepsilon)\beta_p n^{1/p}$, for an appropriate choice of α.

Therefore with the corresponding recovery algorithm $\mathcal{A}$,

$$\Pr_{A\sim\mathcal{L}_1, S}\left\{ \sup_{x\in\mathbb{S}^{n-1}} \mathcal{A}(SAx) \leq (1+\varepsilon)^2\beta_p n^{1/p} \right\} \geq 0.9,$$

$$\Pr_{A\sim\mathcal{L}_2, S}\left\{ \sup_{x\in\mathbb{S}^{n-1}} \mathcal{A}(SAx) \geq (1+4\varepsilon)(1-\varepsilon)\beta_p n^{1/p} \right\} \geq 0.9,$$

which implies that the linear sketch S can be used to distinguish $\mathcal{L}_1$ from $\mathcal{L}_2$ by evaluating $\sup_{x\in\mathbb{S}^{d-1}} \mathcal{A}(SAx)$. It then follows from [28, Theorem 4] that the size of the sketch $md \geq c/\sigma^4 = c'd^2/\varepsilon^2$ for some absolute constants $c, c' > 0$, and thus $m \geq c'd/\varepsilon^2$. ∎

8. Lower bounds for M-estimators. The main theorem of this section is the following.

THEOREM 8.1. *Suppose there exist $\alpha, \lambda > 0$ and $p \in (0,\infty)\backslash 2\mathbb{Z}$ such that $\phi(t/\lambda) \sim \alpha|t|^p$ as $t \to \infty$ or $t \to 0$. When $d \geq C_1 \log(1/\varepsilon)$ and $n \geq C_2 d/(\varepsilon^2 \operatorname{polylog}(1/\varepsilon))$ for some absolute constants $C_1, C_2 > 0$, the subspace sketch problem for $\Phi(x) = \sum_{i=1}^n \phi(x_i)$ requires $\Omega(d/(\varepsilon^2 \operatorname{polylog}(1/\varepsilon)))$ bits.*

Proof. We reduce the problem to the ℓ_p subspace sketch problem. We prove the statement in the case of $t \to \infty$ below. The proof for the case of $t \to 0$ is similar.

For a given $\varepsilon > 0$, there exists M such that $(1-\varepsilon)\alpha|t|^p \leq \phi(t/\lambda) \leq (1+\varepsilon)\alpha|t|^p$ for all $|t| \geq M$. Let A be our hard instance for the ℓ_p subspace sketch problem in Theorem 3.12. Then each row of A is a $\{-1,1\}$-vector scaled by a factor of $\tilde{y}_i \geq \Delta$ for some $\Delta = \Omega\big(\log^{1/(2p)}(1/\varepsilon)\big)$. One can recover a random sign used in the construction of A by querying Ax for a $\{-1,1\}$-vector x. Therefore, if $(Ax)_i \neq 0$, it must hold that $|(Ax)_i| \geq \Delta$. This implies that there exists a scaling factor $\beta = M/\Delta$ such that $(1-\varepsilon)\alpha\|\beta Ax\|_p^p \leq \Phi(\lambda^{-1}\beta Ax) \leq (1+\varepsilon)\alpha\|\beta Ax\|_p^p$, that is, $\alpha^{-1}\beta^{-p}\Phi(\lambda^{-1}\beta Ax)$ is a $(1\pm\varepsilon)$-approximation to $\|Ax\|_p^p$ for $\{-1,1\}$-vectors x. The conclusion follows from Corollary 3.13 (which plants independent copies of hard instance A in diagonal blocks) and a rescaling of ε. ∎

We have the following immediate corollary.

COROLLARY 8.2. *When $d \geq C_1 \log(1/\varepsilon)$ and $n \geq C_2 d/(\varepsilon^2 \operatorname{polylog}(1/\varepsilon))$ for some absolute constants $C_1, C_2 > 0$, for the following functions ϕ, the subspace sketch problem requires $\Omega(d/(\varepsilon^2 \operatorname{polylog}(1/\varepsilon)))$ bits.*

- *(L_1-L_2 estimator) $\phi(t) = 2(\sqrt{1+t^2/2} - 1)$;*
- *(Huber estimator) $\phi(t) = t^2/(2\tau)\cdot \mathbf{1}_{\{|t|\leq\tau\}} + (|t| - \tau/2)\cdot \mathbf{1}_{\{|t|>\tau\}}$;*
- *(Fair estimator) $\phi(t) = \tau^2(|t|/\tau - \ln(1+|t|/\tau))$;*
- *(Tukey loss p-norm) $\phi(t) = |t|^p \cdot \mathbf{1}_{\{|t|\leq\tau\}} + \tau^p \cdot \mathbf{1}_{\{|t|>\tau\}}$.*

Now we prove the $\Omega(d/(\varepsilon^2 \operatorname{polylog}(1/\varepsilon)))$ lower bound for the subspace sketch problem for the Cauchy estimator $\phi(t) = (\tau^2/2)\ln(1 + (t/\tau)^2)$. First consider an auxiliary function $\phi_{\text{aux}}(t) = \ln|x| \cdot \mathbf{1}_{\{|x|\geq 1\}}$, for which we shall also have a lower bound of $\Omega(d/(\varepsilon^2 \operatorname{polylog}(1/\varepsilon))$ by following the approach in section 3 with some changes we highlight below. Instead of $M_{i,j}^{(d,p)} = |\langle i, j\rangle|^p$, we shall define $M_{i,j}^{(d,p)} = \phi_{\text{aux}}(\langle i, j\rangle)$, and we proceed to define $N^{(d)}$ and $\Lambda_0^{(d,p)}$ in the same manner. The following lemma is similar to Corollary 3.4, showing that this new matrix $M^{(d,p)}$ also has large singular values. The proof is postponed to section 8.1.

LEMMA 8.3. *Suppose that $d \in 8\mathbb{Z}$. Then $\Lambda_0^{(d,p)} \geq c2^{d/2}/\sqrt{d}$ for some absolute constant $c > 0$.*

Therefore, the entire lower bound argument in Corollary 3.14 goes through. We can then conclude that the subspace sketch problem for $\Phi_{\text{aux}}(x) = \sum_{i=1}^n \phi_{\text{aux}}(x_i)$ requires $\Omega(d/(\varepsilon^2 \operatorname{polylog}(1/\varepsilon))$ bits. Now, for the Cauchy estimator $\phi(t) = (\tau^2/2)\ln(1+(t/\tau)^2)$, note that $(1-\varepsilon)\tau^2\phi_{\text{aux}}(t) \leq \phi(\tau \cdot t) \leq (1+\varepsilon)\tau^2\phi_{\text{aux}}(t)$ for all sufficiently large t. It follows from a similar argument to the proof of Theorem 8.1 that the same lower bound continues to hold for the subspace sketch problem for the Cauchy estimator.

COROLLARY 8.4. *When $d \geq C_1 \log(1/\varepsilon)$ and $n \geq C_2 d/(\varepsilon^2 \operatorname{polylog}(1/\varepsilon))$ for some absolute constants $C_1, C_2 > 0$, for the Cauchy estimator $\phi(t) = (\tau^2/2)\ln(1 + (t/\tau)^2)$, the subspace sketch problem requires $\Omega(d/(\varepsilon^2 \operatorname{polylog}(1/\varepsilon)))$ bits.*

8.1. Proof of Lemma 8.3. Before starting, let us define a useful integral

$$I_n = \int_0^{\frac{\pi}{2}} \sin^{2n} t dt.$$

We shall repeatedly use the following classical result that (see, e.g., [16, p. 511])

$$I_n = \frac{\sqrt{\pi}\Gamma(n+\frac{1}{2})}{2n\Gamma(n)} = \frac{1}{2}\sqrt{\frac{\pi}{n}} + O\left(\frac{1}{n^{3/2}}\right), \quad n \to \infty. \tag{8.1}$$

Now we begin the proof of Lemma 8.3. Differentiate both sides of (3.3) w.r.t p:

$$\begin{aligned}
-\int_0^\infty \frac{(2\sin t)^{2n}\ln t}{t^{p+1}}dt &= 2\sum_{k=1}^n (-1)^k \binom{2n}{n+k}(2k)^p \ln(2k)\cos\left(\frac{\pi p}{2}\right)\Gamma(-p) \\
&\quad - 2\sum_{k=1}^n (-1)^k \binom{2n}{n+k}(2k)^p \frac{\pi}{2}\sin\left(\frac{\pi p}{2}\right)\Gamma(-p) \\
&\quad - 2\sum_{k=1}^n (-1)^k \binom{2n}{n+k}(2k)^p \cos\left(\frac{\pi p}{2}\right)\Gamma'(-p).
\end{aligned}$$

Invoke the reflection identity (3.4),

$$\begin{aligned}
\frac{\Gamma(p+1)}{\pi}\sin\left(\frac{\pi p}{2}\right)\int_0^\infty \frac{(2\sin t)^{2n}\ln t}{t^{p+1}}dt &= \sum_{k=1}^n (-1)^k \binom{2n}{n+k}(2k)^p \ln(2k) \\
&\quad - 2\sum_{k=1}^n (-1)^k \binom{2n}{n+k}(2k)^p \frac{\pi}{2}\frac{\sin^2\left(\frac{\pi p}{2}\right)}{\sin(p\pi)} \\
&\quad - \sum_{k=1}^n (-1)^k \binom{2n}{n+k}(2k)^p \frac{\Gamma'(-p)}{\Gamma(-p)}.
\end{aligned}$$

Letting $p \to 0^+$, we see that the middle term on the right-hand side vanishes, which implies that

$$\sum_{k=1}^{n}(-1)^k\binom{2n}{n+k}\ln k$$
$$= \lim_{p\to 0^+}\left[\frac{2^{2n}}{\pi}\sin\left(\frac{\pi p}{2}\right)\int_0^\infty \frac{(\sin t)^{2n}\ln t}{t^{p+1}}dt + \sum_{k=1}^{n}(-1)^k\binom{2n}{n+k}\left(\frac{\Gamma'(-p)}{\Gamma(-p)} - \ln 2\right)\right].$$

Note that letting $p \to 0^+$ in Lemma 3.3 leads to

$$\sum_{k=1}^{n}(-1)^{k+1}\binom{2n}{n+k} = \frac{2^{2n}}{\pi}\lim_{p\to 0^+}\sin\left(\frac{\pi p}{2}\right)\int_0^\infty \frac{\sin^{2n} t}{t^{p+1}}dt; \tag{8.2}$$

we thus obtain that

$$\begin{aligned}\sum_{k=1}^{n}(-1)^k\binom{2n}{n+k}\ln k &= \frac{2^{2n}}{\pi}\lim_{p\to 0^+}\sin\left(\frac{\pi p}{2}\right)\int_0^\infty\left(\frac{(\sin^{2n} t)\ln t}{t^{p+1}} - \left(\frac{\Gamma'(-p)}{\Gamma(-p)} - \ln 2\right)\frac{\sin^{2n} t}{t^{p+1}}\right)dt \\ &= \frac{2^{2n}}{2}\lim_{p\to 0^+} p\int_0^\infty\left(\frac{(\sin^{2n} t)\ln t}{t^{p+1}} - \left(\frac{\Gamma'(-p)}{\Gamma(-p)} - \ln 2\right)\frac{\sin^{2n} t}{t^{p+1}}\right)dt.\end{aligned} \tag{8.3}$$

We wish to show that the limit on the rightmost side is at least $c/\sqrt{n}$ for some absolute constant $c > 0$. Note that the left-hand side of (8.2) is

$$that\sum_{k=1}^{n}(-1)^{k+1}\binom{2n}{n+k} = \frac{1}{2}\binom{2n}{n} \sim \frac{1}{2\sqrt{\pi}}\cdot\frac{2^{2n}}{\sqrt{n}}, \quad n\to\infty,$$

thus

$$\lim_{p\to 0^+} p\int_0^\infty \frac{\sin^{2n} t}{t^{p+1}}dt \sim \frac{1}{\sqrt{\pi n}}, \quad n\to\infty.$$

Note the fact that $\Gamma'(x)/\Gamma(x) = -1/x - \gamma + o(1)$ as $x \to 0$ (e.g., plugging $n = 1$ into eq. (1.2.15) in [6, p. 13]), where $\gamma = 0.577\ldots$ is the Euler gamma constant. Thus (8.3) can be rewritten as

$$\begin{aligned}\sum_{k=1}^{n}(-1)^k\binom{2n}{n+k}\ln k &= \frac{2^{2n}}{2}\lim_{p\to 0^+} p\int_0^\infty\left(\frac{(\sin^{2n} t)\ln t}{t^{p+1}} - \left(\frac{1}{p} - \gamma - \ln 2\right)\frac{\sin^{2n} t}{t^{p+1}}\right)dt.\end{aligned} \tag{8.4}$$

Hence, letting

$$f_p(t) = \frac{p\ln t - 1}{t^{p+1}},$$

we see it suffices to show that

$$\lim_{p\to 0^+}\int_0^\infty f_p(t)\sin^{2n} t dt > -\frac{c}{\sqrt{n}} \tag{8.5}$$

for some constant $c \in \left(0, \frac{1}{\sqrt{\pi}}(\gamma + \ln 2)\right)$. Combining (8.4) and (8.2), we know that the limit in (8.5) must exist.

We split the integral into $[0, \pi]$ and $[\pi, \infty)$ and deal with each part separately.

LEMMA 8.5. *It holds that*

$$\lim_{p\to 0^+}\int_0^\pi f_p(t)\sin^{2n}t dt \geq -\frac{2}{\sqrt{\pi n}}+O\left(\frac{1}{n}\right).$$

Proof. Observe that

$$\lim_{p\to 0^+}\int_0^\pi \frac{p\ln t}{t^{p+1}}\sin^{2n}t dt = 0,$$

because the integrands, viewed as functions of (p,t), are bounded on $[0,1]\times[0,\pi]$, since $\sin^{2n}t \sim t^{2n}$ near $t=0$ and so $t=0$ is not a singularity. Furthermore,

$$\lim_{p\to 0^+}\int_0^\pi \frac{\sin^{2n}t}{t^{p+1}}dt = \int_0^\pi \frac{\sin^{2n}t}{t}dt$$

because the integrand is uniformly continuous on $[0,1]\times[0,\pi]$ and we can take the limit under the integral sign.

Hence, for the integral on $[0,\pi]$, we have

$$\lim_{p\to 0^+}\int_0^\pi f_p(t)\sin^{2n}t dt = -\int_0^\pi \frac{\sin^{2n}t}{t}dt.$$

We claim that

$$\int_0^\pi \frac{\sin^{2n}t}{t}dt \leq \frac{2}{\sqrt{\pi n}}+O\left(\frac{1}{n}\right). \tag{8.6}$$

First observe that

$$\int_{\pi/2}^\pi \frac{\sin^{2n}t}{t}dt \leq \frac{2}{\pi}\int_{\pi/2}^\pi \sin^{2n}t dt = \frac{2}{\pi}I_n \sim \frac{1}{\sqrt{\pi n}}$$

and

$$\int_0^1 \frac{\sin^{2n}t}{t}dt \leq \int_0^1 t^{2n-1}dt = \frac{1}{2n}.$$

Letting $\delta = \sqrt{(3\ln n)/(2n)}$, then for $t\in[1,\pi/2-\delta]$,

$$\sin^{2n}t \leq \sin^{2n}\left(\frac{\pi}{2}-\delta\right) \leq \left(1-\frac{\delta^2}{3}\right)^{2n} \leq e^{-\frac{2}{3}n\delta^2} = \frac{1}{n}$$

and, thus,

$$\int_1^{\frac{\pi}{2}-\delta}\frac{\sin^{2n}t}{t}dt \leq \int_1^{\frac{\pi}{2}-\delta}\sin^{2n}t dt \leq \left(\frac{\pi}{2}-1\right)\cdot\frac{1}{n} = O\left(\frac{1}{n}\right).$$

For the last part,

$$\int_{\frac{\pi}{2}-\delta}^{\frac{\pi}{2}}\frac{\sin^{2n}t}{t}dt \leq \frac{1}{\frac{\pi}{2}-\delta}\int_{\frac{\pi}{2}-\delta}^{\frac{\pi}{2}}\sin^{2n}t dt \leq \frac{1}{\frac{\pi}{2}-\delta}\cdot I_n \sim \frac{1}{\sqrt{\pi n}}.$$

The proof of (8.6) is complete. ∎

Now we deal with the integral on $[\pi, \infty)$, for which we have the following approximation.

LEMMA 8.6. *It holds for all small $p > 0$ that*

$$\int_{\pi}^{\infty} f_p(t) \sin^{2n} t dt = \sqrt{\frac{\pi}{n}} \sum_{k=1}^{\infty} f_p\left(\left(k + \frac{1}{2}\right)\pi\right) + o\left(\frac{1}{n}\right).$$

Proof. Consider

$$\begin{aligned}
&\left| \int_{k\pi}^{(k+1)\pi} \left[f_p(t) - f_p\left(\left(k + \frac{1}{2}\right)\pi\right)\right] \sin^{2n} t dt \right| \\
&= \left| \int_0^{\frac{\pi}{2}} \left[f_p\left(\left(k+\frac{1}{2}\right)\pi + t\right) + f_p\left(\left(k+\frac{1}{2}\right)\pi - t\right) - 2f_p\left(\left(k+\frac{1}{2}\right)\pi\right)\right] \sin^{2n}\left(\frac{\pi}{2} - t\right) dt \right| \\
&\leq \max_{t\in[k\pi,(k+1)\pi]} |f_p''(t)| \cdot \int_0^{\frac{\pi}{2}} t^2 \sin^{2n}\left(\frac{\pi}{2} - t\right) dt.
\end{aligned}$$

Hence

$$\left| \int_{\pi}^{\infty} f_p(t) \sin^{2n} t dt - \sum_{k=1}^{\infty} f_p\left(\left(k + \frac{1}{2}\right)\pi\right) \int_0^{\pi} \sin^{2n} t dt \right| \leq \sum_{k=1}^{\infty} \max_{t\in[k\pi,(k+1)\pi]} |f_p''(t)| \cdot J_n,$$

where

$$J_n = \int_0^{\frac{\pi}{2}} t^2 \sin^{2n}\left(\frac{\pi}{2} - t\right) dt.$$

We can calculate that

$$f_p''(t) = \frac{p(1+p)(2+p)\ln t - 3p(2+p) - 2}{t^{3+p}}$$

and thus for all small $p \geq 0$,

$$\sum_{k=1}^{\infty} \max_{t\in[k\pi,(k+1)\pi]} |f_p''(t)|$$

is uniformly bounded. Next we deal with J_n. Let $\delta > 0$ (which could depend on n) to be determined. Then we have

$$\begin{aligned}
J_n &= \int_0^{\frac{\pi}{2}} \left(\frac{\pi}{2} - t\right)^2 \sin^{2n} t dt \\
&\leq \int_0^{\frac{\pi}{2}-\delta} \left(\frac{\pi}{2} - t\right)^2 \sin^{2n}\left(\frac{\pi}{2} - \delta\right) dt + \int_{\frac{\pi}{2}-\delta}^{\frac{\pi}{2}} \delta^2 \sin^{2n} t dt \\
&\leq \sin^{2n}\left(\frac{\pi}{2} - \delta\right) \int_0^{\frac{\pi}{2}} \left(\frac{\pi}{2} - t\right)^2 dt + \delta^2 I_n \\
&\leq \left(1 - \frac{\delta^2}{3}\right)^{2n} \frac{\pi^3}{24} + \delta^2 I_n \\
&\leq \frac{\pi^3}{24} e^{-\frac{2}{3}n\delta^2} + \delta^2 I_n.
\end{aligned}$$

Recall that $I_n = (1/2)\sqrt{\pi/n} + O(1/n^{3/2})$. Taking $\delta = \sqrt{3(\ln n)/n}$, we see that $J_n = O((\ln n)/n^{3/2})$. The conclusion follows. ∎

In view of Lemmas 8.5 and 8.6, in order to show (8.5), it suffices to show that

$$\liminf_{p\to 0^+} A_p > \frac{2-\gamma-\ln 2}{\pi}, \tag{8.7}$$

where

$$A_p = \sum_{k=1}^{\infty} f_p\left(\left(k+\frac{1}{2}\right)\pi\right).$$

Let $N = [e^{1/p}/\pi]$, then $f_p(t) < 0$ and increasing when $t \in [\pi, N\pi]$ and $f_p(t) > 0$ when $t \in [(N+1)\pi, \infty)$. We then have

$$\begin{aligned}
A_p &= f_p\left(\frac{3}{2}\pi\right) + \sum_{k=2}^{N-1} f_p\left(\left(k+\frac{1}{2}\right)\pi\right) + f_p\left(\left(N+\frac{1}{2}\right)\pi\right) + \sum_{k=N+1}^{\infty} f_p\left(\left(k+\frac{1}{2}\right)\pi\right) \\
&\geq f_p\left(\frac{3}{2}\pi\right) + \sum_{k=2}^{N-1} f_p(k\pi) + f_p\left(\left(N+\frac{1}{2}\right)\pi\right) + \sum_{k=N+1}^{\infty} \frac{p\ln(k\pi)-1}{((k+1)\pi)^{p+1}} \\
&\geq f_p\left(\frac{3}{2}\pi\right) + \sum_{k=2}^{\infty} f_p(k\pi) - f_p(N\pi) - f_p((N+1)\pi) + f_p\left(\left(N+\frac{1}{2}\right)\pi\right) \\
&\qquad + \sum_{k=N+1}^{\infty}\left(\frac{p\ln(k\pi)-1}{((k+1)\pi)^{p+1}} - \frac{p\ln((k+1)\pi)-1}{((k+1)\pi)^{p+1}}\right) \\
&= f_p\left(\frac{3}{2}\pi\right) + B_p + C_p + D_p,
\end{aligned}$$

where

$$\begin{aligned}
B_p &= \sum_{k=2}^{\infty} f_p(k\pi), \\
C_p &= -f_p(N\pi) - f_p((N+1)\pi) + f_p\left(\left(N+\frac{1}{2}\right)\pi\right), \\
D_p &= \sum_{k=N+1}^{\infty} \frac{p\ln(1-\frac{1}{k+1})}{((k+1)\pi)^{p+1}}.
\end{aligned}$$

It is clear that $f_p(3\pi/2) \to -2/(3\pi)$ and $C_p \to 0$ as $p \to 0^+$. For D_p, note that the summands

$$\frac{p\ln(1-\frac{1}{k+1})}{((k+1)\pi)^{p+1}} \sim -\frac{p}{(k+1)^{p+2}\pi^{p+1}}, \quad k \to \infty,$$

we also have $D_p \to 0$ as $p \to 0^+$. With the help of Riemann zeta functions $\zeta(s) = \sum_{n=1}^{\infty} n^{-s}$, we can write

$$B_p = \frac{1 - p\ln\pi + (p\ln\pi - 1)\zeta(1+p) - p\zeta'(1+p)}{\pi^{p+1}}.$$

Recall the fact that $\zeta(1+p) = \frac{1}{p} + \gamma + f(p)$ for an analytic function f on $\mathbb{C}$ with $f(0) = 0$ (see, e.g., [6, p. 15]), we see that

$$\lim_{p\to 0^+} B_p = \frac{1-\gamma+\ln\pi}{\pi}.$$

Therefore

$$\liminf_{p\to 0^+} A_p - \frac{2-\gamma-\ln 2}{\pi} \geq \frac{1}{\pi}\left(-\frac{2}{3} + (1-\gamma+\ln\pi) - (2-\gamma-\ln 2)\right)$$
$$= \frac{\ln(2\pi) - \frac{5}{3}}{\pi} > 0,$$

establishing (8.7). The proof of Lemma 8.3 is now complete (recalling that $2n = d/2$ as in Corollary 3.4).

9. Lower bounds on coresets for projective clustering. We shall prove a lower bound of $\widetilde{\Omega}(kj/\varepsilon^2)$ bits for coresets for projective clustering. First we need a lemma which provides codewords to encode the clustering information.

LEMMA 9.1. *For any given integer $L \geq 1$ and even integer $D \geq 2$, there exists a set $S = \{(s_1, t_1), (s_2, t_2), \ldots, (s_m, t_m)\}$ of size $m \geq c2^D/\sqrt{D}$, where $s_i, t_i \in \mathbb{R}^D$ and $c > 0$ is an absolute such that*

- $\langle s_i, t_i\rangle = 0$,
- $\langle s_i, t_j\rangle \geq L^2$ *for* $i \neq j$,
- *all entries of s_i and t_i are in $\{0, L\}$.*

Proof. We first consider the case $L = 1$. Let $\{s_i\}$ be the set of all binary vectors with Hamming weight $D/2$, and $t_i = \mathbf{1}^D - s_i$, i.e., t_i is the complement of s_i. Thus, $\langle s_i, t_i\rangle = 0$ by construction. For any $i \neq j$, since $s_i \neq s_j$, and both s_i and s_j have Hamming weight $D/2$, we have $\langle s_i, t_j\rangle \geq 1$.

For a general L, we replace all entries of value 1 in the construction above with L. ∎

In the rest of the section, we also use an $n \times d$ matrix to represent a point set of size n in $\mathbb{R}^d$, where each row represents a point in $\mathbb{R}^d$.

Below we set up the framework of the hard instance for the projective subspace clustering problem. For a given k, choosing $D = O(\log k)$, we can obtain a set S of size k as guaranteed by Lemma 9.1. Suppose that $j \geq D+1$ and $d \geq j+1$. Without loss of generality we may assume that $d = j+1$, otherwise we just embed our hard instance in $\mathbb{R}^{j+1}$ into $\mathbb{R}^d$ by appending zero coordinates.

For a set $\mathcal{A}$ consisting of k matrices $A^{(1)}, A^{(2)}, \ldots, A^{(k)} \in \mathbb{R}^{n\times(j+1-D)}$, we form a point set $X = X(\mathcal{A}) \in \mathbb{R}^{nk\times d}$, whose rows are indexed by $(i, j) \in [k] \times [n]$ and defined as

$$X_{i,j} = \begin{pmatrix} s_i^T & A_j^{(i)} \end{pmatrix},$$

where $A_j^{(i)}$ denotes the jth row of $A^{(i)}$.

Suppose that $y \in \mathbb{R}^{j+1-D}$. For each $i \in [k]$, let $V_i, W_i \subseteq \mathbb{R}^{j+1}$ be j-dimensional subspaces that satisfy

$$V_i \perp v_i, \quad v_i = \begin{pmatrix} t_i & \mathbf{0}^{j+1-D} \end{pmatrix},$$
$$W_i \perp w_i, \quad w_i = \begin{pmatrix} t_i & y \end{pmatrix},$$

where, for notational simplicity, we write vertical concatenation in a row. Last, for each $\ell \in [k]$, define a center

$$\mathcal{C}_\ell = (V_1, \ldots, V_{\ell-1}, W_\ell, V_{\ell+1}, \ldots, V_k).$$

LEMMA 9.2. *When $\|y\|_2 = 1$ and $L^2 \geq \max_i \|A_i^{(\ell)}\|_2$, it holds that* $\mathrm{cost}(X, \mathcal{C}_\ell) = \Phi(A^{(\ell)}y/\|w_\ell\|_2)$.

Proof. One can readily verify, using Lemma 9.1, that $P_{ij} \perp v_i$ whenever $i \neq \ell$, and thus $P_{ij} \in V_i'$ and $\operatorname{dist}(P_{ij}, X_\ell) = 0$ for $i \neq \ell$.

On the other hand, for $i \neq \ell$,

$$\operatorname{dist}(P_{\ell j}, V_i) = \frac{|\langle P_{\ell j}, v_i \rangle|}{\|v_i\|_2} = \frac{|\langle P_{\ell j}, v_i \rangle|}{L \cdot \sqrt{D/2}} \geq \frac{L}{\sqrt{D/2}}$$

and

$$\operatorname{dist}(P_{\ell j}, W_\ell) = \frac{|\langle P_{\ell j}, w_\ell \rangle|}{\|w_\ell\|_2} = \frac{|\langle A_i^{(\ell)}, y \rangle|}{\|w_\ell\|_2} \leq \frac{\|A_i^{(\ell)}\|_2 \|y\|_2}{\sqrt{\frac{D}{2} L^2 + \|y\|_2^2}}.$$

Hence when $L^2 \geq \|y\|_2 \max_i \|A_i^{(\ell)}\|_2$, it must hold that W_ℓ is the subspace in X_ℓ that is the closest to $P_{\ell j}$ for all j and, therefore,

$$\operatorname{cost}(X, \mathcal{C}_\ell) = \sum_{j=1}^{n} \phi(\operatorname{dist}(P_{\ell j}, W_\ell)) = \Phi\left(\frac{A^{(\ell)} y}{\|w_\ell\|_2}\right).$$

∎

THEOREM 9.3. *Suppose that there exists a function Φ and absolute constants C_0 and ε_0 such that for any $d \geq C_0 \log(k/\varepsilon)$ and $\varepsilon \in (0, \varepsilon_0)$, solving the subspace sketch problem for Φ requires M bits. Then there exists an absolute constant C_1 such that for any $k \geq 1$ and $j \geq C_1 \log(k/\varepsilon)$, any coreset for projective clustering for Φ requires kM bits.*

Proof. We prove this theorem by a reduction from the subspace sketch problem for Φ to coresets for projective clustering for Φ.

Choose $D = O(\log k)$ and $d' := j + 1 - D = C_0 \log(1/\varepsilon)$. Let $A^{(1)}, \ldots, A^{(k)} \in \mathbb{R}^{n \times d'}$ be k independent hard instances for the subspace sketch problem for Φ. Let X be as constructed before Lemma 9.2. If one can compute a projective clustering coreset for X so that one can approximate $\operatorname{cost}(X, \mathcal{C}_\ell)$ up to a $(1 \pm \varepsilon)$-factor, it follows from Lemma 9.2 that one can approximate $\Phi(A^{(\ell)} y / \|w\|_2)$ up to a $(1 \pm \varepsilon)$-factor for every $\ell \in [k]$ and every unit vector $y \in \mathbb{R}^{d'}$. Solving the subspace sketch problem for Φ for each $A^{(\ell)}$ requires M bits. Therefore, solving k independent instances requires kM bits. ∎

We have the following immediate corollary.

COROLLARY 9.4. *Under the assumptions of Theorem 9.3, any coreset for projective clustering requires $\Omega(jM/\log(k/\varepsilon))$ bits.*

Proof. Let $b = j/(C_0 \log(k/\varepsilon))$. Let X' be a block diagonal matrix of b blocks, each diagonal block is an independent copy of the hard instance X in Theorem 9.3. It then follows from Theorem 9.3 that the lower bound is $\Omega(bM)$ bits. ∎

A lower bound of $\Omega(jk/(\varepsilon^2 \log k \cdot \operatorname{polylog}(1/\varepsilon)))$ follows immediately for $\Phi(x) = \|x\|_p^p$ (Theorem 3.12) for $p \in [0, +\infty) \setminus 2\mathbb{Z}^+$, and the M-estimators in Corollaries 8.2 and 8.4.

10. Upper bounds for the Tukey loss p-norm. We shall prove in this section an $\widetilde{O}(1/\varepsilon^2)$ upper bound for estimating a mollified version of the Tukey loss p-norm $\Phi(x)$ for a vector $x \in \mathbb{R}^n$, where $p \in (0, 2]$.

10.1. Jacobi polynomials. Jacobi polynomials $P_n^{(\alpha,\beta)}(x)$ ($\alpha, \beta > -1$, $n = 0, 1, 2, \ldots$) are a class of orthogonal polynomials on $[-1, 1]$ with respect to the weight

function $w_{\alpha,\beta}(x) = (1-x)^\alpha(1+x)^\beta$, that is,

$$\int_{-1}^{1} P_n^{(\alpha,\beta)}(x)P_m^{(\alpha,\beta)}(x)(1-x)^\alpha(1+x)^\beta dx = 0, \quad n \neq m.$$

The convention is to take Jacobi polynomials as the following explicit expression (see, e.g., [39, eq. (4.3.2)]):

$$P_n^{(\alpha,\beta)}(x) = \sum_{\nu=0}^{n} \binom{n+\alpha}{n-\nu}\binom{n+\beta}{\nu}\left(\frac{x-1}{2}\right)^\nu\left(\frac{x+1}{2}\right)^{n-\nu}.$$

Hence $P_n^{(\alpha,\beta)}$ is a polynomial of degree n and

$$P_n^{(\alpha,\beta)}(1) = \binom{n+\alpha}{n}, \quad P_n^{(\alpha,\beta)}(-1) = (-1)^n\binom{n+\beta}{n}.$$

The normalization of Jacobi polynomials is given by

$$\int_{-1}^{1}\left[P_n^{(\alpha,\beta)}(x)\right]^2(1-x)^\alpha(1+x)^\beta dx = \frac{2^{\alpha+\beta+1}}{2n+\alpha+\beta+1}\cdot\frac{\Gamma(n+\alpha+1)\Gamma(n+\beta+1)}{n!\,\Gamma(n+\alpha+\beta+1)}.$$

Let $c_n^{(\alpha,\beta)}$ be the reciprocal of the right-hand side above such that

$$\left\{\sqrt{c_n^{(\alpha,\beta)}}P_n^{(\alpha,\beta)}(x)\right\}_{n\geq 0}$$

is orthonormal with respect to the weight function $w_{\alpha,\beta}(x)$.

The derivatives of a Jacobi polynomial are Jacobi polynomials with different parameters (cf. [39, eq. (4.21.7)]):

$$\frac{d^m}{dx^m}P_n^{(\alpha,\beta)}(x) = \frac{\Gamma(\alpha+\beta+n+1+m)}{2^m\Gamma(\alpha+\beta+n+1)}P_{n-m}^{\alpha+m,\beta+m}(x).$$

The maximum of the Jacobi polynomials on $[-1,1]$ is well known.

LEMMA 10.1 (see [39, Theorem 7.32.1]). *When $\alpha,\beta > -\frac{1}{2}$, $\max_{x\in[-1,1]}|P_n^{(\alpha,\beta)}(x)| = \binom{n+q}{n}$, where $q = \max\{\alpha,\beta\}$.*

A finer upper bound on the Jacobi polynomial on $(-1,1)$ is due to Nevai, Erdélyi, and Magnus [33].

LEMMA 10.2 (see [33]). *There exists an absolute constant $C > 0$ such that for all $\alpha,\beta > -\frac{1}{2}$ and all $n \geq 0$,*

$$\sup_{x\in[-1,1]}(1-x)^{\alpha+\frac{1}{2}}(1-x)^{\beta+\frac{1}{2}}\left(\sqrt{c_n^{\alpha,\beta}}P_n^{(\alpha,\beta)}\right)^2 \leq C(1+\sqrt{\alpha^2+\beta^2}).$$

10.2. Mollification of Tukey loss function. We want to construct a mollifier $\psi \in C_0^s(-1,1)$ which satisfies the moment conditions

$$\int_{-1}^{1} t^k\psi(t)dt = \begin{cases}1, & k=0,\\ 0, & k=1,2,\ldots,s.\end{cases} \tag{10.1}$$

Such mollifier can be constructed using Jacobi polynomials $P_s^{(s+1,\frac{1}{2})}(x)$ as

$$\psi(x) = c_s(1-x^2)^{s+1}P_s^{(s+1,\frac{1}{2})}(2x^2-1), \quad x \in (-1,1).$$

Observe that $\psi(x)$ is a polynomial consisting of even-degree terms only so the moment condition (10.1) is satisfied for all odd k. For even $k = 2\ell$, note that

$$\begin{aligned}\int_{-1}^{1} t^{2\ell}\psi(t)dt &= 2\int_{-1}^{1}(1-u)^{s+1}(1+u)^{\ell-\frac{1}{2}}P_s^{(s+1,\frac{1}{2})}(u)du \\ &= 2\int_{-1}^{1} w_{s+1,\frac{1}{2}}(u)\cdot(1+u)^{\ell-1}\cdot P_s^{(s+1,\frac{1}{2})}(u)du.\end{aligned}$$

When $1 \le \ell \le s$, the polynomial $(1+u)^{\ell-1}$ can be written as a linear combination of $\{P_r^{(s+1,\frac{1}{2})}\}_{r=0,\dots,\ell-1}$, hence the integral above is 0 from the orthogonality of Jacobi polynomials. This implies that the moments condition (10.1) for even $k \ge 2$ is satisfied. (This in fact implies the moment conditions hold up to $k = 2s$.) At last, one can choose the normalization factor c_s such that the moment condition is satisfied for $k = 0$ (see Lemma 10.3 below).

The following are probably classical results but we do not know an appropriate reference and so we produce full proofs here.

LEMMA 10.3. *It holds that* $c_s \sim (-1)^s\sqrt{\frac{2}{\pi}s}$.

Proof.

$$\begin{aligned}&\int_{-1}^{1}(1-x^2)^{s+1}P_s^{(s+1,\frac{1}{2})}(2x^2-1)dx \\ &= \frac{1}{\sqrt{2}\cdot 2^{s+1}}\int_{-1}^{1}(1-u)^{s+1}(1+u)^{-\frac{1}{2}}P_s^{(s+1,\frac{1}{2})}(u)du \\ &= \frac{1}{\sqrt{2}\cdot 2^{s+1}}\cdot(-1)^s\frac{2^{\frac{3}{2}+s}\sqrt{\pi}\Gamma(2s+2)}{\Gamma(2s+\frac{5}{2})} \\ &\sim (-1)^s\sqrt{\frac{\pi}{2s}},\end{aligned}$$

where we used the identity in [2] for the second equality. ∎

LEMMA 10.4. *It holds that*

$$\int_{-1}^{1}|\psi(x)|dx \le C\sqrt{s}\ln s.$$

Proof. Similarly to the proof of the preceding lemma, we have that

$$\int_{-1}^{1}|\psi(x)|dx = \frac{|c_d|}{\sqrt{2}\cdot 2^{s+1}}\int_{-1}^{1}(1-u)^{s+1}(1+u)^{-\frac{1}{2}}\left|P_s^{(s+1,\frac{1}{2})}(u)\right|du.$$

Break the integral on the right-hand side into two pieces on $[-1,-1+1/s]$ and

$[-1+1/s, 1]$. For the integral on $[-1, -1+1/s]$,

$$\begin{aligned}\int_{-1}^{-1+\frac{1}{s}} (1-u)^{s+1}(1+u)^{-\frac{1}{2}} \left|P_s^{(s+1,\frac{1}{2})}(u)\right| du &\le \int_{-1}^{-1+\frac{1}{s}} 2^{s+1}(1+u)^{-\frac{1}{2}} \binom{s+\frac{1}{2}}{s} du \\ &= 2^{s+1}\frac{2}{\sqrt{s}}\binom{s+\frac{1}{2}}{s} \\ &\le C_1 2^{s+1}.\end{aligned}$$

For the integral on $[-1+1/s, 1]$, we use the finer bound in Lemma 10.2, which, in our case, implies that

$$\left|P_s^{s+1,\frac{1}{2}}(u)\right| \le \frac{C2^{s/2}}{(1-u)^{\frac{s}{2}+\frac{3}{4}}(1+u)^{\frac{1}{2}}}, \quad u \in (-1,1),$$

where we used the fact that $c_s^{(s+1,\frac{1}{2})} \sim s^2/2^s$. It follows that

$$\begin{aligned}\int_{-1+\frac{1}{s}}^{1} (1-u)^{s+1}(1+u)^{-\frac{1}{2}} \left|P_s^{(s+1,\frac{1}{2})}(u)\right| du &\le C\int_{-1+\frac{1}{s}}^{1} 2^{s/2}\frac{(1-u)^{\frac{s}{2}+\frac{1}{4}}}{1+u} du \\ &\le C_2 \int_{-1+\frac{1}{s}}^{1} 2^{s/2}\frac{2^{\frac{s}{2}}}{1+u} du \\ &\le C_3 2^s \ln s.\end{aligned}$$

Combining the two parts above with Lemma 10.3, we have

$$\int_{-1}^{1} |\psi(x)| dx \le C_4 \frac{|c_s|}{\sqrt{2}\cdot 2^{s+1}} 2^s \ln s \le C_5\sqrt{s}\ln s.$$

∎

Let $h^\alpha(x) = (1-x^2)^\alpha$. We have the following bound on the derivatives of $h^\alpha(x)$.

LEMMA 10.5. *It holds for $k < \alpha$ that*

$$\max_{x\in[-1,1]} |(h^\alpha)^{(k)}(x)| \le k! 2^k \binom{2\alpha}{k}.$$

Proof. Writing $(1-x^2)^\alpha$ as $(1+x)^\alpha(1-x)^\alpha$, by Leibniz's rule we see that

$$\begin{aligned}(h^\alpha)^{(k)}(x) &= \sum_{\ell=0}^{k}\Bigg[\binom{k}{\ell}\alpha(\alpha-1)\cdots(\alpha-\ell+1)(1+x)^{\alpha-\ell} \\ &\qquad \cdot \alpha(\alpha-1)\cdots(\alpha-(k-\ell)+1)(1-t)^{\alpha-(k-\ell)}(-1)^{k-\ell}\Bigg] \\ &= k!\sum_{\ell=0}^{s}\binom{\alpha}{\ell}\binom{\alpha}{k-\ell}(1+t)^{\alpha-\ell}(1-x)^{\alpha-(k-\ell)}(-1)^{s-\ell}.\end{aligned}$$

Hence

$$|(h^\alpha)^{(s)}(x)| \le k!2^k \sum_{\ell=0}^{k}\binom{\alpha}{\ell}\binom{\alpha}{k-\ell} = k!2^k\binom{2\alpha}{k},$$

as claimed. ∎

Now we are ready to bound the derivatives of $\psi(x)$.

LEMMA 10.6. *Suppose that $k \le s$. It holds that*

$$\max_{x\in[-1,1]} |\psi^{(k)}(x)| \le C_1\sqrt{ks}\left(\frac{C_2 s^2}{k}\right)^k,$$

where $C_1, C_2 > 0$ are absolute constants.

Proof. By Leibniz's rule of differentation,

$$\begin{aligned}\psi^{(k)}(x) &= c_s \sum_{\ell=0}^{k} \binom{k}{\ell} (h^{s+1})^{(k-\ell)}(x)(P_s^{(s+1,\frac{1}{2})})^{(\ell)}(2x^2-1)\cdot(4x)^\ell \\ &= c_s \sum_{\ell=0}^{s} \binom{k}{\ell} (h^{s+1})^{(k-\ell)}(x) \\ &\qquad\qquad \cdot P_{s-\ell}^{(s+\ell+1,\ell+\frac{1}{2})}(2x^2-1)(2x)^\ell \prod_{i=1}^{\ell}\left(2s+\frac{1}{2}+i\right).\end{aligned}$$

Therefore

$$\begin{aligned}|\psi^{(s)}(x)| &\le |c_s| \sum_{\ell=0}^{s} \binom{k}{\ell}(k-\ell)!2^{k-\ell}\binom{2s+2}{\ell}\cdot\binom{2s+1}{k-\ell}2^\ell \prod_{i=2}^{\ell+1}(2s+i) \\ &\le |c_s|2^k \sum_{\ell=0}^{k} \binom{k}{\ell}\binom{2s+2}{\ell}\frac{(2s+\ell+1)!}{(2s+\ell+1-k)!} \\ &= |c_s|2^k k! \sum_{\ell=0}^{k} \binom{k}{k-\ell}\binom{2s+2}{\ell}\binom{2s+\ell+1}{k} \\ &\le |c_s|2^k k! \binom{2s+k+1}{k} \sum_{\ell=0}^{k} \binom{k}{k-\ell}\binom{2s+2}{\ell} \\ &= |c_s|2^k k! \binom{2s+k+1}{k}\binom{2s+k+2}{k} \\ &= |c_s|2^k k! \left(\frac{3s}{k}\right)^{2k} \\ &\le C_1\sqrt{ks}\left(\frac{C_2 s^2}{k}\right)^k\end{aligned}$$

for some absolute constants $C_1, C_2 > 0$. ∎

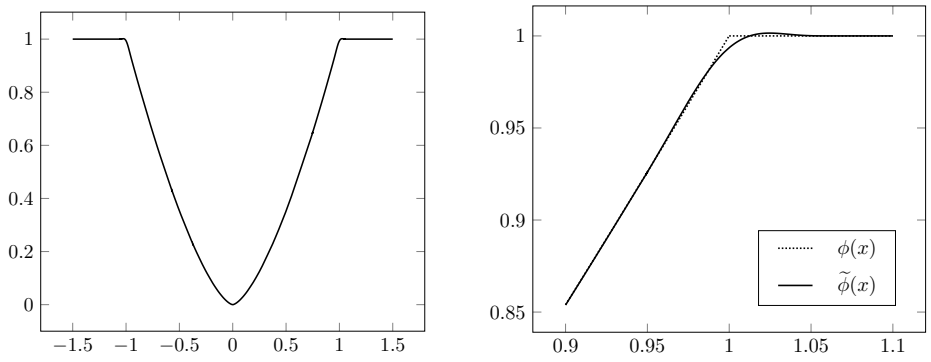
We also define a family of dilated versions of ψ as

$$\psi_t(x) = \frac{1}{t}\psi\left(\frac{x}{t}\right),$$

then $\psi_t \in C_0^{s+1}(-t,t)$ and $\int_{-t}^{t}\psi_t(x)dx = 1$.

Recall that the Tukey p-loss function is

$$\phi(x) = \begin{cases} |x|^p, & |x| \le \tau, \\ \tau^p, & |x| > \tau. \end{cases}$$



(a) Plot of $\widetilde{\phi}$ on $[-1.5, 1.5]$ (b) Plot of $\widetilde{\phi}$ and ϕ on $[0.9, 1.1]$

FIG. 10.1. *Plots of the Tukey loss function ϕ and the mollified version $\widetilde{\phi}$ with $p = 1.5$, $d = 1$, $\delta = 1/16$, $\gamma = 3/16$, and $\tau = 1$. In this case, $\phi(\gamma) \approx 0.081$ and $c_\gamma \approx 7.3 \times 10^{-7}$.*

Let $\delta \in (0, 1/8)$ and $\gamma \in (2\delta, 1/2)$ be constants. We define the mollified version of ϕ to be

$$\widetilde{\phi}(x) = \begin{cases} \phi(x), & |x| \le \gamma\tau, \\ (\phi * \psi_{\delta\tau})(x) + c_\gamma, & |x| > \gamma\tau, \end{cases} \tag{10.2}$$

where the constant c_γ is such that $\widetilde{\phi}(x)$ is continuous at $x = \gamma\tau$. The mollified version $\widetilde{\phi}$ is not necessarily differentiable (unless when $p = 1$, in which case $c_\gamma = 0$ and $\phi \in C^s(0, +\infty)$) but it is differentiable on $[\gamma\tau, +\infty)$ and $\widetilde{\phi}(x) = \phi(x) + c_\gamma = \tau^p + c_\gamma$ for all $|x| \ge (1 + \delta)\tau$. See Figure 10.1 for an example plot.

We are going to choose the parameter s in our polynomial mollifier ψ such that $\widetilde{\phi}$ is close to ϕ. Specifically, we have the following lemma.

LEMMA 10.7. *For $\varepsilon > 0$, when $s = \Omega(\log(1/\varepsilon))$, it holds that $|\widetilde{\phi}(x) - \phi(x)| \le \varepsilon\phi(x)$ on $[\gamma\tau, \tau/2]$. The constant in the Ω-notation depends on p and γ.*

Proof. Observe that

$$\begin{aligned} (\phi * \psi_{\delta\tau} - \phi)(x) &= x^p \int_{-\delta\tau}^{\delta\tau} \left[\left(1 + \frac{t}{x}\right)^p - 1 \right] \psi_{\delta\tau}(t) dx \\ &= x^p \int_{-1}^{1} \left[\left(1 + \delta\tau \cdot \frac{t}{x}\right)^p - 1 \right] \psi(t) dt. \end{aligned}$$

If it holds that

$$\left| \int_{-1}^{1} \left[\left(1 + \frac{t}{x}\right)^p - 1 \right] \psi(t) dt \right| \le \frac{\varepsilon}{2}, \qquad x \in \left[\frac{\gamma}{\delta}, \frac{1}{2\delta} \right], \tag{10.3}$$

then $c_\gamma \le (\varepsilon/2)(\gamma\tau)^p$ and thus

$$|\widetilde{\phi}(x) - \phi(x)| \le |(\phi * \psi_{\tau/8} - \phi)(x)| + |c_\gamma| \le \frac{\varepsilon}{2}(x^p + (\gamma\tau)^p) \le \varepsilon\phi(x)$$

for all $x \in [\gamma\tau, \tau/2]$ as desired.

Now we find s such that (10.3) holds. Note that $|t/x| \le \delta/\gamma < 1/2$. Recall the Taylor expansion with Cauchy remainder,

$$(1+u)^p - 1 = \sum_{i=1}^{s} \binom{p}{i} u^i + r_s(u),$$

where

$$r_s(u) = (s+1)\binom{p}{s+1} u^{s+1} \left(\frac{1-\theta}{1+\theta u}\right)^s (1+\theta u)^{p-1}, \quad \theta \in (0,1).$$

We have bounds (note that $p \le 2$)

$$\left|\binom{p}{s+1}\right| \le \frac{s!}{(s+1)!} = \frac{1}{s+1},$$
$$0 \le \frac{1-\theta}{1+\theta u} \le 1, \quad |u| < 1,$$
$$(1+\theta u)^{p-1} \le \max\{(1+|u|)^{p-1}, (1-|u|)^{p-1}\} =: K_p(u).$$

It follows from the moment conditions (10.1) and Lemma 10.4 that

$$\left|\int_{-1}^{1} \left[\left(1+\frac{t}{x}\right)^p - 1\right] \psi(t)dt\right| \le r_s\left(\frac{t}{x}\right) \le C\sqrt{s}\ln s \left(\frac{\delta}{\gamma}\right)^{s+1} K_p\left(\frac{\delta}{\gamma}\right).$$

Hence (10.3) holds when $s = \Omega(\log(1/\varepsilon))$ (where the constant depends on p, δ, and γ). ∎

A similar argument shows that $\widetilde{\phi}^{(k)}$ can be made $(1 \pm \varepsilon)$-close to $\phi^{(k)}$ on $[\gamma\tau, \tau/2]$ for $k = 1, \dots, d$.

Next we bound the derivatives of $\widetilde{\phi}$.

LEMMA 10.8. *Suppose that $1 \le k \le s$. There exist constants $C_1, C_2 > 0$ that depend only on p, δ, and γ such that*

$$\max_{x \in [\gamma\tau, (1+\delta)\tau]} \left|\widetilde{\phi}^{(k)}(x)\right| \le C_1 \tau^p \sqrt{ks} \left(\frac{C_2 s^2}{k\tau}\right)^k.$$

Proof. Observe that for $x \in [\gamma\tau, (1+\delta)\tau]$,

$$\begin{aligned}\left|\widetilde{\phi}^{(k)}(x)\right| &\le 2\delta\tau \cdot \max_{x \in [(\gamma-\delta)\tau, (1+\delta)\tau]} \phi(x) \cdot \max_{x \in [-\delta\tau, \delta\tau]} \left\|\psi_{\delta\tau}^{(k)}(x)\right\| \\ &\le 2\delta\tau \cdot \tau^p \cdot \max_{x \in [-\delta\tau, \delta\tau]} \left\|\psi_{\delta\tau}^{(k)}(x)\right\| \\ &= \frac{2\tau^p}{(\delta\tau)^k} \max_{x \in [-1,1]} \left\|\psi^{(k)}(x)\right\|.\end{aligned}$$

The result follows from Lemma 10.6. ∎

As a corollary of the preceding lemma, we have the following.

LEMMA 10.9. *Let $a \in (0, \gamma]$ and $b \ge 1+\delta$ be constants. When $s = \Theta(\log \frac{1}{\varepsilon})$, there exists a polynomial $p(x)$ of degree $O(s)$ such that $|p(x) - \widetilde{\phi}(x)| \le \varepsilon\widetilde{\phi}(x)$ on $[a\tau, b\tau]$. The constants in the Θ- and O-notations depend on a, b, δ, γ, and p.*

Proof. Since $\widetilde{\phi}(x) \geq (a\tau)^p$ on $[a\tau, b\tau]$, it is sufficient to consider the uniform approximation $|p(x) - \widetilde{\phi}(x)| \leq \varepsilon(a\tau)^p$ on $[a\tau, b\tau]$. It follows from Lemma 2.8 that when $n > k$,

$$E_n(\widetilde{\phi}; [a\tau, b\tau]) \leq \frac{6^{k+1}e^k}{(k+1)n^k} \cdot \left(\frac{(b-a)\tau}{2}\right)^k \cdot \max_{x\in[a\tau,b\tau]} |\widetilde{\phi}^{(k)}(x)| \cdot \frac{1}{n-k}.$$

Assume that $k > 2$. When $x \in [a\tau, \gamma\tau]$,

$$\left|\widetilde{\phi}^{(k)}(x)\right| = \left|\phi^{(k)}(x)\right| = \left|p(p-1)\cdots(p-k+1)x^{p-k}\right| \leq (k+1)!(a\tau)^{p-k}.$$

When $x > (1+\delta)\tau$, $\widetilde{\phi}(x) = \tau^p + c_\gamma$ and thus $\widetilde{\phi}^{(k)}(x) = 0$. When $x \in [\gamma\tau, (1+\delta)\tau]$, we invoke Lemma 10.8. Combining the three cases, when $s = k$ and $n \geq 2k$, we have that

$$E_n(\widetilde{\phi}; [a\tau, b\tau]) \leq \frac{C_1\tau^p s}{n}\left(\frac{C_2 \cdot \max\{b-a, \frac{1}{a}\} \cdot s}{n}\right)^s,$$

where $C_1, C_2 > 0$ are absolute constants. It is now clear that we can take $s = k = \Omega(\log \frac{1}{\varepsilon})$ and $n = \Omega(s)$ so that $E_n(\widetilde{\phi}; [a\tau, b\tau]) \leq \varepsilon \cdot (a\tau)^p$. ∎

10.3. Estimation algorithm. In this subsection, we let $\delta = 1/16$ and $\gamma = 3/16$ for the definition of $\widetilde{\phi}$ in (10.2). Since $\widetilde{\phi}(x)$ agrees with $|x|^p$ for small $|x|$, it follows from Theorem 8.1 that solving the subspace sketch problem for $\widetilde{\Phi}(x)$ requires $\widetilde{\Omega}(d/\varepsilon^2)$ bits. In this subsection we show that this lower bound is tight up to polylogarithmic factors. Specifically we have the following theorem.

THEOREM 10.10. *Let $p \in (0, 2]$ be a constant and $\widetilde{\Phi}(x)$ be the mollified Tukey loss p-norm of $x \in \mathbb{R}^n$. There exists a randomized algorithm which returns an estimate Z to $\widetilde{\Phi}(x)$ such that $(1-\varepsilon)\widetilde{\Phi}(x) \leq Z \leq (1+\varepsilon)\widetilde{\Phi}(x)$ with probability at least 0.9. The algorithm uses $\widetilde{O}_p(1/\varepsilon^2)$ bits of space.*

This theorem implies an $\widetilde{O}(d/\varepsilon^2)$ upper bound for the corresponding subspace sketch problem. The remainder of the section is devoted to the proof of this theorem.

We shall first sample rows of A with sampling rate $\Theta(\frac{\tau^p}{\widetilde{\Phi}(x)\varepsilon^2})$. However, we do not know $\widetilde{\Phi}(x)$ in advance. To implement this, we sample rows of A using $O(\log n)$ different sampling rates $1, (1.1)^{-1}, (1.1)^{-2}, \ldots, 1.1^{-O(\log n)}$ and, in parallel, estimate $\widetilde{\Phi}(x)$ using a separate data structure of $O(\text{polylog}(n) \cdot d)$ space [11, 10], which gives an estimate F satisfying $0.9\widetilde{\Phi}(x) \leq F \leq 1.1\widetilde{\Phi}(x)$. Then we choose a sampling rate $r = 1.1^{-s}$ for some integer s that is closest to $\frac{\tau^p}{F\varepsilon^2}$. Thus $r \in [\frac{\tau^p}{2\Phi(x)\varepsilon^2}, \frac{2\tau^p}{\widetilde{\Phi}(x)\varepsilon^2}]$ when $\widetilde{\Phi}(x) > \frac{\tau^p}{2\varepsilon^2}$, and $r = 1$ otherwise.

Now we show that for the chosen sampling rate r, the sampled entries give an accurate estimation to $\widetilde{\Phi}(x)$. This is definitely true when $r = 1$, in which case there is no sampling at all. Otherwise, let $X_i = \widetilde{\Phi}(x_i)$ if item i is sampled and $X_i = 0$ otherwise. Let $X = \sum_i X_i$ and $Z = (1/r)X$. It is clear that $\mathbb{E}[Z] = \widetilde{\Phi}(x)$. We

calculate the variance below:

$$\begin{aligned}\operatorname{Var}(Z) = \frac{1}{r^2}\operatorname{Var}(X) = \frac{1}{r^2}\sum_i \operatorname{Var}(X_i)^2 &= \frac{1}{r^2}\sum_i (r - r^2)(\widetilde{\Phi}(x_i))^2 \\ &\le \frac{1}{r}\sum_i (\widetilde{\Phi}(x_i))^2 \\ &= O\left(\frac{\widetilde{\Phi}(x)\varepsilon^2}{\tau}\cdot \sum_i \tilde{\phi}(x_i)\cdot\tau\right) \\ &= O(\varepsilon^2)\cdot(\widetilde{\Phi}(x))^2.\end{aligned}$$

It follows from Chebyshev's inequality that with constant probability,

$$Z = \frac{1}{r}\sum X_i = (1 \pm O(\varepsilon))\widetilde{\Phi}(x).$$

We condition on this event in the rest of the proof. Thus, it suffices to estimate the summation of $\widetilde{\Phi}(x_i)$ for those x_i that are sampled. In the rest of this section, we use L to denote the indices of entries that are sampled at the sampling rate r.

For each $i \in L$ with $|x_i| \ge \tau$, we claim that

$$|x_i| \ge \Omega(\varepsilon^2)\cdot \|(x_L)_{-O(2^p/\varepsilon^2)}\|_p^p, \tag{10.4}$$

where x_L denotes the vector x restricted to the indices in L and v_{-k} denotes the vector v after zeroing out the largest k entries in magnitude.

We first show that $\widetilde{\Phi}(x_L) = O\left(\frac{\tau^p}{\varepsilon^2}\right)$, which is clearly true when $r = 1$, since in this case, $\widetilde{\Phi}(x_L) = \widetilde{\Phi}(x) = O\left(\frac{\tau^p}{\varepsilon^2}\right)$. When $r < 1$, $\sum_{i\in L}\widetilde{\Phi}(x_i) = (1 \pm O(\varepsilon))\cdot r\cdot\widetilde{\Phi}(x) = O\left(\frac{\tau^p}{\varepsilon^2}\right)$.

Let $L' = \{i \in L : |x_i| \ge \tau/2\}$. It follows that $|L'| \le \widetilde{\Phi}(x_L)/(\tau/2)^p = O(2^p/\varepsilon^2)$. Hence

$$\|x_{L\setminus L'}\|_p^p = \sum_{i\in L\setminus L'} |x_i|^p = \widetilde{\Phi}(x_{L\setminus L'}) \le \widetilde{\Phi}(x_L) = O\left(\frac{\tau^p}{\varepsilon^2}\right),$$

establishing (10.4).

Therefore, to find all $i \in L$ with $|x_i| \ge \tau$, we use an ℓ_p-heavy hitter data structure, which can be realized by a CountSketch structure [13] which hashes x_L into $O(1/\beta)$ buckets and finds β-heavy hitters relative to $\|(x_L)_{-1/\beta}\|_1$. Set $\beta = \Theta(\varepsilon^2/2^p)$. In the end we obtain a list $H \subseteq L$ such that every $i \in H$ is a $\beta/2$-heavy hitter relative to $\|(x_L)_{-1/\beta}\|_p^p$, and all β-heavy hitters are in H. Furthermore, for each $i \in H$ the data structure also returns an estimate $\hat{x}_i$ such that $|x_i|/2 \le |\hat{x}_i| \le 2|x_i|$ whenever $|x_i| \ge \tau/2$. The data structure has space complexity $\widetilde{O}_p(1/\varepsilon^2)$.

For each $x_i \in L$ with $|x_i| \ge \tau$, it must hold that $i \in H$. Let $H_1 = \{i \in H : |\hat{x}_i| \ge 5\tau/4\}$ and $H_2 = \{i \in H : 3\tau/8 \le |\hat{x}_i| \le 5\tau/2\}$.

For each $i \in H_1$, by the estimation guarantee it must hold that $|x_i| \ge 5\tau/4$. Hence $S_1 = (\tau^p + c_\gamma)|H_1| = \widetilde{\Phi}(x_{H_1})$.

For each $i \in H_2$ it must hold that $|x_i| \in [\frac{3}{16}\tau, 5\tau]$ and, thus,

$$\|x_{[n]\setminus H_1}\|_p^p \le 5^p\widetilde{\Phi}(x_{[n]\setminus H_1}).$$

Let $p(x)$ be a polynomial such that $|p(x) - \widetilde{\phi}(x)| \le \varepsilon\widetilde{\phi}(x)$ on $[\frac{3}{16}\tau, 5\tau]$. By Lemma 10.9, it is possible to achieve $\deg p = O(\log(1/(\varepsilon\tau)))$. We now use an estimation algorithm

analogous to the HighEnd structure in [23], which uses the same space $\widetilde{O}(1/\varepsilon^2)$. Using the same BasicHighEnd structure in [23], with constant probability, for each $x_i \in H$ we have T estimates $\hat{x}_{i,1}, \ldots, \hat{x}_{i,T} \in \mathbb{C}$ such that $\hat{x}_{i,t} = x_i + \delta_{i,t}$, where each $\delta_{i,t} \in \mathbb{C}$ satisfies $|\delta_{i,t}| \le |x_i|/2$ and $\mathbb{E}(\delta_{i,t})^k = 0$ for $k = 1, \ldots, 3 \deg p$. The estimator is

$$S_2 = \mathrm{Re} \sum_{i \in H_2} \widetilde{\Phi} \left(\frac{1}{T} \sum_{i=1}^{T} p(\hat{x}_{i,t}) \right).$$

It follows from the analysis in [23] (x can be replaced with $x_{[n]\setminus H_1}$ in the analysis of the variance) that the algorithm will output, with a constant probability,

$$S_2 = \widetilde{\Phi}(x_{H_2}) \pm \varepsilon \|x_{[n]\setminus H_1}\|_1 = (1 \pm 5^p \varepsilon) \widetilde{\Phi}(x_{H_2}).$$

For each $i \in [n] \setminus (H_1 \cup H_2)$, it must hold that $|x_i| \le \tau/2$ and thus we can use an ℓ_p sketch algorithm as in [23], and obtain

$$S_3 = (1 \pm \varepsilon) \|x_{[n]\setminus(H_1 \cup H_2)}\|_p^p = (1 \pm 2\varepsilon) \widetilde{\Phi}(x_{[n]\setminus(H_1 \cup H_2)}),$$

where we used Lemma 10.7 in the last step.

Finally, the algorithm returns $S_1 + S_2 + S_3$, which is a $(1 \pm O(\varepsilon))$-approximation to $\widetilde{\Phi}(x)$, where the constant in O-notation depends on p. Rescaling ε proves the correctness of the estimate.

For the part of evaluating S_2 and S_3, the space complexity is the same as the HighEnd and ℓ_p sketch algorithm in [23], which are both $\widetilde{O}(1/\varepsilon^2)$ bits.

11. An upper bound for ℓ_1 subspace sketches in two dimensions. In this section, we prove an $O(\mathrm{polylog}(n)/\varepsilon)$ upper bound for the ℓ_1 subspace sketch problem when $d = 2$. Our plan is to reduce the ℓ_1 subspace sketch problem with $d = 2$ to coresets for the weighted 1-median problem with $d = 1$. For the latter problem, an $O(\mathrm{polylog}(n)/\varepsilon)$ upper bound is known [19].

For the special case where the first column of the A matrix is all ones, the ℓ_1 subspace sketch problem with $d = 2$ is equivalent to coresets for 1-median with $d = 1$. To see this, by homogeneity, we may assume $x_2 = 1$ for the query vector $x \in \mathbb{R}^2$. Thus, $\|Ax\|_1 = \sum_{i=1}^{n} |x_1 + A_{i,2}|$, which is the 1-median cost of using x_2 as the center on $\{-A_{1,2}, -A_{2,2}, \ldots, -A_{n,2}\}$. When entries of the first column of A are positive but not necessarily all ones, we have

$$\|Ax\|_1 = \sum_{i=1}^{n} A_{i,1} \left| x_1 + \frac{A_{i,2}}{A_{i,1}} \right|,$$

which is the weighted 1-median cost of using x_1 as the center on

$$\left\{ -\frac{A_{1,2}}{A_{1,1}}, -\frac{A_{2,2}}{A_{2,1}}, \ldots, -\frac{A_{n,2}}{A_{n,1}} \right\}$$

with weights $\{A_{i,1}, A_{i,2}, \ldots, A_{n,2}\}$. It has been shown in [19, Theorem 2.8] that there exists a coreset of size $O(\mathrm{polylog}(n)/\varepsilon)$ for the weighted 1-median problem when $d = 1$.

For general A, we divide the rows of A into three separate matrices A^+, A^-, and A^0. Here, all entries in the first column of A^+ are positive, all entries in the first column of A^- are negative, and all entries in the first column of A^0 are zeroes. Since $\|Ax\|_1 = \|A^+x\|_1 + \|A^-x\|_1 + \|A^0x\|_1$, we can design subspace sketches separately for

A^+, A^-, and A^0. Our reduction above implies an $O(\text{polylog}(n)/\varepsilon)$ upper bound for A^+ and A^-. For A^0, since all entries in the first column are all zero, we have

$$\|A^0 x\|_1 = |x_2| \sum_i \left|A^0_{i,2}\right|.$$

Thus, it suffices to store $\sum_i \left|A^0_{i,2}\right|$ for A^0.

THEOREM 11.1. *The ℓ_1 subspace sketch problem can be solved using $\widetilde{O}(1/\varepsilon)$ bits when $d = 2$.*

Acknowledgment. We thank the anonymous reviewers for various suggestions that contributed to the exposition of the proofs.

REFERENCES

[1] *Alternate Proof for Weighted Alternating Shifted Central Binomial Sum Relation*, https://math.stackexchange.com/questions/2827591/.

[2] *Jacobi Polynomials: Integration (formula* 05.06.21.0006.01*)*, http://functions.wolfram.com/05.06.21.0006.01.

[3] *Sketching Algorithms for Big Data*, https://www.sketchingbigdata.org/.

[4] A. ANDONI, C. BURNS, Y. LI, S. MAHABADI, AND D. P. WOODRUFF, *Streaming Complexity of SVMs*, in Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2020), J. Byrka and R. Meka, eds., LIPIcs Leibniz Int. Proc. Inform., Schloss Dagstuhl, Wadern, Germany, 2020.

[5] A. ANDONI, J. CHEN, R. KRAUTHGAMER, B. QIN, D. P. WOODRUFF, AND Q. ZHANG, *On sketching quadratic forms*, in Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science, ITCS '16, ACM, New York, 2016, pp. 311–319.

[6] G. E. ANDREWS, R. ASKEY, AND R. ROY, *Special Functions*, Encyclopedia Math. Appl. 71, Cambridge University Press, Cambridge, 1999.

[7] E. D. BOLKER, *A class of convex bodies*, Trans. Amer. Math. Soc., 145 (1969), pp. 323–345, https://doi.org/10.2307/1995073.

[8] S. BOUCHERON, G. LUGOSI, AND O. BOUSQUET, *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford University Press, Oxford, 2013.

[9] J. BOURGAIN, J. LINDENSTRAUSS, AND V. MILMAN, *Approximation of zonoids by zonotopes*, Acta Math., 162 (1989), pp. 73–141, https://doi.org/10.1007/BF02392835.

[10] V. BRAVERMAN, S. R. CHESTNUT, D. P. WOODRUFF, AND L. F. YANG, *Streaming space complexity of nearly all functions of one variable on frequency vectors*, in Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, ACM, New York, 2016, pp. 261–276.

[11] V. BRAVERMAN AND R. OSTROVSKY, *Zero-one frequency laws*, in Proceedings of the Forty-Second ACM Symposium on Theory of Computing, ACM, New York, 2010, pp. 281–290.

[12] C. CARLSON, A. KOLLA, N. SRIVASTAVA, AND L. TREVISAN, *Optimal lower bounds for sketching graph cuts*, in Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '19, SIAM, Philadelphia, 2019, pp. 2565–2569.

[13] M. CHARIKAR, K. CHEN, AND M. FARACH-COLTON, *Finding frequent items in data streams*, Theoret. Comput. Sci., 312 (2004), pp. 3–15, https://doi.org/10.1016/S0304-3975(03)00400-6.

[14] K. L. CLARKSON, P. DRINEAS, M. MAGDON-ISMAIL, M. W. MAHONEY, X. MENG, AND D. P. WOODRUFF, *The fast Cauchy transform and faster robust linear regression*, SIAM J. Comput., 45 (2016), pp. 763–810, https://doi.org/10.1137/1.9781611973105.34.

[15] M. B. COHEN AND R. PENG, *L_p row sampling by Lewis weights*, in Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing, STOC '15, ACM, New York, 2015, pp. 183–192, https://doi.org/10.1145/2746539.2746567.

[16] R. COURANT AND F. JOHN, *Introduction to Calculus and Analysis* II/1, Classics in Mathematics, Springer, Berlin, 1999.

[17] D. DUBHASHI AND A. PANCONESI, *Concentration of Measure for the Analysis of Randomized Algorithms*, Cambridge University Press, New York, 2009.

[18] S. GANGULY AND D. P. WOODRUFF, *High probability frequency moment sketches*, in 45th International Colloquium on Automata, Languages, and Programming, ICALP 2018, Prague,

Czech Republic, LIPIcs Leibniz Int. Proc. Inform. 104, Schloss Dagstuhl, Wadern, Germany, 2018.

[19] S. Har-Peled and A. Kushal, *Smaller coresets for k-median and k-means clustering*, Discrete Comput. Geom., 37 (2007), pp. 3–19, https://doi.org/10.1007/s00454-006-1271-x.

[20] P. Indyk, *Stable distributions, pseudorandom generators, embeddings, and data stream computation*, J. ACM, 53 (2006), pp. 307–323, https://doi.org/10.1145/1147954.1147955.

[21] W. B. Johnson and J. Lindenstrauss, *Basic concepts in the geometry of Banach spaces*, in Handbook of the Geometry of Banach Spaces, Vol. 1, W. B. Johnson and J. Lindenstrauss, eds., Elsevier Science, Amsterdam, 2001, pp. 1–84.

[22] W. B. Johnson and G. Schechtman, *Finite dimensional subspaces of L_p*, in Handbook of the Geometry of Banach Spaces, Vol. 1, W. B. Johnson and J. Lindenstrauss, eds., Elsevier Science, Amsterdam, 2001, pp. 837–870.

[23] D. M. Kane, J. Nelson, E. Porat, and D. P. Woodruff, *Fast moment estimation in data streams in optimal space*, in Proceedings of the Forty-third Annual ACM Symposium on Theory of Computing, STOC '11, ACM, New York, 2011, pp. 745–754.

[24] D. M. Kane, J. Nelson, and D. P. Woodruff, *An optimal algorithm for the distinct elements problem*, in Proceedings of the Twenty-Ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2010, Indianapolis, IN, ACM, New York, 2010, pp. 41–52.

[25] A. Koldobsky and H. König, *Aspects of the isometric theory of Banach spaces*, in Handbook of the Geometry of Banach Spaces, Vol. 1, W. B. Johnson and J. Lindenstrauss, eds., Elsevier Science, Amsterdam, 2001, pp. 837–870.

[26] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*, Springer, Berlin, 1991.

[27] Y. Li, X. Sun, C. Wang, and D. P. Woodruff, *On the communication complexity of linear algebraic problems in the message passing model*, in Distributed Computing, F. Kuhn, ed., Springer, Berlin, 2014, pp. 499–513.

[28] Y. Li and D. P. Woodruff, *Tight bounds for sketching the operator norm, Schatten norms, and subspace embeddings*, in Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2016), K. Jansen, C. Mathieu, J. D. P. Rolim, and C. Umans, eds., LIPIcs Leibniz Int. Proc. Inform. 60, Schloss Dagstuhl, Wadern, Germany, 2016, 39.

[29] J. Matoušek, *Lecture Notes on Metric Embeddings*, https://kam.mff.cuni.cz/~matousek/ba-a4.pdf (2013).

[30] V. D. Milman and G. Schechtman, *Asymptotic Theory of Finite Dimensional Normed Spaces*, Lecture Notes in Math. 1200, Springer, Berlin, 1986.

[31] P. B. Miltersen, N. Nisan, S. Safra, and A. Wigderson, *On data structures and asymmetric communication complexity*, J. Comput. System Sci., 57 (1998), pp. 37–49, https://doi.org/10.1006/jcss.1998.1577.

[32] S. Muthukrishnan, *Data streams: Algorithms and applications*, Found. Trends Theoret. Comput. Sci., 1 (2005), pp. 117–236, https://doi.org/10.1561/0400000002.

[33] P. Nevai, T. Erdélyi, and A. P. Magnus, *Generalized Jacobi weights, Christoffel functions, and Jacobi polynomials*, SIAM J. Math. Anal., 25 (1994), pp. 602–614.

[34] R. Pagh, *Compressed matrix multiplication*, ACM Trans. Comput. Theory, 5 (2013), 9, https://doi.org/10.1145/2493252.2493254.

[35] U. Parampalli, X. Tang, and S. Boztas, *On the construction of binary sequence families with low correlation and large sizes*, IEEE Trans. Inform. Theory, 59 (2013), pp. 1082–1089.

[36] T. J. Rivlin, *An introduction to the approximation of functions*, Blaisdell Book Numer. Anal. Comput. Sci., Blaisdell Publishing Company, Waltham, MA, 1969.

[37] G. Schechtman, *More on embedding subspaces of L_p in l_r^n*, Compos. Math., 61 (1987), pp. 159–169.

[38] G. Schechtman, *Tight embedding of subspaces of L_p in ℓ_p^n for even p*, Proc. Amer. Math. Soc., 139 (2011), pp. 4419–4421.

[39] G. Szegő, *Orthogonal Polynomials*, 4th ed., Amer. Math. Soc. Collaq. Publ. 23, Providence, RI, 1975.

[40] M. Talagrand, *Embedding subspaces of L_1 into ℓ_1^N*, Proc. Amer. Math. Soc., 108 (1990), pp. 363–369, https://doi.org/10.2307/2048283.

[41] M. Talagrand, *Embedding subspaces of L_p in ℓ_p^N*, in Geometric Aspects of Functional Analysis, J. Lindenstrauss and V. Milman, eds., Birkhäuser Basel, Basel, 1995, pp. 311–326.

[42] D. Van Gucht, R. Williams, D. P. Woodruff, and Q. Zhang, *The communication complexity of distributed set-joins with applications to matrix multiplication*, in Proceedings of

the 34th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS '15, ACM, New York, 2015, pp. 199–212.

[43] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*, Camb. Ser. Stat. Probab. Math., Cambridge University Press, Cambridge, 2018.

[44] D. P. Woodruff, *Sketching as a tool for numerical linear algebra*, Found. Trends Theoret. Comput. Sci., 10 (2014), pp. 1–157, https://doi.org/10.1561/0400000060.

[45] D. P. Woodruff and Q. Zhang, *Distributed statistical estimation of matrix products with applications*, in Proceedings of the 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, SIGMOD/PODS '18, ACM, New York, 2018, pp. 383–394.

[46] J. Yang, X. Meng, and M. W. Mahoney, *Quantile regression for large-scale applications*, SIAM J. Sci. Comput., 36 (2014), pp. S78–S110, https://doi.org/10.1137/130919258.