SIAM J. Comput.
Vol. 48, No. 6, pp. 1643–1697

ON APPROXIMATING MATRIX NORMS IN DATA STREAMS[*]

YI LI[†], HUY L. NGUYỄN[‡], AND DAVID P. WOODRUFF[§]

Abstract. This paper presents a systematic study of the space complexity of estimating the Schatten p-norms of an $n \times n$ matrix in the turnstile streaming model. Both kinds of space complexities, bit complexity and sketching dimension, are considered. Furthermore, two sketching models, general linear sketching and bilinear sketching, are considered. When p is not an even integer, we show that any one-pass algorithm with constant success probability requires near-linear space in terms of bits. This lower bound holds even for sparse matrices, i.e., matrices with $O(1)$ nonzero entries per row and per column. However, when p is an even integer, we give for sparse matrices an upper bound which, up to logarithmic factors, is the same as estimating the pth moment of an n-dimensional vector. These results considerably strengthen lower bounds in previous work for arbitrary (not necessarily sparse) matrices. Similar near-linear lower bounds are obtained for Ky Fan norms, SVD entropy, eigenvalue shrinkers, and M-estimators, many of which could have been solvable in logarithmic space prior to this work. The results for general linear sketches give separations in the sketching complexity of Schatten p-norms with the corresponding vector p-norms, and rule out a table-lookup nearest-neighbor search for $p = 1$, making progress on a question of Andoni. The results for bilinear sketches are tight for the rank problem and nearly tight for $p \geq 2$; the latter is the first general subquadratic upper bound for sketching the Schatten norms.

Key words. Schatten norm, matrix norm, streaming algorithm, approximation algorithm, sketching algorithm, numerical linear algebra

AMS subject classifications. 68W25, 68W20, 15A60

DOI. 10.1137/17M1152255

1. Introduction.

In the turnstile data stream model, there is an underlying n-dimensional vector x which is initialized to 0^n and then undergoes a long sequence of additive positive and negative updates of the form $x_i \leftarrow x_i + \Delta$ to its coordinates x_i. 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 f of x with constant probability. A major interest is the space complexity of the summary that the algorithm maintains, for which the following two models have been broadly considered.

Data stream model. In this model it is assumed that $x \in \mathbb{Z}^n$ and that each coordinate update $\Delta \in \{-m, -m+1, \ldots, m\}$. We make the standard simplifying assumption that n, m, and the length of the stream are polynomially related. The goal is to minimize the space usage of the algorithm in bits.

[*] Received by the editors October 16, 2017; accepted for publication (in revised form) August 19, 2019; published electronically November 14, 2019. Parts of the paper appeared in the Proceedings of *SODA*, 2014, pp. 1562–1581; the Proceedings of *APPROX/RANDOM*, 2016, pp. 39:1–39:11; and the Proceedings of *STOC*, 2016, pp. 726–739. Part of this work was done while the first and third authors were visiting the Simons Institute for the Theory of Computing.

https://doi.org/10.1137/17M1152255

Funding: The first author was partially supported by Singapore Ministry of Education (AcRF) Tier 2 grant MOE2018-T2-1-013. The second author was partially supported by NSF CAREER award 1750716. The third author was partially supported by National Science Foundation grant CCF-1815840.

[†] Division of Mathematics, Nanyang Technological University, Singapore, 637371 (yili@ntu.edu.sg).

[‡] Khoury College of Computer Sciences, Northeastern University, Boston, MA 02115 (hu.nguyen@northeastern.edu).

[§] Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 (dwoodruf@cs.cmu.edu).

Sketching model. In this model it is often assumed that $x \in \mathbb{R}^n$ and $\Delta \in \mathbb{R}$. The algorithm chooses a linear sketch $L : \mathbb{R}^n \to \mathbb{R}^k$ (which is usually random) and maintains Lx throughout the process of the data stream, so that for any fixed vector x, one can approximate $f(x)$ from Lx. Given an update of the form $x_i \leftarrow x_i + \Delta$, resulting in a new vector $x' = x + \Delta \cdot e_i$, we can add $L(\Delta \cdot e_i)$ to Lx to obtain Lx' (here e_i is the ith standard unit vector). The goal is to minimize the dimension k.

We will call the model *insertion-only* if each coordinate can be updated at most once.[1] In general, the two models above are not comparable. In the data stream model, if one wants to output a vector $x \in \{0, 1, \ldots, M-1, M\}^n$, one needs $n \log M$ bits of space. On the other hand, if u is the vector $(1, (M+1), (M+1)^2, (M+1)^3, \ldots, (M+1)^n)$, one can recover x from $\langle u, x \rangle$, so the sketching dimension k is only equal to 1. The sketching complexity thus gives a meaningful measure of complexity in the real RAM model. Conversely, lower bounds in the sketching model do not translate into lower bounds in the data stream model. This statement holds even given the work of [55], which characterizes turnstile streaming algorithms as linear sketches. The problem is that lower bounds in the sketching model involve continuous distributions, and after discretizing the distributions it is no longer clear whether the lower bounds hold.

A well-studied problem for both models in the literature is approximating the frequency moments $F_p(x)$, which is equivalent to estimating the ℓ_p-norms[2] $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$, for $p \in [0, \infty]$, dating back to work of Alon, Matias, and Szegedy [2]. For $p \le 2$ it is possible to obtain any constant factor approximation using $\Theta(1)$ sketch dimension or $\tilde{\Theta}(1)$ bits of space [42, 49], while for $p > 2$ the bound is $\tilde{\Theta}(n^{1-2/p})$ sketch dimension or $\tilde{\Theta}(n^{1-2/p})$ bits [18, 12, 44, 5, 33, 56, 15, 34], where $\tilde{f} = f \cdot \mathrm{poly}(\log(f))$. In addition to being of theoretical interest, the problems have several applications. The value $\|x\|_0$, by continuity, is equal to the support size of x, also known as the number of distinct elements [31, 32]. The norm $\|x\|_1$ is the Manhattan norm, which is a robust measure of distance and is proportional to the variation distance between distributions [40, 42, 48]. The Euclidean distance $\|x\|_2$ is important in linear algebra problems [66] and corresponds to the self-join size in databases [2]. Often one wishes to find or approximate the largest coordinates of x, known as the heavy hitters [19, 24], and $\|x\|_\infty$ is defined, by continuity, to equal $\max_i |x_i|$.

In this paper we primarily study the analogous problem of estimating *matrix norms.* In the turnstile streaming model, an underlying $n \times n$ matrix A undergoes a sequence of additive updates to its entries. Each update has the form (i, j, Δ) and indicates that $A_{i,j} \leftarrow A_{i,j} + \Delta$. The analogy of the ℓ_p-norm of vectors is the Schatten p-norm of matrices. Suppose that A is an $n \times n$ rank-r matrix. The Schatten p-norm of A is defined to be[2] $\|A\|_p = (\sum_{i=1}^r \sigma_i^p)^{1/p}$, where $\sigma_1 \ge \cdots \ge \sigma_r$ are the nonzero singular values of A. When $p = 0$, $\|A\|_0$ is defined to be the rank of A. The cases $p = 1, 2$, and ∞ correspond to the trace norm, the Frobenius norm, and the operator norm, respectively. These problems have found applications in several areas; we refer the reader to [22] for graph applications for $p = 0$, to the work on differential privacy [41, 54] and nonconvex optimization [17, 28] for $p = 1$, and to the survey on numerical linear algebra for $p \in \{2, \infty\}$ [57]. Fractional Schatten p-norms of Laplacians were studied by Zhou [74] and Bozkurt and Bozkurt [13]. We refer the reader to [68] for

[1] In some works, insertion-only refers to multiple updates, provided they are all positive. As we will prove lower bounds when each coordinate is updated at most once and by a positive amount, our lower bounds are only stronger by considering this notion of insertion-only.

[2] Technically it is not a norm when $p < 1$, but it is still a well-defined quantity. In this paper we refer to it using the name of "norm" for simplicity.

Table 1.1

Our results for approximating the Schatten p-norm up to a constant factor (except for the $p = \infty$ case) with constant probability in both the bilinear sketching and general sketching models. For the bilinear case, we look at the minimal $r \cdot s$ value, while for general linear sketches we look at the minimal value of k. For $p = \infty$, parameter $\alpha \geq 1$ is the desired approximation factor.

	Bilinear sketches		General sketches	
Schatten p-norm	Lower bound	Upper bound	Lower bound	Upper bound
$p = \infty$	See general sketches		$\Omega(n^2/\alpha^4)$	$\mathcal{O}(n^2/\alpha^4)$ [7]
$p \in [4, \infty)$	See general sketches		$\Omega(n^{2-4/p})$	$\mathcal{O}(n^{2-4/p})$ even p
$p \in (0, 4) \setminus \{2\}$	See general sketches		$\Omega(n)$	$\mathcal{O}(n^2)$
$p = 0$	$\Omega(n^2)$	$\mathcal{O}(n^2)$	$\Omega(\sqrt{n})$	$\mathcal{O}(n^2)$

applications of the case $p = 1/2$, which is the Laplacian-energy-like (LEL) invariant of a graph.

In nearest-neighbor search (NNS), one technique often used is to first replace each of the input objects (points, images, matrices, etc.) with a small sketch, and then build a lookup table for all possible sketches to support fast query time. Alexandr Andoni, in his talks at the Barriers in Computational Complexity II workshop and Mathematical Foundations of Computer Science conference, stated that a goal would be to design an NNS data structure for the Schatten norms, e.g., the trace or Schatten 1-norm (slide 31 of [4]). If a sketch for a norm has small size, building a table lookup is feasible.

It is similarly assumed for the data stream model that $A \in \mathbb{Z}^{n\times n}$ and $\Delta \in \{-m, \dots, m\}$ and for the sketching model that $A \in \mathbb{R}^{n\times n}$ and $\Delta \in \mathbb{R}$. The sketching model can be further categorized into *general linear sketching* and *bilinear sketching.* In the general linear sketching model, the $n \times n$ matrix A is interpreted as a vector in $\mathbb{R}^{n^2}$, and the goal is then to design a distribution over linear maps $L : \mathbb{R}^{n^2} \to \mathbb{R}^k$, such that for any fixed $n \times n$ matrix A, interpreting it as a vector in $\mathbb{R}^{n^2}$, from $L(A)$ one can approximate $\|A\|_p$ up to a factor with constant probability. The goal is to minimize k. In the bilinear sketching model, there is a distribution over pairs of $r \times n$ matrices S and $n \times s$ matrices T such that for any fixed $n \times n$ matrix A, from $S \cdot A \cdot T$ one can approximate $\|A\|_p$ up to an approximation factor with constant probability, where $\|A\|_p$ is a matrix norm. The goal is to minimize $r \cdot s$. This model has been used in several streaming papers for sketching matrices [7, 25, 43], and as far as we are aware, all known sketches in numerical linear algebra applications have this form. It also has the advantage that SAT can be computed quickly if S and T have fast matrix multiplication algorithms.

1.1. Our contributions.

1.1.1. Sketching model. We summarize our results for the two sketching models in Table 1.1. We note that, prior to our work, for all $p \notin \{2, \infty\}$, all upper bounds in the table were a trivial $\mathcal{O}(n^2)$, while all lower bounds for $p \leq 2$ were a trivial $\Omega(1)$, while for $p > 2$ they were a weaker $\Omega(n^{1-2/p} \log n)$.

For the general sketching model, we have the following results. For the spectral norm ($p = \infty$), we prove an $\Omega(n^2/\alpha^4)$ bound for achieving a factor α-approximation with constant probability, matching an upper bound achievable by an algorithm of [7]. This generalizes to Schatten p-norms for $p > 2$, for which we prove an $\Omega(n^{2-4/p})$ lower bound and give a matching $\mathcal{O}(n^{2-4/p})$ upper bound for even integers p. For odd integers p we are only able to achieve this upper bound if we additionally assume that

A is positive semidefinite[3] (PSD). For $p = 0, 1$ our bounds are now weaker $\Omega(\sqrt{n})$ and $\Omega(n)$, respectively. However, they are the first superconstant lower bounds for the rank and the Schatten 1-norm, respectively, which in particular rules out a naïve table-lookup solution to the NNS problem, addressing a question of Andoni.

For the bilinear sketching model, we prove an $\Omega(n^2)$ lower bound for rank approximation ($p = 0$), showing that no nontrivial sketching is possible.

Prior to our work, surprisingly, the only known $o(n^2)$ upper bound for either model was for $p = 2$, in which case one can achieve a bilinear sketch with $r \cdot s = \mathcal{O}(1)$ [43]. Moreover, the only lower bounds known were those for estimating the ℓ_p-norm of a vector x, obtained for $p > 2$ by planting v into the diagonal line of A, and were of the form $k = \Omega(n^{1-2/p} \log n)$ [8, 56, 62]. Thus, it was not even known whether a sketching dimension of $r \cdot s = \mathcal{O}(1)$ was sufficient for bilinear sketches to obtain a constant-factor approximation to the rank or Schatten 1-norm, or if $k = \Omega(n^2)$ was required for general linear sketches.

1.1.2. Data stream model. We show 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 are necessary. Moreover, this holds even for matrices with $\mathcal{O}(1)$ nonzero 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 $\mathcal{O}(1)$ nonzero entries per row and column, which achieves $\tilde{\mathcal{O}}(n^{1-2/p})$ bits of space. Also, $\Omega(n^{1-2/p})$ bits of space are necessary for even integers p, even with $\mathcal{O}(1)$ nonzero entries per row and column, and even if all entries are absolute constants independent of n.

We summarize prior work and its relation to our results in Table 1.2. 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 ℓ_p-norms. For $p \in [1, 2)$, the lower bound was $\Omega(\frac{n^{1/p-1/2}}{\log n})$ [6], based on nonembeddability, and the best lower 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 $\mathcal{O}(\log n)$ bits; further, it is unknown whether the lower bound of [6] holds for sparse matrices [63]. 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 ℓ_p-norms of vectors. For $p = 0$, an $\Omega(n^{1-g(\epsilon)})$ lower bound was shown for $(1+\epsilon)$-approximation [16], where $g(\epsilon) \to 0$ as $\epsilon \to 0^+$.

Our techniques also generalize to other functions of a matrix spectrum, i.e., functions of the form $f(A) = \sum_i f(\sigma_i)$, where σ_i's are the singular values of A. There are a number of other spectrum functions of importance, including the SVD entropy function, which has seen foundational applications in genome processing [3], and optimal eigenvalue shrinkers, which are motivated from regularized low rank approximation. The SVD entropy is defined to be $f(\sigma_i) = (\sigma_i^2/\|A\|_F^2)\log_2(\|A\|_F^2/\sigma_i^2)$. In insertion-only streams, $\|A\|_F^2$ can be computed exactly, so one can set $f(\sigma_i) = \sigma_i^2 \log_2(1/\sigma_i^2)$, from which, given $\|A\|_F^2$ and an approximation to $\sum_{i=1}^n f(\sigma_i)$, one can approximate the SVD entropy. Optimal eigenvalue shrinkers are defined for different loss functions, such as the Frobenius, operator, and nuclear norm losses [36]. 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 $f(x) = 0$

[3] A square matrix is called positive semidefinite if it is symmetric and all its eigenvalues are nonnegative.

Table 1.2

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 logarithmic 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 $\mathcal{O}(1)$.

		Space complexity in bits	
		Previous lower bounds	Our lower bounds
Schatten p-norm	$p \in (2,\infty) \cap 2\mathbb{Z}$	$n^{1-2/p}$ [37, 46]	
	$p \in (2,\infty) \setminus 2\mathbb{Z}$	$n^{1-2/p}$ [37, 46]	$n^{1-g(\epsilon)}$
	$p \in [1,2)$	$\frac{n^{1/p-1/2}}{\log n}$ [6]	$n^{1-g(\epsilon)}$
	$p \in (0,1)$	$\log n$ [49]	$n^{1-g(\epsilon)}$
	$p = 0$	$n^{1-g(\epsilon)}$ [16]	
Ky Fan k-norm		$\max\{\frac{n}{k}, \frac{k^{1/2}}{\log k}\}$ [12, 6]	$n^{1-g(\epsilon)}$ (any k)

otherwise, where α is a given parameter. 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 further discuss these functions in section 13.

We also obtain a similar near-linear lower bound 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 [72, 29]. 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.

1.2. Our techniques.

1.2.1. General linear sketches. A standard technique for proving lower bounds is Yao's minimax principle, which implies that if there exists a distribution on sketches that succeeds on all $n \times n$ input matrices A with large probability, then for any distribution $\mathcal{L}$ on inputs A, there is a fixed sketch $L : \mathbb{R}^{n^2} \to \mathbb{R}^k$, which succeeds with large probability over $A \sim \mathcal{L}$. Moreover, when treating L as a $k \times n^2$ matrix, we can assume the rows of L are orthonormal. The induced distribution of the sketch $L(A)$, where $A \sim \mathcal{L}$, is denoted by $\mathcal{L}'$.

For $p > 2$ we choose $\mathcal{L}_1$ and $\mathcal{L}_2$ such that $\mathcal{L}_1'$ is the distribution of a k-dimensional vector g of independent and identically distributed (i.i.d.) Gaussians and $\mathcal{L}_2'$ is the distribution of $g + h$, where g is as before but $h = \frac{1}{n^{1/2-1/p}}(u^T L^1 v, \ldots, u^T L^k v)$ for random n-dimensional Gaussian vectors u and v, and where L^i is the ith row of the sketching matrix L, viewed as an $n \times n$ matrix. We again use the χ^2-divergence to bound the total variance distance $d_{TV}(\mathcal{L}_1', \mathcal{L}_2')$ with appropriate conditioning; the definitions of χ^2-divergence and total variation distance can be found in subsection 2.3. This idea was previously used in the context of sketching p-norms of vectors [8], improving the previous $\Omega(n^{1-2/p})$ bound to $\Omega(n^{1-2/p} \log n)$.

For $p = 1$ we consider distinguishing a (scaled) $n \times n$ Gaussian matrix $\frac{1}{\sqrt{n}} G$ from a random orthogonal $n \times n$ matrix O. Previously Chatterjee and Meckes showed that the Wasserstein distance between the linear sketches in these two cases is $\mathcal{O}(k/n^{3/2})$ [21]. In general, small Wasserstein distance does not imply small total variation distance, and it seems unlikely that one can obtain a bound on total variation distance following their techniques since the characterization of a multidimensional Gaussian distribution is a second-order differential equation. Instead, we perturb the candidate distributions and consider $\frac{1}{\sqrt{n}}G + \frac{\eta}{\sqrt{n}}G'$ and $O + \frac{\eta}{\sqrt{n}}G'$ for some small constant $\eta > 0$ and an independent copy of Gaussian matrix G'. This perturbs the linear sketches by a random Gaussian vector $N(0, \frac{\eta^2}{n}I_n)$, where I_n denotes the $n \times n$ identity matrix. The total variation distance between the new linear sketches can now be bounded in terms of the Wasserstein distance between the original linear sketches.

For $p = 0$ we look at distinguishing an $n \times n$ Gaussian matrix G from a matrix $\begin{pmatrix} G' & G'O \end{pmatrix}$, where G' is an $n \times n/2$ Gaussian random matrix and O is a random $n/2 \times n/2$ orthogonal matrix. It is clear that the ranks in the two cases are different by a constant factor. Applying our sketching matrix L, we have $\mathcal{L}_1'$ distributed as $N(0, I_k)$, but $\mathcal{L}_2'$ is the distribution of $(Z_1, \ldots, Z_k)$, where $Z_i = \langle A^i, G' \rangle + \langle B^i, G'O \rangle$, and each L^i is written as the adjoined matrix $\begin{pmatrix} A^i & B^i \end{pmatrix}$ for $(n \times n/2)$-dimensional matrices A^i and B^i. For each fixed O, we can view Z as a k-dimensional Gaussian vector formed from linear combinations of entries of G'. Thus the problem amounts to bounding the variation distance between two zero-mean k-dimensional Gaussian vectors with different covariance matrices. For $\mathcal{L}_1'$ the covariance matrix is the identity I_k, while for $\mathcal{L}_2'$ it is $I_k + P$ for some perturbation matrix P. We show that with constant probability over O, the Frobenius norm $\|P\|_F$ is small enough to give us an $k = \Omega(\sqrt{n})$ bound, and so it suffices to fix O with this property. One may worry that fixing O reduces the variation distance—in this case one can show that with $k = \mathcal{O}(\sqrt{n})$, distributions $\mathcal{L}_1'$ and $\mathcal{L}_2'$ already have constant variation distance.

1.2.2. Bilinear sketches. We follow the same framework for the general sketches, and similarly we may assume, when proving the lower bounds, that the sketching matrix S has orthonormal rows and T has orthonormal columns. Thus, it suffices to give two distributions $\mathcal{L}_1$ and $\mathcal{L}_2$ on A for which the $\|A\|_p$ values differ by a factor α with high probability (w.h.p.) in the two distributions, but for any matrix S with orthonormal rows and T with orthonormal columns, the induced distributions $\mathcal{L}_1'$ and $\mathcal{L}_2'$ on SAT, when $A \sim \mathcal{L}_1$ and $A \sim \mathcal{L}_2$, respectively, have low total variation distance $d_{TV}(\mathcal{L}_1', \mathcal{L}_2')$.

Since S has orthonormal rows and T has orthonormal columns, if $\mathcal{L}_1$ and $\mathcal{L}_2$ are rotationally invariant distributions, then SAT is equal in distribution to an $r \times s$ submatrix of A. For our $\Omega(n^2)$ bound for $p = 0$, we consider the following rotationally invariant distributions: $\mathcal{L}_1 = UV^T$ for $n \times n/2$ i.i.d. Gaussian U and V, while $\mathcal{L}_2 = UV^T + \gamma G$ for the same U and V and G an $n \times n$ i.i.d. Gaussian matrix with variance $\gamma \leq 1/\operatorname{poly}(n)$.

The problem of bounding $d_{TV}(\mathcal{L}_1', \mathcal{L}_2')$ amounts to distinguishing an $r \times s$ submatrix Q of UV^T from an $r \times s$ submatrix of $UV^T + \gamma G$. Working directly with the density function of UV^T is intractable. Instead, we provide an algorithmic proof to bound the variation distance. See Theorem 6.1 for details. The proof also works for an arbitrary subset $Q \subseteq [n] \times [n]$ of cardinality $\mathcal{O}(n^2)$, implying a lower bound of $\Omega(n^2)$ to decide whether an $n \times n$ matrix is of rank at most $n/2$ or ϵ-far from rank $n/2$ (for constant ϵ), showing that an algorithm of Krauthgamer and Sasson is optimal [52].

1.2.3. Sketching algorithm. Due to these negative results, a natural question is whether nontrivial sketching is possible for any Schatten p-norm other than the Frobenius norm. To show that this is possible, given an $n \times n$ matrix A, we left multiply by an $n \times n$ matrix G of i.i.d. Gaussians and right multiply by an $n \times n$ matrix H of i.i.d. Gaussians, resulting in a matrix A' of the form $G'\Sigma H'$, where G', H' are i.i.d. Gaussian and Σ is diagonal with the singular values of A on the diagonal. We then look at *cycles* in a submatrix of $G'\Sigma H'$. The (i,j)th entry of A' is $\sum_{\ell=1}^{n} \sigma_\ell G'_{i,\ell} H'_{\ell,j}$. Interestingly, for even p, for any distinct $i_1,\ldots,i_{p/2}$ and distinct $j_1,\ldots,j_{p/2}$,

$$\mathbb{E}[(A'_{i_1,j_1}A'_{i_2,j_1})\cdot(A'_{i_2,j_2}A'_{i_3,j_2})\cdots(A'_{i_{p/2},j_{p/2}}A'_{i_1,j_{p/2}})] = \|A\|_p^p.$$

The row indices of A' read from left to right to form a cycle $(i_1, i_2, i_2, i_3, i_3, , \ldots, i_{p/2}, i_{p/2}, i_1)$ which, since also each column index occurs twice, results in an unbiased estimator. We need to average over many cycles to reduce the variance, and one way to obtain the associated estimates is to store a submatrix of A' and average over all cycles in it. While some of the cycles are dependent, their covariance is small, and we show that storing an $n^{1-2/p} \times n^{1-2/p}$ submatrix of A' suffices.

1.2.4. Bit lower bound. The starting point of our work is [16], 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 [69].

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; or (2) block diagonal, in which each block is the concatenation of an all-1s matrix and a diagonal matrix with an *odd number* of 1s. 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 [11, 35, 69], we can group the blocks into independent groups of four matrices and then apply a Chernoff bound across the groups to conclude that w.h.p., $\sum_i f(\sigma_i)$ 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 9.1.

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. Our choice of hard instance allows us to consider only the contribution from the singular values $r(k)$, which are the square roots of the roots of a quadratic equation. 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

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

where $\left\{{s \atop m}\right\}$ is the Stirling number of the second kind, the second equality is a combinatorial identity (see, e.g., [39, p. 265]), and the third equality follows from the fact that $\left\{{s \atop m}\right\} = 0$ for $s < m$. 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, for which we use machinery from hypergeometric functions.

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.

1.2.5. Upper bound for even p and sparse matrices. 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 ith 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 the Cauchy–Schwarz inequality, $\alpha^4 = \|a_i\|^4 \geq |\langle a_i, a_j\rangle|^2$ 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 $\Omega(n)$ nonzero terms. It follows that if we sample $\mathcal{O}(\sqrt{n})$ rows uniformly and in their entirety, by looking at all $\mathcal{O}(n)$ pairs $|\langle a_i, a_j\rangle|^2$ for sampled rows a_i and a_j, we obtain $\Omega(1)$ samples of the "contributing" pairs $i \neq j$. Using the fact that each row and column has $\mathcal{O}(1)$ nonzero entries, this can be shown to be enough to obtain a good estimate to $\|A\|_4^4$, and it uses $\mathcal{O}(\sqrt{n}\log n)$ 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 [60, 5, 47] and using our conditions that each row and each column of A have $\mathcal{O}(1)$ nonzero 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 $\mathcal{O}(\sqrt{n})$ rows need to be sampled, and we can look at all $\mathcal{O}(n)$ pairs of inner products between such rows.

1.2.6. Open questions. We believe our work raises a number of intriguing open questions.

1. Is it possible that for every odd integer $p < \infty$, sketching the Schatten p-norm requires $k = \Omega(n^2)$? Interestingly, odd and even p behave very differently since for even p, we have $\|A\|_p = \|A^2\|_{p/2}$, where A^2 is PSD. Note that estimating Schatten norms of PSD matrices A can be much easier: in the extreme case of $p = 1$, the Schatten norm $\|A\|_1$ is equal to the trace of A, which can be computed with $k = 1$, while we show $k = \Omega(\sqrt{n})$ for estimating $\|A\|_1$ for non-PSD A.
2. Both the bit lower bound and the sketching lower bound show that even integers p are a special case. Is it possible that the even integers p are the only special case in which the sketching lower bound for all $p \notin 2\mathbb{Z}$ requires $k = \Omega(n^2)$?

3. Is it possible to prove an $\Omega(n^2)$, or even an $\Omega(n)$, bit lower bound for dense matrices? Recently the bit lower bound of estimating rank was improved to $\Omega(n^{2-g(\epsilon)})$ by Assadi, Khanna, and Li using the Erdős–Szemerédi graph [10], which is a dense graph. The new construction, however, does not yield a gap in Schatten norms.

1.2.7. Follow-up work. After the publication of our work [55], Kong and Valiant improved the running time for estimation of Schatten p-norms for even integers p in [51], while achieving the same sketching dimension; see also follow-up work on sampling matrices to estimate their spectrum by Khetan and Oh [50]. In other follow-up work, Braverman et al. [14] extended our results to multipass algorithms, showing that better space complexity for estimating Schatten p-norms is possible with a constant (depending on p) number of passes. They also obtained faster algorithms based on more structured random sketches.

1.2.8. Organization. In section 2 we review basic notions and concepts which are used in subsequent sections. In sections 3–8 we discuss the sketching model, and sections 9–13 explore the data stream model.

2. Preliminaries.

2.1. Notation. Let $\mathbb{R}^{n\times d}$ be the set of $n \times d$ real matrices, and let O_n be the orthogonal group of degree n (i.e., the set of $n \times n$ orthogonal matrices).[4] We write $X \sim \mathcal{D}$ for a random variable X subject to a probability distribution $\mathcal{D}$. Let $N(\mu, \Sigma)$ denote the (multivariate) normal distribution of mean μ and covariance matrix Σ, and let $\chi^2(n)$ denote the χ^2-distribution with n degrees of freedom. Denote the uniform distribution on a set S (if it exists) by $\mathrm{Unif}(S)$. We also use O_n to denote the uniform distribution over the orthogonal group of order n (i.e., endowed with the normalized Haar measure). We denote by $\mathcal{G}(m, n)$ the ensemble of $m \times n$ random matrices with i.i.d. $N(0, 1)$ entries.

We write the n-dimensional vector full of 1s as $\mathbf{1}^n$ and write the n-dimensional zero vector as $\mathbf{0}^n$. We also adopt the conventional notation $[n]$ as shorthand for the set $\{1, 2, \ldots, n\}$.

For two $n \times n$ matrices X and Y, we define $\langle X, Y\rangle$ as $\langle X, Y\rangle = \mathrm{tr}(X^T Y) = \sum_{i,j} X_{ij}Y_{ij}$, i.e., the entrywise inner product of X and Y.

For two distributions μ and ν, we denote by $\mu * \nu$ the convolution of μ and ν.

We write $X = a \pm b$ if $a - b \le X \le a + b$. Similarly, we write $X = (1 \pm \epsilon)Y$ if $(1 - \epsilon)X \le Y \le (1 + \epsilon)X$ and say that X is a $(1 \pm \epsilon)$-approximation to Y in this case. For $\beta > 1$, we also say X is a β-approximation to Y if $Y \le X \le \beta Y$. It is clear that if X is a $(1 \pm \epsilon)$-approximation to Y, $(1 + \epsilon)X$ is a $(1 + 3\epsilon)$-approximation to Y when $\epsilon \in (0, 1)$; consequently, it is conventional not to distinguish between $(1 + \epsilon)$-approximation algorithms and $(1 \pm \epsilon)$-approximation algorithms, provided that $\epsilon > 0$ is sufficiently small.

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

[4] Throughout the paper we use the italic letter O for orthogonal matrices and the script letter $\mathcal{O}$ for the big-O notation in asymptotics.

2.2. Singular values and Schatten norms. Consider a square matrix $A \in \mathbb{R}^{n\times n}$; its eigenvalues are denoted by $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_n(A)$ in decreasing order. The set of the eigenvalues of A is called the spectrum of A.

Now consider a matrix $A \in \mathbb{R}^{n\times d}$ (not necessarily a square matrix); then $A^T A$ is a positive semidefinite matrix. The eigenvalues of $\sqrt{A^T A}$ are called the singular values of A, denoted by $\sigma_i(A) = \sqrt{\lambda_i(A^T A)}$, sorted in decreasing order again. Let $r = \operatorname{rank}(A)$. It is clear that $\sigma_{r+1}(A) = \cdots = \sigma_d(A) = 0$. The matrix A also has the following *singular value decomposition* (SVD) $A = U\Sigma V^T$, where $U \in O_n$, $V \in O_d$, and Σ is an $n \times d$ diagonal matrix with diagonal entries $\sigma_1(A), \ldots, \sigma_{\min\{n,d\}}(A)$.

Define

$$\|A\|_p = \left(\sum_{i=1}^{r} (\sigma_i(A))^p\right)^{\frac{1}{p}}, \quad p > 0;$$

then $\|A\|_p$ is a norm called the pth *Schatten norm*, over $\mathbb{R}^{n\times d}$ for $p \geq 1$. When $p = 1$, it is also called the trace norm or Ky Fan norm. When $p = 2$, it is exactly the Frobenius norm $\|A\|_F$, recalling that $\sigma_i(A)^2 = \lambda_i(A^T A)$ and thus $\|A\|_F^2 = \operatorname{tr}(A^T A) = \sum_i \lambda_i(A^T A)$.

Let $\|A\|_{op}$ denote the operator norm of A when treating A as a linear operator from ℓ_2^d to ℓ_2^n. Additionally, it holds that $\lim_{p\to\infty} \|A\|_p = \sigma_1(A) = \|A\|_{op}$ and $\lim_{p\to 0^+} \|A\|_p = \operatorname{rank}(A)$. We define $\|A\|_\infty$ and $\|A\|_0$ accordingly in this limit sense.

Finally, note that A and A^T have the same nonzero singular values, so $\|A\|_p = \|A^T\|_p$ for all p.

2.3. Distance between probability measures. Suppose μ and ν are two probability measures over some Borel algebra $\mathcal{B}$ on $\mathbb{R}^n$ such that μ is absolutely continuous with respect to ν. For a convex function $\phi : \mathbb{R} \to \mathbb{R}$ such that $\phi(1) = 0$, we define the ϕ-divergence

$$D_\phi(\mu||\nu) = \int \phi\left(\frac{d\mu}{d\nu}\right) d\nu.$$

In general, $D_\phi(\mu||\nu)$ is not a distance because it is not symmetric. For more details on ϕ-divergence, the reader may refer to any standard text, e.g., [26].

The *total variation distance* between μ and ν, denoted by $d_{TV}(\mu, \nu)$, is defined as $D_\phi(\mu||\nu)$ for $\phi(x) = |x - 1|$.

The χ^2*-divergence* between μ and ν, denoted by $\chi^2(\mu||\nu)$, is defined as $D_\phi(\mu||\nu)$ for $\phi(x) = (x - 1)^2$ or $\phi(x) = x^2 - 1$. It can be verified that these two choices of ϕ give exactly the same value of $D_\phi(\mu||\nu)$.

Proposition 2.1 (see [64, p. 99]). $d_{TV}(\mu, \nu) \leq \sqrt{\chi^2(\mu||\nu)}$.

Proposition 2.2 (see [45, p. 97]). $\chi^2(N(0, I_n) * \mu||N(0, I_n)) \leq \mathbb{E}\, e^{\langle x,x'\rangle} - 1$, *where* $x, x' \sim \mu$ *are independent.*

In the case of $n = 1$, if $F(x)$ and $G(x)$ are the cumulative distribution functions of μ and ν, respectively, the *Kolmogorov distance* is defined as

$$d_K(\mu, \nu) = \sup_x |F(x) - G(x)|.$$

It follows easily that for continuous and bounded f,

$$\left|\int f d\mu - \int f d\nu\right| \leq \|f\|_\infty \cdot d_K(\mu, \nu). \tag{2.1}$$

If both μ and ν are compactly supported, it suffices to have f continuous and bounded on the union of the supports of μ and ν.

The *Wasserstein* distance (or the earth mover distance) between μ and ν is defined as

$$d_W(\mu,\nu) = \inf_{\pi} \int d(x,y) d\pi(x,y),$$

where the infimum is taken over all distributions π on $\mathbb{R}^n \times \mathbb{R}^n$ with marginals μ and ν on the first and the second factors, respectively. Usually, the Wasserstein distance and the total variation distance have no clear connection. The following seems to be a folklore result and has been used implicitly in the literature; nevertheless, for completeness we include a proof from [1].

PROPOSITION 2.3. *Let μ and ν be two distributions in $\mathbb{R}^n$. Suppose that γ is a probability distribution in $\mathbb{R}^n$ symmetric around* 0 *and that γ_x denotes the probability distribution that shifts the center of γ to $x \in \mathbb{R}^n$. Then*

$$d_{TV}(\mu * \gamma, \nu * \gamma) \leq \left(\sup_{x \neq y} \frac{d_{TV}(\gamma_x, \gamma_y)}{\|x-y\|_2} \right) \cdot d_W(\mu,\nu).$$

Proof. Let π be any measure with marginals μ and ν so that $\mu = \int \delta_x d\pi(x,y)$ and $\nu = \int \delta_y d\pi(x,y)$. It follows that

$$\mu - \nu = \int (\delta_x - \delta_y) d\pi(x,y)$$

and

$$d_{TV}(\mu * \gamma, \nu * \gamma) \leq \int d_{TV}(\gamma_x, \gamma_y) d\pi(x,y) \leq K \int \|x-y\|_2 d\pi(x,y),$$

where $K = \sup_{x \neq y} d_{TV}(\gamma_x, \gamma_y)/\|x-y\|_2$. Taking infimum over $\pi(x,y)$ yields the claimed result. ∎

2.4. Distribution of singular values. We need the following two lemmata.

LEMMA 2.4 (Marčenko–Pastur law [58, 38]). *Suppose that X is an $m \times n$ matrix with i.i.d. $N(0,1/m)$ entries. Consider the probability distribution $F_X(x)$ associated with the spectrum of $X^T X$ as*

$$F_X(x) = \frac{1}{n} \left| \left\{ i : \lambda_i(X^T X) \leq x \right\} \right|.$$

For $\gamma \in (0,1]$, define a distribution $G_\gamma(x)$ with density function $p_\gamma(x)$ as

$$p_\gamma(x) = \frac{\sqrt{(b-x)(x-a)}}{2\pi\gamma x}, \quad x \in [a,b],$$

where

$$a = (1-\sqrt{\gamma})^2, \quad b = (1+\sqrt{\gamma})^2.$$

Then when $n \to \infty$, $m \to \infty$, and $m/n \to \gamma \in (0,1)$, it holds that the expected Kolmogorov distance is

$$\mathop{\mathbb{E}}_{X} \sup_x |F_X(x) - G_\gamma(x)| = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right).$$

LEMMA 2.5 (operator norm of Gaussian random matrix [70]). *Suppose that $X \sim \mathcal{G}(m,n)$. Then with probability at least $1 - e^{-t^2/2}$, it holds that $\sigma_1(X) \leq \sqrt{m} + \sqrt{n} + t$.*

2.5. Communication complexity. We shall use a problem called Boolean Hidden Hypermatching, denoted by $\mathrm{BHH}^0_{t,n}$, defined in [69]. In this subsection, we denote the exclusive-or by $\oplus$ and the Hamming weight of a Boolean vector x by $w_H(x)$.

Definition 2.6 (see [69]). *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 positive integers* r *and* t*, 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* $(\bigoplus_{1\le i\le t} x_{M_{1,i}}, \ldots, \bigoplus_{1\le i\le t} x_{M_{n/t,i}})$*, where* $\{(M_{j,1},\ldots,M_{j,t})\}_{j=1}^{n/t}$ *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.*

Verbin and Yu [69] 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 [16], 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 2.7. *In the* Boolean Hidden Hypermatching 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* $w_H(x) = n/2$*, and 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* $\left(\bigoplus_{i=1}^t x_{M_{1,i}}, \ldots, \bigoplus_{i=1}^t x_{M_{n/t,i}}\right)$*, where* $\{(M_{j,1},\ldots,M_{j,t})\}_{j=1}^{n/t}$ *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 almost identical) version of the proof in [16] shows that this problem also has an $\Omega(n^{1-1/t})$ randomized one-way communication lower bound. We include the proof below for completeness.

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, and let M be a perfect t-hypermatching on the n coordinates of x and $x \in \{0,1\}^{n/t}$. 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 $\{x_1,\ldots,x_t\} \in M$ be the lth hyperedge of M. We include two hyperedges in M', the input of Bob's input in the $\mathrm{BHH}^0_{t,2n}$, as follows. When $w_l = 0$, include $\{x_1,\ldots,x_t\}$ and $\{\overline{x_1},\overline{x_2},\ldots,\overline{x_t}\}$ in M; when $w_l = 1$, include $\{\overline{x_1},x_2,\ldots,x_t\}$ and $\{x_1,\overline{x_2},\ldots,\overline{x_t}\}$ in M'. Observe 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 $M'x' = \mathbf{0}^{2n}$ if $Mx + w = \mathbf{0}^{n/2}$ and $M'x' = \mathbf{1}^{2n}$ if $Mx + w = \mathbf{1}^{n/2}$. The lower bound then follows from the lower bound for $\mathrm{BHH}_{t,n}$. □

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

2.6. Special functions. The gamma function $\Gamma(x)$ is defined as

$$\Gamma(x) = \int_0^\infty x^{t-1} e^{-x}\, dx.$$

For positive integer n, it holds that $\Gamma(n) = (n-1)!$. The definition above can be extended by analytic continuation to all complex numbers except nonpositive integers.

The hypergeometric function ${}_pF_q(a_1,a_2,\ldots,a_p;b_1,b_2,\ldots,b_q;z)$ of p upper parameters and q lower parameters is defined as

$$ {}_pF_q\left(\begin{matrix} a_1, a_2, \dots, a_p \\ b_1, b_2, \dots, b_q \end{matrix}; z\right) = \sum_{n=0}^{\infty} \frac{(a_1)_n (a_2)_n \cdots (a_p)_n}{(b_1)_n (b_2)_n \cdots (b_q)_n} \cdot \frac{z^n}{n!}, $$

where

$$ (a)_n = \begin{cases} 1, & n = 0, \\ a(a+1)\cdots(a+n-1), & n > 0. \end{cases} $$

The parameters must be set such that the denominator factors are never 0. When $a_i = -n$ for some $1 \le i \le p$ and nonnegative integer n, the series above becomes terminating and the hypergeometric function is in fact a polynomial in x.

3. Sketching lower bound for $p > 2$.

LEMMA 3.1.[5] *Suppose that $x \sim N(0, I_m)$ and $y \sim N(0, I_n)$ are independent and $A \in \mathbb{R}^{m\times n}$ satisfies $\|A\|_F < 1$. It holds that*

$$ \mathop{\mathbb{E}}_{x,y} e^{x^T A y} \le \frac{1}{\sqrt{1 - \|A\|_F^2}}. $$

Proof. First, it is easy to verify that

$$ \begin{aligned} \mathop{\mathbb{E}}_{x,y\sim N(0,1)} e^{axy} &= \frac{1}{2\pi} \iint_{\mathbb{R}\times\mathbb{R}} e^{axy - \frac{x^2+y^2}{2}} dxdy \\ &= \frac{1}{2\pi} \int_{\mathbb{R}} \int_{\mathbb{R}} e^{-\frac{1}{2}(x-ay)^2} e^{-\frac{1}{2}(1-a^2)y^2} dxdy \\ &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{1}{2}(1-a^2)y^2} dy \\ &= \frac{1}{\sqrt{1-a^2}}, \quad a \in [0,1). \end{aligned} $$

Without loss of generality, assume that $m \ge n$. Consider the singular value decomposition $A = U\Sigma V^T$, where U and V are orthogonal matrices of dimension m and n, respectively, and $\Sigma = \operatorname{diag}\{\sigma_1, \dots, \sigma_n\}$, with $\sigma_1, \dots, \sigma_n$ being the nonzero singular values of A. We know that $\sigma_i \in [0, 1)$ for all i by the assumption that $\|A\|_F < 1$. By rotational invariance of the Gaussian distribution, we may assume that $m = n$ and thus that

$$ \begin{aligned} \mathop{\mathbb{E}}_{x,y\sim N(0,I_n)} e^{x^T A y} &= \mathop{\mathbb{E}}_{x,y\sim N(0,I_n)} e^{x^T \Sigma y} \\ &= \frac{1}{(2\pi)^n} \iint_{\mathbb{R}^n\times\mathbb{R}^n} \exp\left\{ \sum_{i=1}^n \left(\sigma_i x_i y_i - \frac{x_i^2 + y_i^2}{2} \right) \right\} dxdy \\ &= \prod_{i=1}^n \frac{1}{\sqrt{1-\sigma_i^2}} \\ &\le \frac{1}{\sqrt{1 - \sum_{i=1}^n \sigma_i^2}} \\ &= \frac{1}{\sqrt{1 - \|A\|_F^2}}. \end{aligned} $$

□

[5] A similar result holds for x and y of i.i.d. centered sub-Gaussian coordinates, where the right-hand side is replaced with $\exp(c\|A\|_F^2)$ for some constant $c > 0$ that only depends on the sub-Gaussian distribution. The proof requires heavier machinery, but we only need the elementary variant here by our choice of hard instance.

Next we consider the problem of distinguishing two distributions $\mathcal{D}_1 = \mathcal{G}(m,n)$ and $\mathcal{D}_2$ as defined below. Let $u_1,\ldots,u_r$ be i.i.d. $N(0,I_m)$ vectors, let $v_1,\ldots,v_r$ be i.i.d. $N(0,I_n)$ vectors, and further suppose that $\{u_i\}$ and $\{v_i\}$ are independent. Let $s \in \mathbb{R}^r$, and define the distribution $\mathcal{D}_2$ as $\mathcal{G}(m,n) + \sum_{i=1}^r s_i u^i (v^i)^T$. We take k linear measurements and denote the corresponding rows (measurements) of the sketching matrix by $L^1,\ldots,L^k$. Without loss of generality, we may assume that $\operatorname{tr}((L^i)^T L^i) = 1$ and $\operatorname{tr}((L^i)^T L^j) = 0$ for $i \neq j$ since these correspond to the rows of the sketching matrix being orthonormal, which we can assume since we can always change the basis of the row space of the sketching matrix in a postprocessing step. Let $\mathcal{L}_1$ and $\mathcal{L}_2$ be the corresponding distribution of the linear sketch of dimension k on $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. The main result is the following theorem.

THEOREM 3.2. *There exists an absolute constant $c > 0$ such that $d_{TV}(\mathcal{L}_1,\mathcal{L}_2) \leq 1/10$ whenever $k \leq c/\|s\|_2^4$.*

Proof. It is not difficult to verify that $\mathcal{L}_1 = N(0,I_k)$ and $\mathcal{L}_2 = N(0,I_k) + \mu$, where μ is the distribution of

$$\begin{pmatrix} \sum_{i=1}^r s_i (u^i)^T L^1 v^i \\ \sum_{i=1}^r s_i (u^i)^T L^2 v^i \\ \vdots \\ \sum_{i=1}^r s_i (u^i)^T L^k v^i \end{pmatrix}.$$

Consider a random variable (we shall see in a moment where it comes from)

$$\xi = \sum_{i=1}^k \sum_{j,l=1}^r \sum_{a,c=1}^m \sum_{b,d=1}^n s_j s_l (L^i)_{ab} (L^i)_{cd} (u^j)_a (v^j)_b (u^l)_c (v^l)_d.$$

Take expectation on both sides, and notice that the nonvanishing terms on the right-hand side must have $j = l$, $a = c$, and $b = d$,

$$\mathbb{E}\,\xi = \sum_{i=1}^k \sum_{j=1}^r \sum_{a=1}^m \sum_{b=1}^n s_j^2 (L^i)_{ab}^2 \,\mathbb{E}(u^j)_a^2 \,\mathbb{E}(v^j)_a^2 = k\|s\|_2^2.$$

Define an event $\mathcal{E} = \{\|s\|^2 \xi < 1/2\}$, and it follows from our assumption and Markov's inequality that $\Pr(\mathcal{E}) \geq 1 - 2c$. Restrict μ to this event, and denote the induced distribution by $\tilde{\mu}$. Let $\tilde{\mathcal{L}}_2 = N(0,I_n) + \tilde{\mu}$.

Then the total variation distance between $\mathcal{L}_1$ and $\mathcal{L}_2$ can be upper bounded, using Propositions 2.1 and 2.2, as

$$\begin{aligned} d_{TV}(\mathcal{L}_1,\mathcal{L}_2) &\leq d_{TV}(\mathcal{L}_1,\tilde{\mathcal{L}}_2) + d_{TV}(\mathcal{L}_2,\tilde{\mathcal{L}}_2) \\ &\leq \sqrt{\mathop{\mathbb{E}}_{z_1,z_2\sim\tilde{\mu}} e^{\langle z_1,z_2\rangle} - 1} + d_{TV}(\mu,\tilde{\mu}) \\ &\leq \sqrt{\frac{1}{\Pr(\mathcal{E})}\left(\mathop{\mathbb{E}}_{z_1\sim\tilde{\mu},z_2\sim\mu} e^{\langle z_1,z_2\rangle} - 1\right)} + \Pr(\mathcal{E}^c), \end{aligned}$$

and we shall bound $\mathbb{E}\, e^{\langle z_1, z_2\rangle}$ in the rest of the proof.

$$\begin{aligned}\mathop{\mathbb{E}}_{z_1\sim\tilde{\mu}, z_2\sim\mu} e^{\langle z_1,z_2\rangle} &= \mathbb{E}\exp\left\{\sum_{i=1}^{k}\sum_{j,a,b}\sum_{j',a',b'} s_j (L^i)_{ab}(u^j)_a(v^j)_b\cdot s_{j'}(L^i)_{a'b'}(x^{j'})_{a'}(y^{j'})_{b'}\right\}\\ &= \mathop{\mathbb{E}}_{u^1,\dots,u^r,v^1,\dots,v^r|\tilde{\mu}} \prod_{j'=1}^{r} \mathop{\mathbb{E}}_{\substack{x_{j'}\sim N(0,I_m)\\ y_{j'}\sim N(0,I_n)}} \exp\left\{\sum_{a',b'} Q^{j'}_{a',b'}(x^{j'})_{a'}(y^{j'})_{b'}\right\},\end{aligned}$$

where $Q^{j'}$ is an $m\times n$ matrix whose (a',b')th entry is defined as

$$Q^{j'}_{a',b'} = s_{j'}\sum_{i=1}^{k}\sum_{j,a,b}(L^i)_{ab}(L^i)_{a'b'}\cdot s_j(u^j)_a(v^j)_b.$$

In order to apply the preceding lemma, we need to verify that $\|Q^{j'}\|_F^2 < 1$. Indeed,

$$\begin{aligned}&\|Q^{j'}\|_F^2\\ &= \sum_{a',b'}(Q^{j'})^2_{a',b'}\\ &= s_{j'}^2\sum_{a',b'}\sum_{i,i'}\sum_{j,a,b}\sum_{\ell,c,d} s_j(L^i)_{ab}(L^i)_{a'b'}(u^j)_a(v^j)_b\cdot s_\ell(L^{i'})_{cd}(L^{i'})_{a'b'}(u^\ell)_c(v^\ell)_d\\ &= s_{j'}^2\sum_{a',b'}\sum_{i}(L^i)^2_{a'b'}\sum_{j,a,b}\sum_{\ell,c,d} s_j(L^i)_{ab}(u^j)_a(v^j)_b\cdot s_\ell(L^i)_{cd}(u^\ell)_c(v^\ell)_d \text{ (i must be equal to i')}\\ &= s_{j'}^2\sum_{i}\sum_{j,a,b}\sum_{\ell,c,d} s_j(L^i)_{ab}(u^j)_a(v^j)_b\cdot s_\ell(L^i)_{cd}(u^\ell)_c(v^\ell)_d\\ &= s_{j'}^2\xi < 1\end{aligned}$$

since we have conditioned on $\mathcal{E}$. Now it follows from the preceding lemma that

$$\begin{aligned}\mathop{\mathbb{E}}_{u^1,\dots,u^r,v^1,\dots,v^r}\prod_{i=1}^{r}\mathop{\mathbb{E}}_{x_{j'},y_{j'}}\exp\left\{\sum_{a',b'}Q^{j'}_{a',b'}(x^{j'})_{a'}(y^{j'})_{b'}\right\} &\le \mathop{\mathbb{E}}_{u^1,\dots,u^r,v^1,\dots,v^r}\prod_{j'=1}^{r}\frac{1}{\sqrt{1-s_{j'}^2\xi}}\\ &\le \mathop{\mathbb{E}}_{u^1,\dots,u^r,v^1,\dots,v^r}\frac{1}{\sqrt{1-\|s\|^2\xi}}\\ &\le 1+\|s\|^2\,\mathbb{E}\,\xi\\ &\le 1+k\|s\|^4,\end{aligned}$$

where, in the third inequality, we used the fact that $1/\sqrt{1-x}\le 1+x$ for $x\in[0,1/2]$. Therefore,

$$d_{TV}(\mathcal{L}_1,\mathcal{L}_2)\le\sqrt{\frac{k\|s\|^4}{1-2c}}+2c\le\sqrt{\frac{c}{1-2c}}+2c\le\frac{1}{10}$$

when $c>0$ is small enough. □

We will apply the preceding theorem to obtain our lower bounds for the applications. To do so, we note that by Yao's minimax principle, we can fix the rows of our sketching matrix and show that the resulting distributions $\mathcal{L}_1$ and $\mathcal{L}_2$ above have small total variation distance. By standard properties of the variation distance, this implies that no estimation procedure can be used to distinguish the two distributions with sufficiently large probability, thereby establishing our lower bound.

COROLLARY 3.3 (α-approximation to operator norm). *Let $c > 0$ be an arbitrarily small constant. For $\alpha \geq 1 + c$, any sketching algorithm that estimates $\|X\|_{op}$ for $X \in \mathbb{R}^{n\times n}$ within a factor of α with error probability $\leq 1/6$ requires sketching dimension $\Omega(n^2/\alpha^4)$.*

Proof. Let $m = n$, take $r = 1$ and $s_1 = 5\alpha/\sqrt{n}$ in $\mathcal{D}_2$, and apply the preceding theorem. For notational convenience, we write u^1 and v^1 in the definition of $\mathcal{D}_2$ as u and v, respectively. We claim that $\|G\|_{op}$ and $\|G + \frac{5\alpha}{\sqrt{n}}uv^T\|_{op}$ differ by a factor of α with high probability; then the preceding theorem yields the desirable lower bound.

It follows from Lemma 2.5 that $\|G\|_{op} \leq 2.1\sqrt{n}$ with probability $\geq 1 - e^{-n/200}$. On the other hand, let $X = \frac{C\alpha}{\sqrt{n}}uv^T$. Note that X is of rank one, and the only nonzero singular value $\sigma_1(X) = \|X\|_F \geq 4.9\alpha\sqrt{n}$ with w.h.p., since $\|uv^T\|_F^2 = \|u\|_2^2\|v\|_2^2 \sim (\chi^2(n))^2$, which is tightly concentrated around n^2. It follows that

$$\|G + X\|_{op} \geq \|X\|_{op} - \|G\|_{op} \geq (4.9\alpha - 2.1)\sqrt{n} \geq 2.1\alpha\sqrt{n}.$$

Thus $\|G\|_{op}$ and $\|G + X\|_{op}$ differ by a factor of at least α. ◻

COROLLARY 3.4 (Schatten norm for $p > 2$). *Suppose that $p > 2$. There exists a constant $c = c(p) > 0$ such that any sketching algorithm that estimates $\|X\|_p^p$ for $X \in \mathbb{R}^{n\times n}$ within a factor of $1 + c$ with error probability $\leq 1/6$ requires sketching dimension $\Omega(n^{2(1-2/p)})$.*

Proof. Let $m = n$, and take $r = 1$ and $s_1 = 5/n^{1/2-1/p}$ in $\mathcal{D}_2$. For notational convenience, we write u^1 and v^1 in the definition of $\mathcal{D}_2$ as u and v, respectively. We shall show that $\|G\|_p$ and $\|G + 5n^{1/p-1/2}uv^T\|_p$ differ by a constant factor w.h.p.

Let $X = 5n^{1/p-1/2}uv^T$. Since X is of rank one, the only nonzero singular value $\sigma_1(X) = \|X\|_F \geq 4.9 \cdot n^{1/p+1/2}$ w.h.p, since $\|uv^T\|_F^2 \sim (\chi^2(n))^2$, which is tightly concentrated around n^2.

On the other hand, combining Lemmas 2.4 and 2.5 as well as (2.1) with $f(x) = x^p$ on $[0, 4]$, we can see that with probability $1 - o(1)$ it holds for $X \sim \frac{1}{\sqrt{n}}\mathcal{G}(n, n)$ (note the normalization!) that

$$\|X\|_p^p = (I_p + o(1))\, n, \quad p > 0, \tag{3.1}$$

where

$$I_p = \int_0^4 x^{\frac{p}{2}} \cdot \frac{\sqrt{(4-x)x}}{2\pi x}dx = \frac{2^p\Gamma(\frac{1+p}{2})}{\sqrt{\pi}\Gamma(2 + \frac{p}{2})}, \quad p > 0. \tag{3.2}$$

We claim that $I_p \leq 2^p$. To prove the claim, it suffices to show that $\Gamma(x) \leq \sqrt{\pi}\Gamma(x + \frac{3}{2})$ for $x \geq \frac{1}{2}$. It is a well-known fact that $\Gamma(x)$ is decreasing on $[0, x_0]$ and increasing on $[x_0, +\infty)$ for some $x_0 \in (1.46, 1.47)$ (see, e.g., [27]), and thus it suffices to show the inequality for $x \in [\frac{1}{2}, x_0]$, in which case $\Gamma(x) \leq \Gamma(\frac{1}{2}) = \sqrt{\pi} = \sqrt{\pi}\Gamma(2) \leq \sqrt{\pi}\Gamma(x + \frac{3}{2})$. This proves the claim.

Hence $\|G\|_p \leq 1.1 \cdot I_p^{1/p} n^{1/2+1/p} \leq 1.1 \cdot 2 \cdot n^{1/2+1/p}$ w.h.p. On the other hand, by the triangle inequality,

$$\|G + X\|_p \geq \|X\|_p - \|G\|_p \geq (4.9 - 2.2)n^{1/p+1/2} \geq 1.2 \cdot 2.2n^{1/p+1/2} \geq 1.2\|G\|_p$$

w.h.p.

The result follows from the preceding theorem, where c can be chosen as $c = c(p) = \frac{|I_p-1|}{2(I_p+1)}$. ◻

4. Sketching lower bound for $p > 0$ and $p \neq 2$. Let $G, G' \sim \mathcal{G}(n,n)$ and $O \sim O_n$ be independent. When $p > 0$, it follows from (3.2) that $I_p = 1$ if and only if $p = 2$, and it then follows from (3.1) that $\|\frac{1}{\sqrt{n}}G\|_p^p$ and $\|O\|_p^p$ differ by a constant factor (depending on p) for $p \neq 2$ with probability $1 - o(1)$; therefore we can pick $\eta \in (0,1)$ to be a constant sufficiently small such that $\|\frac{1}{\sqrt{n}}G + \frac{\eta}{\sqrt{n}}G'\|_p^p$ and $\|O + \frac{\eta}{\sqrt{n}}G'\|_p^p$ differ by a constant factor of $(1 + c_p)$ with probability $1 - o(1)$ for some small constant c_p that depends only on p.

Let $\mathcal{D}_1$ be the distribution of $\frac{1}{\sqrt{n}}G + \frac{\eta}{\sqrt{n}}G'$, and let $\mathcal{D}_2$ be the distribution of $O + \frac{\eta}{\sqrt{n}}G$. We consider the problem of distinguishing the two distributions $\mathcal{D}_1$ and $\mathcal{D}_2$. For a distribution $\mathcal{D}$, we denote by $\mathcal{L}(\mathcal{D})$ the distribution of the induced linear sketch, where, as in the previous section, we assume that the linear sketch L has orthonormal rows.

Theorem 4.1 (Schatten norm for $p \neq 2$). *Let $p > 0$ and $p \neq 2$. There exists a constant $c = c(p) > 0$ such that any sketching algorithm that estimates $\|X\|_p^p$ for $X \in \mathbb{R}^{n\times n}$ ($n \geq 2$) within a factor of $1 + c$ with error probability $\leq 1/6$ requires sketching dimension $\Omega(n)$.*

Proof. It was proved by Chatterjee and Meckes [21, Theorem 4.4] (note that $\mathcal{L}(\frac{1}{\sqrt{n}}G) = N(0, \frac{1}{n}I_n)$ and that our scaling of sketches L_i differs from that in [21] by a factor of $\sqrt{n}$) that

$$d_W\left(\mathcal{L}\left(\frac{1}{\sqrt{n}}G\right), \mathcal{L}(O)\right) \leq \frac{\sqrt{2}k}{\sqrt{n}(n-1)}.$$

For notational convenience, let $\gamma = N(0, \frac{\eta^2}{n}I_n)$, and we consider $d_{TV}(\gamma_x, \gamma_y)$ for $x, y \in \mathbb{R}^n$, where γ_x and γ_y are as defined in Proposition 2.3. It is well known that (see, e.g. [64, p. 146])

$$d_{TV}(N(x, I_n), N(y, I_n)) \leq \min\left\{1, \frac{1}{\sqrt{2}}\|x-y\|_2\right\}, \quad x, y \in \mathbb{R}^n.$$

Rescaling the covariance matrix, we see that

$$d_{TV}(\gamma_x, \gamma_y) \leq \min\left\{1, \frac{\frac{1}{\sqrt{2}}\|x-y\|_2}{\eta/\sqrt{n}}\right\} \leq \min\left\{1, \frac{\eta}{\sqrt{2}}\sqrt{n}\|x-y\|_2\right\}.$$

In Proposition 2.3, taking η as in the current argument, $\mu = \mathcal{L}(\frac{1}{\sqrt{n}}G)$ and $\nu = \mathcal{L}(O)$, we obtain that

$$d_{TV}(\mathcal{L}(\mathcal{D}_1), \mathcal{L}(\mathcal{D}_2)) \leq \frac{\eta k}{n-1} < \frac{k}{n-1}.$$

Hence when $k \leq (n-1)/10$, it will hold that $d_{TV} \leq 1/10$. □

5. Sketching lower bound for rank ($p = 0$). Consider a random matrix $\begin{pmatrix} G & GO \end{pmatrix}$, where $G \sim \mathcal{G}(n, n/2)$ and $O \sim O_{n/2}$.

As before, consider k orthonormal sketches $L^1, \ldots, L^k$. For each i, write L^i as $L^i = \begin{pmatrix} A^i & B^i \end{pmatrix}$. Then by orthonormality, $\langle A^i, A^j\rangle + \langle B^i, B^{(j)}\rangle = \delta_{i,j}$. Define $Z_i = \langle A^i, G\rangle + \langle B^i, GO\rangle$ and $\mathcal{D}_{n,k}$ to be the distribution of $(Z_1, \ldots, Z_k)$.

Theorem 5.1. *Let $\mathcal{D}_{n,k}$ be defined as above, and let $\zeta \in (0,1)$. Then it holds for $k \leq (\zeta/3)^{3/2}\sqrt{n}$ that $d_{TV}(\mathcal{D}_{n,k}, N(0, I_k)) \leq \zeta$.*

Proof. The sketch can be written as a matrix Φg, where $\Phi \in \mathbb{R}^{k\times n^2/2}$ is a random matrix that depends on A^i, B^i, and O, and $g \sim N(0, I_{n^2/2})$. Assume that Φ has full row rank (we shall justify this assumption below). Fix Φ (by fixing O). Then $\Phi g \sim N(0, \Phi\Phi^T)$. It is known that [64, p. 146]

$$d_{TV}(N(0,\Phi\Phi^T), N(0,I_k)) \le \sqrt{\operatorname{tr}(\Phi\Phi^T) - k - \ln\det(\Phi\Phi^T)}.$$

Write $\Phi\Phi^T = I + P$. Define an event $E = \{O : \|P\|_F^2 \le \frac{12}{\zeta}\cdot\frac{k^2}{n}\}$. When E happens, the eigenvalues of P are bounded by $\sqrt{\frac{12}{\zeta}}\cdot\frac{k}{\sqrt{n}} \le \frac{2}{3}$. Let $\mu_1, \ldots, \mu_k$ be the eigenvalues of P, then $\lambda_i(\Phi\Phi^T) = 1 + \mu_i$ with $|\mu_i| \le \frac{2}{3}$. Hence

$$\begin{aligned} d_{TV}(N(0,\Phi\Phi^T), N(0,I_k)) &\le \sqrt{\sum_{i=1}^k (\mu_i - \ln(1+\mu_i))} \le \sqrt{\sum_{i=1}^k \mu_i^2} = \sqrt{\|P\|_F^2} \\ &\le \sqrt{\frac{12}{\zeta}}\cdot\frac{k}{\sqrt{n}} \le \frac{2}{3}\zeta, \end{aligned}$$

where we use the fact that $x - \ln(1+x) \le x^2$ for $x \ge -2/3$. Therefore, when E happens, Φ is of full rank, and we can apply the total variation bound above. We claim that $\mathbb{E}\, P_{ij}^2 \le 4/n$ for all i,j and thus $\mathbb{E}\,\|P\|_F^2 \le 4k^2/n$; it then follows from Markov's inequality that $\Pr(E) \ge 1 - \zeta/3$ and thus

$$d_{TV}\left(\mathcal{D}_{n,k}, N(0,I_k)\right) \le \frac{2}{3}\zeta + \Pr(E^c) \le \frac{2}{3}\zeta + \frac{1}{3}\zeta = \zeta$$

as stated.

Now we show that $\mathbb{E}\, P_{ij}^2 \le 4/n$ for all i,j. Suppose that $O = (o_{ij})$. Notice that the rth row of Φ is

$$A_{i\ell}^{(r)} + \sum_j B_{ij}^{(r)} o_{\ell j}, \quad i = 1, \ldots, n, \quad \ell = 1, \ldots, \frac{n}{2}.$$

Hence by a straightforward calculation, the inner product of the rth and sth rows is

$$\begin{aligned} \langle \Phi_{r\cdot}, \Phi_{s\cdot}\rangle &= \delta_{rs} + \sum_{i,j,\ell} A_{i\ell}^{(r)} B_{ij}^{(s)} o_{\ell j} + \sum_{i,j,\ell} A_{i\ell}^{(s)} B_{ij}^{(r)} o_{\ell j} \\ &= \delta_{rs} + \sum_{j,\ell}\left(\langle A_\ell^{(r)}, B_j^{(s)}\rangle + \langle A_\ell^{(s)}, B_j^{(r)}\rangle\right) o_{\ell j}, \end{aligned}$$

where $A_i^{(r)}$ denotes the ith column of $A^{(r)}$. Then

$$P_{rs} = \operatorname{tr}(UO),$$

where the matrix U is defined by

$$u_{j\ell} = \langle A_\ell^{(r)}, B_j^{(s)}\rangle + \langle A_\ell^{(s)}, B_j^{(r)}\rangle.$$

Since

$$u_{jk}^2 \le 2\left\{\left(\sum_i |A_{ik}^{(r)}|^2\right)\left(\sum_i |B_{ij}^{(s)}|^2\right) + \left(\sum_i |A_{ik}^{(s)}|^2\right)\left(\sum_i |B_{ij}^{(r)}|^2\right)\right\}$$

and thus

$$\begin{aligned}\|U\|_F^2 &\le 2\sum_{j,k}\left\{\left(\sum_i |A_{ik}^{(r)}|^2\right)\left(\sum_i |B_{ij}^{(s)}|^2\right)+\left(\sum_i |A_{ik}^{(s)}|^2\right)\left(\sum_i |B_{ij}^{(r)}|^2\right)\right\}\\ &\le 2\left(\|A^{(r)}\|_F^2\|B^{(s)}\|_F^2+\|A^{(s)}\|_F^2\|B^{(r)}\|_F^2\right)\le 2,\end{aligned}$$

we conclude that

$$\mathbb{E}[P_{rs}^2]=\sum_{j,k}u_{jk}^2\,\mathbb{E}\,o_{kj}^2+\sum_{(j,k)\neq(i,\ell)}u_{jk}u_{i\ell}\,\mathbb{E}(o_{kj}o_{i\ell})=\frac{2}{n}\sum_{j,k}u_{jk}^2=\frac{2\|U\|_F^2}{n}\le\frac{4}{n}.$$

This completes the proof. □

Let $X\sim\mathcal{G}(n,n)$ be independent of $Y=\begin{pmatrix}G & GO\end{pmatrix}$, where $G\sim\mathcal{G}(n,n/2)$ and $O\sim O_{n/2}$ are independent. It is clear that w.h.p., $\operatorname{rank}(X)=n$ and $\operatorname{rank}(Y)=n/2$, which differ by a constant factor. The preceding theorem immediately yields the following corollary.

Corollary 5.2 (rank). *There exists an absolute constant $c>0$ such that any sketching algorithm that estimates $\operatorname{rank}(X)$ for $X\in\mathbb{R}^{n\times n}$ within a factor of $1+c$ with error probability $\le 1/6$ requires sketching dimension $\Omega(\sqrt{n})$.*

We remark that this hard instance also gives an $\Omega(\sqrt{n})$ lower bound for all $p>0$ and $p\neq 2$.

6. Bilinear sketch lower bound for rank ($p=0$).

Let $S\subset[n]\times[n]$ be a set of indices of an $n\times n$ matrix. For a distribution $\mathcal{L}$ over $\mathbb{R}^{n\times n}$, the entries of S induce a marginal distribution $\mathcal{L}(S)$ on $\mathbb{R}^{|S|}$ as

$$(X_{p_1,q_1},X_{p_2,q_2},\ldots,X_{p_{|S|},q_{|S|}}),\quad X\sim\mathcal{L}.$$

Theorem 6.1. *Let $U,V\sim\mathcal{G}(n,d)$, and let $G\sim\gamma\mathcal{G}(n,n)$ for $\gamma=n^{-14}$. Consider two distributions $\mathcal{L}_1$ and $\mathcal{L}_2$ over $\mathbb{R}^{n\times n}$ defined by UV^T and UV^T+G, respectively. Let $S\subset[n]\times[n]$. When $|S|\le d^2$, it holds that*

$$d_{TV}(\mathcal{L}_1(S),\mathcal{L}_2(S))\le C|S|\left(n^{-2}+dc^d\right), \tag{6.1}$$

where $C>0$ and $0<c<1$ are absolute constants.

Proof. (Sketch; see the next subsection for the full proof.) We give an algorithm which gives a bijection $f:\mathbb{R}^{|S|}\to\mathbb{R}^{|S|}$ with the property that for all but a subset of $\mathbb{R}^{|S|}$ of measure $o(1)$ under both $\mathcal{L}_1(S)$ and $\mathcal{L}_2(S)$, the probability density functions of the two distributions are equal up to a multiplicative factor of $(1\pm 1/\operatorname{poly}(n))$. The idea is to start with the row vectors $U_1,\ldots,U_n$ of U and $V_1,\ldots,V_n$ of V, and to iteratively perturb them by adding $\gamma G_{i,j}$ to UV^T for each $(i,j)\in S$. We find new vectors $U_1',\ldots,U_n'$ and $V_1',\ldots,V_n'$ of $n\times d$ matrices U' and V' so that $(U')(V')^T$ and $UV^T+\gamma G$ are equal on S. We do this in a way such that $\|U_i\|_2=(1\pm 1/\operatorname{poly}(n))\|U_i'\|_2$ and $\|V_i\|_2=(1\pm 1/\operatorname{poly}(n))\|V_i'\|_2$ for all i, and so the marginal density function evaluated on U_i (or V_j) is close to that evaluated on U_i' (or V_j'), by definition. Moreover, our mapping is bijective, so the joint distribution of $(U_1',\ldots,U_n',V_1',\ldots,V_n')$ is the same as that evaluated on $(U_1,\ldots,U_n,V_1,\ldots,V_n)$ up to a $(1\pm 1/\operatorname{poly}(n))$-factor. The bijection we create depends on properties of S; e.g., if the entry $(UV^T)_{i,j}=\langle U_i,V_j\rangle$ is perturbed, and more than d entries of the ith row of A appear in S, this places more than d constraints on U_i, but U_i is only d-dimensional. Thus, we must also change some of the vectors V_j. We change those V_j for which $(i,j) \in S$, and there are fewer than d rows $i' \neq i$ for which $(i',j) \in S$; in this way there are fewer than d constraints on V_j, so it is not yet fixed. We can find enough V_j with this property by the assumption that $|S| \leq d^2$. ☐

In the theorem above, choose $d = n/2$ so that $\operatorname{rank}(UV^T) \leq n/2$, while $\operatorname{rank}(G) = n$ with probability 1. Note that both distributions are rotationally invariant, and so the lower bound on bilinear sketches follows immediately.

THEOREM 6.2. *Let $A \in \mathbb{R}^{n\times n}$ be an arbitrary matrix. Suppose that an algorithm takes a bilinear sketch SAT ($S \in \mathbb{R}^{s\times n}$ and $T \in \mathbb{R}^{n\times t}$) and computes Y with $(1-c)\operatorname{rank}(A) \leq Y \leq (1+c)\operatorname{rank}(A)$ with probability at least $3/4$, where $c \in (0,1/3)$ is a constant. It must hold that $st = \Omega(n^2)$.*

As an aside, we note that given that w.h.p. over $A \sim \mathcal{L}_2$ in Theorem 6.1 the matrix A requires modifying $\Theta(n^2)$ of its entries to reduce its rank to at most d if $d \leq n/2$, this implies that we obtain an $\Omega(d^2)$ bound on the nonadaptive query complexity of deciding if an $n \times n$ matrix is of rank at most d or ϵ-far from rank d (for constant ϵ), showing that an algorithm of Krauthgamer and Sasson is optimal [52].

6.1. Proof of Theorem 6.1. We shall overload the notation p for different meanings in this subsection, since this is the proof for a lower bound on rank estimation, independent of the parameter p in the Schatten p-norm.

Proof. Suppose that $|S| = k$ and $S = \{(p_i,q_i)\}_{i=1}^k$. By symmetry, without loss of generality, we can assume that S does not contain a pair of symmetric entries. Throughout this proof, we rewrite $\mathcal{L}_1(S)$ as $\mathcal{L}_1$ and $\mathcal{L}_2(S)$ as $\mathcal{L}_2$. Now, using the new notation, let us denote the marginal distribution of $\mathcal{L}_2$ with fixed G by $\mathcal{L}_2|G$. We also denote the canonical Borel algebra on $\mathbb{R}^k$ by $\mathcal{M}(\mathbb{R}^k)$.

Then

$$
\begin{aligned}
d_{TV}(\mathcal{L}_1,\mathcal{L}_2) &= \sup_{A\in\mathcal{M}(\mathbb{R}^k)} \left|\Pr_{\mathcal{L}_1}(A) - \Pr_{\mathcal{L}_2}(A)\right| \\
&= \sup_{A\in\mathcal{M}(\mathbb{R}^k)} \left|\Pr_{\mathcal{L}_1}(A) - \int_{\mathbb{R}^{n^2}} \Pr_{\mathcal{L}_2}(A|G)p(G)dG\right| \\
&\leq \sup_{A\in\mathcal{M}(\mathbb{R}^k)} \int_{\mathbb{R}^{n^2}} \left|\Pr_{\mathcal{L}_1}(A) - \Pr_{\mathcal{L}_2}(A|G)\right| p(G)dG \\
&\leq \sup_{A\in\mathcal{M}(\mathbb{R}^k)} \left(\int_{F(\delta)} \left|\Pr_{\mathcal{L}_1}(A) - \Pr_{\mathcal{L}_2}(A|G)\right| p(G)dG \right. \\
&\qquad\qquad \left. + \int_{F(\delta)^c} \left|\Pr_{\mathcal{L}_1}(A) - \Pr_{\mathcal{L}_2}(A|G)\right| p(G)dG\right) \\
&\leq \sup_{G\in F(\delta)} d_{TV}(\mathcal{L}_1,\mathcal{L}_2|G) + 2\Pr\{F(\delta)^c\},
\end{aligned}
\tag{6.2}
$$

where

$$F(\delta) = \{G \in \mathbb{R}^{n\times n} : |G_{p_i,q_i}| \leq \delta \ \forall i = 1,\ldots,k\},$$

and $\Pr\{F(\delta)^c\}$ is the probability of the complement of $F(\delta)$ under the distribution on G, and we choose $\delta = n^{1/4}\gamma$. Recalling the probability density function of a Gaussian random variable and that $k \leq n^2$, it follows from a union bound that

$$\Pr\{F(\delta)^c\} \leq ke^{-\delta^2/(2\gamma^2)} = ke^{-n^{1/2}/2} \leq n^{-3}. \tag{6.3}$$

Now we examine $d_{TV}(\mathcal{L}_1, \mathcal{L}_2|G)$ with $G \in F(\delta)$. For notational convenience, let $\xi = (\xi_1, \ldots, \xi_k)^T = (G_{p_1,q_1}, \ldots, G_{p_k,q_k})^T$, and define $\xi^{(i)} = (\xi_1, \ldots, \xi_i, 0, \ldots, 0)^T$. Applying the triangle inequality to a telescoping sum, we obtain

$$\begin{aligned} d_{TV}(\mathcal{L}_1, \mathcal{L}_2|G) &= \sup_{A \in \mathcal{M}(\mathbb{R}^k)} \left| \Pr_{\mathcal{L}_1}(A) - \Pr_{\mathcal{L}_2|G}(A) \right| \\ &= \sup_{A \in \mathcal{M}(\mathbb{R}^k)} \left| \Pr_{\mathcal{L}_1}(A) - \Pr_{\mathcal{L}_1}(A - \xi) \right| \qquad (6.4) \\ &\le \sup_{A \in \mathcal{M}(\mathbb{R}^k)} \sum_{i=1}^{k} \left| \Pr_{\mathcal{L}_1}(A - \xi^{(i-1)}) - \Pr_{\mathcal{L}_1}(A - \xi^{(i)}) \right|. \end{aligned}$$

To bound (6.4), we need a way of bounding $|\Pr_{\mathcal{L}_1}(A) - \Pr_{\mathcal{L}_1}(A - te_i)|$ for a value t with $|t| \le \delta$. In this case, we say that we *perturb* a single entry $(p, q) := (p_i, q_i)$ of UV^T by t while fixing the remaining $k - 1$ entries. We claim that there exists a mapping $T_t : \mathbb{R}^{n \times d} \times \mathbb{R}^{n \times d} \to \mathbb{R}^{n \times d} \times \mathbb{R}^{n \times d}$ with $(U, V) \mapsto (U', V')$ for which the following three properties hold:

1. $(U'(V')^T)_{pq} = (UV)_{pq} + t$, and for all $(p', q') \in S \setminus \{(p, q)\}$ it holds that $(U'(V')^T)_{p'q'} = (UV^T)_{p'q'}$.
2. $\|U - U'\|_F \le t'$, $\|V - V'\|_F \le t'$ with probability $1 - \mathcal{O}(1/n^2 + dc^d)$, over the randomness of U and V. When this holds, we say that U and V are *good*; otherwise, we say that they are *bad*.
3. $T_{-t} \circ T_t = \text{id}$.

The last property shows that T_t is a bijection. We defer the construction of T_t to the end of the proof and now examine the implications on the total variation distance. Define

$$\begin{aligned} E(x) &= \{(U, V) : UV^T|_S = x\}, \\ E_{\text{good}}(x) &= \{(U, V) \in E(x) : (U, V) \text{ is good}\}, \\ E_{\text{bad}}(x) &= \{(U, V) \in E(x) : (U, V) \text{ is bad}\}. \end{aligned}$$

Then, using these three properties about T_t, as well as the triangle inequality, and letting $p(U), p(V)$ be the p.d.f.'s of U and V so that

$$p(U) = \frac{1}{(2\pi)^{nd/2}} \exp\left(-\frac{\|U\|_F^2}{2}\right), \quad p(V) = \frac{1}{(2\pi)^{nd/2}} \exp\left(-\frac{\|V\|_F^2}{2}\right), \qquad (6.5)$$

we have that

$$\begin{aligned}&\left|\Pr_{\mathcal{L}_1}(A) - \Pr_{\mathcal{L}_1}(A - te_i)\right| \\ &= \left|\int_{E(A)} p(U)p(V)dUdV - \int_{E(A-te_i)} p(U)p(V)dUdV\right| \\ &\leq \left|\int_{E_{\text{good}}(A)} p(U)p(V)dUdV - \int_{E_{\text{good}}(A)} p(U')p(V')dUdV\right| \\ &\quad + \int_{E_{\text{bad}}(A)} p(U)p(V)dUdV + \int_{E_{\text{bad}}(A)} p(U')p(V')dUdV \\ &\leq \int_{E_{\text{good}}(A)} |p(U) - p(U')|p(V)dUdV \\ &\quad + \int_{E_{\text{good}}(A)} p(U')|p(V) - p(V')|dUdV + \mathcal{O}\left(\frac{1}{n^2} + dc^d\right). \end{aligned} \tag{6.6}$$

Using (6.5), we obtain

$$|p(U) - p(U')| = p(U) \cdot \left|1 - \exp\left(\frac{\|U\|_F^2 - \|U'\|_F^2}{2}\right)\right|. \tag{6.7}$$

Notice that $\|U\|_F^2 \sim \chi^2(nd)$, and so by a tail bound for the χ^2-distribution [53, Lemma 1], $\|U\|_F^2 \leq 6nd - t'$ (recall that $t' = 1/n^4$) with probability at least $1 - e^{-nd} \geq 1 - n^{-3}$. When this happens, for good U it follows from the triangle inequality, the second property of T_t above, and the fact that $t' = 1/n^4$ that

$$\begin{aligned}\left|\|U\|_F^2 - \|U'\|_F^2\right| &= (\|U\|_F + \|U'\|_F)\left|\|U\|_F - \|U'\|_F\right| \\ &\leq (2\|U\|_F + \|U - U'\|_F)\,\|U - U'\|_F \leq 2\sqrt{6nd} \cdot t' \leq 6n^{-3}.\end{aligned}$$

Using $|1 - e^{|x|}| \leq 2|x|$ for $|x| < 1$ and combining with (6.7), we have

$$|p(U) - p(U')| \leq p(U) \cdot 12n^{-3}. \tag{6.8}$$

Similarly, it holds that $\|V\|_F^2 \leq 6nd - t'$ with probability $\geq 1 - n^{-3}$, and when this happens,

$$|p(V) - p(V')| \leq p(V) \cdot 12n^{-3}.$$

It then follows that

$$\int_{E_{\text{good}}(A)} |p(U) - p(U')|p(V)dUdV = \mathcal{O}\left(\frac{1}{n^3}\right), \tag{6.9}$$

$$\int_{E_{\text{good}}(A)} p(U')|p(V) - p(V')|dUdV = \mathcal{O}\left(\frac{1}{n^3}\right). \tag{6.10}$$

Plugging (6.9) and (6.10) into (6.6) yields that

$$\left|\Pr_{\mathcal{L}_1}(A) - \Pr_{\mathcal{L}_1}(A - te_i)\right| = \mathcal{O}\left(\frac{1}{n^2} + dc^d\right), \tag{6.11}$$

which, combined with (6.4), (6.2), and (6.3), finally leads to

$$d_{TV}(\mathcal{L}_1, \mathcal{L}_2) \leq d_{TV}(\mathcal{L}_1, \mathcal{L}_2 | G) + 2\Pr\{F(\delta)^c\} = \mathcal{O}\left(k\left(\frac{1}{n^2} + dc^d\right)\right).$$

Construction of T_t. Now we construct T_t. Suppose the entry to be perturbed is (p,q).

Case 1a. Suppose that the pth row contains $s \leq d$ entries read, say at columns $q_1,\dots,q_s$. Without loss of generality, we can assume that $s = d$, as we can preserve the entries in S, perturbing one of them, and preserve $d-s$ arbitrary additional entries in the pth row.

Then we have (U_i denotes the ith row of U, and V_i denotes the ith column of V)

$$U_i \begin{pmatrix} V_{q_1} & V_{q_2} & \cdots & V_{q_d} \end{pmatrix} = x,$$

and we want to construct U_i' such that

$$U_i' \begin{pmatrix} V_{q_1} & V_{q_2} & \cdots & V_{q_d} \end{pmatrix} = x + \Delta x.$$

Hence **property 1** automatically holds, and

$$(U_i' - U_i) \begin{pmatrix} V_{q_1} & V_{q_2} & \cdots & V_{q_d} \end{pmatrix} = \Delta x. \tag{6.12}$$

With probability 1, the matrix $\tilde{V} := \begin{pmatrix} V_{q_1} & V_{q_2} & \cdots & V_{q_d} \end{pmatrix}$ has rank d. Hence we can solve $U_i' - U_i$ uniquely, and

$$\|U_i' - U_i\|_2 \leq \frac{\|\Delta x\|_2}{\sigma_d(\tilde{V})} \leq \frac{\delta}{\sigma_d(\tilde{V})}.$$

Using the tail bound on the minimum singular value given in [65], we have $\Pr\{\sigma_d(\tilde{V}) \leq \epsilon/\sqrt{d}\} \leq C\epsilon + c^d$, where $C > 0$ and $0 < c < 1$ are absolute constants. Choosing $\epsilon = n^{-4}$ and recalling that $d \leq n$, we see that with probability at least $1 - Cn^{-4} - c^d$, it holds that $\sigma_d(\tilde{V}) \geq 1/n^{9/2}$ and thus that

$$\|U_i' - U_i\|_2 \leq n^{9/2}\delta = n^{9/2+1/4}\gamma \leq n^{-5}.$$

This proves **property 2** of this case. This step is invertible, because if we replace Δx with $-\Delta x$ in (6.12), the solution $U_i' - U_i$ will be of the opposite sign, too. This shows **property 3** of this case.

Case 1b. Suppose that the qth column contains $s \leq d$ entries read. Similarly to Case 1a, we have $U_i' = U_i$ and $\|V_i' - V_i\|_2 \leq n^{-5}$ with probability $\geq 1 - Cn^{-4} - c^d$. The invertibility is similar to that in Case 1a and therefore holds.

Case 2. Suppose that there are more than d entries read in both the pth row and the qth column. Define

$$J = \{i \in [n] : i\text{th column has} \leq d \text{ entries contained in } S\},$$
$$\mathsf{Col}_r = \{i \in [n] : (r,i) \in S\}, \quad \mathsf{Row}_c = \{i \in [n] : (i,c) \in S\}.$$

Call the columns with index in J good columns and those with index in J^c bad columns.

Note that $|J^c| \leq d$ since the total number of entries in S is at most d^2. Take the columns in $J^c \cap \mathsf{Col}_p$, and note that $q \in J^c \cap \mathsf{Col}_p$. As in Case 1a, we can change U_p to U_p' such that $U'V^T$ agrees with the perturbed entry of UV^T at (p,q) and keeps the entries of UV^T the same for all (p,q') for $q' \in (J^c \cap \mathsf{Col}_p) \setminus \{q\}$, since $|(J^c \cap \mathsf{Col}_p) \setminus \{q\}| \leq d - 1$.

However, this new choice of U' possibly causes $(U'V^T)_{p,i} \neq (UV^T)_{p,i}$ for $i \in \mathsf{Col}_p \cap J$. For each $i \in \mathsf{Col}_p \cap J$, we also need to change V_i to a vector V_i' without affecting the entries read in any bad column. Now for each good column, the matrix $\tilde{U}$ used in Case 1b applied to each i in $\mathsf{Col}_p \cap J$ is no longer i.i.d. Gaussian because one row of $\tilde{V}$ has been changed, and this change has ℓ_2-norm at most n^{-5} (the guarantee on property 2 in Case 1b) with probability at least $1 - 4n^{-3} - c^d$, and since the minimum singular value is a 1-Lipschitz function of matrices, the minimum singular value is perturbed by at most n^{-5}. Hence for each good column i, with probability at least $1-4n^{-3}-c^d$, we have $\|V_i - V_i'\|_2 \le n^{-5}$. Since there are at most d good columns, by a union bound, with probability at least $1-4/n^2-dc^d$ we have $\|V-V'\|_F \le n^{-4}$. **This concludes the proof of properties 1 and 2 in Case 2**.

The final step is to verify **property 3**, i.e., that this step is invertible in this case. Suppose that $T_t(U,V) = (U',V')$ and $T_{-t}(U',V') = (U'',V'')$; we want to show that $U'' = U$ and $V'' = V$. Observe that $V_i = (V')_i = (V'')_i$ for $i \in J^c \cap \mathsf{Col}_p = \{q_1, \ldots, q_d\}$; we have

$$\begin{aligned} (U_p' - U_p)\begin{pmatrix} V_{q_1} & V_{q_2} & \cdots & V_{q_d}\end{pmatrix} &= \Delta x, \\ (U_p'' - U_p')\begin{pmatrix} V_{q_1} & V_{q_2} & \cdots & V_{q_d}\end{pmatrix} &= -\Delta x. \end{aligned}$$

Summing the two equations yields

$$(U_p'' - U_p)\begin{pmatrix} V_{q_1} & V_{q_2} & \cdots & V_{q_d}\end{pmatrix} = 0,$$

and thus $U_p'' = U_p$ provided that $(V_{q_1}, V_{q_2}, \ldots, V_{q_d})$ is invertible. Since $U_i = U_i' = U_i''$ for all $i \neq p$, we have $U = U''$. Next we show that $V_i'' = V_i$ for each $i \in \mathsf{Col}_p \cap J$. Suppose that $\mathsf{Row}_i = \{p_1, \ldots, p_d\} \ni p$. Similarly to the above, we have that

$$\begin{pmatrix} U'_{p_1} \\ \vdots \\ U'_p \\ \vdots \\ U'_{p_d} \end{pmatrix} (V_i' - V_i) = \begin{pmatrix} 0 \\ \vdots \\ -(U_p' - U_p)V_i \\ \vdots \\ 0 \end{pmatrix}$$

and

$$\begin{pmatrix} U''_{p_1} \\ \vdots \\ U''_p \\ \vdots \\ U''_{p_d} \end{pmatrix} (V_i'' - V_i') = \begin{pmatrix} 0 \\ \vdots \\ -(U_p'' - U_p')V_i' \\ \vdots \\ 0 \end{pmatrix}.$$

Summing the two equations and recalling that $U_p'' = U_p$ and $U_i = U_i' = U_i''$ for all $i \neq p$, we obtain that

$$\begin{pmatrix} U_{p_1} \\ \vdots \\ U_p \\ \vdots \\ U_{p_d} \end{pmatrix} (V_i'' - V_i) + \begin{pmatrix} 0 \\ \vdots \\ U_p' - U_p \\ \vdots \\ 0 \end{pmatrix} (V_i' - V_i) = \begin{pmatrix} 0 \\ \vdots \\ (U_p' - U_p)(V_i' - V_i) \\ \vdots \\ 0 \end{pmatrix},$$

i.e.,

$$\begin{pmatrix} U_{p_1} \\ \vdots \\ U_p \\ \vdots \\ U_{p_d} \end{pmatrix} (V_i'' - V_i) = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

whence it follows immediately that $V_i'' = V_i$ provided that $(U_{p_1}^T, \dots, U_{p_d}^T)$ is invertible. Together with $V_i = V_i' = V_i''$ for all $i \in \mathsf{Col}_p \cap J^c$, we conclude that $V'' = V$. ∎

7. Reduction to square matrices for the upper bound.

Before presenting our bilinear sketching algorithms for the Schatten p-norm, we make a reduction from general matrices to square matrices based on the Johnson–Lindenstrauss Transform, and henceforth we assume that the input matrices are square in our bilinear sketching algorithms.

Suppose that $A \in \mathbb{R}^{n\times d}$ $(n > d)$. When $n = \mathcal{O}(d/\epsilon^2)$, let $\tilde{A} = \begin{pmatrix} A & 0 \end{pmatrix}$ with zero padding so that $\tilde{A}$ is a square matrix of dimension n. Then $\|\tilde{A}\|_p = \|A\|_p$ for all $p > 0$. Otherwise, we can sketch the matrix with $\mathcal{O}(d/\epsilon^2)$ rows while roughly maintaining the singular values as follows. Call Φ a (d, δ)-subspace embedding matrix if with probability $\geq 1 - \delta$ it holds that

$$(1 - \epsilon)\|x\|_2 \leq \|\Phi x\|_2 \leq (1 + \epsilon)\|x\|_2$$

for all x in a fixed d-dimensional subspace. In [66], the following lemma is proved.

Lemma 7.1. *Suppose that $H \subset \mathbb{R}^n$ is a d-dimensional subspace. Let Φ be an r-by-n random matrix with i.i.d. $N(0, 1/r)$ entries, where $r = \Theta(d/\epsilon^2 \log(1/\delta))$. Then it holds with probability $\geq 1 - \delta$ that*

$$(1 - \epsilon)\|x\|_2 \leq \|\Phi x\|_2 \leq (1 + \epsilon)\|x\|_2 \quad \forall x \in H.$$

In fact we can use more modern subspace embeddings [66, 23, 59, 61] to improve the time complexity, though since our focus is on the sketching dimension, we defer a thorough study of the time complexity to future work.

Now we are ready to show the subspace embedding transform on singular values, which follows from the min-max principle for singular values.

Lemma 7.2. *Let Φ be a (d, δ)-subspace embedding matrix. Then, with probability $\geq 1 - \delta$, it holds that $(1 - \epsilon)\sigma_i(\Phi A) \leq \sigma_i(A) \leq (1 + \epsilon)\sigma_i(\Phi A)$ for all $1 \leq i \leq d$.*

Proof. The min-max principle for singular values says that

$$\sigma_i(A) = \max_{S_i} \min_{\substack{x \in S_i \\ \|x\|_2 = 1}} \|Ax\|_2,$$

where S_i runs through all i-dimensional subspace. Observe that the range of A is a subspace of dimension at most d. It follows from Lemma 7.1 that with probability $\geq 1 - \delta$,

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

The claimed result follows immediately from the min-max principle for singular values. ∎

Algorithm 8.1. The sketching algorithm for odd $p \geq 3$.

```
Input: $n$, $\epsilon > 0$, odd integer $p \geq 3$, and PSD $A \in \mathbb{R}^{n\times n}$
$N \leftarrow \Omega(\epsilon^{-2})$
Let $\{G_i\}$ be independent $n^{1-2/p} \times n$ matrices with i.i.d. $N(0,1)$ entries
Maintain each $G_i A G_i^T$, $i = 1, \ldots, N$
Compute $Z$ as defined in (8.1)
return $Z$
```

Let $\tilde{A} = \begin{pmatrix}\Phi A & 0\end{pmatrix}$ with zero padding so that $\tilde{A}$ is a square matrix of dimension $\mathcal{O}(d/\epsilon^2)$. Then by the preceding lemma, with probability $\geq 1-\delta$, $\|\tilde{A}\|_p = \|\Phi A\|_p$ is a $(1 \pm \epsilon)$-approximation of $\|A\|_p$ for all $p > 0$. Therefore we have reduced the problem to the case of square matrices.

8. Bilinear sketch algorithms. We present two sketching algorithms to compute a $(1 \pm \epsilon)$-approximation of $\|A\|_p^p$ for $A \in \mathbb{R}^{n\times n}$ using linear sketches, which can thus be implemented in the most general turnstile data stream model (an arbitrary number of positive and negative additive updates to entries given in an arbitrary order). The first algorithm works for PSD matrices A when $p \geq 3$ is an odd integer, and the second algorithm works for arbitrary A when $p \geq 4$ is an even integer.

8.1. PSD matrices and odd *p*. In this subsection, we assume that $A \in \mathbb{R}^{n\times n}$ is PSD and $p \geq 3$ is an odd integer.

Given integers k and $p < k$, call a sequence $(i_1, \ldots, i_p)$ a cycle if $i_j \in [k]$ for all j and $i_{j_1} \neq i_{j_2}$ for all $j_1 \neq j_2$. On a $k \times k$ matrix A, each cycle σ defines a product

$$A_\sigma = \prod_{i=1}^{p} A_{\sigma_i, \sigma_{i+1}},$$

where we adopt the convention that $i_{p+1} = i_1$. Let Cyc denote the set of cycles. Call two cycles $\sigma, \tau \in \mathsf{Cyc}$ k-disjoint if $|\sigma \Delta \tau| = 2k$, where σ and τ are viewed as multisets.

THEOREM 8.1. *With probability* $\geq 3/4$*, the output* X *returned by Algorithm* 8.1 *satisfies* $(1-\epsilon)\|A\|_p^p \leq X \leq (1+\epsilon)\|A\|_p^p$ *when* A *is PSD. The algorithm is a bilinear sketch with* $r \cdot s = \mathcal{O}_p(\epsilon^{-2} n^{2-4/p})$.

Proof. Since A is symmetric, it can be written as $A = O \Lambda O^T$, where Λ is a diagonal matrix and O is an orthogonal matrix. Let G be a random matrix with i.i.d. $N(0,1)$ entries. By rotational invariance, GAG^T is identically distributed as $G\Lambda G^T$. Let $k = n^{1-2/p}$, and let $\tilde{A}$ be the upper-left $k \times k$ block of $G \Lambda G^T$. It is clear that

$$\tilde{A}_{s,t} = \sum_{i=1}^{n} \lambda_i G_{i,s} G_{i,t}.$$

Define

$$Y = \frac{1}{|\mathsf{Cyc}|} \sum_{\sigma \in \mathsf{Cyc}} \tilde{A}_\sigma.$$

Suppose that $\sigma = (i_1, \ldots, i_p)$; then

$$\tilde{A}_\sigma = \sum_{j_1,\ldots,j_p=1}^{n} \lambda_{j_1} \cdots \lambda_{j_p} \prod_{\ell=1}^{p} G_{j_\ell, i_\ell} G_{j_\ell, i_{\ell+1}}.$$

It is easy to see that $\mathbb{E}\,\tilde{A}_\sigma = \sum \lambda_i^p = \|A\|_p^p$ (all j_ℓ's are the same) and thus that $\mathbb{E}\,Y = \|X\|_p^p$. Now we compute $\mathbb{E}\,Y^2$:

$$\mathbb{E}\,Y^2 = \frac{1}{|\mathsf{Cyc}|^2}\sum_{m=0}^{p}\sum_{\substack{\sigma,\tau\in\mathsf{Cyc}\\ \sigma,\tau\text{ are }m\text{-disjoint}}}\mathbb{E}(\tilde{A}_\sigma\tilde{A}_\tau).$$

It is not difficult to see that (see subsection 8.3 for details), when $m \le p-2$, it holds for m-disjoint σ and τ that

$$\mathbb{E}(\tilde{X}_\sigma\tilde{X}_\tau) \lesssim_{m,p} \|A\|_\kappa^{\kappa n_1}\|A\|_{\kappa+2}^{(\kappa+2)n_2},$$

where $\kappa = 2\lceil\frac{p}{2(p-m)}\rceil$ and (n_1,n_2) is the solution to

$$\begin{gathered}\kappa n_1 + (\kappa+2)n_2 = 2p,\\ n_1 + n_2 = p-m.\end{gathered}$$

By Hölder's inequality, $\|A\|_q^q \le n^{1-\frac{q}{p}}\|A\|_p^q$ for $q<p$, and thus

$$\mathbb{E}(\tilde{A}_\sigma\tilde{A}_\tau) \lesssim_{m,p} n^{p-m-2}, \quad m \le p-2.$$

When σ, τ are $(p-1)$- or p-disjoint, we obtain

$$\mathbb{E}(\tilde{A}_\sigma\tilde{A}_\tau) \le \|A\|_p^{2p}.$$

There are $\mathcal{O}_p(k^{p+m})$ pairs of m-disjoint cycles, and $|\mathsf{Cyc}| = \Theta(k^p)$,

$$\begin{aligned}
\frac{1}{|\mathsf{Cyc}|^2}\sum_{\substack{\sigma,\tau\in\mathsf{Cyc}\\ |\sigma\Delta\tau|=2m}}\mathbb{E}(\tilde{A}_\sigma\tilde{A}_\tau) &\le C_{m,p}\frac{n^{p-m-2}}{k^{p-m}}\|A\|_p^{2p} \le C_{m,p}\|A\|_p^{2p}, \quad m \le p-2,\\
\frac{1}{|\mathsf{Cyc}|^2}\sum_{\substack{\sigma,\tau\in\mathsf{Cyc}\\ |\sigma\Delta\tau|=2m}}\mathbb{E}(\tilde{A}_\sigma\tilde{A}_\tau) &\le \frac{1}{k}\|A\|_p^{2p}, \quad m = p-1,\\
\frac{1}{|\mathsf{Cyc}|^2}\sum_{\substack{\sigma,\tau\in\mathsf{Cyc}\\ |\sigma\Delta\tau|=2m}}\mathbb{E}(\tilde{A}_\sigma\tilde{A}_\tau) &\le \|A\|_p^{2p}, \quad m = p.
\end{aligned}$$

Therefore $\mathbb{E}\,Y^2 \le C_p\|A\|_p^{2p}$ for some constant C_p dependent on p only. Hence if we take multiple copies of this distribution as stated in Algorithm 8.1 and define

$$Z = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathsf{Cyc}|}\sum_{\sigma\in\mathsf{Cyc}}(G_iAG_i^T)_\sigma =: \frac{1}{N}\sum_{i=1}^{N}Y_i, \tag{8.1}$$

then Y_i's are i.i.d. copies of Y. It follows that

$$\mathbb{E}\,Z = \mathbb{E}\,Y = \|A\|_p^p$$

and

$$\mathrm{Var}(Z) = \frac{\mathrm{Var}(Y)}{N} \le \frac{\mathbb{E}\,Y^2}{N} = \frac{1}{4}\epsilon^2\|A\|_p^{2p}.$$

Algorithm 8.2. The sketching algorithm for even $p \geq 4$.

```
Input: $n$, $\epsilon > 0$, even integer $p \geq 4$, and $A \in \mathbb{R}^{n\times n}$
$N \leftarrow \Theta_p(\epsilon^{-2})$
Let $\{G_i\}_{i\in[N]}$ and $\{H_i\}_{i\in[N]}$ be independent $n^{1-2/p} \times n$ matrices with i.i.d. $N(0,1)$ entries
Maintain each $G_i A H_i^T$, $i = 1, \ldots, N$
Compute $Z$ as defined in (8.2)
return $Z$
```

Finally, by Chebyshev's inequality,

$$\Pr\left\{\left|Z - \|A\|_p^p\right| > \epsilon\|A\|_p^p\right\} \leq \frac{Var(Z)}{\epsilon^2\|A\|_p^{2p}} \leq \frac{1}{4},$$

which implies the correctness of the algorithm. It is easy to see that the algorithm only reads the upper-left $k \times k$ block of each $G_i A G_i^T$, and thus it can be maintained in $\mathcal{O}(Nk^2) = \mathcal{O}_p(\epsilon^{-2}n^{2-4/p})$ space. □

8.2. Arbitrary matrices and even *p*.

In this subsection, we assume that $p \geq 4$ is an even integer. We redefine the notion of cycles and the estimator below, using the same letters as in the preceding subsection. The algorithm and the analysis are similar to those in the preceding subsection.

Suppose that $p = 2q$. We define a cycle σ to be an ordered pair of a sequence of length q: $\sigma = ((i_1, \ldots, i_q), (j_1, \ldots, j_q))$ such that $i_r, j_r \in [k]$ for all r, $i_r \neq i_s$ and $j_r \neq j_s$ for $r \neq s$. Now we associate with σ

$$A_\sigma = \prod_{\ell=1}^{q} A_{i_\ell, j_\ell} A_{i_{\ell+1}, j_\ell},$$

where we adopt the convention that $i_{k+1} = i_1$.

For two sequences $i = (i_1, \ldots, i_q)$ and $i' = (i'_1, \ldots, i'_q)$, each having distinct elements, we denote by $i\Delta i'$ the symmetric difference between i and i', treating i and i' as multisets.

Let Cyc denote the set of cycles. We say that two cycles $\sigma = (\{i\}, \{j\})$ and $\tau = (\{i'\}, \{j'\})$ are (m_1, m_2)-disjoint if $|i\Delta i'| = 2m_1$ and $|j\Delta j'| = 2m_2$, denoted by $|\sigma\Delta\tau| = (m_1, m_2)$.

We define

$$Z = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{|\mathsf{Cyc}|} \sum_{\sigma\in\mathsf{Cyc}} (G_i A H_i^T)_\sigma \tag{8.2}$$

for even p, where $G_1, \ldots, G_N, H_1, \ldots, H_N$ are independent $n^{1-2/p} \times n$ Gaussian random matrices with entries i.i.d. $N(0,1)$.

THEOREM 8.2. *With probability* $\geq 3/4$*, the output* Z *returned by Algorithm* 8.2 *satisfies* $(1-\epsilon)\|A\|_p^p \leq Z \leq (1+\epsilon)\|A\|_p^p$ *when* p *is even. The algorithm is a bilinear sketch with* $r \cdot s = \mathcal{O}_p(\epsilon^{-2}n^{2-4/p})$.

Proof. Let $A = U\Sigma V$ be the SVD of A. Let G and H be random matrices with i.i.d. $N(0,1)$ entries. By rotational invariance, GAH^T is identically distributed as

$G\Sigma H^T$. Let $k = n^{1-2/p}$, and let $\tilde{A}$ be the upper-left $k \times k$ block of $G\Sigma H^T$. It is clear that[6]

$$\tilde{A}_{s,t} = \sum_{i=1}^{n} \sigma_i G_{i,s} H_{i,t}.$$

Define

$$Y = \frac{1}{|\mathsf{Cyc}|} \sum_{\sigma \in \mathsf{Cyc}} \tilde{A}_\sigma.$$

Suppose that $\sigma = (\{i_s\}, \{j_s\})$; then

$$\tilde{A}_\sigma = \sum_{\substack{\ell_1,\ldots,\ell_q \\ m_1,\ldots,m_q}} \prod_{s=1}^{q} \sigma_{\ell_s}\sigma_{m_s} \prod_{s=1}^{q} G_{\ell_s,i_s} H_{\ell_s,j_s} G_{m_s,i_{s+1}} H_{m_s,j_s}.$$

It is easy to see that $\mathbb{E}\,\tilde{X}_\sigma = \sum \sigma_i^{2q} = \|A\|_p^p$ (all $\{\ell_s\}$ and $\{m_s\}$ are the same) and thus that $\mathbb{E}\,Y = \|X\|_p^p$. Now we compute $\mathbb{E}\,Y^2$:

$$\mathbb{E}\,Y^2 = \frac{1}{|\mathsf{Cyc}|^2} \sum_{m_1=0}^{q} \sum_{m_2=0}^{q} \sum_{\substack{\sigma,\tau \in \mathsf{Cyc} \\ |\sigma \Delta \tau| = (m_1,m_2)}} \mathbb{E}(\tilde{A}_\sigma \tilde{A}_\tau).$$

Suppose that $|\sigma \Delta \tau| = (m_1, m_2)$,

$$\mathbb{E}(\tilde{A}_\sigma \tilde{A}_\tau) = \sum_{\substack{\ell_1,\ldots,\ell_q \\ \ell'_1,\ldots,\ell'_q \\ m_1,\ldots,m_q \\ m'_1,\ldots,m'_q}} \left(\prod_{i=1}^{q} \sigma_{\ell_i}\sigma_{m_i}\sigma_{\ell'_i}\sigma_{m'_i} \right),$$

$$\mathbb{E}\left\{ \prod_{s=1}^{q} G_{\ell_s,i_s} G_{m_s,i_{s+1}} G_{\ell'_s,i'_s} G_{m'_s,i'_{s+1}} \right\} \mathbb{E}\left\{ \prod_{s=1}^{q} H_{\ell_s,j_s} H_{m_s,j_s} H_{\ell'_s,j'_s} H_{m'_s,j'_s} \right\}.$$

An argument similar to that in the preceding section gives that

$$\mathbb{E}(\tilde{A}_\sigma \tilde{A}_\tau) \lesssim_{m,p} \begin{cases} \|A\|_\kappa^{\kappa n_1} \|A\|_{\kappa+2}^{(\kappa+2)n_2}, & m_1, m_2 \le q-1, \\ \|A\|_4^{4(q-m_2-1)} \|A\|_{2(m_2+1)}^{4(m_2+1)}, & m_1 = q, m_2 \le q-1, \\ \|A\|_4^{4(q-m_1-1)} \|A\|_{2(m_1+1)}^{4(m_1+1)}, & m_2 = q, m_1 \le q-1, \\ \|A\|_{2q}^{4q}, & m_1 = m_2 = q, \end{cases}$$

where $\kappa = 2\lceil \frac{p}{p-m} \rceil$ and (n_1, n_2) is the solution to

$$\kappa n_1 + (\kappa+2) n_2 = 2p,$$
$$n_1 + n_2 = p - m_1 - m_2.$$

By Hölder's inequality, $\|A\|_r^r \le n^{1-\frac{r}{p}} \|A\|_p^r$ for $r < p$. Thus,

$$\mathbb{E}(\tilde{A}_\sigma \tilde{A}_\tau) \lesssim_{m,p} \begin{cases} n^{p-m_1-m_2-2} \|A\|_p^{2p}, & m_1, m_2 \le q-1, \\ n^{q-m_1-1} \|A\|_p^{2p}, & p_2 = q, p_1 \le q-1, \\ n^{q-m_2-1} \|A\|_p^{2p}, & p_1 = q, p_2 \le q-1, \\ \|A\|_p^{2p}, & m_1 = m_2 = q. \end{cases}$$

[6] In this proof we use σ_i (with subscript) for singular values and use σ (without subscripts) for cycles.

There are $\mathcal{O}_q(k^{q+m_1}k^{q+m_2}) = \mathcal{O}_p(k^{p+m_1+m_2})$ pairs of (m_1, m_2)-disjoint cycles, and $|\mathsf{Cyc}| = \Theta(k^p)$,

$$\begin{aligned}
\frac{1}{|\mathsf{Cyc}|^2} \sum_{\substack{\sigma,\tau\in\mathsf{Cyc}\\ |\sigma\Delta\tau|=(m_1,m_2)}} \mathbb{E}(\tilde{A}_\sigma\tilde{A}_\tau) &\le C_{m_1+m_2,p}\frac{n^{p-m_1-m_2-2}}{k^{p-m_1-m_2}}\|A\|_p^{2p}\\
&\le C_{m_1+m_2,p}\|A\|_p^{2p}, \qquad m_1, m_2 \le q-1,\\
\frac{1}{|\mathsf{Cyc}|^2} \sum_{\substack{\sigma,\tau\in\mathsf{Cyc}\\ |\sigma\Delta\tau|=(m_1,m_2)}} \mathbb{E}(\tilde{A}_\sigma\tilde{A}_\tau) &\le C_{m_1+m_2,p}\frac{n^{q-m_1-1}}{k^{q-m_1}}\|A\|_p^{2p}\\
&\le C_{m_1,p}\|A\|_p^{2p}, \qquad \frac{p}{2} \le m_2 = q, m_1 \le q-1,\\
\frac{1}{|\mathsf{Cyc}|^2} \sum_{\substack{\sigma,\tau\in\mathsf{Cyc}\\ |\sigma\Delta\tau|=(m_1,m_2)}} \mathbb{E}(\tilde{A}_\sigma\tilde{A}_\tau) &\le C_{m_1+m_2,p}\frac{n^{p-m_2-1}}{k^{q-m_2}}\|A\|_p^{2p}\\
&\le C_{m_1,p}\|A\|_p^{2p}, \qquad \frac{p}{2} \le m_1 = q, m_2 \le q-1,\\
\frac{1}{|\mathsf{Cyc}|^2} \sum_{\substack{\sigma,\tau\in\mathsf{Cyc}\\ |\sigma\Delta\tau|=(m_1,m_2)}} \mathbb{E}(\tilde{A}_\sigma\tilde{A}_\tau) &\le \|A\|_p^{2p}, \qquad m_1 + m_2 = p.
\end{aligned}$$

Therefore $\mathbb{E}\,Y^2 \le C_p\|A\|_p^{2p}$ for some constant C_p dependent on p only. Since

$$Z = \frac{1}{N}\sum_{i=1}^{N} Y_i,$$

then Y_i's are i.i.d. copies of Y. It follows that

$$\mathbb{E}\,Z = \mathbb{E}\,Y = \|A\|_p^p$$

and

$$\mathrm{Var}(Z) = \frac{\mathrm{Var}(Y)}{N} \le \frac{\mathbb{E}\,Y^2}{N} = \frac{1}{4}\epsilon^2\|A\|_p^{2p}.$$

Finally, by Chebyshev's inequality,

$$\Pr\left\{\left|Z - \|A\|_p^p\right| > \epsilon\|A\|_p^p\right\} \le \frac{\mathrm{Var}(Z)}{\epsilon^2\|A\|_p^{2p}} \le \frac{1}{4},$$

which implies the correctness of the algorithm. It is easy to see that the algorithm only reads the upper-left $k \times k$ block of each $G_iAH_i^T$, and thus it can be maintained in $\mathcal{O}(Nk^2) = \mathcal{O}_p(\epsilon^{-2}n^{2-4/p})$ space. ∎

8.3. Omitted details in the proof of Theorem 8.1. Suppose that $\sigma = (i_1, \ldots, i_p)$ and $\tau = (j_1, \ldots, j_p)$ are m-disjoint. Then

$$\mathbb{E}(\tilde{A}_\sigma\tilde{A}_\tau) = \sum_{\substack{\ell_1,\ldots,\ell_p\\ \ell_1',\ldots,\ell_p'}} \lambda_{\ell_1}\cdots\lambda_{\ell_p}\lambda_{\ell_1'}\cdots\lambda_{\ell_p'}\,\mathbb{E}\left\{\prod_{s=1}^{p} G_{\ell_s,i_s}G_{\ell_s,i_{s+1}}G_{\ell_s',j_s}G_{\ell_s',j_{s+1}}\right\}.$$

For the expectation to be nonzero, each appearing entry must be repeated an even number of times. Hence, if some i_s appears only once among the indices $\{i_s\}$ and $\{j_s\}$, it must hold that $\ell_s = \ell_{s+1}$. Thus for each of the summation terms, the indices $\{\ell_s\}$ break into a few blocks, in each of which all ℓ_s take the same value; the same holds for $\{\ell'_s\}$. We also need to piece together the blocks of $\{\ell_s\}$ with those of $\{\ell'_s\}$. Hence the whole sum breaks into sums corresponding to different block configurations. For a certain kind of configuration, in which $\ell_1, \ldots, \ell_w$ are free variables with multiplicity $r_1, \ldots, r_w$, respectively, the sum is bounded by

$$C \cdot \sum_{\ell_1,\ldots,\ell_w} \lambda_{\ell_1}^{r_1} \cdots \lambda_{\ell_w}^{r_w} \leq C\|A\|_{r_1}^{r_1} \cdots \|A\|_{r_w}^{r_w},$$

where the constant C depends on the configuration only, and thus it can be made dependent on m and p only by taking the maximum constant over all possible block configurations. Notice that in a configuration, all r_w's are even, $r_1 + \cdots + r_w = 2p$, and $w \leq p - m$. Note that

$$\begin{aligned} \|A\|_r^r \|A\|_s^s &\leq \|A\|_{r-1}^{r-1} \|A\|_{s+1}^{s+1}, \quad r > s, \\ \|A\|_{r+s}^{r+s} &\leq \|A\|_r^r \|A\|_s^s, \end{aligned}$$

it is easy to see that the worst case of configuration is when $w = p - m$ and

$$(r_1, \ldots, r_w) = (\underbrace{\kappa, \ldots, \kappa}_{n_1 \text{ times}}, \underbrace{\kappa + 2, \ldots, \kappa + 2}_{n_2 \text{ times}}),$$

where

$$\kappa = 2 \left\lceil \frac{p}{2(p-m)} \right\rceil,$$

and (n_1, n_2) is the solution to

$$\begin{aligned} \kappa n_1 + (\kappa + 2) n_2 &= 2p, \\ n_1 + n_2 &= p - m. \end{aligned}$$

This gives the bound

$$C\|A\|_\kappa^{\kappa n_1} \|A\|_{\kappa+2}^{(\kappa+2)n_2}.$$

Finally, observe that the number of configurations is a constant that depends on p and m only, giving the variance claim.

9. Bit lower bound for Schatten norms.

Let m be an even integer, and let $D_{m,k}$ $(0 \leq k \leq m)$ be an $m \times m$ diagonal matrix with the first k diagonal elements equal to 1 and the remaining diagonal entries 0. Define

$$M_{m,k} = \begin{pmatrix} \mathbf{1}_m \mathbf{1}_m^T & 0 \\ \sqrt{\gamma} D_{m,k} & 0 \end{pmatrix}, \tag{9.1}$$

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

Our starting point is the following theorem. Let $m \geq 2$, and let $p_m(k) = \binom{m}{k}/2^{m-1}$ for $0 \leq k \leq m$. Let $\mathsf{Even}(m)$ be the probability distribution on even integers $\{0, 2, \ldots, m\}$ with probability density function $p_m(k)$, and let $\mathsf{Odd}(m)$ be the distribution on odd integers $\{1, 3, \ldots, m-1\}$ 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 $X_1, \ldots, X_s$. It is clear that $f(X) = \sum_i f(\sigma_i(X))$ is diagonally block-additive.

THEOREM 9.1. *Let t be an even integer, and let N be sufficiently large (independent of t). Suppose that $X \in \mathbb{R}^{N\times N}$. Let f be a function of square matrices that is diagonally block-additive. If there exists $m = m(t)$ such that*

$$\mathop{\mathbb{E}}_{q\sim\mathsf{Even}(t)} f(M_{m,q}) - \mathop{\mathbb{E}}_{q\sim\mathsf{Odd}(t)} f(M_{m,q}) \neq 0, \tag{9.2}$$

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 $\mathrm{BHH}^0_{t,n}$ problem, construct a graph G as follows. The graph contains n vertices $v_1, \ldots, v_n$, 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 jth clique of size m, we fix t vertices, denoted by $w_{j,1}, \ldots, w_{j,t}$. Whenever $x_i = 1$ for $i = (j-1)(n/t) + r$ $(0 \le r < n/t)$, 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:

$$\mathcal{M}_{n,m,t} = \begin{pmatrix} M_{m,q_1} & & & \\ & M_{m,q_2} & & \\ & & \ddots & \\ & & & M_{m,q_{n/t}} \end{pmatrix}, \tag{9.3}$$

where $q_1, \ldots, q_{n/t}$ 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: $q_1, \ldots, q_{n/t}$ 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 $\mathsf{Even}(t)$ or $\mathsf{Odd}(t)$ 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 $F_q = f(M_{m,q})$, $q = 0, \ldots, t$. Suppose that the gap in (9.2) is positive. Let $A = \mathbb{E}_{q\sim\mathsf{Even}(t)} 2(F_q + F_{t-q})$ and $B = \mathbb{E}_{q\sim\mathsf{Odd}(t)} 2(F_q + F_{t-q})$; then $A - B > 0$. Summing up $(n/2t)$ independent groups and applying a Chernoff bound, w.h.p., $f(\mathcal{M}) \ge (1-\delta)\frac{n}{2t}A$ in the even case and $f(\mathcal{M}) \le (1+\delta)\frac{n}{2t}A$, 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 $X \ge (1-c)(1-\delta)\frac{n}{2t}A$, and in the odd case $X \le (1+c)(1+\delta)\frac{n}{2t}A$. 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 the gap in (9.2) is negative. □

Our main theorem in this section is the following, which shows the near linear lower bound for estimating Schatten-p norms.

THEOREM 9.2. *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 9.1 to $f(x) = x^p$ and $m = t$ and verifying that (9.2) is satisfied. The proof is technical and thus postponed to section 10.

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 11.

LEMMA 9.3. *For $f(x) = x^p$ and integer $p \geq 2$, the gap condition* (9.2) *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 [37, 46], 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 that our lower bound is tight up to $\mathrm{poly}(\log n)$ factors for matrices in which every row and column has $\mathcal{O}(1)$ nonzero elements by providing an algorithm in section 12 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 13.

10. Proof of Theorem 9.2.

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 ith 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 $\mathrm{rank}(M_{m,k}) = k+1$, there are two more nonzero 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 nonzero 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 nonzero 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 (9.2) becomes

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

where the $\gamma^{p/2}$ terms cancel and

$$G_i = \sum_k (-1)^k \binom{m}{k} r_i^{\frac{p}{2}}(k), \quad i = 1,2. \tag{10.1}$$

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, in the remainder of this section we replace $p/2$ with p in (10.1), and hence G_1 and G_2 are redefined as

$$G_i = \sum_k (-1)^k \binom{m}{k} r_i^p(k), \quad i = 1,2, \tag{10.2}$$

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

Next we choose

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

We claim that they admit the following power series expansion in k (the proof deferred to subsection 10.2):

$$r_1^p(k) = \sum_{s \geq 0} A_s k^s, \quad r_2^p(k) = \sum_{s \geq 0} B_s k^s, \tag{10.3}$$

where for $s \geq 2$,

$$A_s = \frac{(-1)^{s-1}\gamma^s m^{2p-s}}{s!(m^2-\gamma)^{2s-1}} \sum_{i=0}^{s-1} (-1)^i \binom{s-1}{i} F_{p,s,i} \gamma^{s-i-1} m^{2i}, \tag{10.4}$$

$$B_s = \frac{(-1)^s \gamma^p m^s}{s!(m^2-\gamma)^{2s-1}} \sum_{i=0}^{s-1} (-1)^i \binom{s-1}{i} F_{p,s,i} \gamma^i m^{2(s-i-1)}, \tag{10.5}$$

and

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

We analyze 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

$$\sum_{k=0}^{m} \binom{m}{k} k^s (-1)^k = \left\{ {s \atop m} \right\} (-1)^m m!, \tag{10.6}$$

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 subsection 10.1.

LEMMA 10.1. *For any fixed noninteger $p > 0$, one can choose γ_0 and m such that B_s has 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 (10.6) 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

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

It follows from Lemma 10.1 that the above is also a lower bound for G_2. Therefore G_1 is negligible compared with G_2 and $G_1 + G_2 \neq 0$. This ends the proof of Theorem 9.2. □

10.1. Proof of Lemma 10.1. 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 10.1. Let $x = \gamma/m^2$ and write B_s as

$$B_s = (-1)^{s-1} \frac{\gamma^p m^{3s-2}}{s!(m^2-\gamma)^{2s-1}} \sum_{i=0}^{s-1} (-1)^{i+1} \binom{s-1}{i} F_{s,i} x^i. \tag{10.7}$$

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

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

Invoking Euler's Transformation (see, e.g., [9, p. 78])

$${}_2F_1\left(\begin{matrix} a, b \\ c \end{matrix}; x\right) = (1-x)^{c-a-b}\ {}_2F_1\left(\begin{matrix} c-a, c-b \\ c \end{matrix}; x\right)$$

gives

$${}_2F_1\left(\begin{matrix} 1-s, 1+p-2s \\ 1+p-s \end{matrix}; x\right) = (1-x)^{2s-1}\ {}_2F_1\left(\begin{matrix} p, s \\ p-s+1 \end{matrix}; x\right). \tag{10.8}$$

Therefore

$$B_s m^s = (-1)^s \frac{\gamma^p \Gamma(1+p)}{s!\ \Gamma(1+p-s)}\ {}_2F_1\left(\begin{matrix} p, s \\ p-s+1 \end{matrix}; x\right). \tag{10.9}$$

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 $s \geq s^*$, 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 $n \geq s - \lceil p \rceil$ and has alternating signs for $n \leq s - \lceil p \rceil$. 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 $n \geq 2s$ and $x \leq 1/10$, $|b_n/b_{n-1}| < 3x \leq 1/3$ and thus $|\sum_{n \geq 2s} b_n| \leq \frac{3}{2}|b_{2s}|$. Also, when $s \geq 3p$ is large enough, $x \leq 1/10$ and $n \leq s/2$.

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

$$0 \leq \sum_{2\leq n\leq s/2} b_n \leq b_2. \tag{10.10}$$

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

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

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

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

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 $|b_{2s}| \leq x^3$. Therefore

$$\left|\sum_{n>s/2} b_n\right| \leq Kx^3 \tag{10.11}$$

for some K and sufficiently large s and small x, all of which depend only on p.

It follows that

$$
\begin{aligned}
{}_2F_1\left(\begin{matrix} p, s \\ p-s+1 \end{matrix}; x\right) &\ge 1 - \frac{ps}{s-p-1}x - \sum_{2\le n\le s/2} b_n - \left|\sum_{n>s/2} b_n\right| \\
&\ge 1 - \frac{ps}{s-p-1}x - b_2 - Kx^3 \quad \text{(using (10.10) and (10.11))} \\
&\ge 1 - \frac{ps}{s-p-1}x - \frac{p(p+1)s(s+1)}{2(s-p-1)(s-p-2)}x^2 - Kx^3 \\
&> 0
\end{aligned}
$$

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

The proof of Lemma 10.1 is now complete. ◻

10.2. Proof of power series expansion. In this subsection, our main task is to prove the claim made in section 10 that $r_1^p(k)$ and $r_2^p(k)$ admit the series expansion in (10.3).

Proof of (10.3). We first verify the series expansion of $r_1(k)$. It is a standard result that for $|x| \le 1/4$ (see, e.g., [39, p. 203]),

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

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

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

where we replace $\sum_j n_j$ with s. It is a known result using the Lagrange inversion formula that (see, e.g., [67, p. 128])

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

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

$$A_s = \frac{(-1)^{s+1}\gamma^s m^{2p}}{(m^2-\gamma)^{2s-1}} \sum_{i=0}^{s-1} (-1)^i \binom{p}{i+1} \frac{i+1}{s} \binom{2s-i-2}{s-1} m^{s-2(i+1)} (m^2-\gamma)^i. \tag{10.12}$$

To see that (10.12) agrees with (10.4), it suffices to show that

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

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

$$(-1)^{s-1} \binom{s-1}{j} F_{p,s,j,s-j-1} = s! \sum_{i=j}^{s-1} (-1)^i \binom{p}{i+1} \frac{i+1}{s} \binom{2s-i-2}{s-1} \binom{i}{j}.$$

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

$$S_0 \ {}_2F_1\left(\begin{matrix} 1+j-p, 1+j-s \\ 2+j-2s \end{matrix} ; 1 \right),$$

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 $1 \leq k \leq s-j-1$. This holds by the Chu–Vandermonde identity (see, e.g., [9, Corollary 2.2.3]), which states, in our case, that

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

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

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

and then show it agrees with (10.5). The whole proof is almost identical to that for $r_1(k)$.

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

We remark that one can continue from (10.9) to bound $\sum_s B_s m^s \lesssim_p 1/m^p$. Hence $G_1 + G_2 \simeq 1/m^p$, and thus the gap in (9.2) is $\Theta_p(1/2^m m^p)$ with the constant dependent on p only.

11. Proofs related to even *p*. Now we prove Lemma 9.3. 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 $\operatorname{rank}(M_{m,k}) = m - k + 1$, and there are two more nonzero eigenvalues. Using the trace and Frobenius norm as in the case of the $M_{m,k}$ for noneven p, 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 $k = m$ happen 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 (10.1) (we do not need to replace $p/2$ with p here). It remains to show again that $G_1 + G_2 \neq 0$ if and only if $m \leq [p/2]$.

Proof of Lemma 9.3. Applying the binomial theorem, we obtain

$$\begin{aligned} r_1^p(k) + r_2^p(k) &= \frac{1}{2^{p-1}} \sum_{i:2|(p-i)} \binom{p}{i} (m-1)^i ((m-1)^2 + 4k)^{\frac{p-i}{2}} \\ &= \frac{1}{2^{p-1}} \sum_{i:2|(p-i)} \binom{p}{i} (m-1)^i \sum_{j=0}^{\frac{p-i}{2}} \binom{\frac{p-i}{2}}{j} (m-1)^{2j} 4^{\frac{p-i}{2}-j} k^{\frac{p-i}{2}-j}. \end{aligned}$$

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 $m > [\frac{p}{2}]$. □

Although when p is even we have $G_1 + G_2 = 0$, we can, however, show that $G_1, G_2 \neq 0$, which will be useful for some applications in section 13.

LEMMA 11.1. *When p is even, $G_1 \neq 0$ and $G_2 \neq 0$, 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 (10.6), we obtain that

$$G_2 = -\frac{(m-1)^p m!}{2^p} \sum_{s \geq m} \left\{ \begin{matrix} s \\ m \end{matrix} \right\} \frac{4^s}{(m-1)^{2s}} \sum_{\substack{\text{odd } i \\ 1 \leq i \leq p-1}} \binom{p}{i} \binom{i/2}{j}.$$

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

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

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

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

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. □

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 nonzero.

12. Algorithm for even p and sparse matrices.

First, we recall the classic result on Count-Sketch [20].

Theorem 12.1 (Count-Sketch). *There is a randomized linear function $M : \mathbb{R}^n \to \mathbb{R}^S$ with $S = \mathcal{O}(w \log(n/\delta))$, and a recovery algorithm A satisfying the following. For any $x \in \mathbb{R}^n$, with probability $\ge 1 - \delta$, A reads Mx and outputs $\tilde{x} \in \mathbb{R}^n$ such that $\|\tilde{x} - x\|_\infty^2 \le \|x\|_2^2/w$.*

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

Theorem 12.2 (precision sampling [5]). *Fix $0 < \epsilon < 1/3$. There is a randomized linear function $M : \mathbb{R}^n \to \mathbb{R}^S$, with $S = \mathcal{O}(\epsilon^{-2} \log^3 n)$, and an "$\ell_p$-sampling algorithm A" satisfying the following. For any nonzero $x \in \mathbb{R}^n$, 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 an $(\epsilon, 0)$-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 $y_i \ge C\|x\|_2^2/\epsilon$ 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 the 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 $\mathcal{O}(1)$ nonzero entries in each row and each column. Assume $p \ge 4$. We 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 12.1 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 [5], which uses $\mathcal{O}(\log n)$ seeds.

Algorithm 12.1. Algorithm for even p and sparse matrices.

```
Assume that matrix $A \in \mathbb{R}^{n\times n}$ has at most $k = \mathcal{O}(1)$ nonzero entries per row and per column.
$T \leftarrow \Theta(n^{1-2/p}/\epsilon^2)$
$R \leftarrow \Theta(\log n)$
$w \leftarrow \mathcal{O}(\epsilon^{-1}\log n + \epsilon^{-2})$
$I_s \leftarrow \emptyset$ is a multiset for $s = 1,\ldots,p/2$
Choose i.i.d. $u_1,\ldots,u_n$ with $u_i \sim \mathrm{Unif}(0,1)$
$D \leftarrow \mathrm{diag}\{1/\sqrt{u_1},\ldots,1/\sqrt{u_n}\}$
In parallel, maintain $p/2$ COUNT-SKETCH structures $\mathcal{S}_s$ ($s \in [p]$) of size $\Theta(\epsilon^{-1}T\log n)$
Maintain a sketch for estimating $\|A\|_F^2$ and obtain a $(1\pm\epsilon)$-approximation $L$ as in [2]
In parallel, maintain $pT/2$ structures $\mathcal{P}_{s,t}$ ($(s,t) \in [p/2]\times[T]$); 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 $\mathcal{O}(w\log n)$
Maintain a sketch for estimating $\|B\|_F^2$ and obtain a $(1\pm\epsilon)$-approximation $L'$ as in [2]
for $s \leftarrow 1$ to $p/2$ do
    for $t \leftarrow 1$ to $T$ do
        for $r \leftarrow 1$ to $R$ do
            Use the $r$th repetition of the $\mathcal{P}_{s,t}$ to obtain estimates $\tilde{b}_{i'1},\ldots,\tilde{b}_{i'n}$ for all $i'$ and form rows $\tilde{b}_{i'} = (b_{i'1},\ldots,b_{i'n})$.
            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
        end for
        Retain only entries of $b_{i'}$ that are at least $2L'/\sqrt{w}$.
        $\tilde{a}_{i'} \leftarrow \sqrt{u_{i'}}\tilde{b}_{i'}$
        $I_s \leftarrow I_s \cup \{i'\}$
    end for
end for
for $s \leftarrow 1$ to $p/2$ do
    Use $\mathcal{S}_s$ to obtain estimates $\tilde{a}'_{i'1},\ldots,\tilde{a}'_{i'n}$ for all $i'$ and form rows $\tilde{a}'_{i'} = (a'_{i'1},\ldots,a'_{i'n})$
    Find all $i$ such that $\|\tilde{a}'_{i'}\|_2^2 \geq L/(10T)$, and retain $\mathcal{O}(T)$ of them corresponding to the largest $\|\tilde{a}'_{i'}\|_2^2$, making a set $K_s$
    $\tilde{a}_i \leftarrow \tilde{a}'_i$ for all $i \in K_s$
    $I_s = I_s \cup K_s$
end for
Return $Y$ as defined in (12.4)
```

THEOREM 12.3. *Let $\epsilon \in (0,1)$. For a sparse matrix $A \in \mathbb{R}^{n\times n}$ with $\mathcal{O}(1)$ nonzero entries per row and per column, Algorithm* 12.1 *returns a value that is a $(1+\epsilon)$-approximation to $\|A\|_p^p$ with constant probability, using space $\mathcal{O}_p(n^{1-2/p}\,\mathrm{poly}(1/\epsilon,\log n))$.*

Proof. It is the guarantee from the underlying COUNT-SKETCH structure of size

$\Theta(w \log n)$ (where $w = \mathcal{O}(\epsilon^{-1} \log n + \epsilon^{-2})$) that

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

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

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

and thus

$$\|\tilde{b}_{i'}\|_2^2 = \|b_{i'}\|_2^2 \pm \frac{\|B\|_F^2}{w}. \tag{12.1}$$

Since each row i is scaled by the same factor $1/\sqrt{u_i}$, we can apply the proof of Theorem 12.2 to the vector of row norms $\{\|a_i\|_2\}$ and $\{\|b_i\|_2\}$, which still remains valid because of the error guarantee (12.1), 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

$$\Pr\{i' = i\} = (1 \pm \epsilon)\frac{\|a_i\|_2^2}{\|A\|_F^2} \pm \frac{1}{\operatorname{poly}(n)}. \tag{12.2}$$

Next we analyze the estimation error. It holds w.h.p. that $\|B\|_F^2 \leq w\|A\|_F^2$. Since a_i (and thus b_i) has $\mathcal{O}(1)$-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 $w' = \mathcal{O}(w)$ 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 $\mathcal{O}(1)$ nonzero entries $a_{i\ell}$ such that $|a_{i\ell}| \leq \eta|a_{ij}|$. and each of them has at most a $\Theta(\epsilon\eta a_{ij})$ additive error by the threshold in step 17; the approximation $\tilde{a}_i$ to a_i therefore satisfies

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

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

Similarly, by a standard heavy hitter argument, with probability $\Omega(1)$, the set K_s contains all i such that $\|a_i\|_2^2 \geq \epsilon\|A\|_F^2/T$ if we choose the size of the COUNT-SKETCH structure with a large enough heading constant. This implies that if $i \in I_s \setminus K_s$, then $\|a_i\|_2^2 \leq \epsilon\|A\|_F^2/T$, where (and henceforth in the proof) I_s is taken to be its value at the beginning of step 28.

Next we show that our estimate is desirable. First, we observe that the additive $1/\operatorname{poly}(n)$ term in (12.2) 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

$$\Pr\{i' = i\} = (1 \pm \epsilon)\frac{\|a_i\|_2^2}{\|A\|_F^2}. \tag{12.3}$$

For notational simplicity, let $q = p/2$, and let $\ell_i = \|\tilde{a}_i\|_2^2$ if i is a sampled row. For $i_s \in I_s$, define

$$\tau(i_s) = \begin{cases} 1, & i_s \in K_s, \\ \|A\|_F^2/\|a_{i_s}\|_2^2, & \text{otherwise}, \end{cases} \qquad \tilde{\tau}(i_s) = \begin{cases} 1, & i_s \in K_s, \\ L/\ell_{i_s}, & \text{otherwise}, \end{cases}$$

and for $(i_1, \ldots, i_q) \in I_1 \times \cdots \times I_q$, define

$$\tau(i_1, \ldots, i_q) = \tau(i_1) \cdots \tau(i_q),$$
$$\tilde{\tau}(i_1, \ldots, i_q) = \tilde{\tau}(i_1) \cdots \tilde{\tau}(i_q),$$

and

$$X(i_1, \ldots, i_q) = \prod_{j=1}^{q} \langle a_{i_i}, a_{i_{j+1}} \rangle \tau(i_1) \cdots \tau(i_q),$$
$$\tilde{X}(i_1, \ldots, i_q) = \prod_{j=1}^{q} \langle \tilde{a}_{i_i}, \tilde{a}_{i_{j+1}} \rangle \tilde{\tau}(i_1) \cdots \tilde{\tau}(i_q),$$

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

$$p_{s,t}(i) = \Pr\{\text{row } i \text{ gets sampled in } (s,t)\text{th precision sampling}\}.$$

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,

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

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.

Our estimator is

$$Y = \sum_{i_1 \in I_1, \ldots, i_q \in I_q} \frac{1}{T^{\sigma(i_1, \ldots, i_q)}} \tilde{X}(i_1, \ldots, i_q), \tag{12.4}$$

where

$$\sigma(i_1, \ldots, i_q) = |\{s : i_s \notin K_s\}|.$$

Then

$$\left|\mathbb{E}\,Y - \|A\|_p^p\right| \leq \sum_{i_1,\ldots,i_q} \left| \frac{\Pr\{i_1 \in I_1\}\cdots\Pr\{i_q \in I_q\}}{T^\sigma}\tilde{X}(i_1,\ldots,i_q) - \prod_{j=1}^{q}\langle a_{i_i}, a_{i_{j+1}}\rangle \right|.$$

For $i \in I_s \setminus K_s$, we have

$$\Pr\{i \in I_s\} = 1 - \prod_{t=1}^{T}(1 - p_{s,t}(i)) = (1 \pm \mathcal{O}(\epsilon))\frac{T}{\tau(i)},$$

where we used the fact that $p_s(i_s) = (1 \pm \epsilon)/\tau(i)$ and $1/\tau(i) \leq \epsilon/T$ for $i \in I_s \setminus K_s$. For $i_s \in K_s$, we have

$$\Pr\{i \in I_s\} = \frac{1}{\tau(i)} = 1.$$

Hence

$$\frac{\Pr\{i_1 \in I_1\}\cdots\Pr\{i_q \in I_q\}}{T^\sigma} = (1 \pm \mathcal{O}(\epsilon))\frac{1}{\tau(i_1)\cdots\tau(i_s)} = (1 \pm \mathcal{O}(\epsilon))\frac{1}{\tilde{\tau}(i_1)\cdots\tilde{\tau}(i_s)}$$

and

$$\frac{\Pr\{i_1 \in I_1\}\cdots\Pr\{i_q \in I_q\}}{T^\sigma}\tilde{X}(i_1,\ldots,i_q) = (1 \pm \mathcal{O}(\epsilon))\prod_{j=1}^{q}\langle \tilde{a}_{i_j}, \tilde{a}_{i_{j+1}}\rangle.$$

It follows that

$$\left|\mathbb{E}\,Y - \|A\|_p^p\right| \leq \sum_{i_1,\ldots,i_q}\left\{\left|\prod_{j=1}^{q}\langle \tilde{a}_{i_j}, \tilde{a}_{i_{j+1}}\rangle - \prod_{j=1}^{q}\langle a_{i_i}, a_{i_{j+1}}\rangle\right| + \mathcal{O}(\epsilon)\left|\prod_{j=1}^{q}\langle \tilde{a}_{i_j}, \tilde{a}_{i_{j+1}}\rangle\right|\right\}.$$

The key observation is that each a_i has only $\mathcal{O}(1)$ rows with overlapping support, since each row and each column has only $\mathcal{O}(1)$ nonzero entries. The same claim holds for $\tilde{a}_i$, which is due to our threshold in step 17: 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 $\mathcal{O}(1)$ contributing summands. Each contributing summand is bounded by

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

Therefore

$$\left|\mathbb{E}\,Y - \|A\|_p^p\right| \lesssim \epsilon\sum_i \|a_i\|_2^{2q} \leq \epsilon\|A\|_{2q}^{2q}, \tag{12.5}$$

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

Next we bound the variance:

$$\mathbb{E}\,Y^2 = \mathbb{E}\sum_{\substack{i_1\in I_1,\ldots,i_q\in I_q\\ j_1\in I_1,\ldots,j_q\in I_q}}\frac{1}{T^{\sigma(i_1,\ldots,i_q)}T^{\sigma(j_1,\ldots,j_q)}}\tilde{X}(i_1,\ldots,i_q)\tilde{X}(j_1,\ldots,j_q).$$

Similarly to before, the right-hand side can be simplified as

$$\sum_{r=0}^{q} \sum_{\substack{i_1\in I_1,\ldots,i_q\in I_q\\ j_1\in I_1,\ldots,j_q\in I_q\\ |\{s:i_s=j_s\notin K_s\}|=r}} (1+\mathcal{O}(\epsilon))\frac{\prod_{s:i_s=j_s\notin K_s}\tau(i_s)}{T^r}\prod_{s=1}^{q}\langle\tilde{a}_{i_s},\tilde{a}_{i_{s+1}}\rangle\prod_{s=1}^{q}\langle\tilde{a}_{j_s},\tilde{a}_{j_{s+1}}\rangle. \tag{12.6}$$

We can upper bound each individual summand as

$$\begin{aligned}
&\left|(1+\mathcal{O}(\epsilon))\frac{\prod_{s:i_s=j_s\notin K_s}\tau(i_s)}{T^r}\prod_{s=1}^{q}\langle\tilde{a}_{i_s},\tilde{a}_{i_{s+1}}\rangle\prod_{s=1}^{q}\langle\tilde{a}_{j_s},\tilde{a}_{j_{s+1}}\rangle\right|\\
&\lesssim \frac{1}{T^r}\frac{\|A\|_F^{2r}}{\prod_{s:i_s=j_s\notin K_s}\|a_{i_s}\|_2^2}\prod_{s=1}^{q}\|a_{i_s}\|_2^2\cdot\prod_{s=1}^{q}\|a_{j_s}\|_2^2\\
&\le \frac{1}{T^r}\|A\|_F^{2r}\left(\max_i\|a_i\|_2^2\right)^{2q-r}.
\end{aligned}$$

Now, note that the terms corresponding to $r=0$ in (12.6) are covered by the expansion of $(\mathbb{E}\,Y)^2$. Also, by the same argument as before, each i_s or j_s appears in $\mathcal{O}(1)$ contributing summands, and we have that

$$\mathbb{E}\,Y^2-(\mathbb{E}\,Y)^2\lesssim\sum_{r=1}^{q}\frac{1}{T^r}\|A\|_F^{2r}\|A\|_p^{2p-2r}\le\sum_{r=1}^{q}\frac{1}{T^r}n^{r(1-\frac{2}{p})}\|A\|_p^{2p},$$

which implies that

$$\mathbb{E}\,Y^2-(\mathbb{E}\,Y)^2\le\epsilon^2\|A\|_p^{2p}$$

if the constant C in $T=Cn^{1-2/p}/\epsilon^2$ is large enough. □

13. General functions and applications.

The following is a direct corollary of Theorem 9.2.

THEOREM 13.1. *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 9.2 for the Schatten p-norm. Let ϵ be sufficiently small (it could depend on t and thus m) such that the singular values of $\epsilon\mathcal{M}$ are at most $\delta(c_0/3)$, where $\mathcal{M}$ is the hard instance matrix used in Theorem 9.2. Then $\alpha(1-c_0/3)f(\epsilon\mathcal{M})\le\|\epsilon\mathcal{M}\|_p^p\le\alpha(1+c_0/3)f(\epsilon\mathcal{M})$. Therefore, any algorithm that approximates $f(\epsilon\mathcal{M})$ 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 9.2.

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 9.3.

THEOREM 13.2. *Suppose that f admits a Taylor expansion near 0 that has infinitely many even-order terms of nonzero coefficients. Then for any arbitrary large m, there exists $c=c(m)\in(0,1)$ 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.*

Table 13.1

Application of Theorem 13.1 *and Theorem* 13.2 *to some M-estimators from* [73].

Function $\rho(x)$	Apply	Function $\rho(x)$	Apply
$2(\sqrt{1+\frac{1}{2}x^2}-1)$	Theorem 13.1	$\frac{x^2/2}{1+x^2}$	Theorem 13.1
$c^2(\frac{x}{c}-\ln(1+\frac{x}{c}))$	Theorem 13.1	$\frac{c^2}{2}(1-\exp(-\frac{x^2}{c^2}))$	Theorem 13.2
$\begin{cases} x^2/2, & x \le k; \\ k(x-k/2), & x > k \end{cases}$	Theorem 13.1	$\begin{cases} \frac{c^2}{6}(1-(1-\frac{x^2}{c^2})^3), & x \le c; \\ \frac{c^2}{6}, & x > c \end{cases}$	Remark after Lemma 11.1
$\frac{c^2}{2}\ln(1+\frac{x^2}{c^2})$	Theorem 13.2		

Proof. If the expansion has an odd-order term with a nonzero coefficient, apply Theorem 13.1 with the lowest nonzero 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 nonzero coefficient a_p. Let p be the lowest order of such a term, and write

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

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 9.3. Lemma 9.3 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 the 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 [36]:

$$\eta_1(x) = \begin{cases} \frac{1}{x}\sqrt{(x^2-\alpha-1)^2-4\alpha}, & x \ge 1+\sqrt{\alpha}, \\ 0, & x < 1+\sqrt{\alpha}, \end{cases}$$

$$\eta_2(x) = \begin{cases} \frac{1}{\sqrt{2}}\sqrt{x^2-\alpha-1+\sqrt{(x^2-\alpha-1)^2-4\alpha}}, & x \ge 1+\sqrt{\alpha}, \\ 0, & x < 1+\sqrt{\alpha}, \end{cases}$$

$$\eta_3(x) = \frac{1}{x\eta_2^2(x)}\max\left\{\eta_2^4(x)-\alpha-\alpha x\eta_2(x), 0\right\},$$

where we assume that $0 \cdot \infty = 0$. Since $\eta_i(x) \simeq x$ when x is large, the lower bound follows from Theorem 13.1.

Some commonly used influence functions $\rho(x)$ can be found in [73]; we summarize them in Table 13.1. Several are asymptotically linear when x is large and Theorem 13.1 applies. Some are covered by Theorem 13.2. 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 the proof of Lemma 11.1.

Finally, we consider functions of the form

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

and prove the following theorem.

THEOREM 13.3. *Let $\alpha > 0$ be a small constant. Suppose that f is strictly increasing. There exists N_0 and c_0 such that for all $N \geq N_0$, $k \leq \alpha N$, 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 $\Omega_\alpha(N^{1-\Theta(\alpha^{2/3})})$ bits of space.*

Proof. Similarly to Theorem 9.1, we reduce the problem from the BHH_n^0 problem. Let $m = t$ be the largest integer such that $c_1/(2t^{3/2}) \geq \alpha$, where $c_1 > 0$ is some constant to be specified later. Then $m = t = \Theta(\alpha^{2/3})$. We analyze the largest k singular values of $\mathcal{M}$ as defined in (9.3). 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 at least $c_1/\sqrt{m}$, $X_i = m$ in the even case and $X_i = m/2 - 1$ in the odd case. It immediately follows from a Chernoff bound that w.h.p., it holds that $X_i = m$ (resp., $X_i = m-1$) for at least $(N/2m)(c_1/\sqrt{m})(1-\delta) = (1-\delta)c_1 N/(m\sqrt{m})$ 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 $k \leq \alpha N \leq (1-\delta)c_1 N/(m\sqrt{m})$, w.h.p., exhibits a gap of size at least $c_0 \cdot k$ for some constant c_0 between the even and the odd cases. Since $F_k(\mathcal{M}) = \Theta(k)$ w.h.p., the lower bound for the Ky Fan k-norm follows from the lower bound for BHH_n^0. ◻

The lower bound for the Ky Fan k-norm 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.

As the final result of the paper, we show an $\Omega(n^{1-1/t})$ lower bound for SVD entropy function of matrices in the following subsection.

13.1. SVD entropy. Let $h(x) = x^2 \ln x^2$. For $X \in \mathbb{R}^{N\times N}$, we define its SVD entropy $H(X)$ as

$$H(X) = \sum_i h\left(\frac{\sigma_i(X)}{\|X\|_F}\right).$$

For notational convenience, we also write $h(X) = \sum_i h(\sigma_i(X))$ (singular values unnormalized by $\|X\|_F$).

In this subsection, our goal is to show the following theorem.

THEOREM 13.4. *Let t be an even integer, and let $X \in \mathbb{R}^{N\times N}$, where N is sufficiently large. There exists $c = c(t)$ such that estimating the matrix entropy $H(X)$ within an additive error of c requires $\Omega_t(N^{1-1/t})$ bits of space.*

This theorem will follow easily from the next lemma, whose proof is postponed to later in this subsection. It is based on Theorem 9.1, with the same hard instance used by Theorem 9.2.

LEMMA 13.5. *Let t be an even integer, and let $X \in \mathbb{R}^{N\times N}$, where N is sufficiently large. There exists a constant $c = c(t) > 0$ such that any algorithm that approximates $H(X)$ within a factor $1 \pm c$ with constant probability in the streaming model must use $\Omega_t(N^{1-1/t})$ bits of space.*

Proof of Theorem 13.4. Let X be the matrix in the hard instance for estimating $h(X)$, which is the same hard instance used by Theorem 9.2. Then X consists of smaller diagonal blocks of size $m = m(t)$, and $\|X\|_F^2 = CN$, where $C = C(m,t)$ is a constant depending on t and m only. It is also easy to see that $K_1 N \le h(X) \le K_2 N$ for some constants K_1, K_2 depending only on m and t.

Now we show that an additive c-approximation to $H(X)$ can yield a multiplicative $(1 \pm c')$-approximation to $H(X)$. Recall that $H(X) = \ln\|X\|_F^2 - h(X)/\|X\|_F^2 = \ln(CN) - h(X)/(CN)$. Suppose that Z is an additive c-approximation to $H(X)$; then we compute $\hat{X} = CN\ln(CN) - CNZ$. Since $Z \le \mathcal{H}(Y) + c$,

$$\hat{X} \ge CN\ln(CN) - CN(h(X)+c) = h(X) - cCN \ge \left(1 - \frac{cC}{K_1}\right) h(Y).$$

Similarly, it can be shown that $\hat{X} \le (1+\frac{cC}{K_2})h(X)$, and thus choosing $c' = Cc/\max\{K_1, K_2\}$ suffices.

The lower bound follows from Lemma 13.5. □

We devote the rest of this subsection to the proof of Lemma 13.5, for which we apply Theorem 9.1 to $h(x)$.

Proof of Lemma 13.5. Following the same argument as in the proof of Theorem 9.2, our goal is to show that

$$G_1 + G_2 \neq 0,$$

where

$$G_i = \sum_k (-1)^k \binom{m}{k} h(\sqrt{r_i(k)}) = \sum_k (-1)^k \binom{m}{k} r_1(k) \ln r_1(k), \quad i = 1, 2.$$

Taking $\gamma = 1$ in the definition of $M_{m,k}$ in (9.1), we obtain (see subsection 10.2) that

$$r_1(k) = m^2 + \sum_{j=1}^{\infty} (-1)^{j-1} \frac{C_{j-1} m^j k^j}{(m^2-1)^{2j-1}}, \tag{13.1}$$

$$r_2(k) = 1 + \sum_{j=1}^{\infty} (-1)^{j} \frac{C_{j-1} m^j k^j}{(m^2-1)^{2j-1}}, \tag{13.2}$$

where C_j denotes the jth Catalan number. Plugging (13.2) into

$$(1+x)\ln(1+x) = x + \sum_{n\ge 2} (-1)^n \frac{x^n}{n(n-1)}, \quad |x| \le 1,$$

and arranging the terms as in subsection 10.2 yields that

$$r_2(k)\ln r_2(k) = \sum_s B_s k^s,$$

where

$$B_s = \frac{(-1)^s m^s}{s(m^2-1)^{2s}} \left((m^2-1)\binom{2s-2}{s-1} + \sum_{i=2}^{s} \frac{(m^2-1)^i(-1)^i}{(i-1)} \binom{2s-i-1}{s-1} \right), \ s \geq 2.$$

Let

$$D_i = \frac{x^i}{i-1}\binom{2s-i-1}{s-1};$$

then

$$\frac{D_{i+1}}{D_i} = \frac{(i-1)(s-i)}{i(2s-i-1)}x.$$

Let $i^* = \max_i D_i$. One can obtain by solving $D_{i+1}/D_i \geq 1$ that $i^* = \lceil \frac{x-2}{x-1}s \rceil$. Hence $i^* = s$ for $s \leq m^2-3$, and $\sum(-1)^i D_i \simeq (-1)^s D_s$ when $s < \alpha m^2$ for some $\alpha \in (0,1)$. Note that B_s has the same sign for $s \leq \alpha m^2$ because $(m^2-1)\binom{2s-2}{s-1}$ is negligible compared with D_s. The partial sum (choosing even m)

$$\sum_{s=m}^{\alpha m^2} B_s \left\{ {s \atop m} \right\} m! \gtrsim \sum_{s=m}^{\alpha m^2} \frac{m^s}{s(m^2-1)^{2s}} \cdot \frac{(m^2-1)^s}{s-1} \left\{ {s \atop m} \right\} m! \gtrsim \frac{1}{m^2 e^m}.$$

Next we show that $G_2 \gtrsim 1/m^{m+2}$.

Write

$$\sum_{i=2}^{s} \frac{(m^2-1)^i(-1)^i}{i-1}\binom{2s-i-1}{s-1} = (m^2-1)^2\binom{2s-3}{s-1} \ {}_3F_2\left({1,1,2-s \atop 2,3-2s} ; 1-m^2 \right).$$

We can write (see, e.g., [71])

$${}_3F_2\left({1,1,2-s \atop 2,3-2s} ; 1-m^2 \right) = \frac{1}{m^2-1}\int_{1-m^2}^{0} {}_2F_1\left({1,2-s \atop 3-2s} ; x \right) dx.$$

Arranging the terms, we can write

$$B_s = \frac{(-1)^s m^s}{s(m^2-1)^{2s}}(m^2-1)\binom{2s-2}{s-1} B_s',$$

where

$$B_s' = 1 + \frac{1}{2}\int_{1-m^2}^{0} {}_2F_1\left({1,2-s \atop 3-2s} ; x \right) dx.$$

It is shown in [30, Theorem 3.1] that ${}_2F_1(1,2-s;3-2s;x)$ has no real root on $(-\infty,0]$ when s is even and has a single root on $(-\infty,0]$ when s is odd. Therefore $B_s' > 0$ when s is even, and thus $B_s > 0$. Note that (see, e.g., [9, eq. (2.5.1)])

$$\frac{d}{dx} \ {}_2F_1\left({1,2-s \atop 3-2s} ; x \right) = \frac{s-2}{2s-3} \ {}_2F_1\left({2,3-s \atop 4-2s} ; x \right).$$

Again applying [30, Theorem 3.1] gives that ${}_2F_1(2,3-s;4-2s;x) > 0$ on $(-\infty,0]$ when s is odd. Hence ${}_2F_1(1,2-s;3-2s;x)$ is increasing on $(-\infty,0]$ when s is odd, and

$$B_s' \leq 1 + \frac{1}{2}\int_{1-m^2}^{0} 1 \ dx = \frac{m^2+1}{2}.$$

Let $I = \{\text{odd } s : B'_s > 0\}$. Since we showed above that B_s has the same sign for $s \le \alpha m^2$, we know that

$$\left|\sum_{s\in I} B_s \left\{ {s \atop m} \right\} m!\right| \le \sum_{s\in I} |B_s| m^s \le \sum_{s\in I} \frac{m^2+1}{2} \frac{m^{2s}}{(m^2-1)^{2s}} (m^2-1)4^s \lesssim \frac{1}{m^{\alpha m^2 - 6}}.$$

For odd $s \notin I$ it holds that $B_s > 0$. Therefore

$$\begin{aligned} G_2 &= \sum_{s\ge m} B_s \left\{ {s \atop m} \right\} m! \\ &= \sum_{s=m}^{\alpha m^2} B_s \left\{ {s \atop m} \right\} m! + \sum_{\substack{s>\alpha m^2 \\ s\notin I}} B_s \left\{ {s \atop m} \right\} m! + \sum_{s\in I} B_s \left\{ {s \atop m} \right\} m! \\ &\gtrsim \frac{1}{m^2 e^m} + 0 - \frac{1}{m^{\alpha m^2-6}} \\ &\gtrsim \frac{1}{m^2 e^m}, \end{aligned}$$

provided that m is large enough.

Next we analyze the contribution from $r_1(k)$. Plugging (13.1) into

$$(m^2+x)\ln(m^2+x) = m^2 \ln m^2 + x \ln m^2 + x + \sum_{i=2}^{\infty} (-1)^i \frac{x^i}{i(i-1)m^{2(i-1)}}$$

gives that

$$r_1(k) \ln r_1(k) = 2m^2 \ln m^2 + (\ln m^2) \sum_j \frac{(-1)^{j-1} C_{j-1} m^j k^j}{(m^2-1)^{2j-1}} + \sum A_s k^s,$$

where

$$A_s = \frac{(-1)^s m^s}{s(m^2-1)^{2s}} \left((m^2-1)\binom{2s-2}{s-1} + \sum_{i=2}^{s} \frac{(m^2-1)^i}{(i-1)m^{2(i-1)}} \binom{2s-i-1}{s-1} \right).$$

Therefore

$$G_1 = (-1)^m m! (\ln m^2) \left[\sum_{j\ge m} \left\{ {j \atop m} \right\} \frac{(-1)^j m^j}{(m^2-1)^{2j-1}} + \sum_{s\ge m} A_s \left\{ {s \atop m} \right\} \right],$$

whence it follows that

$$|G_1| \le \ln(m^2) \left[\sum_{j=m}^{\infty} \frac{(2m)^{2j}}{(m^2-1)^{2j-1}} + \sum_{s\ge m} \frac{m^s m^2 2^s}{s(m^2-1)^{2s}} \cdot m^s \right] \lesssim \frac{\ln(m^2)}{m^{2m-2}},$$

which is negligible compared with G_2. We conclude that $G_1 + G_2 \ne 0$. □

REFERENCES

[1] 1-*Wasserstein Distance vs. Total Variation Distance*, https://mathoverflow.net/questions/194803/1-wasserstein-distance-v-s-total-variation-distance, 25 January 2015 (accessed 1 January 2019).

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

[3] O. Alter, P. O. Brown, and D. Botstein, *Singular value decomposition for genome-wide expression data processing and modeling*, Proc. Natl. Acad. Sci. USA, 97 (2000), pp. 10101–10106.

[4] A. Andoni, *Nearest neighbor search in high-dimensional spaces*, presented at the Barriers in Computational Complexity Workshop II, 2010, http://www.mit.edu/~andoni/nns-barriers.pdf.

[5] A. Andoni, R. Krauthgamer, and K. Onak, *Streaming algorithms via precision sampling*, in Proceedings of the 52nd IEEE Symposium on Foundations of Computer Science (FOCS), 2011, pp. 363–372.

[6] A. Andoni, R. Krauthgamer, and I. P. Razenshteyn, *Sketching and embedding are equivalent for norms*, in Proceedings of the 47th ACM Symposium on Theory of Computing (STOC), 2015, pp. 479–488.

[7] A. Andoni and H. L. Nguyễn, *Eigenvalues of a matrix in the streaming model*, in Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '13), 2013, pp. 1729–1737, https://doi.org/10.1137/1.9781611973105.124.

[8] A. Andoni, H. L. Nguyen, Y. Polyanskiy, and Y. Wu, *Tight lower bound for linear sketches of moments*, in Proceedings of the 40th International Colloquium on Automata, Languages, and Programming (ICALP 2013), Part I, Riga, Latvia, F. V. Fomin, R. Freivalds, M. Kwiatkowska, and D. Peleg, eds., Springer, Berlin, 2013, pp. 25–32.

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

[10] S. Assadi, S. Khanna, and Y. Li, *On estimating maximum matching size in graph streams*, in Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '17), 2017, pp. 1723–1742, https://doi.org/10.1137/1.9781611974782.113.

[11] 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 Symposium on Theory of Computing (STOC), 2004, pp. 128–137.

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

[13] Ş. B. Bozkurt and D. Bozkurt, *On the sum of powers of normalized Laplacian eigenvalues of graphs*, MATCH Commun. Math. Comput. Chem., 68 (2012), pp. 817–930.

[14] V. Braverman, S. R. Chestnut, R. Krauthgamer, and L. F. Yang, *Sketches for Matrix Norms: Faster, Smaller and More General*, preprint, https://arxiv.org/abs/1609.05885v1, 2016.

[15] 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 Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM), LIPIcs 28, K. Jansen et al., eds., Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2014, pp. 531–544.

[16] M. Bury and C. Schwiegelshohn, *Sublinear estimation of weighted matchings in dynamic data streams*, in Proceedings of Algorithms–ESA 2015, 23rd Annual European Symposium, Patras, Greece, Lecture Notes in Comput. Sci., 9294, Springer, Berlin, 2015, pp. 263–274.

[17] E. J. Candès and B. Recht, *Exact matrix completion via convex optimization*, Comm. ACM, 55 (2012), pp. 111–119.

[18] 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 Annual IEEE Conference on Computational Complexity, 2003, pp. 107–117.

[19] M. Charikar, K. Chen, and M. Farach-Colton, *Finding frequent items in data streams*, in Proceedings of the 29th International Colloquium on Automata, Languages and Programming (ICALP), Springer-Verlag, Berlin, 2002, pp. 693–703.

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

[21] S. Chatterjee and E. Meckes, *Multivariate normal approximation using exchangeable pairs*, ALEA Lat. Am. J. Probab. Math. Stat., 4 (2008), pp. 257–283.

[22] H. Y. Cheung, L. C. Lau, and K. M. Leung, *Graph connectivities, network coding, and expander graphs*, in Proceedings of the 52nd Annual IEEE Symposium on Foundations of Computer Science (FOCS '11), 2011, pp. 190–199.

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

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

[25] M. S. Crouch and A. McGregor, *Periodicity and cyclic shifts via linear sketches*, in Proceedings of Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques: 14th International Workshop (APPROX 2011) and 15th International Workshop (RANDOM 2011), Princeton, NJ, L. A. Goldberg, K. Jansen, R. Ravi, and J. D. P. Rolim, eds., Springer, Berlin, 2011, pp. 158–170.

[26] I. Csiszár and P. C. Shields, *Information theory and statistics: A tutorial*, Commun. Inf. Theory, 1 (2004), pp. 417–528.

[27] W. E. Deming and C. G. Colcord, *The minimum in the gamma function*, Nature, 135 (1935), p. 917.

[28] A. Deshpande, M. Tulsiani, and N. K. Vishnoi, *Algorithms and hardness for subspace approximation*, in Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '11), 2011, pp. 482–496, https://doi.org/10.1137/1.9781611973082.39.

[29] X. V. Doan and S. Vavasis, *Finding the largest low-rank clusters with Ky Fan 2-k-norm and ℓ_1-norm*, SIAM J. Optim., 26 (2016), pp. 274–312, https://doi.org/10.1137/140962097.

[30] K. Driver and K. Jordan, *Zeroes of the hypergeometric polynomials $F(-n, b; c; z)$*, in Algorithms for Approximations IV: 2001 International Symposium, Huddersfield, UK, 2002, pp. 436–445.

[31] P. Flajolet and G. N. Martin, *Probabilistic counting*, in Proceedings of the 24th Annual IEEE Symposium on Foundations of Computer Science (FOCS), 1983, pp. 76–82.

[32] P. Flajolet and G. N. Martin, *Probabilistic counting algorithms for data base applications*, J. Comput. System Sci., 31 (1985), pp. 182–209.

[33] S. Ganguly, *A Lower Bound for Estimating High Moments of a Data Stream*, preprint, https://arxiv.org/abs/1201.0253, 2011.

[34] S. Ganguly, *Taylor polynomial estimator for estimating frequency moments*, in Proceedings of Automata, Languages, and Programming: 42nd International Colloquium (ICALP 2015), Part II, Kyoto, Japan, M. M. Halldórsson et al., eds., Springer, Berlin, 2015, pp. 542–553.

[35] 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 Annual ACM Symposium on Theory of Computing (STOC), 2007, pp. 516–525.

[36] M. Gavish and D. L. Donoho, *Optimal Shrinkage of Singular Values*, Tech. report 2014-08, Department of Statistics, Stanford University, 2014.

[37] C. Gentry, *Fully homomorphic encryption using ideal lattices*, in Proceedings of the 41st ACM Symposium on Theory of Computing (STOC), 2009, pp. 169–178.

[38] F. Götze and A. Tikhomirov, *Rate of convergence in probability to the Marchenko-Pastur law*, Bernoulli, 10 (2004), pp. 503–548, http://www.jstor.org/stable/3318771.

[39] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete Mathematics: A Foundation for Computer Science*, 2nd ed., Addison-Wesley Longman Publishing Co., Inc., Boston, MA, 1994.

[40] S. Guha, P. Indyk, and A. McGregor, *Sketching information divergences*, Machine Learning, 72 (2008), pp. 5–19.

[41] M. Hardt, K. Ligett, and F. McSherry, *A simple and practical algorithm for differentially private data release*, in Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS '12), Curran Associates Inc., 2012, pp. 2339–2347.

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

[43] P. Indyk and A. McGregor, *Declaring independence via the sketching of sketches*, in Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '08), Philadelphia, 2008, pp. 737–745.

[44] P. Indyk and D. P. Woodruff, *Optimal approximations of the frequency moments of data streams*, in Proceedings of the 37th Annual ACM Symposium on Theory of Computing (STOC0), 2005, pp. 202–208.

[45] Y. Ingster and I. A. Suslina, *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*, Springer, Berlin, 2002.

[46] T. S. Jayram, *Hellinger strikes back: A note on the multi-party information complexity of AND*, in Proceedings of Approximation, Randomization, and Combinatorial Optimization (RANDOM/APPROX), Lecture Notes in Comput. Sci. 5687, Springer, Berlin, 2009, pp. 562–573.

[47] 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 Conference, 2011, pp. 49–58.

[48] D. M. Kane, J. Nelson, E. Porat, and D. P. Woodruff, *Fast moment estimation in data streams in optimal space*, in Proceedings of the 43rd Annual ACM Symposium on Theory of Computing (STOC '11), 2011, pp. 745–754.

[49] D. M. Kane, J. Nelson, and D. P. Woodruff, *On the exact space complexity of sketching and streaming small norms*, in Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '10), 2010, pp. 1161–1178, https://doi.org/10.1137/1.9781611973075.93.

[50] A. Khetan and S. Oh, *Spectrum estimation from a few entries*, J. Mach. Learn. Res., 20 (2019), pp. 1–55.

[51] W. Kong and G. Valiant, *Spectrum estimation from samples*, Ann. Statist., 45 (2017), pp. 2218–2247.

[52] R. Krauthgamer and O. Sasson, *Property testing of data dimensionality*, in Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '03), 2003, pp. 18–27.

[53] B. Laurent and P. Massart, *Adaptive estimation of a quadratic functional by model selection*, Ann. Statist., 28 (2000), pp. 1302–1338.

[54] C. Li and G. Miklau, *Optimal error of query sets under the differentially-private matrix mechanism*, in Proceedings of the 16th International Conference on Database Theory, (ICDT '13), ACM, New York, 2013, pp. 272–283.

[55] Y. Li, H. L. Nguyen, and D. P. Woodruff, *Turnstile streaming algorithms might as well be linear sketches*, in Proceedings of the 2014 ACM Symposium on Theory of Computing (STOC '14), 2014, pp. 174–183.

[56] Y. Li and D. P. Woodruff, *A tight lower bound for high frequency moment estimation with small error*, in Proceedings of Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques: 16th International Workshop (APPROX 2013) and 17th International Workshop (RANDOM 2013), Berkeley, CA, P. Raghavendra, S. Raskhodnikova, K. Jansen, and J. D. P. Rolim, eds., Springer, Berlin, 2013, pp. 623–638.

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

[58] V. A. Marčenko and L. A. Pastur, *Distribution of eigenvalues in certain sets of random matrices*, Mat. Sb. (N.S.), 72 (114) (1967), pp. 507–536.

[59] X. Meng and M. W. Mahoney, *Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression*, in Proceedings of the 2013 ACM Symposium on Theory of Computing (STOC '13), 2013, pp. 91–100.

[60] M. Monemizadeh and D. P. Woodruff, *1-pass relative-error L_p-sampling with applications*, in Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '10), 2010, pp. 1143–1160, https://doi.org/10.1137/1.9781611973075.92.

[61] J. Nelson and H. L. Nguyen, *OSNAP: Faster Numerical Linear Algebra Algorithms via Sparser Subspace Embeddings*, preprint, https://arxiv.org/abs/1211.1002, 2012.

[62] E. Price and D. P. Woodruff, *Applications of the Shannon-Hartley theorem to data streams and sparse recovery*, in Proceedings of the 2012 IEEE International Symposium on Information Theory Proceedings (ISIT), 2012, pp. 2446–2450.

[63] I. Razenshteyn, *Personal communication*, 2015.

[64] R.-D. Reiss, *Approximate Distributions of Order Statistics: With Applications to Nonparametric Statistics*, Springer Texts in Statistics, Springer-Verlag, New York, 2008.

[65] M. Rudelson and R. Vershynin, *The Littlewood-Offord problem and invertibility of random matrices*, Adv. Math., 218 (2008), pp. 600–633.

[66] T. Sarlós, *Improved approximation algorithms for large matrices via random projections*, in Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS), 2006, pp. 143–152.

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

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

[69] E. Verbin and W. Yu, *The streaming complexity of cycle counting, sorting by reversals, and other problems*, in Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '11), 2011, pp. 11–25, https://doi.org/10.1137/1.9781611973082.2.

[70] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, in Compressed Sensing: Theory and Applications, Y. C. Eldar and G. Kutyniok, eds., Cambridge University Press, Cambridge, UK, 2011.

[71] Wolfram Research, *Generalized Hypergeometric Function $_3F_2$: Integral Representation*, http://functions.wolfram.com/HypergeometricFunctions/Hypergeometric3F2/07/01/01/0001/, 29 October 2001 (accessed 29 March 2015).

[72] D. Xia, *Optimal Schatten-q and Ky-Fan-k Norm Rate of Low Rank Matrix Estimation*, preprint, https://arxiv.org/abs/1403.6499, 2014.

[73] Z. Zhang, *Parameter estimation techniques: A tutorial with application to conic fitting*, Image Vis. Comput., 15 (1997), pp. 59–76.

[74] B. Zhou, *On sum of powers of the Laplacian eigenvalues of graphs*, Linear Algebra Appl., 429 (2008), pp. 2239–2246.