On Approximating Functions of the Singular Values in a Stream

Yi Li*
Facebook, Inc.
leeyi@umich.edu

David P. Woodruff†
IBM Almaden Research Center
dpwoodru@us.ibm.com

ABSTRACT

For any real number $p > 0$, we nearly completely characterize the space complexity of estimating $\|A\|_p^p = \sum_{i=1}^n \sigma_i^p$ for $n \times n$ matrices A in which each row and each column has $O(1)$ non-zero entries and whose entries are presented one at a time in a data stream model. Here the σ_i are the singular values of A, and when $p \geq 1$, $\|A\|_p^p$ is the p-th power of the Schatten p-norm. We show that when p is not an even integer, to obtain a $(1+\epsilon)$-approximation to $\|A\|_p^p$ with constant probability, any 1-pass algorithm requires $n^{1-g(\epsilon)}$ bits of space, where $g(\epsilon) \to 0$ as $\epsilon \to 0$ and $\epsilon > 0$ is a constant independent of n. However, when p is an even integer, we give an upper bound of $n^{1-2/p} \operatorname{poly}(\epsilon^{-1} \log n)$ bits of space, which holds even in the turnstile data stream model. The latter is optimal up to $\operatorname{poly}(\epsilon^{-1} \log n)$ factors.

Our results considerably strengthen lower bounds in previous work for arbitrary (not necessarily sparse) matrices A: the previous best lower bound was $\Omega(\log n)$ for $p \in (0,1)$, $\Omega(n^{1/p-1/2}/\log n)$ for $p \in [1,2)$ and $\Omega(n^{1-2/p})$ for $p \in (2,\infty)$. We note for $p \in (2,\infty)$, while our lower bound for even integers is the same, for other p in this range our lower bound is $n^{1-g(\epsilon)}$, which is considerably stronger than the previous $n^{1-2/p}$ for small enough constant $\epsilon > 0$. We obtain similar near-linear lower bounds for Ky-Fan norms, eigenvalue shrinkers, and M-estimators, many of which could have been solvable in logarithmic space prior to our work.

Categories and Subject Descriptors

F.2 [**Theory of Computation**]: Analysis of Algorithms and Problem Complexity

*Supported by ONR grant N00014-15-1-2388 when the author was at Harvard University.

†Supported in part by the XDATA program of the Defense Advanced Research Projects Agency (DARPA), administered through Air Force Research Laboratory contract FA8750-12-C-0323.

STOC '16, June 18-21, 2016, Cambridge, MA, USA

DOI: http://dx.doi.org/10.1145/2897518.2897581

Keywords

data streams, complexity, algorithms, Schatten norms, matrix norms

1. INTRODUCTION

In the data stream model, there is an underlying vector $x \in \mathbb{Z}^n$ which undergoes a sequence of additive updates to its coordinates. Each update has the form $(i, \delta) \in [n] \times \{-m, -m+1, \ldots, m\}$ (where $[n]$ denotes $\{1, \ldots, n\}$), and indicates that $x_i \leftarrow x_i + \delta$. The algorithm maintains a small summary of x while processing the stream. At the end of the stream it should succeed in approximating a prespecified function of x with constant probability. The goal is often to minimize the space complexity of the algorithm while processing the stream. We make the standard simplifying assumption that n, m, and the length of the stream are polynomially related.

A large body of work has focused on characterizing which functions f it is possible to approximate $f(x) = \sum_{i=1}^n f(x_i)$ using a polylogarithmic (in n) amount of space. The first class of functions studied were the ℓ_p norms $f(x_i) = |x_i|^p$, dating back to work of Alon, Matias, and Szegedy [1]. For $p \leq 2$ it is possible to obtain any constant factor approximation using $\tilde{\Theta}(1)$ bits of space [30, 37], while for $p > 2$ the bound is $\tilde{\Theta}(n^{1-2/p})$ [17, 7, 32, 2, 23, 44, 12, 24], where $\tilde{f} = f \cdot \operatorname{poly}(\log(f))$. Braverman and Ostrovsky later developed a zero-one law for monotonically non-decreasing f for which $f(0) = 0$, showing that if f has at most quadratic growth and does not have large "local jumps", then a constant factor approximation to $f(x)$ can be computed in $\tilde{O}(1)$ space [13]. Moreover, if either condition is violated, then there is no polylogarithmic space algorithm. This was extended by Braverman and Chestnut to periodic and to decreasing f [9, 10]. Characterizations were also given in the related sliding window model [14]. Recently, Bravermen et al. gave conditions nearly characterizing all f computable in a constant number of passes using $n^{o(1)}$ space [11].

Despite a nearly complete understanding of which functions f one can approximate $\sum_{i=1}^n f(x_i)$ for a vector x using small space in a stream, little is known about estimating functions of an $n \times n$ *matrix* A presented in a stream. Here, an underlying $n \times n$ matrix A undergoes a sequence of additive updates to its entries. Each update has the form $(i, j, \delta) \in [n] \times [n] \times \{-m, -m+1, \ldots, m\}$ and indicates that $A_{i,j} \leftarrow A_{i,j} + \delta$. Every matrix A can be expressed in its singular value decomposition as $A = U\Sigma V^T$, where U and V are orthogonal $n \times n$ matrices, and Σ is a non-negative diagonal matrix with diagonal entries $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$, which are the singular values of A. We are interested in functions which do not depend on the bases U and V, but rather only on the spectrum (singular values) of A. These functions have the same value under any (orthogonal) change of basis.

The analogue of the functions studied for vectors are functions of the form $\sum_{i=1}^n f(\sigma_i)$. Here, too, we can take $f(\sigma_i) = \sigma_i^p$, in which case $\sum_{i=1}^n f(\sigma_i)$ is the p-th power of *Schatten* p-norm $\|A\|_p^p$ of A. When $p = 0$, interpreting 0^0 as 0 this is the *rank* of A, which has been studied in the data stream [19, 15] and property testing models [39, 43]. When $p = 1$, this is the nuclear or trace norm[1], with applications to differential privacy [29, 40] and non-convex optimization [16, 21]. When $p = 2$ this is the Frobenius norm, while for large p, this sum approaches the p-th power of the operator norm $\sup_{x:\|x\|_2=1} \|Ax\|_2$. Such norms are useful in geometry and linear algebra, see, e.g., [53]. The Schatten p-norms also arise in the context of estimating p-th moments of a multivariate Gaussian matrix in which the components are independent but not of the same variance, see, e.g., [38]. The Schatten p-norms have been studied in the sketching model [41], and upper bounds there imply upper bounds for streaming. Fractional Schatten-p norms of Laplacians were studied by Zhou [57] and Bozkurt et al. [8]. We refer the reader to [50] for applications of the case $p = 1/2$, which is the Laplacian-energy-like (LEL) invariant of a graph.

There are a number of other functions $\sum_{i=1}^n f(\sigma_i)$ of importance, for example, functions motivated from regularized low rank approximation, where one computes the optimal *eigenvalue shrinkers* for different loss functions, such as the Frobenius, operator, and nuclear norm losses [27]. For example, for Frobenius norm loss, $f(x) = \frac{1}{x}\sqrt{(x^2 - \alpha - 1)^2 - 4\alpha}$ for $x \geq 1 + \sqrt{\alpha}$, and $f(x) = 0$ otherwise, for a given parameter α.

Other applications include low rank approximation with respect to functions on the singular values that are not norms, such as Huber or Tukey loss functions, which could find more robust low dimensional subspaces as solutions; we discuss these functions more in Section 7.

Our Contributions.

The aim of this work is to obtain the first sufficient criteria in the streaming model for functions of a matrix spectrum. Prior to our work we did not even know the complexity of most of the problems we study even in the *insertion-only* data stream model in which each coordinate is updated at most once in the stream, and even when A is promised to be *sparse*, i.e., it has only $O(1)$ non-zero entries per row and column. Sparse matrices have only a constant factor more entries than diagonal matrices, and the space complexity of diagonal matrices is well-understood since it corresponds to that for vectors. As a main application, we considerably strengthen the known results for approximating Schatten p-norms. We stress that the difficulty with functions of a matrix spectrum is that updates to the matrix entries often affect the singular values in subtle ways.

The main qualitative message of this work is that for approximating Schatten p-norms up to a sufficiently small constant factor, for any positive real number p which is not an even integer, almost n bits of space is necessary. Moreover, this holds even for matrices with $O(1)$ non-zero entries per row and column, and consequently is tight for such matrices. It also holds even in the insertion-only model. Furthermore, for even integers p, we present an algorithm achieving an arbitrarily small constant factor approximation for any matrix with $O(1)$ non-zero entries per row and column which achieves $\tilde{O}(n^{1-2/p})$ bits of space. Also, $\Omega(n^{1-2/p})$ bits of space is necessary for even integers p, even with $O(1)$ non-zero entries per row and column and even if all entries are absolute constants independent of n. Thus, for p-norms, there is a substantial difference in the complexity in the vector and matrix cases: in the vector case the complexity is logarithmic for $p \leq 2$ and grows as $n^{1-2/p}$ for $p \geq 2$, while in the matrix case the complexity is always almost n bits unless p is an even integer! Furthermore, for each even integer p the complexity is $\tilde{\Theta}(n^{1-2/p})$, just as in the vector case. Note that our results show a "singularity" at $p = 2 \pm o(1)$, which are the only values of p for which $O(\log n)$ bits of space is possible.

We now state our improvements over prior work more precisely. Henceforth in this section, the approximation parameter ϵ is a constant (independent of n), and $g(\epsilon) \to 0$ as $\epsilon \to 0$. The number of non-zero entries of A is denoted by $\mathrm{nnz}(A)$.

THEOREM 1. *(Lower Bound for Schatten p-Norms) Let $p \in [0, \infty) \setminus 2\mathbb{Z}$. Any randomized data stream algorithm which outputs, with constant error probability, a $(1+\epsilon)$-approximation to the Schatten p-norm of an $n \times n$ matrix A requires $\Omega(n^{1-g(\epsilon)})$ bits of space. This holds even if $nnz(A) = O(n)$.*

We obtain similar lower bounds for estimating the Ky-Fan k-norm, which is defined to be the sum of the k largest singular values, and has applications to clustering and low rank approximation [55, 22]. Interestingly, these norms do not have the form $\sum_{i=1}^n f(\sigma_i)$ but rather have the form $\sum_{i=1}^k f(\sigma_i)$, yet our framework is robust enough to handle them. In the latter case, we have the following general result for strictly monotone f:

THEOREM 2. *Let $\alpha \in (0, 1/2)$ and f be strictly monotone with $f(0) = 0$. There exists a constant $\epsilon > 0$ such that for all sufficiently large n and $k \leq \alpha n$, any data stream algorithm which outputs a $(1+\epsilon)$-approximation to $\sum_{i=1}^k f(\sigma_i(A))$ of an $n \times n$ matrix A requires $\Omega(n^{1+\Theta(1/\ln \alpha)})$ space. This holds even if $nnz(A) = O(n)$.*

We summarize prior work on Schatten p-norms and Ky-Fan k-norms and its relation to our results in Table 1. The previous bounds for Ky-Fan norms come from planting a hard instances of the set disjointness communication problem on the diagonal of a diagonal matrix (where each item is copied k times) [35, 47], or from a Schatten 1-lower bound on $k \times k$ matrices padded with zeros [3].

The best previous lower bound for estimating the Schatten p-norm up to an arbitrarily small constant factor for $p \geq 2$ was $\Omega(n^{1-2/p})$, which is the same for vector p-norms. In [41], an algorithm for *even integers* $p \geq 2$ was given, and it works in the data stream model using $O(n^{2-4/p})$ bits of space. See also [4] for finding large eigenvalues, which can be viewed as an additive approximation to the case $p = \infty$. For $p \in [1, 2)$, the lower bound was $\Omega(\frac{n^{1/p-1/2}}{\log n})$ [3]. Their approach is based on non-embeddability, and the best lower

[1]The trace norm is not to be confused with the trace. These two quantities only coincide if A is positive semidefinite.

		Space complexity in bits	
		Previous lower bounds	Our lower bounds
Schatten p-norm	$p \in (2,\infty) \cap 2\mathbb{Z}$	$n^{1-2/p}$ [28, 33]	
	$p \in (2,\infty) \setminus 2\mathbb{Z}$	$n^{1-2/p}$ [28, 33]	$n^{1-g(\epsilon)}$
	$p \in [1,2)$	$\frac{n^{1/p-1/2}}{\log n}$ [3]	$n^{1-g(\epsilon)}$
	$p \in (0,1)$	$\log n$ [37]	$n^{1-g(\epsilon)}$
	$p = 0$	$n^{1-g(\epsilon)}$ [15]	
Ky-Fan k-norm		$\max\{\frac{n}{k}, \frac{k^{1/2}}{\log k}\}$ [7, 3]	$n^{1-g(\epsilon)}$ (any k)

Table 1: A summary of existing and new lower bounds for $(1+\epsilon)$-approximating Schatten p-norms and Ky-Fan k-norms, where ϵ is an arbitrarily small constant. The Ω-notation is suppressed. The function $g(\epsilon) \to 0$ as $\epsilon \to 0$ and could depend on the parameters p or k and be different in different rows. We show that the lower bound $n^{1-2/p}$ is tight up to log factors by providing a new upper bound for even integers p and sparse matrices. For even integers we also present a new proof of an $n^{1-2/p}$ lower bound in which all entries of the matrix are bounded by $O(1)$.

bound obtainable via this approach is $\Omega(n^{1/p-1/2})$, since the identity map is an embedding of the Schatten p-norm into the Schatten-2 norm with $n^{1/p-1/2}$ distortion, and the latter can be sketched with $O(\log n)$ bits; further it is unknown if the lower bound of [3] holds for sparse matrices [48]. For $p \in (0,1)$, which is not a norm but still a well-defined quantity, the prior bound is only $\Omega(\log n)$, which follows from lower bounds for p-norms of vectors. For $p = 0$, an $\Omega(n^{1-g(\varepsilon)})$ lower bound was shown for $(1+\varepsilon)$-approximation [15]. We note that lower bounds for Schatten-p norms in the sketching model, as given in [41], do not apply to the streaming model, even given work which characterizes "turnstile" streaming algorithms as linear sketches[2] [42].

One feature of previous work is that it rules out constant factor approximation for a large constant factor, whereas our work focuses on small constant factor approximation. For vector norms, the asymptotic complexity in the two cases is the same [37, 12] or the same up to a logarithmic factor [23, 44]. Given the many motivations and extensive work on obtaining $(1+\eta)$-approximation for vector norms for arbitrarily small η [31, 52, 20, 25, 46, 37, 36, 23, 54, 44], we do not view this as a significant shortcoming. Nevertheless, this is an interesting open question, which could exhibit another difference between matrix and vector norms.

Although Theorem 1 makes significant progress on Schatten p-norms, and is nearly optimal for sparse matrices (i.e., matrices with $O(1)$ non-zero entries per row and column), for dense matrices our bounds are off by a quadratic factor. That is, for p not an even integer, we achieve a lower bound which is almost n bits of space, while the upper bound is a trivial $O(n^2)$ words of space used to store the matrix. When p is an even integer, $\tilde{\Theta}(n^{1-2/p})$ is an upper and lower bound for sparse matrices, while for dense matrices the best upper bound is $O(n^{2-4/p})$ given in [41]. Thus, in both cases the upper bound is the square of the current lower bound. Resolving this gap is an intriguing open question.

The Schatten p-norms capture a wide range of possibilities of growths of more general functions, and we are able to obtain lower bound for a general class of functions by considering their growth near 0 (by scaling down our hard instance) or their growth for large inputs (by scaling up our hard instance). If in either case the function "behaves" like a Schatten p-norm (up to low order terms), then we can apply our lower bounds for Schatten p-norms to obtain lower bounds for the function.

[2] In short, in the sketching model one has a matrix S and one distinguishes $S \cdot X$ from $S \cdot Y$ where X, Y are vectors (or vectorized matrices) with $X \sim \mu_1$ and $Y \sim \mu_2$ for distributions μ_1 and μ_2. One argues if S has too few rows, then $S \cdot X$ and $S \cdot Y$ have small statistical distance, but such a statement is not true if we first discretize X and Y.

Technical Overview.

Lower Bound. The starting point of our work is [15], which showed an $\Omega(n^{1-g(\epsilon)})$ lower bound for estimating the rank of A up to a $(1+\epsilon)$-factor by using the fact that the rank of the Tutte matrix equals twice the size of the maximum matching of the corresponding graph, and there are lower bounds for estimating the maximum matching size in a stream [51].

This suggests that lower bounds for approximating matching size could be used more generally for establishing lower bounds for estimating Schatten p-norms. We abandon the use of the Tutte matrix, as an analysis of its singular values turns out to be quite involved. Instead, we devise simpler families of hard matrices which are related to hard graphs for estimating matching sizes. Our matrices are block diagonal in which each block has constant size (depending on ϵ). For functions $f(x) = |x|^p$ for $p > 0$ not an even integer, we show a constant-factor multiplicative gap in the value of $\sum_i f(\sigma_i)$ in the case where the input matrix is (1) block diagonal in which each block is the concatenation of an all-1s matrix and a diagonal matrix with an *even number* of 1s versus (2) block diagonal in which each block is the concatenation of an all-1s matrix and a diagonal matrix with an *odd number* of ones. We call these Case 1 and Case 2. We also refer to the 1s on a diagonal matrix inside a block as *tentacles.*

The analysis proceeds by looking at a block in which the number of tentacles follows a binomial distribution. We show that the expected value of $\sum_i f(\sigma_i)$ restricted to a block given that the number of tentacles is even, differs by a constant factor from the expected value of $\sum_i f(\sigma_i)$ restricted to a block given that the number of tentacles is odd. Using the hard distributions for matching [6, 26, 51], we can group the blocks into independent groups of four matrices and then apply a Chernoff bound across the groups to conclude that with high probability, $\sum_i f(\sigma_i)$ of the entire matrix in Case 1 differs by a $(1+\epsilon)$-factor from $\sum_i f(\sigma_i)$ of the entire matrix in Case 2. This is formalized in Theorem 3.

The number k of tentacles is subject to a binomial distribution supported on even or odd numbers in Case 1 or 2 respectively. Proving a "gap" in expectation for a random even value of k in a block versus a random odd value of k in a block is intractable if the expressions for the singular values are sufficiently complicated. For example, the singular values of the adjacency matrix of the instance in [15] for $p = 0$ involve roots of a cubic equation, which poses a great obstacle. Instead our hard instance has the advantage that the singular values $r(k)$ are the square roots of the roots of a quadratic equation, which are more tractable. The function value $f(r(k))$, viewed as a function of the number of tentacles k, can be expanded into a power series $f(r(k)) = \sum_{s=0}^{\infty} c_s k^s$, and the difference in expectation in the even and odd cases subject to a binomial distribution is

$$\begin{aligned}\sum_{k=0}^{m}(-1)^k \binom{m}{k} f(r(k)) &= \sum_{s=0}^{\infty} c_s \sum_{k=0}^{m}(-1)^i \binom{m}{k} k^s \\ &= \sum_{s=0}^{\infty} c_s (-1)^m m! \left\{ {s \atop m} \right\} \\ &= (-1)^m m! \sum_{s=m}^{\infty} c_s \left\{ {s \atop m} \right\}\end{aligned}$$

where $\left\{ {s \atop m} \right\}$ is the Stirling number of the second kind and $\left\{ {s \atop m} \right\} = 0$ for $s < m$, and where the second equality is a combinatorial identity. The problem reduces to analyzing the last series of s. For $f(x) = |x|^p$ ($p > 0$ not an even integer), with our choice of hard instance which we can parameterize by a small constant $\gamma > 0$, the problem reduces to showing that $c_s = c_s(\gamma) > 0$ for a small γ, and for all large s. However, $c_s(\gamma)$ is complicated and admits the form of a hypergeometric polynomial, which can be transformed to a different hypergeometric function $c_s'(\gamma)$ of simpler parameters. It has a standard infinite series expansion $c_s'(\gamma) = 1 + \sum_{n=0}^{\infty} a_n \gamma^n$. By analyzing the series coefficients, the infinite series can be split into three parts: a head part, a middle part, and a tail; each can be analyzed separately. Roughly speaking, the head term is alternating and decreasing and thus dominated by its first term, the tail term has geometrically decreasing terms and is also dominated by its first term, which is much smaller than the head, and finally the middle term is dominated by the head term.

The result for $f(x) = x^p$ generalizes to functions which are asymptotically x^p near 0 or infinity, by first scaling the input matrix by a small or a large constant.

A simple $\sqrt{n}$ lower bound. To illustrate our ideas, here we give a very simple proof of an $\Omega(\sqrt{n})$ lower bound for any real $p \neq 2$. Consider the three possible blocks

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},\ B = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix},\ C = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

A simple computation shows the two singular values are $(1,1)$ for A, are $((\sqrt{5}+1)/2, (\sqrt{5}-1)/2)$ for B, and $(2,0)$ for C. Our reduction above from the Boolean Hidden Matching Problem implies for Schatten p-norm estimation, we get an $\Omega(\sqrt{n})$ lower bound for c-approximation for a small enough constant $c > 1$, provided

$$\frac{1}{2}\cdot(1^p + 1^p) + \frac{1}{2}\cdot 2^p \neq \left(\left(\frac{\sqrt{5}+1}{2}\right)^p + \left(\frac{\sqrt{5}-1}{2}\right)^p \right),$$

which holds for any real $p \neq 2$.

Upper Bound. We illustrate the ideas of our upper bound with $p = 4$, in which case, $\|A\|_4^4 = \sum_{i,j} |\langle a_i, a_j\rangle|^2$, where a_i is the i-th row of A. Suppose for the moment that every row a_i had the same norm $\alpha = \Theta(1)$. It would then be easy to estimate $n\alpha^4 = \sum_i |\langle a_i, a_i\rangle|^2 = \Theta(n)$ just by looking at the norm of a single row. Moreover, by Cauchy-Schwarz, $\alpha^4 = \|a_i\|^4 \geq |\langle a_i, a_j\rangle|^2$ for all $j \neq i$. Therefore in order for $\sum_{i\neq j} |\langle a_i, a_j\rangle|^2$ to "contribute" to $\|A\|_4^4$, its value must be $\Omega(n\alpha^4)$, but since each summand is upper-bounded by α^4, there must be $\Omega(n)$ non-zero terms. It follows that if we sample $O(\sqrt{n})$ rows uniformly and in their entirety, by looking at all $O(n)$ pairs $|\langle a_i, a_j\rangle|^2$ for sampled rows a_i and a_j, we shall obtain $\Omega(1)$ samples of the "contributing" pairs $i \neq j$. Using that each row and column has $O(1)$ non-zero entries, this can be shown to be enough to obtain a good estimate to $\|A\|_4^4$ and it uses $O(\sqrt{n}\log n)$ bits of space.

In the general situation where the rows of A have differing norms, we need to sample them proportional to their squared 2-norm. Also, it is not possible to obtain the sampled rows a_i in their entirety, but we can obtain noisy approximations to them. We achieve this by adapting known algorithms for ℓ_2-sampling in a data stream [45, 2, 34] , and using our conditions that each row and each column of A have $O(1)$ non-zero entries. Given rows a_i and a_j, one can verify that $|\langle a_i, a_j\rangle|^2 \frac{\|A\|_F^4}{\|a_i\|_2^2\|a_j\|_2^2}$ is an unbiased estimator of $\|A\|_4^4$, and in fact, this is nothing other than importance sampling. It turns out that also in this more general case, only $O(\sqrt{n})$ rows need to be sampled, and we can look at all $O(n)$ pairs of inner products between such rows.

2. PRELIMINARIES

Notation. Let $\mathbb{R}^{n\times d}$ be the set of $n \times d$ real matrices. We write $X \sim \mathcal{D}$ for a random variable X subject to a probability distribution $\mathcal{D}$. Denote the uniform distribution on a set S (if it exists) by $\mathrm{Unif}(S)$.

We write $f \gtrsim g$ (resp. $f \lesssim g$) if there exists a constant $C > 0$ such that $f \geq Cg$ (resp. $f \leq Cg$). Also we write $f \simeq g$ if there exist constants $C_1 > C_2 > 0$ such that $C_2 g \leq f \leq C_1 g$. For the notations hiding constants, such as $\Omega(\cdot)$, $O(\cdot)$, $\lesssim$, $\gtrsim$, we may add subscripts to highlight the dependence, for example, $\Omega_a(\cdot)$, $O_a(\cdot)$, $\lesssim_a$, $\gtrsim_a$ mean that the hidden constant depends on a.

Singular values and matrix norms. Consider a matrix $A \in \mathbb{R}^{n\times n}$. Then $A^T A$ is a positive semi-definite matrix. The eigenvalues of $\sqrt{A^T A}$ are called the singular values of A, denoted by $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_n(A)$ in decreasing order. Let $r = \mathrm{rank}(A)$. It is clear that $\sigma_{r+1}(A) = \cdots = \sigma_n(A) = 0$. Define $\|A\|_p = (\sum_{i=1}^r (\sigma_i(A))^p)^{1/p}$ ($p > 0$). For $p \geq 1$, it is a norm over $\mathbb{R}^{n\times d}$, called the p-th *Schatten norm*, over $\mathbb{R}^{n\times n}$ for $p \geq 1$. When $p = 1$, it is also called the trace norm or nuclear norm. When $p = 2$, it is exactly the Frobenius norm $\|A\|_F$. Let $\|A\|_{op}$ denote the operator norm of A when treating A as a linear operator from ℓ_2^n to ℓ_2^n. It holds that $\lim_{p\to\infty} \|A\|_p = \sigma_1(A) = \|A\|_{op}$.

The Ky-Fan k-norm of A, denoted by $\|A\|_{F_k}$, is defined as the sum of the largest k singular values: $\|A\|_{F_k} = \sum_{i=1}^k \sigma_i(A)$. Note that $\|A\|_{F_1} = \|A\|_{op}$ and $\|A\|_{F_k} = \|A\|_1$ for $k \geq r$.

Communication Complexity. We shall use a problem called Boolean Hidden Hypermatching, denoted by $\mathrm{BHH}_{t,n}^0$, defined in [51].

Definition 1. In the Boolean Hidden Hypermatching Problem $\mathrm{BHH}_{t,n}$, Alice gets a Boolean vector $x \in \{0,1\}^n$ with

$n = 2rt$ for some integer r and Bob gets a perfect t-hypermatching M on the n coordinates of x, i.e., each edge has exactly t coordinates, and a binary string $w \in \{0,1\}^{n/t}$. Let Mx denote the vector of length n/t defined as $(\bigoplus_{1\le i\le t} x_{M_{1,i}}$, ..., $\bigoplus_{1\le i\le t} x_{M_{n/t,i}})$, where $\{(M_{j,1},\dots,M_{j,t})\}_{j=1}^{n/t}$ are edges of M. It is promised that either $Mx \oplus w = \mathbf{1}^{n/t}$ or $Mx \oplus w = \mathbf{0}^{n/t}$. The problem is to return 1 in the first case and 0 otherwise.

They proved that this problem has an $\Omega(n^{1-1/t})$ randomized one-way communication lower bound by proving a lower bound for deterministic protocols with respect to the hard distribution in which x and M are independent and respectively uniformly distributed, and $w = Mx$ with probability 1/2 and $w = \overline{Mx}$ (bitwise negation of Mx) with probability 1/2. In [15], Bury and Schwiegelshohn defined a version without w and with the constraint that $w_H(x) = n/2$, for which they also showed an $\Omega(n^{1-1/t})$ lower bound. We shall use this version, with a slight modification.

Definition 2. In the *Boolean Hidden Hyper-matching Problem* $\mathrm{BHH}^0_{t,n}$, Alice gets a Boolean vector $x \in \{0,1\}^n$ with $n = 4rt$ for some $r \in \mathbb{N}$ and even integer t and $w_H(x) = n/2$, Bob gets a perfect t-hypermatching M on the n coordinates of x, i.e., each edge has exactly t coordinates. We denote by Mx the Boolean vector of length n/t given by $\left(\bigoplus_{i=1}^t x_{M_{1,i}},\dots,\bigoplus_{i=1}^t x_{M_{n/t,i}}\right)$, where $\{(M_{j,1},\dots,M_{j,t})\}_{j=1}^{n/t}$ are the edges of M. It is promised that either $Mx = \mathbf{1}^{n/t}$ or $Mx = \mathbf{0}^{n/t}$. The problem is to return 1 in the first case and 0 otherwise.

A slightly modified (yet remaining almost identical) proof as in [15] shows that this problem also has an $\Omega(n^{1-1/t})$ randomized one-way communication lower bound. We include the proof here.

Proof. We reduce $\mathrm{BHH}_{t,n}$ to $\mathrm{BHH}^0_{t,2n}$. Let $x \in \{0,1\}^n$ with $n = 2rt$ for some r, M be a perfect t-hypermatching on the n coordinates of x and $x \in \{0,1\}^{w/t}$. Define $x' = \begin{pmatrix} x^T & \bar{x}^T \end{pmatrix}^T$ to be the concatenation of x and $\bar{x}$ (bitwise negation of x).

Let $\{x_1,\dots,x_t\} \in M$ be the l-th hyperedge of M. We include two hyperedges in M' as follows. When $w_l = 0$, include $\{x_1,\dots,x_t\}$ and $\{\overline{x_1},\overline{x_2},\dots,\overline{x_t}\}$ in M; when $w_l = 1$, include $\{\overline{x_1},x_2,\dots,x_t\}$ and $\{x_1,\overline{x_2}\dots,\overline{x_t}\}$ in M'. The observation is that we flip an even number of bits in the case $w_l = 0$ and an odd number of bits when $w_l = 1$, and since t is even, this does not change the parity of each set. Therefore $M'x' = 0^{2n}$ if $Mx + w = 0^{n/2}$ and $M'x' = 1^{2n}$ if $Mx + w = 1^{n/2}$. The lower bound then follows from the lower bound for $\mathrm{BHH}_{t,n}$. □

When t is clear from context, we shorthand $\mathrm{BHH}^0_{t,n}$ as BHH^0_n.

3. SCHATTEN NORMS

Let $D_{m,k}$ $(0 \le k \le m)$ be an $m \times m$ diagonal matrix with the first k diagonal elements equal to 1 and the remaining diagonal entries 0, and let $\mathbf{1}_m$ be an m dimensional vector full of 1s. Define

$$M_{m,k} = \begin{pmatrix} \mathbf{1}_m \mathbf{1}_m^T & 0 \\ \sqrt{\gamma} D_{m,k} & 0 \end{pmatrix},$$

where $\gamma > 0$ is a constant (which may depend on m).

Our starting point is the following theorem. Let $m \ge 2$ and $p_m(k) = \binom{m}{k}/2^{m-1}$ for $0 \le k \le m$. Let $\mathcal{E}(m)$ be the probability distribution on even integers $\{0,2,\dots,m\}$ with probability density function $p_m(k)$, and $\mathcal{O}(m)$ be the distribution on odd integers $\{1,3,\dots,m-1\}$ with density function $p_m(k)$. We say a function f on square matrices is diagonally block-additive if $f(X) = f(X_1) + \cdots + f(X_s)$ for any block diagonal matrix X with square diagonal blocks $X_1,\dots,X_s$. It is clear that $f(X) = \sum_i f(\sigma_i(X))$ is diagonally block-additive.

Theorem 3. *Let t be an even integer and $X \in \mathbb{R}^{N\times N}$, where N is sufficiently large. Let f be a function of square matrices that is diagonally block-additive. If there exists $m = m(t)$ such that*

$$\mathop{\mathbb{E}}_{q\sim\mathcal{E}(t)} f(M_{m,q}) - \mathop{\mathbb{E}}_{q\sim\mathcal{O}(t)} f(M_{m,q}) \neq 0, \tag{1}$$

then there exists a constant $c = c(t) > 0$ such that any streaming algorithm that approximates $f(X)$ within a factor $1 \pm c$ with constant error probability must use $\Omega_t(N^{1-1/t})$ bits of space.

Proof. We reduce the problem from the $\mathrm{BHH}^0_{t,n}$ problem. Let $n = Nt/(2m)$. For the input of the problem $\mathrm{BHH}^0_{t,n}$, construct a graph G as follows. The graph contains n vertices $v_1,\dots,v_n$, together with n/t cliques of size m, together with edges connecting v_i's with the cliques according to Alice's input x. These latter edges are called 'tentacles'. In the j-th clique of size m, we fix t vertices, denoted by $w_{j,1},\dots,w_{j,t}$. Whenever $x_i = 1$ for $i = (j-1)(n/t) + r$, we join v_i and $w_{j,r}$ in the graph G.

Let $\mathcal{M}$ be constructed from G as follows: both the rows and columns are indexed by nodes of G. For every pair w, v of clique nodes in G, let $\mathcal{M}_{w,v} = 1$, where we allow $w = v$. For every 'tentacle' (u,w), where w is a clique node, let $\mathcal{M}(u,w) = \sqrt{\gamma}$. Then $\mathcal{M}$ is an $N \times N$ block diagonal matrix of the following form after permuting the rows and columns

$$\mathcal{M}_{n,m,t} = \begin{pmatrix} M_{m,q_1} & & & \\ & M_{m,q_2} & & \\ & & \ddots & \\ & & & M_{m,q_{n/t}} \end{pmatrix}, \tag{2}$$

where $q_1,\dots,q_{n/t}$ satisfy the constraint that $q_1 + q_2 + \cdots + q_{n/t} = n/2$ and $0 \le q_i \le t$ for all i. It holds that $f(\mathcal{M}_{n,m,t}) = \sum_i f(M_{m,q_i})$.

Alice and Bob will run the following protocol. Alice keeps adding the matrix entries corresponding to 'tentacles' while running the algorithm for estimating $f(\mathcal{M})$. Then she sends the state of the algorithm to Bob, who will continue running the algorithm while adding the entries corresponding to the cliques defined by the matching he owns. At the end, Bob outputs which case the input of BHH^0_n belongs to based upon the final state of the algorithm.

From the reduction for $\mathrm{BHH}^0_{t,n}$ and the hard distribution of $\mathrm{BHH}_{t,n}$, the hard distribution of $\mathrm{BHH}^0_{t,n}$ exhibits the following pattern: $q_1,\dots,q_{n/t}$ can be divided into $n/(2t)$ groups. Each group contains two q_i's and has the form $(q, t-q)$, where q is subject to distribution $\mathcal{E}(t)$ or $\mathcal{O}(t)$ depending on the promise. Furthermore, the q's across the $n/(2t)$ groups are independent. The two cases to distinguish are that all q_i's are even (referred to as the *even case*) and that all q_i's are odd (referred to as the *odd case*).

For notational simplicity, let $F_q = f(M_{m,q})$. Suppose that the gap in (1) is positive. Let $A = \mathbb{E}_{q\sim\mathcal{E}(t)} 2(F_q + F_{t-q})$ and $B = \mathbb{E}_{q\sim\mathcal{O}(t)} 2(F_q + F_{t-q})$, then $A - B > 0$. Summing up $(n/2t)$ independent groups and applying a Chernoff bound, with high probability, $f(\mathcal{M}) \geq (1-\delta)\frac{n}{2t}A$ in the even case and $f(\mathcal{M}) \leq (1+\delta)\frac{n}{2t}A$, where δ is a small constant to be determined. If we can approximate $f(\mathcal{M})$ up to a $(1\pm c)$-factor, say X, then with constant probability, in the even case we have an estimate $X \geq (1-c)(1-\delta)\frac{n}{2t}A$ and in the odd case $X \leq (1+c)(1+\delta)\frac{n}{2t}A$. Choose $\delta = c$ and choose $c < \frac{B-A}{3(B+A)}$. Then there will be a gap between the estimates in the two cases. The conclusion follows from the lower bound for the BHH_n^0 problem.

A similar argument works when (1) is negative. $\square$

Our main theorem in this section is the following, a restatement of Theorem 1 advertised in the introduction.

Theorem 4. *Let $p \in (0,\infty)\setminus 2\mathbb{Z}$. For every even integer t, there exists a constant $c = c(t) > 0$ such that any algorithm that approximates $\|X\|_p^p$ within a factor $1\pm c$ with constant probability in the streaming model must use $\Omega_t(N^{1-1/t})$ bits of space.*

The theorem follows from applying Theorem 3 to $f(x) = x^p$ and $m = t$ and verifying that (1) is satisfied. The proof is technical and thus postponed to Section 4.

For even integers p, we change our hard instance to

$$M_{m,k} = \mathbf{1}_m\mathbf{1}_m^T - I_m + D_{m,k},$$

where I_m is the $m\times m$ identity matrix. We then have the following lemma, whose proof is postponed to the end of Section 5.

Lemma 1. *For $f(x) = x^p$ and integer $p \geq 2$, the gap condition (1) is satisfied if and only if $t \leq p/2$, under the choice that $m = t$.*

This yields an $\Omega(n^{1-2/p})$ lower bound, which agrees with the lower bound obtained by injecting the F_p moment problem into the diagonal elements of the input matrix [28, 33], but here we have the advantage that the entries are bounded by a constant independent of n. In fact, for even integers p, we show our lower bound is tight up to $\mathrm{poly}(\log n)$ factors for matrices in which every row and column has $O(1)$ non-zero elements by providing an algorithm in Section 6 for the problem. Hence our matrix construction $M_{m,k}$ will not give a substantially better lower bound. Our lower bound for even integers p also helps us in the setting of general functions f in Section 7.

4. PROOF OF THEOREM 4

Proof. First we find the singular values of $M_{m,k}$. Assume that $1 \leq k \leq m-1$ for now.

$$M_{m,k}^T M_{m,k} = \begin{pmatrix} m\mathbf{1}_m\mathbf{1}_m^T + \gamma D_{m,k} & 0 \\ 0 & 0 \end{pmatrix}.$$

Let e_i denote the i-th vector of the canonical basis of $\mathbb{R}^{2m}$. It is clear that $e_1 - e_i$ $(i = 2,\ldots,k)$ are the eigenvectors with corresponding eigenvalue γ, which means that $M_{m,k}$ has $k-1$ singular values of $\sqrt{\gamma}$. Since $M_{m,k}$ has rank $k+1$, there are two more non-zero singular values, which are the square roots of another two eigenvalues, say $r_1(k)$ and $r_2(k)$, of $M_{m,k}^T M_{m,k}$. It follows from $\mathrm{tr}(M_{m,k}^T M_{m,k}) = m + \gamma k$ that $r_1(k) + r_2(k) = m^2 + \gamma$ and from $\|M_{m,k}M_{m,k}\|_F^2 = (m+\gamma)^2 k + (m^2-k)m^2$ that $r_1^2(k) + r_2^2(k) = m^4 + 2\gamma km + \gamma^2$. Hence $r_1(k)r_2(k) = m^2\gamma - km\gamma$. In summary, the non-zero singular values of $M_{m,k}$ are: $\sqrt{\gamma}$ of multiplicity $k-1$, $\sqrt{r_1(k)}$ and $\sqrt{r_2(k)}$, where $r_{1,2}(k)$ are the roots of the following quadratic equation:

$$x^2 - (m^2+\gamma)x + (m^2 - km)\gamma = 0.$$

The conclusion above remains formally valid for $k = 0$ and $k = m$. In the case of $k = 0$, the matrix $M_{m,0}$ has a single non-zero singular value m, while $r_1(k) = m^2$ and $r_2(k) = \gamma$. In the case of $k = m$, the matrix $M_{m,m}$ has singular values $\sqrt{m^2+\gamma}$ of multiplicity 1 and $\sqrt{\gamma}$ of multiplicity $m-1$, while $r_1(k) = m^2+\gamma$ and $r_2(k) = 0$. Hence the left-hand side of (1) becomes

$$\begin{aligned} &\frac{1}{2^{m-1}}\sum_{\text{even } k}\binom{m}{k}\left((k-1)\gamma^{p/2} + r_1^{p/2}(k) + r_2^{p/2}(k)\right) \\ &\quad - \frac{1}{2^{m-1}}\sum_{\text{odd } k}\binom{m}{k}\left((k-1)\gamma^{p/2} + r_1^{p/2}(k) + r_2^{p/2}(k)\right) \\ &= \frac{1}{2^{m-1}}(G_1 + G_2), \end{aligned}$$

where γ^p on both sides cancel and

$$G_i = \sum_k (-1)^k \binom{m}{k} r_i^{\frac{p}{2}}(k), \quad i = 1,2. \tag{3}$$

Our goal is to show that $G_1 + G_2 \neq 0$ when p is not an even integer. To simplify and to abuse notation, hereinafter in this section, we replace $p/2$ with p in (3) and hence G_1 and G_2 are redefined to be

$$G_i = \sum_k (-1)^k \binom{m}{k} r_i^p(k), \quad i = 1,2, \tag{4}$$

and our goal becomes to show that $G_1 + G_2 \neq 0$ for *non-integers p*.

Next we choose

$$\begin{aligned} r_1(k) &= \frac{1}{2}\left(m^2 + \gamma + \sqrt{m^4 - 2\gamma m^2 + \gamma^2 + 4\gamma km}\right), \\ r_2(k) &= \frac{1}{2}\left(m^2 + \gamma - \sqrt{m^4 - 2\gamma m^2 + \gamma^2 + 4\gamma km}\right). \end{aligned}$$

We claim that they admit the following power series expansion in k (proof deferred to Section 4.2),

$$r_1^p(k) = \sum_{s\geq 0} A_s k^s, \quad r_2^p(k) = \sum_{s\geq 0} B_s k^s,$$

where for $s \geq 2$,

$$A_s = \frac{(-1)^{s-1}\gamma^s m^{2p-s}}{s!(m^2-\gamma)^{2s-1}}\sum_{i=0}^{s-1}(-1)^i\binom{s-1}{i}F_{p,s,i}\gamma^{s-i-1}m^{2i} \tag{5}$$

and

$$B_s = \frac{(-1)^s\gamma^p m^s}{s!(m^2-\gamma)^{2s-1}}\sum_{i=0}^{s-1}(-1)^i\binom{s-1}{i}F_{p,s,i}\gamma^i m^{2(s-i-1)}, \tag{6}$$

and

$$F_{p,s,i} = \prod_{j=0}^{s-i-1}(p-j)\cdot\prod_{j=1}^{i}(p-2s+j).$$

We analyse A_s first. It is easy to see that $|F_{s,i}| \le (2s)^s$ for $s > 2p$, and hence

$$\begin{aligned}|A_s| &\le \frac{\gamma^s m^{2p-s}}{s!(m^2-\gamma)^{2s-1}}\sum_{i=0}^{s-1}\binom{s-1}{i}|F_{s,i}|\gamma^{s-i-1}m^{2i}\\ &\le \frac{\gamma^s m^{2p-s}}{\sqrt{2\pi s}(\frac{s}{e})^s(m^2-\gamma)^{2s-1}}(2s)^s m^{2(s-1)}\left(1+\frac{\gamma}{m^2}\right)^{s-1}\\ &\le m^{2p}\frac{2e}{\sqrt{2\pi s}}\frac{\gamma}{m^2(m^2-\gamma)}\left(\frac{4em\gamma}{(m^2-\gamma)^2}\right)^{s-1},\end{aligned}$$

whence it follows immediately that $\sum_s A_s k^s$ is absolutely convergent. We can apply term after term the identity

$$\sum_{k=0}^{m}\binom{m}{k}k^s(-1)^k = \left\{{s \atop m}\right\}(-1)^m m!, \tag{7}$$

where $\left\{{s \atop m}\right\}$ is the Stirling number of the second kind, and obtain that (since m is even)

$$G_1 = \sum_{s\ge m}\left\{{s \atop m}\right\}m!A_s,$$

which, using the fact that $\left\{{s \atop m}\right\}m! \le m^s$, can be bounded as

$$|G_1| \le \sum_{s\ge m} m^s|A_s| \le c_1 m^{2p}\left(\frac{c_2}{m^2}\right)^{m-1}$$

for some absolute constants $c_1, c_2 > 0$.

Bounding G_2 is more difficult, because B_s contains an alternating sum. However, we are able to prove the following critical lemma, whose proof is postponed to Section 4.1.

Lemma 2. *For any fixed non-integer $p > 0$, one can choose γ_0 and m such that B_s have the same sign for all $s \ge m$ and all $0 < \gamma < \gamma_0$.*

Since $\sum_{s\ge m} B_s m^s$ is a convergent series with positive terms, we can apply (7) to $\sum_s B_s k^s$ term after term, giving the gap contribution from $r_2(k)$ as

$$G_2 = \sum_{s\ge m}\left\{{s \atop m}\right\}m!B_s.$$

Let $a_{m,i}$ be the summand in B_m, that is,

$$a_{m,i} = \binom{s-1}{i}F_{s,i}\gamma^i m^{2(s-i-1)}.$$

Since p is not an integer, $a_{m,i} \ne 0$ for all i. Then

$$r_{m,i} := \frac{a_{m,i}}{a_{m,i-1}} = \frac{m-i-1}{i+1}\cdot\frac{p-2m+i}{p-m+i}\cdot\frac{\gamma}{m^2}.$$

If we choose m such that $m^2/\gamma \gtrsim ([p]-1)/(p-[p])$ when $p > 1$ or $m^2/\gamma \gtrsim 1/(p-[p])$ when $p < 1$, it holds that $|r_{m,i}| \le 1/3$ for all i and thus the sum is dominated by $a_{m,0}$. It follows that

$$\begin{aligned}G_2 \ge B_m &\gtrsim \frac{\gamma^p m^m}{s!(m^2-\gamma)^{2m-1}}|a_{m,0}|\\ &\gtrsim (p-[p])^2[p]!\frac{\gamma^p}{(m-\lceil p\rceil-1)^{p-[p]}m^{[p]}}\end{aligned}$$

It follows from Lemma 2 that the above is also a lower bound for G_2. Therefore G_1 is negligible compared with G_2 and $G_1 + G_2 \ne 0$. This ends the proof of Theorem 4. $\square$

4.1 Proof of Lemma 2

The difficulty is due to the fact that the sum in B_s is an alternating sum. However, we notice that the sum in B_s is a hypergeometric polynomial with respect to γ/m^2. This is our starting point.

Proof of Lemma 2. Let $x = \gamma/m^2$ and write B_s as

$$B_s = (-1)^{s-1}\frac{\gamma^p m^{3s-2}}{s!(m^2-\gamma)^{2s-1}}\sum_{i=0}^{s-1}(-1)^{i+1}\binom{s-1}{i}F_{s,i}x^i, \tag{8}$$

Then

$$B_s m^s = (-1)^{s-1}\gamma^p\frac{m^{4s-2}}{s!(m^2-\gamma)^{2s-1}}\sum_{i=0}^{s-1}(-1)^{i+1}\binom{s-1}{i}F_{s,i}x^i.$$

Observe that the sum can be written using a hypergeometric function and the series above becomes

$$\begin{aligned}B_s m^s = (-1)^s\gamma^p&\frac{1}{s!(1-x)^{2s-1}}\cdot\frac{\Gamma(1+p)}{\Gamma(1+p-s)}\\ &\cdot {}_2F_1(1-s, 1+p-2s; 1+p-s; x),\end{aligned}$$

where $\Gamma(x)$ is the gamma function and ${}_2F_1(a,b;c;x)$ the hypergeometric function with 2 upper parameters and 1 lower parameter, defined as

$$\begin{aligned}&{}_2F_1(a,b;c;x)\\ &= \sum_{n=0}^{\infty}\frac{a(a+1)\cdots(a+n-1)\cdot b(b+1)\cdots(b+n-1)}{c(c+1)\cdots(c+n-1)}\frac{x^n}{n!}.\end{aligned}$$

Invoking Euler's Transformation (see, e.g., [5, p78])

$${}_2F_1(a,b;c;x) = (1-x)^{c-a-b}{}_2F_1(c-a, c-b; c; x)$$

gives

$${}_2F_1(1-s,1+p-2s;1+p-s;x) = (1-x)^{2s-1}{}_2F_1(p,s;p-s+1;x). \tag{9}$$

Therefore

$$B_s m^s = (-1)^s\frac{\gamma^p\Gamma(1+p)}{s!\ \Gamma(1+p-s)}{}_2F_1(p,s;p-s+1;x). \tag{10}$$

Since $\Gamma(1+p-s)$ has alternating signs with respect to s, it suffices to show that ${}_2F_1(p,s;p-s+1;x) > 0$ for all $x \in [0,x^*]$ and all $s \ge s^*$, where both x^* and s^* depend only on p.

Now, we write ${}_2F_1(p,s;p-s+1;x) = \sum_n b_n$, where

$$b_n = \frac{p(p+1)\cdots(p+n-1)\cdot s(s+1)\cdots(s+n-1)}{(1+p-s)(2+p-s)\cdots(n+p-s)n!}x^n.$$

It is clear that b_n has the same sign for all $n \ge s-[p]$, and has alternating signs for $n \le s-\lceil p\rceil$. Consider

$$\left|\frac{b_n}{b_{n-1}}\right| = \frac{(p+n-1)(s+n-1)}{(p-s+n)n}x.$$

One can verify that when $n \ge 2s$ and $x \le 1/10$, $|b_n/b_{n-1}| < 3x \le 1/3$ and thus $|\sum_{n\ge 2s} b_n| \le \frac{3}{2}|b_{2s}|$. Also, when $s \ge 3p$ is large enough, $x \le 1/10$ and $n \le s/2$. It holds that

$|b_n/b_{n-1}| < 1$ and thus $\{|b_n|\}$ is decreasing when $n \le s/2$. (In fact, $\{|b_n|\}$ is decreasing up to $n = \frac{1-x}{1+x}s + O(1)$.) Recall that $\{b_n\}$ has alternating signs for $n \le s/2$, and it follows that

$$0 \le \sum_{2 \le n \le s/2} b_n \le b_2.$$

Next we bound $\sum_{s/2<n<2s} b_n$. Let $n^* = \mathrm{argmax}_{s/2<n<2s} |b_n|$. When $n^* \le s - \lceil p \rceil$,

$$\begin{aligned}
&\left|\sum_{s/2<n<2s} b_n\right| \\
&\le \frac{3}{2}s|b_{n^*}| \\
&\le \frac{3}{2}s\frac{p(p+1)\cdots(p+n^*)}{n^*!}\frac{(s-[p]-n^*)!}{(s-[p]-1)!}\frac{(s+n^*-1)!}{s!}x^{n^*} \\
&\le \frac{3}{2}s(n^*)^p\frac{\binom{s+n^*-1}{s}}{\binom{s-[p]-1}{s-[p]-n^*}}x^{n^*} \\
&\le \frac{3}{2}s\cdot es^p\cdot 4^s\cdot x^{s/2} \\
&\le x^3,
\end{aligned}$$

provided that x is small enough (independent of s) and s is big enough. When $n^* > s - \lceil p \rceil$,

$$\begin{aligned}
&\left|\sum_{s/2<n<2s} b_n\right| \\
&\le \frac{3}{2}s|b_{n^*}| \\
&\le \frac{3}{2}s\frac{p(p+1)\cdots(p+n^*)}{n^*!}\frac{(s+n^*-1)!}{(s-[p]-1)!(n^*-s+[p]-1)!s!}x^{n^*} \\
&\le \frac{3}{2}s^2(n^*)^p\binom{s+n^*-1}{s-[p],n^*-s+[p]-1,s}x^{n^*} \\
&\le \frac{3}{2}s^2\cdot e(2s)^p\cdot 3^{3s-1}\cdot x^{s-\lceil p\rceil} \\
&\le x^3.
\end{aligned}$$

provided that x is small enough (independent of s) and s is big enough. Similarly we can bound, under the same assumption on x and s as above, that $|b_{2s}| \le x^3$. Therefore $|\sum_{n>s/2} b_n| \le Kx^3$ for some K and sufficiently large s and small x, all of which depend only on p.

It follows that

$$\begin{aligned}
&{}_2F_1(p,s;p-s+1;x) \\
&\ge 1 - \frac{ps}{s-p-1}x - \sum_{2\le n\le s/2} b_n - \left|\sum_{n>s/2} b_n\right| \\
&\ge 1 - \frac{ps}{s-p-1}x - b_2 - Kx^3 \\
&\ge 1 - \frac{ps}{s-p-1}x - \frac{p(p+1)s(s+1)}{2(s-p-1)(s-p-2)}x^2 - Kx^3 \\
&> 0
\end{aligned}$$

for sufficiently large s and small x (independent of s).

The proof of Lemma 2 is now complete. □

4.2 Proof of Power Series Expansion

Proof of Claim. We first verify the series expansion of $r_1(k)$. It is a standard result that for $|x| \le 1/4$,

$$\frac{1+\sqrt{1-4x}}{2} = 1 - \sum_{n=1}^{\infty} C_{n-1}x^n,\ \ \frac{1-\sqrt{1-4x}}{2} = \sum_{n=1}^{\infty} C_{n-1}x^n,$$

where $C_n = \frac{1}{n+1}\binom{2n}{n}$ is the n-th Catalan number. Let $x = -\gamma km/(m^2-\gamma)^2$, we have

$$\begin{aligned}
r_1(k) &= m^2\frac{1+\sqrt{1-4x}}{2} + \gamma\frac{1-\sqrt{1-4x}}{2} \\
&= m^2 - (m^2-\gamma)\sum_{n=1}^{\infty} C_{n-1}x^n \\
&= m^2 - \sum_{n=1}^{\infty} C_{n-1}\frac{(-1)^n\gamma^n k^n m^n}{(m^2-\gamma)^{2n-1}}
\end{aligned}$$

Applying the generalized binomial theorem,

$$\begin{aligned}
&r_1(k)^p \\
&= m^{2p} + \sum_{i=1}^{\infty}\binom{p}{i}(-1)^i m^{2(p-i)}\left(\sum_{n=1}^{\infty} C_{n-1}\frac{(-1)^n\gamma^n k^n m^n}{(m^2-\gamma)^{2n-1}}\right)^i \\
&= m^{2p} + \sum_{i=1}^{\infty}\binom{p}{i}(-1)^i m^{2(p-i)}\sum_{n_1,\ldots,n_i\ge 1}\frac{\prod_{j=1}^{i} C_{n_j-1}(-k\gamma m)^{\sum_j n_j}}{(m^2-\gamma)^{2\sum_j n_j - i}} \\
&= m^{2p} + \sum_{s=1}^{\infty}\sum_{i=1}^{s}\binom{p}{i}m^{2(p-i)}\frac{(-k\gamma m)^s}{(m^2-\gamma)^{2s-i}}\sum_{\substack{n_1,\ldots,n_i\ge 1\\ n_1+\cdots+n_i=s}}\prod_{j=1}^{i} C_{n_j-1},
\end{aligned}$$

where we replace $\sum_j n_j$ with s. It is a known result using the Lagrange inversion formula that (see, e.g., [49, p128])

$$\sum_{\substack{n_1,\ldots,n_i\ge 1\\ n_1+\cdots+n_i=s}}\prod_{j=1}^{i} C_{n_j-1} = \frac{i}{s}\binom{2s-i-1}{s-1}$$

Hence (replacing i with $i+1$ in the expression above)

$$\begin{aligned}
A_s = \frac{(-1)^{s+1}\gamma^s m^{2p}}{(m^2-\gamma)^{2s-1}}\sum_{i=0}^{s-1}(-1)^i\binom{p}{i+1} \\
\cdot\frac{i+1}{s}\binom{2s-i-2}{s-1}\cdot m^{s-2(i+1)}(m^2-\gamma)^i \quad (11)
\end{aligned}$$

To see that (11) agrees with (5), it suffices to show that

$$\begin{aligned}
&\sum_{i=0}^{s-1}(-1)^i\binom{s-1}{i}F_{p,s,i}\gamma^{s-i-1}m^{2i-s} = \\
&s!\sum_{i=0}^{s-1}(-1)^i\binom{p}{i+1}\frac{i+1}{s}\binom{2s-i-2}{s-1}\cdot m^{s-2(i+1)}(m^2-\gamma)^i
\end{aligned}$$

Comparing the coefficients of γ^j, we need to show that

$$\begin{aligned}
&(-1)^{s-1}\binom{s-1}{j}F_{p,s,j,s-j-1} \\
&\qquad = s!\sum_{i=j}^{s-1}(-1)^i\binom{p}{i+1}\frac{i+1}{s}\binom{2s-i-2}{s-1}\binom{i}{j}
\end{aligned}$$

Note that both sides are a degree-s polynomial in p with head coefficient $(-1)^{s-1}$, so it suffices to verify they have the same roots. It is clear that $0, \ldots, j$ are roots. When $r > j$, each summand on the right-hand is non-zero, and the right-hand side can be written as, using the ratio of successive summands,

$$S_0 \cdot {}_2F_1(1+j-p, 1+j-s; 2+j-2s; 1),$$

where $S_0 \neq 0$. Hence it suffices to show that ${}_2F_1(1+j-p, 1+j-s; 2+j-2s; 1) = 0$ when $p = 2s-k$ for $1 \le k \le s-j-1$. This holds by the Chu-Vandermonde identity (see, e.g., [5, Corollary 2.2.3]), which states, in our case, that

$$ {}_2F_1(1+j-p, 1+j-s; 2+j-2s; 1) = \frac{(1+p-2s)(2+p-2s)\cdots(-1+p-s-j)}{(2+j-2s)(3+j-2s)\cdots(-s)}.$$

The proof of expansion of $r_1(k)$ is now complete. Similarly, starting from $r_2(k) = \gamma(\frac{1+\sqrt{1-4x}}{2}) + m^2(\frac{1-\sqrt{1-4x}}{2})$, we can deduce as an intermediate step that

$$B_s = \frac{(-1)^s \gamma^p m^s}{(m^2-\gamma)^{2s-1}} \sum_{i=0}^{s-1} (-1)^i \binom{p}{i+1} \cdot \frac{i+1}{s} \binom{2s-i-2}{s-1} \gamma^{s-i-1} (m^2-\gamma)^i$$

and then show it agrees with (6). The whole proof is almost identical to that for $r_1(k)$.

The convergence of both series for $0 \le k \le m$ follows from the absolute convergence of series expansion of $(1+z)^p$ for $|z| \le 1$. Note that $r_2(m)$ corresponds to $z = -1$. $\square$

We remark that one can continue from (10) to bound that $\sum_s B_s m^s \lesssim 1/m^p$, where the constant depends on p. Hence $G_1 + G_2 \simeq 1/m^p$ and thus the gap in (1) is $\Theta(1/2^m m^p)$ with constant dependent on p only.

5. PROOFS RELATED TO EVEN p

Now we prove Lemma 1 below. Since our new $M_{m,k}$ is symmetric, the singular values are the absolute values of the eigenvalues. For $0 < k < m$, $-e_i + e_m$ ($i = k+1, \ldots, m-1$) are eigenvectors of eigenvalue -1. Hence there are $m-k-1$ singular values of 1. Observe that the bottom $m-k+1$ rows of $M_{m,k}$ are linearly independent, so the rank of $M_{m,k}$ is $m-k+1$ and there are two more non-zero eigenvalues. Using the trace and Frobenius norm as in the case of the old $M_{m,k}$, we find that the other two eigenvalues $\lambda_1(k)$ and $\lambda_2(k)$ satisfy $\lambda_1(k) + \lambda_2(k) = m-1$ and $\lambda_1^2(k) + \lambda_2^2(k) = (m-1)^2 + 2k$. Therefore, the singular values $r_{1,2}(k) = |\lambda_{1,2}(k)| = \frac{1}{2}(\sqrt{(m-1)^2+4k} \pm (m-1))$. Formally define $r_{1,2}(k)$ for $k = 0$ and $k = m$. When $k = 0$, the singular values are actually $r_1(0)$ and $r_2(m)$ and when $k = m$, the singular values are $r_1(m)$ and $r_2(0)$. Since $k = 0$ and $k = m$ happens with the same probability, this 'swap' of singular values does not affect the sum. We can proceed pretending that $r_{1,2}(k)$ are correct for $k = 0$ and $k = m$.

Recall that the gap is $\frac{1}{2^{m-1}}(G_1 + G_2)$, where G_1 and G_2 are as defined in (4) (we do not need to replace $p/2$ with p here). It remains the same to show that $G_1 + G_2 \neq 0$ if and only if $m \le [p/2]$.

Proof of Lemma 1. Applying the binomial theorem,

$$r_1^p(k) + r_2^p(k)$$

$$= \frac{1}{2^{p-1}} \sum_{i:2|(p-i)} \binom{p}{i} (m-1)^i ((m-1)^2 + 4k)^{\frac{p-i}{2}}$$

$$= \frac{1}{2^{p-1}} \sum_{i:2|(p-i)} \binom{p}{i} (m-1)^i \sum_{j=0}^{\frac{p-i}{2}} \binom{\frac{p-i}{2}}{j} (m-1)^{2j} 4^{\frac{p-i}{2}-j} k^{\frac{p-i}{2}-j}.$$

Therefore

$$G_1 + G_2 = (-1)^m m! \sum_{i:2|(p-i)} \binom{p}{i} (m-1)^i \cdot \sum_{j=0}^{\frac{p-i}{2}} \binom{\frac{p-i}{2}}{j} (m-1)^{2j} 4^{\frac{p-i}{2}-j} \left\{ \begin{matrix} \frac{p-i}{2} \\ m \end{matrix} \right\}.$$

Note that all terms are of the same sign (interpreting 0 as any sign) and the sum vanishes only when $\left\{ \begin{smallmatrix} \frac{p-i}{2} \\ m \end{smallmatrix} \right\} = 0$ for all i, that is, when $m > [\frac{p}{2}]$. $\square$

Although when p is even, we have $G_1 + G_2 = 0$, however, we can show that $G_1, G_2 \neq 0$, which will be useful for some applications in Section 7. It suffices to show the following lemma.

Lemma 3. *When p is even, the contribution from individual $r_i(k)$ $(i = 1, 2)$ is not zero, provided that m is large enough.*

Proof. First we have

$$r_2^p(k) = \frac{(m-1)^p}{2^p} \sum_{s=0}^{\infty} \sum_{i=0}^{p} \binom{p}{i} (-1)^i \binom{i/2}{s} \frac{4^s}{(m-1)^{2s}} k^s.$$

When $s > p/2$, the binomial coefficient $\binom{i/2}{s}$ vanishes if i is an even integer. Plugging in (7) we obtain the gap contribution

$$-\frac{(m-1)^p m!}{2^p} \sum_{s \ge m} \left\{ \begin{matrix} s \\ m \end{matrix} \right\} \frac{4^s}{(m-1)^{2s}} \sum_{\substack{\text{odd } i \\ 1 \le i \le p-1}} \binom{p}{i} \binom{i/2}{j}.$$

Hence it suffices to show that $\sum_s B_s \neq 0$, where

$$B_s = \left\{ \begin{matrix} s \\ m \end{matrix} \right\} \frac{4^s}{(m-1)^{2s}} \sum_{\substack{\text{odd } i \\ 1 \le i \le p-1}} \binom{p}{i} \binom{i/2}{s}.$$

Note that B_s has alternating signs, so it suffices to show that $|B_{s+1}| < |B_s|$. Indeed,

$$\frac{\left\{ \begin{smallmatrix} s+1 \\ m \end{smallmatrix} \right\} \frac{4^{s+1}}{(m-1)^{2(s+1)}}}{\left\{ \begin{smallmatrix} s \\ m \end{smallmatrix} \right\} \frac{4^s}{(m-1)^{2s}}} = \frac{4}{(m-1)^2} \cdot \frac{\left\{ \begin{smallmatrix} s+1 \\ m \end{smallmatrix} \right\}}{\left\{ \begin{smallmatrix} s \\ m \end{smallmatrix} \right\}} \le \frac{8m}{(m-1)^2} < 1$$

when m is large enough, and

$$\left| \frac{\binom{i/2}{s+1}}{\binom{i/2}{s}} \right| = \left| \frac{s - \frac{i}{2}}{s+1} \right| < 1.$$

The proof is now complete. $\square$

It also follows from the proof that for the same large m, the gap from $r_i(k)$ has the same sign for all even p up to some p_0 depending on m. This implies that when f is an even polynomial, the gap contribution from $r_i(k)$ is non-zero.

6. ALGORITHM FOR EVEN p

We first recall the classic result on Count-Sketch [18].

THEOREM 5 (COUNT-SKETCH). *There is a randomized linear function $M : \mathbb{R}^n \to \mathbb{R}^S$ with $S = O(w \log(n/\delta))$ and a recovery algorithm A satisfying the following. For any $x \in \mathbb{R}^n$, with probability $\geq 1 - \delta$, A reads Mx and outputs $\tilde{x} \in \mathbb{R}^n$ such that $\|\tilde{x} - x\|_\infty^2 \leq \|x\|_2^2/w$.*

We also need a result on ℓ_2-sampling. We say x is an (c, δ)-approximator to y if $(1-c)y - \delta \leq x \leq (1+c)y + \delta$.

THEOREM 6 (PRECISION SAMPLING [2]). *Fix $0 < \epsilon < 1/3$. There is a randomized linear function $M : \mathbb{R}^n \to \mathbb{R}^S$, with $S = O(\epsilon^{-2} \log^3 n)$, and an "$\ell_p$-sampling algorithm A" satisfying the following. For any non-zero $x \in \mathbb{R}^n$, there is a distribution D_x on $[n]$ such that $D_x(i)$ is an $(\epsilon, 1/\operatorname{poly}(n))$-approximator to $|x_i|^2/\|x\|_2^2$. Then A generates a pair (i, v) such that i is drawn from D_x (using the randomness of the function M only), and v is a $(\epsilon, 0)$-approximator to $|x_i|^2$.*

The basic idea is to choose $u_1, \ldots, u_n$ with $u_i \sim \operatorname{Unif}(0,1)$ and hash $y_i = x_i/\sqrt{u_i}$ using a COUNT-SKETCH structure of size $\Theta(w \log n)$ (where $w = \Theta(\epsilon^{-1} \log n + \epsilon^{-2})$), and recover the heaviest y_i and thus x_i if y_i is the unique entry satisfying $y_i \geq C\|x\|_2^2/\epsilon$ for some absolute constant C, which happens with the desired probability $|x_i|^2/\|x\|_2^2 \pm 1/\operatorname{poly}(n)$. The estimate error of x_i follows from COUNT-SKETCH guarantee.

Now we turn to our algorithm. Let $A = (a_{ij})$ be an integer matrix and suppose that the rows of A are $a_1, a_2, \ldots$. There are $O(1)$ non-zero entries in each row and each column. Assume $p \geq 4$. We shall use the structure for ℓ_2 sampling on n rows while using a bigger underlying COUNT-SKETCH structure to hash all n^2 elements of a matrix.

For simplicity, we present our algorithm in Algorithm 1 with the assumption that $u_1, \ldots, u_n$ are i.i.d. $\operatorname{Unif}(0,1)$. The randomness can be reduced using the same technique in [2] which uses $O(\log n)$ seeds.

Algorithm 1 Algorithm for even p and sparse matrices

Assume that matrix A has at most $k = O(1)$ non-zero entries per row and per column.

1: $T \leftarrow \Theta(n^{1-2/p}/\epsilon^2)$
2: $R \leftarrow \Theta(\log n)$
3: $w \leftarrow O(\epsilon^{-1} \log n + \epsilon^{-2})$
4: $I \leftarrow \emptyset$ is a multiset
5: Choose i.i.d. $u_1, \ldots, u_n$ with $u_i \sim \operatorname{Unif}(0,1)$.
6: $D \leftarrow \operatorname{diag}\{1/\sqrt{u_1}, \ldots, 1/\sqrt{u_n}\}$
7: Maintain a sketch for estimating $\|A\|_F^2$ and obtain a $(1+\epsilon)$-approximation L as in [1]
8: In parallel, maintain T structures, each has R repetitions of the Precision Sampling structure for all n^2 entries of $B = DA$, $t = 1, \ldots, T$. The Precision Sampling structure uses a COUNT-SKETCH structure of size $O(w \log n)$.
9: Maintain a sketch for estimating $\|B\|_F^2$ and obtain an $(1+\epsilon)$-approximation L' as in [1]
10: **for** $t \leftarrow 1$ to T **do**
11:     **for** $r \leftarrow 1$ to R **do**
12:         Use the r-th repetition of the t-th structure to obtain estimates $\tilde{b}_{i'1}, \ldots, \tilde{b}_{i'n}$ for all i' and form rows $\tilde{b}_{i'} = (b_{i'1}, \ldots, b_{i'n})$.
13:         If there exists a unique i' such that $\|\tilde{b}'_i\|_2^2 \geq C'L/\epsilon$ for some appropriate absolute constant C', return i' and exit the inner loop
14:     **end for**
15:     Retain only entries of $b_{i'}$ that are at least $2L'/\sqrt{w}$.
16:     $\tilde{a}_{i'} \leftarrow \sqrt{u_{i'}}\tilde{b}_{i'}$
17:     $I \leftarrow I \cup \{i'\}$
18: **end for**
19: Return Y as defined in (17)

THEOREM 7. *For sparse matrices A with $O(1)$ non-zero entries per row and per column, Algorithm 1 returns a value that is a $(1+\epsilon)$-approximation to $\|A\|_p^p$ with constant probability, using space $O(n^{1-2/p} \operatorname{poly}(1/\epsilon, \log n))$.*

PROOF. It is the guarantee from the underlying COUNT-SKETCH structure of size $\Theta(w \log n)$ (where $w = O(\epsilon^{-1} \log n + \epsilon^{-2})$) that

$$\tilde{b}_{i'j} = b_{i'j} \pm \sqrt{\frac{\|B\|_F^2}{w}}$$

for all j. Since there are only $O(1)$ non-zero entries in $b_{i'}$, we can use a constant-factor larger size $w' = O(w)$ for COUNT-SKETCH such that

$$\tilde{b}_{i'j} = b_{i'j} \pm \sqrt{\frac{\|B\|_F^2}{w'}}$$

and thus

$$\|\tilde{b}_{i'}\|_2^2 = \|b_{i'}\|_2^2 \pm \sqrt{\frac{\|B\|_F^2}{w}}. \tag{12}$$

Since each row i is scaled by the same factor $1/\sqrt{u_i}$, we can apply the proof of Theorem 6 to the vector of row norms $\{\|a_i\|_2\}$ and $\{\|b_i\|_2\}$, which remains still valid because of the error guarantee (12) which is analogous to the 1-dimensional case. It follows that with probability $\geq 1 - 1/n$ (since there are $\Theta(\log n)$ repetitions in each of the T structures), an i' is returned from the inner for-loop such that

$$\Pr\{i' = i\} = (1 \pm \epsilon)\frac{\|a_i\|_2^2}{\|A\|_F^2} \pm \frac{1}{\operatorname{poly}(n)}. \tag{13}$$

Next we analyse estimation error. It holds with high probability that $\|B\|_F^2 \leq w\|A\|_F^2$. Since a_i (and thus b_i) has $O(1)$-elements, the heaviest element $a_{i'j'}$ (resp. $b_{i'j'}$) has weight at least a constant fraction of $\|a_i\|_2$ (resp. $\|b_i\|_2$). It follows from the thresholding condition of the returned $\|b_i\|_2$ that we can use a constant big enough for $w' = O(w)$ to obtain

$$\tilde{a}_{i'j'} = \sqrt{u_{i'}} \cdot \tilde{b}_{i'j'} = (1 \pm \epsilon)a_{i'j'},$$

Suppose that the heaviest element is b_{ij}. Similarly, if $|a_{i\ell}| \geq \eta|a_{ij}|$ (where η is a small constant to be determined later), making $w' = \Omega(w/\eta)$, we can recover

$$\tilde{a}_{i\ell} = \sqrt{u_i} \cdot \tilde{b}_{i\ell} = a_{i\ell} \pm \epsilon\eta a_{ij} = (1 \pm \epsilon)a_{i\ell}.$$

Note that there are $O(1)$ non-zero entries $a_{i\ell}$ such that $|a_{i\ell}| \leq \eta|a_{ij}|$ and each of them has at most $\Theta(\epsilon\eta a_{ij})$ additive error by the threshold in Step 15, the approximation $\tilde{a}_i$ to a_i therefore satisfies

$$\|\tilde{a}_i - a_i\|_2^2 \leq \epsilon^2\|a_i\|_2^2 + O(1) \cdot \epsilon^2\eta^2\|a_i\|_2^2 \leq 2\epsilon^2\|a_i\|_2^2$$

by choosing an η small enough. It follows that $\|\tilde{a}_i\|_2$ is a $(1+\Theta(\epsilon))$-approximation to $\|a_i\|_2$, and $|\langle\tilde{a}_i, \tilde{a}_j\rangle| = |\langle a_i, a_j\rangle| \pm \Theta(\epsilon)\|a_i\|_2\|a_j\|_2$.

Next we show that our estimate is desirable. First, we observe that the additive $1/\operatorname{poly}(n)$ term in (13) can be dropped at the cost of increasing the total failure probability by $1/\operatorname{poly}(n)$. Hence we may assume in our analysis that

$$\Pr\{i' = i\} = (1 \pm \epsilon)\frac{\|a_i\|_2^2}{\|A\|_F^2}. \tag{14}$$

For notational simplicity let $q = p/2$ and $\ell_i = \|\tilde{a}_i\|_2^2$ if $i \in I$. Let $\tilde{a}_{i_1}, \ldots, \tilde{a}_{i_q}$ be q sampled rows. Define

$$\tilde{X}(i_1, \ldots, i_q) = \prod_{j=1}^{q} \langle \tilde{a}_{i_j}, \tilde{a}_{i_{j+1}} \rangle \cdot \frac{L^q}{\ell_{i_1}\ell_{i_2}\cdots\ell_{i_q}},$$

where it is understood that $a_{i_{q+1}} = a_{i_1}$. Let

$$X(i_1, \ldots, i_q) = \prod_{j=1}^{q} \langle a_{i_j}, a_{i_{j+1}} \rangle \cdot \frac{\|A\|_F^{2q}}{\|a_{i_1}\|_2^2\|a_{i_2}\|_2^2\cdots\|a_{i_q}\|_2^2}$$

Then

$$\begin{aligned}|X(i_1, \ldots, i_q) - \tilde{X}(i_1, \ldots, i_q)| &\lesssim \epsilon \prod_{j=1}^{q} \|a_{i_j}\|_2^2 \ \frac{\|A\|_F^{2q}}{\prod_{j=1}^{q}\|a_{i_j}\|_2^2} \\ &= \epsilon\|A\|_F^{2q}. \end{aligned} \tag{15}$$

Also let

$$p(i) = \Pr\{\text{row } i \text{ gets sampled}\},$$

then

$$\left| p(i_1)p(i_2)\cdots p(i_q) - \frac{\prod_{j=1}^{q}\|a_{i_j}\|_2^2}{\|A\|_F^2} \right| \lesssim \epsilon \frac{\prod_{j=1}^{q}\|a_{i_j}\|_2^2}{\|A\|_F^2}.$$

We claim that

$$\|A\|_p^p = \sum_{1 \le i_1, \ldots, i_q \le n} \prod_{j=1}^{q} \langle a_{i_j}, a_{i_{j+1}} \rangle.$$

When $q = p/2$ is odd,

$$\begin{aligned}&\|A\|_p^p = \|(A^TA)\cdots(A^TA)A^T\|_F^2 \\ &= \sum_{k,\ell}\sum_{i_1,\ldots,i_{q-1}} (A^T_{k,i_1}A_{i_1,i_2}A^T_{i_2,i_3}A_{i_3,i_4}\cdots A_{i_{q-2},i_{q-1}}A^T_{i_{q-1},\ell})^2 \\ &= \sum_{k,\ell}\sum_{\substack{i_1,\ldots,i_{q-1}\\ j_1,\ldots,j_{q-1}}} A_{i_1,k}A_{j_1,k}A_{i_1,i_2}A_{j_1,j_2}\cdots A_{\ell,i_{q-1}}A_{\ell,j_{q-1}} \\ &= \sum \langle a_{i_1}, a_{j_1}\rangle \cdot \prod_{\substack{\text{odd } t\\ 1\le t\le q-2}} \langle a_{i_t}, a_{i_{t+2}}\rangle\langle a_{j_t}, a_{j_{t+2}}\rangle \cdot \langle a_{i_{q-2}}, a_\ell\rangle\langle a_{j_{q-2}}, a_\ell\rangle,\end{aligned}$$

which is a 'cyclic' form of inner products and the rightmost sum is taken over all appearing variables (i_t, j_t and ℓ) in the expression. A similar argument works when q is even. It follows that

$$\begin{aligned}&\left|\mathbb{E}\,\tilde{X}(i_1, \ldots, i_q) - \|A\|_p^p\right| \le \\ &\qquad \sum_{i_1,\ldots,i_q} \left| p(i_1)\cdots p(i_q)\tilde{X}(i_1, \ldots, i_q) - \prod_{j=1}^{q}\langle a_{i_j}, a_{i_{j+1}}\rangle \right|,\end{aligned}$$

where $i_1, \ldots, i_q$ are sampled according to the density function $p(i)$. It is clear that each a_i has only $O(1)$ rows with overlapping support, since each row and each column has only $O(1)$ non-zero entries. The observation is that the same result holds for $\tilde{a}_i$. This is due to our threshold in Step 15: for an entry to be retained, it must be larger than $\|B\|_F/\sqrt{w}$ (the uniform additive error from COUNT-SKETCH), which is impossible for zero entries. Therefore, each row i appears in $O(1)$ contributing summands. Each contributing summand is bounded by

$$\Theta(1)\cdot\epsilon\prod_{j=1}^{q}\|a_{i_j}\|_2^2 \le \Theta(1)\cdot\epsilon\max\{\|a_{i_1}\|_2^{2q}, \ldots, \|a_{i_q}\|_2^{2q}\}.$$

Then

$$\left|\mathbb{E}\,\tilde{X}(i_1, \ldots, i_q) - \|A\|_p^p\right| \lesssim \epsilon\sum_i \|a_i\|_2^{2q} \le \epsilon\|A\|_{2q}^{2q}. \tag{16}$$

as desired, where the last inequality follows from the fact of Schatten r-norms ($r \ge 1$) that $\|M\|_r^r \ge \sum_{i=1}^n |M_{ii}|^r$ and choosing $M = A^TA$ and $r = q$. Our estimator is

$$\tilde{Y} = \frac{1}{|I|^q}\sum_{i_1,\ldots,i_q\in I}\tilde{X}(i_1, \ldots, i_q), \tag{17}$$

then $|\mathbb{E}\,Y - \|A\|_p^p| \lesssim \varepsilon\|A\|_p^p$. Next we bound the variance. We will show below that the following bound holds when $\{i_1, \ldots, i_q\} \cap \{j_1, \ldots, j_q\} = r$,

$$\mathbb{E}(\tilde{X}(i_1, i_2, \ldots, i_q)\tilde{X}(j_1, j_2, \ldots, j_q)) \lesssim \|A\|_F^{2r}\|A\|_p^{2p-2r}. \tag{18}$$

Given this,

$$\begin{aligned}&\mathbb{E}\,Y^2 - (EY)^2 \\ &\le \frac{1}{|I|^{2q}}\sum_{r\ge1}\sum_{\substack{i_1,\ldots,i_q,j_1,\ldots,j_q\in I\\ \{i_1,\ldots,i_q\}\cap\{j_1,\ldots,j_q\}=r}} \mathbb{E}(\tilde{X}(i_1, \ldots, i_q)\tilde{X}(j_1, \ldots, j_q)) \\ &\lesssim \frac{1}{|I|^{2q}}\sum_{r\ge1}\sum_{\substack{i_1,\ldots,i_q,j_1,\ldots,j_q\in I\\ \{i_1,\ldots,i_q\}\cap\{j_1,\ldots,j_q\}=r}} n^{r(1-\frac{2}{p})}\|A\|_p^{2p} \\ &\le \frac{1}{|I|^{2q}}\sum_{r\ge1}|I|^{2q-r}n^{r(1-\frac{2}{p})}\|A\|_p^{2p} \lesssim \epsilon^2\|A\|_p^{2p},\end{aligned}$$

where the constant in the last $\lesssim$ can be made arbitrarily small if we choose $|I| = T = Cn^{1-2/p}/\epsilon^2$ for C large enough. It follows from Chebyshev's inequality that $X = (1+\Theta(\epsilon))\|A\|_p^p$ with constant probability.

Next we prove (18). It is similar to (15) that

$$\begin{aligned}&\Big|\tilde{X}(i_1, \ldots, i_q)\tilde{X}(j_1, \ldots, j_q)) \\ &\qquad - X(i_1, \ldots, i_q)X(j_1, \ldots, j_q))\Big| \lesssim \varepsilon\|A\|_F^{4q},\end{aligned} \tag{19}$$

then

$$\begin{aligned}&\Big|\mathbb{E}(\tilde{X}(i_1, i_2, \ldots, i_q)\tilde{X}(j_1, j_2, \ldots, j_q)) \\ &\qquad - \mathbb{E}(X(i_1, i_2, \ldots, i_q)X(j_1, j_2, \ldots, j_q))\Big| \\ &\lesssim \sum_{\substack{i_1,\ldots,i_q,j_1,\ldots,j_q\\ \{i_1,\ldots,i_q\}\cap\{j_1,\ldots,j_q\}=\{s_1,\ldots,s_r\}}} \frac{p(i_1)\cdots p(i_q)p(j_1)\cdots p(j_q)}{p(s_1)\cdots p(s_r)}\cdot\varepsilon\|A\|_F^{4q} \\ &\lesssim \sum_{\substack{i_1,\ldots,i_q,j_1,\ldots,j_q\\ \{i_1,\ldots,i_q\}\cap\{j_1,\ldots,j_q\}=\{s_1,\ldots,s_r\}}} \epsilon\|A\|_F^{2r}\left(\max_i\|a_i\|_2^2\right)^{2q-r} \qquad (20) \\ &\lesssim \epsilon\|A\|_F^{2r}\sum_i\|a_i\|_2^{4q-2r} \\ &\le \epsilon\|A\|_F^{2r}\|A\|_p^{2p-2r},\end{aligned}$$

where (20) follows from column sparsity and the last inequality follows from the property of Schatten norms as in (16). Now it suffices to show that

$$\mathbb{E}(X(i_1, i_2, \dots, i_q)X(j_1, j_2, \dots, j_q)) \lesssim \|A\|_F^{2r}\|A\|_p^{2p-2r}. \quad (21)$$

We write

$$\mathbb{E}(X(i_1, i_2, \dots, i_q)X(j_1, j_2, \dots, j_q)) = \Sigma + \Delta,$$

where

$$\Sigma = \sum_{\substack{\{i_t\},\{j_t\} \\ |\{i_t\}\cap\{j_t\}|=r}} \frac{\prod_{t=1}^q \|a_{i_t}\|_2^2 \|a_{j_t}\|_2^2}{\prod_{t=1}^r \|a_{s_t}\|_2^2} \|A\|_F^{2r-4q} \cdot X(i_1, i_2, \dots, i_q)X(j_1, j_2, \dots, j_q))$$

$$= \sum_{\substack{\{i_t\},\{j_t\} \\ |\{i_t\}\cap\{j_t\}|=r}} \frac{\|A\|_F^{2r}}{\prod_{t=1}^r \|a_{s_t}\|_2^2} \prod_{t=1}^q \langle a_{i_t}, a_{i_{t+1}} \rangle \langle a_{j_t}, a_{j_{t+1}} \rangle$$

and

$$\Delta = \sum_{\substack{\{i_t\},\{j_t\} \\ |\{i_t\}\cap\{j_t\}|=r}} \left(\frac{\prod_{t=1}^q p(i_t)p(j_t)}{\prod_{t=1}^r p(s_t)} - \frac{\prod_{t=1}^q \|a_{i_t}\|_2^2 \|a_{j_t}\|_2^2}{\prod_{t=1}^r \|a_{s_t}\|_2^2} \|A\|_F^{2r-4q} \right) \cdot X(i_1, i_2, \dots, i_q)X(j_1, j_2, \dots, j_q)),$$

where the sum is over all choices of $i_1, \dots, i_q, j_1, \dots, j_q \in I$ such that $|\{i_1, \dots, i_q\} \cap \{j_1, \dots, j_q\}| = r$. It follows from a similar argument as before that

$$|\Sigma| \le \sum_{\substack{\{i_t\},\{j_t\} \\ |\{i_t\}\cap\{j_t\}|=r}} \|A\|_F^{2r} \left(\max_i \|a_i\|_2^2 \right)^{2q-r} \lesssim \|A\|_F^{2r} \|A\|_p^{2p-2r}$$

and

$$|\Delta| \le \sum_{\substack{\{i_t\},\{j_t\} \\ |\{i_t\}\cap\{j_t\}|=r}} \epsilon \|A\|_F^{2r} \left(\max_i \|a_i\|_2^2 \right)^{2q-r} \lesssim \epsilon \|A\|_F^{2r} \|A\|_p^{2p-2r},$$

establishing (21). We used the assumption of column sparsity for both bounds, c.f. (20). □

7. GENERAL FUNCTIONS AND APPLICATIONS

The following is a direct corollary of Theorem 4.

THEOREM 8. *Let f be a diagonally block-additive function. Suppose that $f(x) \simeq x^p$ for x near 0 or x near infinity, where $p > 0$ is not an even integer. For any even integer t, there exists a constant $c = c(t) > 0$ such that any streaming algorithm that approximates $f(X)$ within a factor $1 \pm c$ with constant error probability must use $\Omega_t(N^{1-1/t})$ bits of space.*

PROOF. Suppose that $f(x) \sim \alpha x^p$ for x near 0, that is, for any $\eta > 0$, there exists $\delta = \delta(\eta) > 0$ such that $\alpha(1-\eta)f(x) \le x^p \le \alpha(1+\eta)f(x)$ for all $x \in [0, \delta)$.

Let c_0 be the approximation ratio parameter in Theorem 4 for Schatten p-norm. Let ϵ be sufficiently small (it could depend on t and thus m) such that the singular values of $\varepsilon\mathcal{M}$ are at most $\delta(c_0/3)$, where $\mathcal{M}$ is the hard instance matrix used in Theorem 4. Then $\alpha(1 - c_0/3)f(\varepsilon\mathcal{M}) \le \|\varepsilon\mathcal{M}\|_p^p \le \alpha(1 + c_0/3)f(\varepsilon\mathcal{M})$. Therefore, any algorithm that approximates $f(\varepsilon\mathcal{M})$ within a factor of $(1 \pm c_0/3)$ can produce a $(1 \pm c_0)$-approximation of $\|\epsilon\mathcal{M}\|_p^p$. The lower bound follows from Theorem 4.

When $f(x) \simeq x^p$ for x near infinity, a similar argument works for $\lambda\mathcal{M}$ where λ is sufficiently large. □

The following is a corollary of Lemma 1.

THEOREM 9. *Suppose that f admits a Taylor expansion near 0 that has infinitely many even-order terms of non-zero coefficient. Then for any arbitrary large m, there exists $c = c(m)$ such that any data stream algorithm which outputs, with constant error probability, a $(1 + c)$-approximation to $\|X\|_p^p$ requires $\Omega(N^{1-1/m})$ bits of space.*

PROOF. If the expansion has a odd-order term with non-zero coefficient, apply Theorem 8 with the lowest non-zero odd-order term. Hence we may assume that all terms are of even order. For any given m, there exists $p > 2m$ such that the x^p term in the Taylor expansion of f has a non-zero coefficient a_p. Let p be the lowest order of such a term, and write

$$f(x) = \sum_{i=0}^{p-1} a_i x^{p-1} + a_p x^p + O(x^{p+1}).$$

Let $\epsilon > 0$ be a small constant to be determined later and consider the matrix $\epsilon\mathcal{M}$, where $\mathcal{M}$ is our hard instance matrix used in Lemma 1. Lemma 1 guarantees a gap of $f(\epsilon\mathcal{M})$, which is then $a_p\epsilon^p G + R(\epsilon)$, where G is the gap for x^p on unscaled hard instance $\mathcal{M}$ and $|R(\epsilon)| \le K\epsilon^{p+1}$ for some constant K depending only on $f(x)$, m and p. Choosing $\epsilon < a_p G/K$ guarantees that the gap $a_p\epsilon^p G + R(\epsilon) \neq 0$. □

Now we are ready to prove the lower bound for some eigenvalue shrinkers and M-estimators. The following are the three optimal eigenvalue shrinkers from [27]:

- $\eta_1(x) = x^{-1}\sqrt{(x^2 - \alpha - 1)^2 - 4\alpha}$ for $x \ge 1 + \sqrt{\alpha}$ and $\eta_1(x) = 0$ for $x < 1 + \sqrt{\alpha}$,
- $\eta_2(x) = \frac{1}{\sqrt{2}}\sqrt{x^2 - \alpha - 1 + \sqrt{(x^2-\alpha-1)^2 - 4\alpha}}$ for $x \ge 1 + \sqrt{\alpha}$ and $\eta_2(x) = 0$ for $x < 1 + \sqrt{\alpha}$,
- $\eta_3(x) = (x\eta_2^2(x))^{-1} \max\left\{\eta_2^4(x) - \alpha - \alpha x \eta_2(x), 0\right\}$.

where we assume that $0 \cdot \infty = 0$. Since $\eta_i(x) \simeq x$ when x is large, the lower bound follows from Theorem 8.

Some commonly used influence functions $\rho(x)$ can be found in [56], summarized in Table 2. Several of them are asymptotically linear when x is large and Theorem 8 applies. Some are covered by Theorem 9. For the last function, notice that it is a constant on $[c, +\infty)$, we can rescale our hard instance matrix $\mathcal{M}$ such that the larger root $r_1(k)$ falls in $[c, +\infty)$ and the smaller root $r_2(k)$ in $[0, c]$. The larger root $r_1(k)$ therefore has no contribution to the gap. The contribution from the smaller root $r_2(k)$ is nonzero by the remark following Lemma 3.

Finally we consider functions of the form

$$F_k(X) = \sum_{i=1}^k f(\sigma_i(X))$$

and prove (a slightly rephrased) Theorem 2 in the introduction.

THEOREM 10. *Let $\alpha \in (0, 1/2)$. Suppose that f is strictly increasing. There exists N_0 and c_0 such that for all $N \ge N_0$, $k \le \alpha N$ and $c \in (0, c_0)$, any data stream algorithm which outputs, with constant error probability, a $(1 + c)$-approximation to $F_k(X)$ of $X \in \mathbb{R}^{N\times N}$ requires $\Omega_\alpha(N^{1+\Theta(1/\ln\alpha)})$ bits of space.*

Function $\rho(x)$	Apply	Function $\rho(x)$	Apply
$2(\sqrt{1+x^2/2}-1)$	Theorem 8	$\frac{x^2/2}{1+x^2}$	Theorem 8
$c^2(\frac{x}{c}-\ln(1+\frac{x}{c}))$	Theorem 8	$\frac{c^2}{2}(1-\exp(-x^2/c^2))$	Theorem 9
$\begin{cases} x^2/2, & x \le k; \\ k(x-k/2), & x > k \end{cases}$	Theorem 8	$\begin{cases} \frac{c^2}{6}(1-(1-x^2/c^2)^3), & x \le c; \\ c^2/6, & x > c \end{cases}$	Remark after Lemma 3
$\frac{c^2}{2}\ln(1+\frac{x^2}{c^2})$	Theorem 9		

Table 2: Application of Theorem 8 and Theorem 9 to some M-estimators from [56].

PROOF. Similarly to Theorem 3 we reduce the problem from the BHH_n^0 problem. Let $m = t$ be the largest integer such that $1/(t2^t) \geq \alpha$. Then $m = t = \Theta(\ln(1/\alpha))$. We analyse the largest k singular values of $\mathcal{M}$ as defined in (2). Recall that $q_1, \ldots, q_{n/m}$ are divided into $N/(2m)$ groups. Let $X_1, \ldots, X_{N/(2m)}$ be the larger q_i's in each group, then $X_1, \ldots, X_{N/(2m)}$ are i.i.d. random variables. In the even case, they are defined on $\{m/2, m/2+2, \ldots, m\}$ subject to the distribution

$$\Pr\left\{X_1 = \frac{m}{2} + j\right\} = \begin{cases} p_m(\frac{m}{2}), & j = 0; \\ 2p_m(\frac{m}{2}+j), & j > 0, \end{cases} \quad j = 0, 2, \ldots, \frac{m}{2}.$$

In the odd case, they are defined on $\{m/2+1, m/2+3, \ldots, m-1\}$ with probability density function

$$\Pr\left\{X_1 = \frac{m}{2} + j\right\} = 2p_m\left(\frac{m}{2} + j\right), \quad j = 1, 3, \ldots, \frac{m}{2} - 1.$$

With probability $1/2^{m-2}$, $X_i = m$ in the even case and with probability $m/2^{m-2}$, $X_i = m/2 - 1$ in the odd case. It immediately follows from a Chernoff bound that with high probability, it holds that $X_i = m$ (resp. $X_i = m-1$) for at least $(N/2m)\cdot(1/2^{m-2})(1-\delta) = (1-\delta)N/(m2^{m-1})$ different i's in the even case (resp. odd case). Since $r_1(m-1) < r_1(m)$ and f is strictly increasing, the value $F_k(X)$, when $k \leq \alpha N \leq (1-\delta)N/(m2^{m-1})$, with high probability, exhibits a gap of size at least $c \cdot k$ for some constant c between the even and the odd cases. Since $F_k(\mathcal{M}) = \Theta(k)$ with high probability, the lower bound for Ky-Fan k-norm follows from the lower bound for BHH_n^0. $\square$

The lower bound for Ky-Fan k-norms follows immediately. For $k \leq \alpha N$ it follows from the preceding theorem with $f(x) = x$; for $k > \alpha N$, the lower bound follows from our lower bound for the Schatten 1-norm by embedding the hard instance of dimension $\alpha N \times \alpha N$ into the $N \times N$ matrix X, padded with zeros.

8. REFERENCES

[1] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *J. Comput. Syst. Sci.*, 58(1):137–147, 1999.

[2] A. Andoni, R. Krauthgamer, and K. Onak. Streaming algorithms via precision sampling. In *Proceedings of the IEEE 52nd FOCS*, pages 363–372, 2011.

[3] A. Andoni, R. Krauthgamer, and I. P. Razenshteyn. Sketching and embedding are equivalent for norms. In *Proceedings of the 47th ACM STOC*, pages 479–488, 2015.

[4] A. Andoni and H. L. Nguyen. Eigenvalues of a matrix in the streaming model. In *Proceedings of the 24th ACM-SIAM SODA*, pages 1729–1737, 2013.

[5] G. A. Andrews, R. Askey, and R. Roy. *Special Functions*. Cambridge University Press, 1999.

[6] Z. Bar-Yossef, T. S. Jayram, and I. Kerenidis. Exponential separation of quantum and classical one-way communication complexity. In *Proceedings of the 36th ACM STOC*, pages 128–137, 2004.

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

[8] Ş. B. Bozkurt and D. Bozkurt. On the sum of powers of normalized laplacian eigenvalues of graphs. *MATCH Communications in Mathematical and in Computer Chemistry*, 68(3), 2012.

[9] V. Braverman and S. R. Chestnut. Streaming sums in sublinear space. arXiv:1408.5096, 2014.

[10] V. Braverman and S. R. Chestnut. Universal sketches for the frequency negative moments and other decreasing streaming sums. In *Proceedings of APPROX/RANDOM*, pages 591–605, 2015.

[11] 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 PODS*, 2016.

[12] V. Braverman, J. Katzman, C. Seidell, and G. Vorsanger. An optimal algorithm for large frequency moments using $O(n^{1-2/k})$ bits. In *Proceedings of APPROX/RANDOM*, pages 531–544, 2014.

[13] V. Braverman and R. Ostrovsky. Zero-one frequency laws. In *Proceedings of the 42nd ACM STOC*, pages 281–290, 2010.

[14] V. Braverman, R. Ostrovsky, and A. Roytman. Zero-one laws for sliding windows and universal sketches. In *Proceedings of APPROX/RANDOM*, pages 573–590, 2015.

[15] M. Bury and C. Schwiegelshohn. Sublinear estimation of weighted matchings in dynamic data streams. In *Proceedings of 23rd ESA*, pages 263–274, 2015.

[16] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Commun. ACM*, 55(6):111–119, 2012.

[17] A. Chakrabarti, S. Khot, and X. Sun. Near-optimal lower bounds on the multi-party communication complexity of set disjointness. In *Proceedings of the 18th IEEE CCC*, pages 107–117, 2003.

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

[19] K. L. Clarkson and D. P. Woodruff. Numerical linear

algebra in the streaming model. In *Proceedings of the 41st ACM STOC*, pages 205–214, 2009.

[20] G. Cormode and S. Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. *J. Algorithms*, 55(1):58–75, 2005.

[21] A. Deshpande, M. Tulsiani, and N. K. Vishnoi. Algorithms and hardness for subspace approximation. In *Proceedings of SODA*, pages 482–496, 2011.

[22] X. V. Doan and S. Vavasis. Finding the largest low-rank clusters with Ky Fan 2-k-norm and ℓ_1-norm. arXiv:1403.5901, 2014.

[23] S. Ganguly. A lower bound for estimating high moments of a data stream. arXiv:1201.0253, 2012.

[24] S. Ganguly. Taylor polynomial estimator for estimating frequency moments. In *Proceedings of 42th ICALP*, pages 542–553, 2015.

[25] S. Ganguly and G. Cormode. On estimating frequency moments of data streams. In *Proceedings of RANDOM/APPROX*, pages 479–493, 2007.

[26] D. Gavinsky, J. Kempe, I. Kerenidis, R. Raz, and R. de Wolf. Exponential separations for one-way quantum communication complexity, with applications to cryptography. In *Proceedings of the 39th ACM STOC*, pages 516–525, 2007.

[27] M. Gavish and D. L. Donoho. Optimal shrinkage of singular values. Technical Report 2014-08, Department of Statistics, Stanford University, May 2014.

[28] A. Gronemeier. Asymptotically optimal lower bounds on the NIH-multi-party information complexity of the AND-function and disjointness. In *Proceedings of the 26th STACS*, pages 505–516, 2009.

[29] M. Hardt, K. Ligett, and F. McSherry. A simple and practical algorithm for differentially private data release. In *Advances in Neural Information Processing Systems 25*, pages 2348–2356. 2012.

[30] P. Indyk. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *J. ACM*, 53(3):307–323, 2006.

[31] P. Indyk and D. P. Woodruff. Tight lower bounds for the distinct elements problem. In *Proceedings of the 44th IEEE FOCS*, pages 283–288, 2003.

[32] P. Indyk and D. P. Woodruff. Optimal approximations of the frequency moments of data streams. In *Proceedings of the 37th ACM STOC*, pages 202–208, 2005.

[33] T. S. Jayram. Hellinger strikes back: A note on the multi-party information complexity of AND. In *Proceedings of RANDOM/APPROX*, pages 562–573, 2009.

[34] H. Jowhari, M. Saglam, and G. Tardos. Tight bounds for L_p samplers, finding duplicates in streams, and related problems. In *Proceedings of the 30th ACM SIGMOD PODS*, pages 49–58, 2011.

[35] B. Kalyanasundaram and G. Schnitger. The probabilistic communication complexity of set intersection. *SIAM J. Discrete Math.*, 5(4):545–557, 1992.

[36] D. M. Kane, J. Nelson, E. Porat, and D. P. Woodruff. Fast moment estimation in data streams in optimal space. In *STOC*, pages 745–754, 2011.

[37] D. M. Kane, J. Nelson, and D. P. Woodruff. On the exact space complexity of sketching and streaming small norms. In *Proceedings of the 21st ACM-SIAM SODA*, pages 1161–1178, 2010.

[38] W. Kong and G. Valiant. Spectrum estimation from samples. arXiv:1602.00061, 2016.

[39] R. Krauthgamer and O. Sasson. Property testing of data dimensionality. In *Proceedings of the Fourteenth ACM-SIAM SODA*, pages 18–27, 2003.

[40] C. Li and G. Miklau. Optimal error of query sets under the differentially-private matrix mechanism. In *Proceedings of ICDT*, pages 272–283, 2013.

[41] Y. Li, H. L. Nguyen, and D. P. Woodruff. On sketching matrix norms and the top singular vector. In *Proceedings of the 25th ACM-SIAM SODA*, pages 1562–1581, 2014.

[42] Y. Li, H. L. Nguyen, and D. P. Woodruff. Turnstile streaming algorithms might as well be linear sketches. In *Proceedings of the 46th ACM STOC*, pages 174–183, 2014.

[43] Y. Li, Z. Wang, and D. P. Woodruff. Improved testing of low rank matrices. In *Proceedings of the 20th ACM SIGKDD*, pages 691–700, 2014.

[44] Y. Li and D. P. Woodruff. A tight lower bound for high frequency moment estimation with small error. In *Proceedings of RANDOM/APPROX*, pages 623–638, 2013.

[45] M. Monemizadeh and D. P. Woodruff. 1-pass relative-error L_p-sampling with applications. In *SODA*, pages 1143–1160, 2010.

[46] J. Nelson and D. P. Woodruff. Fast Manhattan sketches in data streams. In *PODS*, pages 99–110, 2010.

[47] A. A. Razborov. On the distributional complexity of disjointness. *Theor. Comput. Sci.*, 106(2):385–390, 1992.

[48] I. Razenshteyn. Personal communication, 2015.

[49] R. Sedgewick and P. Flajolet. *An Introduction to the Analysis of Algorithms*. Addison-Wesley, 1996.

[50] D. Stevanovic, A. Ilic, C. Onisor, and M. V. Diudea. LEL–a newly designed molecular descriptor. *Acta Chim. Slov*, 56:410–417, 2009.

[51] E. Verbin and W. Yu. The streaming complexity of cycle counting, sorting by reversals, and other problems. In *Proceedings of the 22nd ACM-SIAM SODA*, pages 11–25, 2011.

[52] D. P. Woodruff. Optimal space lower bounds for all frequency moments. In *SODA*, pages 167–175, 2004.

[53] D. P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1-2):1–157, 2014.

[54] D. P. Woodruff and Q. Zhang. Tight bounds for distributed functional monitoring. In *Proceedings of the 44th ACM STOC*, pages 941–960, 2012.

[55] D. Xia. Optimal schatten-q and Ky-Fan-k norm rate of low rank matrix estimation, 2014.

[56] Z. Zhang. Parameter estimation techniques: a tutorial with application to conic fitting. *Image and Vision Computing*, 15(1):59–76, 1997.

[57] B. Zhou. On sum of powers of the Laplacian eigenvalues of graphs. *Linear Algebra and its Applications*, 429(8–9):2239–2246, 2008.