On Sketching Matrix Norms and the Top Singular Vector

Yi Li
University of Michigan, Ann Arbor
leeyi@umich.edu

Huy L. Nguyễn
Princeton University
hlnguyen@princeton.edu

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

Abstract

Sketching is a prominent algorithmic tool for processing large data. In this paper, we study the problem of sketching *matrix norms*. We consider two sketching models. The first is bilinear sketching, in which 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 $\alpha \geq 1$ with constant probability, where $\|A\|_p$ is a matrix norm. The second is general linear sketching, in which there is 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 α.

We study some of the most frequently occurring matrix norms, which correspond to Schatten p-norms for $p \in \{0, 1, 2, \infty\}$. The p-th Schatten norm of a rank-r matrix A is defined to be $\|A\|_p = (\sum_{i=1}^r \sigma_i^p)^{1/p}$, where $\sigma_1, \ldots, \sigma_r$ are the 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, Frobenius, and operator norms, respectively. For bilinear sketches we show:

1. For $p = \infty$ any sketch must have $r \cdot s = \Omega(n^2/\alpha^4)$ dimensions. This matches an upper bound of Andoni and Nguyen (SODA, 2013), and implies one cannot approximate the top right singular vector v of A by a vector v' with $\|v' - v\|_2 \leq \frac{1}{2}$ with $r \cdot s = \tilde{o}(n^2)$.

2. For $p \in \{0, 1\}$ and constant α, any sketch must have $r \cdot s \geq n^{1-\epsilon}$ dimensions, for arbitrarily small constant $\epsilon > 0$.

3. For even integers $p \geq 2$, we give a sketch with $r \cdot s = O(n^{2-4/p}\epsilon^{-2})$ dimensions for obtaining a $(1 + \epsilon)$-approximation. This is optimal up to logarithmic factors, and is the first general subquadratic upper bound for sketching the Schatten norms.

For general linear sketches our results, though not optimal, are qualitatively similar, showing that for $p = \infty$, $k = \Omega(n^{3/2}/\alpha^4)$ and for $p \in \{0, 1\}$, $k = \Omega(\sqrt{n})$. These 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.

1 Introduction

Sketching is an algorithmic tool for handling big data. A linear skech is a distribution over $k \times n$ matrices S, $k \ll n$, so that for any fixed vector v, one can approximate a function $f(v)$ from Sv with high probability. The goal is to minimize the dimension k.

The sketch Sv thus provides a compression of v, and is useful for compressed sensing and for achieving low-communication protocols in distributed models. One can perform complex procedures on the sketch which would be too expensive to perform on v itself, and this has led to the fastest known approximation algorithms for fundamental problems in numerical linear algebra and nearest neighbor search. Sketching is also the only technique known for processing data in the turnstile model of data streams [22, 38], in which there is an underlying vector v initialized to 0^n which undergoes a long sequence of additive positive and negative updates of the form $v_i \leftarrow v_i + \Delta$ to its coordinates v_i. Given an update of the form $v_i \leftarrow v_i + \Delta$, resulting in a new vector $v' = v + \Delta \cdot e_i$, we can add $S(\Delta \cdot e_i)$ to Sv to obtain Lv' (here e_i is the i-th standard unit vector).

A well-studied problem in the sketching literature is approximating the frequency moments $F_p(v)$, which is equivalent to estimating the p-norms $\|v\|_p = (\sum_{i=1}^n |v_i|^p)^{1/p}$, for $p \in [0, \infty]$. This problem was introduced by Alon, Matias, and Szegedy [1], and many insights in information complexity [6, 9] and sketching [1, 3, 7, 24] were made on the path to obtaining optimal bounds. In addition to being of theoretical interest, the problems have several applications. The value $\|v\|_0$, by continuity, is equal to the support size of x, also known as the number of distinct elements [16, 17]. The norm $\|v\|_1$ is the Manhattan norm, which is a robust measure of distance and is proportional to the variation distance between distributions [19, 21, 28]. The Euclidean distance $\|v\|_2$ is important in linear algebra problems [44], and corresponds to the self-join size in databases [1].

Often one wishes to find or approximate the largest coordinates of x, known as the heavy hitters [10, 13], and $\|v\|_\infty$ is defined, by continuity, to equal $\max_i |v_i|$.

In this paper, we are interested in the analogous problem of sketching *matrix norms*. We study some of the most frequently occurring matrix norms, which correspond to Schatten p-norms for $p \in \{0, 1, 2, \infty\}$. The p-th Schatten norm of an $n \times n$ rank-r matrix A is defined to be $\|A\|_p = (\sum_{i=1}^r \sigma_i^p)^{1/p}$, where $\sigma_1, \ldots, \sigma_r$ are the 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 [11] for graph applications for $p = 0$, to differential privacy [20, 33] and non-convex optimization [8, 15] for $p = 1$, and to the survey on numerical linear algebra for $p \in \{2, \infty\}$ [35].

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, then build a lookup table for all possible sketches to support fast query time. In his talks at the Barriers in Computational Complexity II workshop and Mathematical Foundations of Computer Science conference, Alexandr Andoni states that a goal would be to design a NNS data structure for the Schatten norms, e.g., the trace or Schatten 1-norm (slide 31 of [2]). If a sketch for a norm has small size, then building a table lookup is feasible.

Sketching Model. We give the first formal study of the sketching complexity of the Schatten-p norms. A first question is what does it mean to have a linear sketch of a matrix, instead of a vector as is typically studied. We consider two sketching models.

The first model we consider is *bilinear sketching*, in which 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 $\alpha \geq 1$ 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 [4, 14, 23], 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.

The second model is more general, which we dub *general linear sketching*, and interprets the $n \times n$ matrix A as a vector in $\mathbb{R}^{n^2}$. 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 α with constant probability. The goal is to minimize k.

Previous Results. Somewhat surprisingly the only known $o(n^2)$ upper bound for either model is for $p = 2$, in which case one can achieve a bilinear sketch with $r \cdot s = O(1)$ [23]. Moreover, the only lower bounds known were those for estimating the p-norm of a vector v, obtained for $p > 2$ by setting $A = \text{diag}(v)$ and are of the form $k = \Omega(n^{1-2/p} \log n)$ [5, 34, 40]. We note that the bit complexity lower bounds of [6, 9, 43] do not apply, since a single linear sketch $(1, 1/M, 1/M^2, 1/M^3, \ldots, 1/M^{n-1})^T v$ is enough to recover v if its entries are bounded by M. The sketching model thus gives a meaningful measure of complexity in the real RAM model.

Thus, it was not even known if a sketching dimension of $r \cdot s = O(1)$ 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.

Our Results. We summarize our results for the two sketching models in Table 1. We note that, prior to our work, for all $p \notin \{2, \infty\}$, all upper bounds in the table were a trivial $O(n^2)$ 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 bilinear 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 [4]. 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 $O(n^{2-4/p})$ 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 semi-definite (PSD). For the rank, $p = 0$, we prove an $\Omega(n^2)$ lower bound, showing that no non-trivial sketching is possible. Finally for $p = 1$, we prove an $n^{1-\varepsilon}$ lower bound for arbitrarily small constant $\varepsilon > 0$. Note that our bounds are optimal in several cases, e.g., for $p = \infty$, for even integers $p > 2$, and for $p = 0$.

For the general sketching model, we show the following. For the spectral norm, our bound is $\Omega(n^{3/2}/\alpha^3)$, which although is a bit weaker than the bilinear case, is super-linear in n. It implies, for example, that any general sketching algorithm for approximating the top right (or left) singular vector v of A by a vector v' with $\|v - v'\|_2 \leq \frac{1}{2}$ requires $k = \Omega(n^{3/2})$. Indeed, otherwise, with a sketch of $O(n)$ dimensions one could store gA, for a random Gaussian vector g, and together with v' obtain a constant approximation to $\|A\|_\infty$, which our lower bound rules out. This bound naturally gen-

	Bilinear sketches		General Sketches	
Schatten p-norm	Lower Bound	Upper Bound	Lower Bound	Upper Bound
$p = \infty$	$\Omega(n^2/\alpha^4)$	$O(n^2/\alpha^4)$ [4]	$\Omega(n^{3/2}/\alpha^3)$	$O(n^2/\alpha^4)$ [4]
$p > 2$	$\Omega(n^{2-4/p})$	$O(n^{2-4/p})$ even p	$\Omega(n^{(3/2)(1-2/p)})$	$O(n^{2-4/p})$ even p
$p = 0$	$\Omega(n^2)$	$O(n^2)$	$\Omega(\sqrt{n})$	$O(n^2)$
$p = 1$	$n^{1-\varepsilon}$ for any $\varepsilon > 0$	$O(n^2)$	$\Omega(\sqrt{n})$	$O(n^2)$

Table 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 $p = \infty$, $\alpha \geq 1$ is the desired approximation factor.

eralizes to an $\Omega(n^{(3/2)(1-2/p)})$ bound for $p > 2$. For $p \in \{0, 1\}$ our bound is now a weaker $\Omega(\sqrt{n})$. However, it is the first super-constant lower bound for rank and Schatten-1 norm, which in particular rules out a naïve table lookup solution to the NNS problem, addressing a question of Andoni.

Our Techniques for Bilinear Sketches. A standard technique in proving lower bounds is Yao's minimax principle which implies if there exists a distribution on sketches that succeeds on all $n \times n$ inputs matrices A with large probability, then for any distribution $\mathcal{L}$ on inputs A, there is a fixed pair S and T of $r \times n$ and $n \times s$ matrices, respectively, which succeeds with large probability over $A \sim \mathcal{L}$. Moreover, we can assume the rows of S are orthonormal, as well as the columns of T. This is because, given SAT, we can compute $USATV$, where U and V are arbitrary invertible $r \times r$ and $s \times s$ matrices, respectively. 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 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, then 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. This observation already suffices to get an $\Omega(\sqrt{n})$ bound on $r \cdot s$ for all Schatten p-norms for a fixed constant $p \neq 2$, using a result of Jiang [26] which shows that square $o(\sqrt{n}) \times o(\sqrt{n})$ submatrices of an $n \times n$ matrix of i.i.d. Gaussians and an $n \times n$ orthonormal matrix have $o(1)$ variation distance. Note that both distributions are rotationally invariant and by the Marčenko-Pastur Law have constant factor difference in Schatten-p norm w.h.p. We slightly generalize Jiang's proof to show the variation distance is $o(1)$ for any $r \times s$ submatrix of these distributions provided $r \cdot s < n^{1-\varepsilon}$ for arbitrarily small $\varepsilon > 0$.

For our $\Omega(n^2)$ bound for $p = 0$ and $\Omega(n^{2-4/p})$ bound for $p > 2$, we propose the following rotationally-invariant distributions with constant factor gap in Schatten norm:

- For $p = 0$, $\mathcal{L}_1 = UV^T$ 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 $\gamma \leq 1/\text{poly}(n)$.
- For $p > 2$, $\mathcal{L}_1 = G$ for an $n \times n$ i.i.d. Gaussian matrix, while $\mathcal{L}_2 = G + \frac{1}{n^{1/2-1/p}} uv^T$ for the same G and random n-dimensional Gaussian vectors u and v.

The major difficulty is bounding $d_{TV}(\mathcal{L}_1', \mathcal{L}_2')$. For $p > 2$ this amounts to distinguishing g from $g + h$, where g is an $r \times s$ matrix of Gaussians and h is a *random* $r \times s$ matrix with (i, j)-th entry equal to $u_i v_j$. The fact that h is random makes the probability density function of $g + h$ intractable. Moreover, for each fixed h, the variation distance of g and $g+h$ is much larger than for a random h, and the best lower bound we could obtain by fixing h is $\Omega(n^{1-2/p})$ which would just match the vector lower bound (up to a $\log n$ factor). Instead of bounding $d_{TV}(\mathcal{L}_1', \mathcal{L}_2')$ directly, we bound the χ^2-divergence of $\mathcal{L}_1'$ and $\mathcal{L}_2'$, which if $o(1)$ implies $d_{TV}(\mathcal{L}_1', \mathcal{L}_2') = o(1)$. This idea was previously used in the context of sketching p-norms of vectors [5], improving the previous $\Omega(n^{1-2/p})$ bound to $\Omega(n^{1-2/p} \log n)$. Surprisingly, for the Schatten p-norm, it improves a simple $\Omega(n^{1-2/p})$ bound by a quadratic factor to $\Omega(n^{2-4/p})$, which can similarly be used to show an $\Omega(n^2/\alpha^4)$ bound for α-approximating the spectral norm. One caveat is that if we were to directly compute the χ^2-divergence between $\mathcal{L}_1'$ and $\mathcal{L}_2'$, it would be infinite once $r \cdot s \geq n^{1-2/p}$. We fix this by conditioning $\mathcal{L}_1'$ and $\mathcal{L}_2'$ on a constant probability event, resulting in distributions $\widetilde{\mathcal{L}}_1'$ and $\widetilde{\mathcal{L}}_2'$ for which the χ^2-divergence is small.

For $p = 0$, the problem amounts to distinguishing an $r \times c$ 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. We instead provide an algorithmic proof to bound the variation distance. See Theorem 3.5 for details. The proof also works for arbitrary Q of size $O(n^2)$, implying a lower bound of $\Omega(n^2)$ to decide if an $n \times n$ matrix is of rank at most $n/2$ or ε-far from rank $n/2$ (for constant ε), showing an algorithm of Krauthgamer and Sasson is optimal [29].

Our Algorithm. Due to these negative results, a natural question is whether non-trivial sketching is possible for any Schatten p-norm, other than the Frobenius norm. To show 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 form a cycle $(i_1, i_2, i_2, i_3, i_3, , ..., i_{p/2}, i_{p/2}, i_1)$, 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 these 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.

Our Techniques for General Linear Sketches: We follow the same framework for bilinear sketches. The crucial difference is that since the input A is now viewed as an n^2-dimensional vector, we are not able to design two rotationally invariant distributions $\mathcal{L}_1$ and $\mathcal{L}_2$, since unlike rotating A in $\mathbb{R}^n$, rotating A in $\mathbb{R}^{n^2}$ does not preserve its Schatten p-norm. Fortunately, for both of our lower bounds ($p > 2$) and ($p \in \{0, 1\}$) we can choose $\mathcal{L}_1$ and $\mathcal{L}_2$ so that $\mathcal{L}'_1$ is the distribution of a k-dimensional vector g of i.i.d. Gaussians. For $p > 2$, $\mathcal{L}'_2$ 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 u and v and where L^i is the i-th row of the the sketching matrix L, viewed as an $n \times n$ matrix. We again use the χ^2-divergence to bound $d_{TV}(\mathcal{L}'_1, \mathcal{L}'_2)$ with appropriate conditioning. In this case the problem reduces to finding tail bounds for Gaussian chaoses of degree 4, namely, sums of the form $\sum_{a,b,c,d} A_{a,b,c,d}\, u_a u'_b v_c v'_d$ for a 4th-order array P of n^4 coefficients and independent n-dimensional Gaussian vectors u, u', v, v'. We use a tail bound of Latała [30], which generalizes the more familiar Hanson-Wright inequality for second order arrays P.

For $p \in \{0, 1\}$ we look at distinguishing an $n \times n$ Gaussian matrix G from a matrix $(G', G'M)$, where G' is an $n \times n/2$ Gaussian random matrix and M is a random $n/2 \times n/2$ orthogonal matrix. For all constants $p \neq 2$, the Schatten p-norms differ by a constant factor in the two cases. 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 $Z_i = \langle A^i, G' \rangle + \langle B^i, G'M \rangle$ and each L^i is written as the adjoined matrix (A^i, B^i) 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 M, the Frobenius norm $\|P\|_F$ is small enough to give us an $k = \Omega(\sqrt{n})$ bound, and so it suffices to fix M with this property. One may worry that fixing M reduces the variation distance—in this case one can show that with $k = O(\sqrt{n})$, distributions $\mathcal{L}'_1$ and $\mathcal{L}'_2$ already have constant variation distance.

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

Open Question 1: Is it possible that for every odd integer $p < \infty$, 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.

Open Question 2: For general linear sketches our lower bound for the operator norm is $\Omega(n^{3/2}/\alpha^3)$ for α-approximation. Can this be improved to $\Omega(n^2/\alpha^4)$, which would match our lower bound for bilinear sketches and the upper bound of [4]? Using the tightness of Latała's bounds for Gaussian chaoses, this would either require a new conditioning of distributions $\mathcal{L}'_1$ and $\mathcal{L}'_2$, or bounding the variation distance without using the χ^2-divergence.

2 Preliminaries

Notation. Let $\mathbb{R}^{n \times d}$ be the set of $n \times d$ real matrices and O_n the orthogonal group of degree n (*i.e.*, the set of $n \times n$ orthogonal matrices). Let $N(\mu, \Sigma)$ denote the (multi-variate) normal distribution of mean μ and covariance matrix Σ. We write $X \sim \mathcal{D}$ for a random variable X subject to a probability distribution $\mathcal{D}$. 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 random matrices with entries i.i.d. $N(0,1)$.

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

Singular values and Schatten norms. Consider a matrix $A \in \mathbb{R}^{n\times d}$. 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 $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_d(A)$ in decreasing order. 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 $\sigma_1(A), \dots, \sigma_{\min\{n,d\}}(A)$. 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 over $\mathbb{R}^{n\times d}$, called the p-th *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 $\sigma_i(A)^2$ are the eigenvalues of $A^T A$ and thus $\|A\|_F^2 = \operatorname{tr}(A^T A)$. Let $\|A\|$ denote the operator norm of A when treating A as a linear operator from ℓ_2^d to ℓ_2^n. Besides, it holds that $\lim_{p\to\infty} \|A\|_p = \sigma_1(A) = \|A\|$ 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 non-zero singular values, so $\|A\|_p = \|A^T\|_p$ for all p.

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.

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

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. ([45, p90]) *It holds that* $d_{TV}(\mu,\nu) \leq \sqrt{\chi^2(\mu||\nu)}$.

Proposition 2.2. ([25, p97]) *It holds that* $\chi^2(N(0,I_n) * \mu||N(0,I_n)) \leq \mathbb{E}e^{\langle x,x'\rangle} - 1$, *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.

Hanson-Wright Inequality. Suppose that μ is a distribution over $\mathbb{R}$. We say μ is subgaussian if there exists a constant $c > 0$ such that $\Pr_{x\sim\mu}\{|x| > t\} \leq e^{1-ct^2}$ for all $t \geq 0$.

The following form of the Hanson-Wright inequality on the tail bound of a quadratic form is contained in the proof of [41, Theorem 1.1] due to Rudelson and Vershynin.

Theorem 2.1. (Hanson-Wright inequality) *Let* $u = (u_1, \dots, u_n), v = (v_1, \dots, v_n) \in \mathbb{R}^n$ *be random vectors such that* $u_1, \dots, u_n$, $v_1, \dots, v_n$ *are i.i.d. symmetric subgaussian variables. Let* $A \in \mathbb{R}^{n\times n}$ *be a fixed matrix, then*

$$\Pr\{|u^T Av - \mathbb{E}u^T Av| > t\} \leq 2\exp\left(-c\min\left\{\frac{t}{\|A\|}, \frac{t^2}{\|A\|_F^2}\right\}\right)$$

for some constant $c > 0$*, which depends only on the constant of the subgaussian distribution.*

Latała's Tail Bound. Suppose that $g_{i_1}, \dots, g_{i_d}$ are i.i.d. $N(0,1)$ random variables. The following result, due to Latała [30], bounds the tails of Gaussian chaoses $\sum a_{i_1} \cdots a_{i_d} g_{i_1} \cdots g_{i_d}$. The Hanson-Wright inequality above, when restricted to Gaussian random variables, is a special case ($d = 2$) of this tail bound. The proof of Latała's tail bound was later simplified by Lehec [32].

Suppose that $A = (a_{\mathbf{i}})_{1\leq i_1,\dots,i_d\leq n}$ is a finite multi-indexed matrix of order d. For $\mathbf{i} \in [n]^d$ and $I \subseteq [d]$, define $i_I = (i_j)_{j\in I}$. For disjoint nonempty subsets $I_1, \dots, I_k \subseteq [d]$ define

$$\|A\|_{I_1,\dots,I_k} = \sup\left\{\sum_{\mathbf{i}} a_{\mathbf{i}} x^{(1)}_{i_{I_1}} \cdots x^{(k)}_{i_{I_k}} : \right.$$

$$\left.\sum_{i_{I_1}}\left(x^{(1)}_{i_{I_1}}\right)^2 \le 1,\ldots,\sum_{i_{I_k}}\left(x^{(1)}_{i_{I_k}}\right)^2 \le 1\right\}.$$

Also denote by $S(k,d)$ the set of all partitions of $\{1,\ldots,d\}$ into k nonempty disjoint sets $I_1,\ldots,I_k$. It is easy to see that if a partition $\{I_1,\ldots,I_k\}$ is finer than another partition $\{J_1,\ldots,J_\ell\}$, then $\|A\|_{I_1,\ldots,I_k} \le \|A\|_{J_1,\ldots,J_\ell}$.

Theorem 2.2. *For any $t>0$ and $d \ge 2$,*

$$\Pr\left\{\left|\sum_{\mathbf{i}} a_{\mathbf{i}} \prod_{j=1}^{d} g^{(j)}_{i_j}\right| \ge t\right\}$$
$$\le C_d \exp\left\{-c_d \min_{1\le k\le d}\min_{(I_1,\ldots,I_k)\in S(k,d)}\left(\frac{t}{\|A\|_{I_1,\ldots,I_k}}\right)^{\frac{2}{k}}\right\},$$

where $C_d, c_d > 0$ are constants depending only on d.

Distribution of Singular Values We shall need the following two lemmata.

Lemma 2.1. (Marčenko-Pastur Law [36, 18]) *Suppose that X is a $p\times m$ matrix with i.i.d $N(0,1/p)$ entries. Consider the probability distribution $F_X(x)$ associated with the spectrum of X^TX as*

$$F_X(x) = \frac{1}{m}\left|\{i : \lambda_i(X^TX) \le x\}\right|.$$

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

$$\mathbb{E}\sup_x |F_X(x) - G_\gamma(x)| = O(n^{-1/2}).$$

Lemma 2.2. ([46]) *Suppose that $X \sim \mathcal{G}(p,m)$. Then with probability at least $1-e^{-t^2/2}$, it holds that $\sigma_1(X) \le \sqrt{p}+\sqrt{m}+t$.*

3 Lower Bounds for Bilinear Sketches

3.1 The case of $p > 2$ Fix $r_n \le n$ and $s_n \le n$. Let $\mathcal{L}_1 = \mathcal{G}(r_n,s_n)$ and $\mathcal{L}_2$ denote the distribution of the upper-left $r_n \times s_n$ block of $G + 5n^{-1/2+1/p}uv^T$, where $G \sim \mathcal{G}(n,n)$, $u,v \sim N(0,I_n)$ and G,u,v are independent.

Theorem 3.1. *Suppose that $p > 2$, $\zeta \in (0,1)$ and $r_n, s_n \ge 4$. Whenever $r_ns_n \le c\zeta^2 n^{2(1-2/p)}$, it holds that $d_{TV}(\mathcal{L}_1,\mathcal{L}_2) \le \zeta + 0.009$, where $c>0$ is an absolute constant.*

Proof. For simplicity we write r_n and s_n as r and s, respectively. The distribution $\mathcal{L}_1$ is identical to $N(0,I_{rs})$, and the distribution $\mathcal{L}_2$ is a Gaussian mixture $N(z,I_{rs})$ with shifted mean $z \in \mathbb{R}^{rs}$, where

$$z_{ij} = 5n^{-1/2+1/p}u_iv_j \quad 1\le i\le r, 1\le j\le s.$$

For a given $t > 0$, define the truncated Gaussian distribution, denoted by $N_t(0,I_n)$, as the marginal distribution of $N(0,I_n)$ conditioned on the event that $\|x\| \le t$, where $x \sim N(0,I_n)$. Since $\|x\|^2 \sim \chi^2(n)$, the probability $p_n := \Pr\{\|x\| > 2\sqrt{n}\} < 0.004$ by evaluating an integral of the p.d.f. of $\chi^2(n)$ distribution. Consider an auxiliary random vector $\tilde{z} \in \mathbb{R}^{rs}$ defined as

$$\tilde{z}_{ij} = 5n^{-1/2+1/p}\tilde{u}_i\tilde{v}_j \quad 1\le i\le r, 1\le j\le s,$$

where $(\tilde{u}_1,\ldots,\tilde{u}_r)$ is drawn from $N_{2\sqrt{r}}(0,I_r)$ and $(\tilde{v}_1,\ldots,\tilde{v}_s)$ is drawn from $N_{2\sqrt{s}}(0,I_s)$. Define an auxiliary distribution $\widetilde{\mathcal{L}}_2$ as $N(\tilde{z},I_{rs})$. It is not difficult to see that

$$d_{TV}(\mathcal{L}_2,\widetilde{\mathcal{L}}_2) \le \max\left\{\frac{1}{1-p_r-p_s} - 1, p_r+p_s\right\} < 0.009.$$

So we need only bound $d_{TV}(\mathcal{L}_1,\widetilde{\mathcal{L}}_2)$. It follows from Proposition 2.1 and Proposition 2.2 that

$$d_{TV}(\mathcal{L}_1,\widetilde{\mathcal{L}}_2) \le \sqrt{\mathbb{E}e^{\langle\tilde{z}_1,\tilde{z}_2\rangle}-1},$$

where the expectation is taken over independent $\tilde{z}_1$ and $\tilde{z}_2$, which are both identically distributed as $\tilde{z}$. Next we compute $\mathbb{E}e^{\langle\tilde{z}_1,\tilde{z}_2\rangle}$.

$$\begin{aligned}&\mathbb{E}_{\tilde{z}_1,\tilde{z}_2}\exp(\langle\tilde{z}_1,\tilde{z}_2\rangle)\\ &= \mathbb{E}_{\tilde{u},\tilde{v},\tilde{u}',\tilde{v}'}\exp\left(5n^{-1+2/p}\sum\nolimits_{i,j}\tilde{u}_i\tilde{v}_j\tilde{u}'_i\tilde{v}'_j\right)\\ &= \mathbb{E}_{\tilde{u},\tilde{v},\tilde{u}',\tilde{v}'}\exp\left(5n^{-1+2/p}\sum\nolimits_i\tilde{u}_i\tilde{u}'_i\sum\nolimits_j\tilde{v}_j\tilde{v}'_j\right).\end{aligned} \tag{3.2}$$

Now let us bound $\mathbb{E}|\sum_i\tilde{u}\tilde{u}'|^{2k}$. Note that $|\sum_i\tilde{u}\tilde{u}'| \le \|\tilde{u}\|_2\|\tilde{u}'\|_2 \le 4r$. By our assumption that $r\ge 4$ it holds that $r^{2k-1}\ge 4$, so $t^2/r \le t$ for $t \le 4r$. Applying the Hanson-Wright inequality (Theorem 2.1) to the identity matrix I_r, we have that

$$\begin{aligned}\mathbb{E}\left|\sum_i\tilde{u}\tilde{u}'\right|^{2k} &\le \int_0^{4r}\Pr\left\{\left|\sum_i\tilde{u}_i\tilde{u}'_i\right|^{2k} > t\right\}dt\\ &\le 2\int_0^{4r}e^{-c_1t^{1/k}/r}dt \le 2\left(\frac{r}{c_1}\right)^k k!\end{aligned}$$

for some absolute constant $c_1 > 0$, where the integral is evaluated by variable substitution. We continue bounding (3.2) using the Taylor series as below:

$$\begin{aligned}
&\mathbb{E}_{\tilde{z}_1,\tilde{z}_2}\exp(\langle \tilde{z}_1,\tilde{z}_2\rangle)\\
&\le \sum_{k=0}^{\infty} \frac{(5n^{-1+\frac{2}{p}})^{2k}\mathbb{E}|\sum_i \tilde{u}\tilde{u}'|^{2k}\mathbb{E}|\sum_i \tilde{v}\tilde{v}'|^{2k}}{(2k)!}\\
&\le 1+2\sum_{k=1}^{\infty}\left(\frac{5n^{2(-1+\frac{2}{p})}rs}{c_1^2}\right)^k\cdot\frac{(k!)^2}{(2k)!}\\
&\le 1+2\sum_{k=1}^{\infty}\left(\frac{\zeta^2}{3}\right)^k \le 1+\zeta^2,
\end{aligned}$$

provided that $5rs/c_1^2n^{2(1-2/p)} \le \zeta^2/3$. Therefore

$$d_{TV}(\mathcal{L}_1,\mathcal{L}_2) \le (\ln \mathbb{E}e^{\langle \tilde{z}_1,\tilde{z}_2\rangle})^{1/2} + d_{TV}(\mathcal{L}_2,\widetilde{\mathcal{L}}_2) \le \zeta + 0.009,$$

as claimed. The absolute constant c in the statement can be taken as $c = c_1^2/15$. □

It is not difficult to see that $\|G\|_p$ and $\|G + 5n^{1/p-1/2}uv^T\|_p$ differ by a constant factor with high probability. The lower bound on the dimension of the bilinear sketch is immediate, using $\mathcal{L}_1$ and $\mathcal{L}_2$ as the hard pair of distributions and the observations that (i) both distributions are rotationally invariant, and (ii) by increasing the dimension by at most a constant factor, one can assume that $r_n \ge 4$ and $s_n \ge 4$.

Theorem 3.2. *Let $A \in \mathbb{R}^{n\times n}$ and $p > 2$. Suppose that an algorithm takes a bilinear sketch SAT ($S \in \mathbb{R}^{r\times n}$ and $T \in \mathbb{R}^{n\times s}$) and computes Y with $(1-c_p)\|A\|_p^p \le Y \le (1+c_p)\|A\|_p^p$ for any $A \in \mathbb{R}^{n\times n}$ with probability at least $3/4$, where $c_p = (1.2^p-1)/(1.2^p+1)$. Then $rs = \Omega(n^{2(1-2/p)})$.*

Proof. It suffices to show that $\|G\|_p$ and $\|G + 5n^{1/p-1/2}uv^T\|_p$ differ by a constant factor with high probability. Let $X = 5n^{1/p-1/2}uv^T$. Since X is of rank one, the only non-zero singular value $\sigma_1(X) = \|X\|_F \ge 4.9 \cdot n^{1/p+1/2}$ with high probability, since $\|uv^T\|_F^2 \sim (\chi^2(n))^2$, which is tightly concentrated around n^2.

On the other hand, combining Lemma 2.1 and Lemma 2.2 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

$$\|X\|_p^p = (I_p + o(1))\,n, \tag{3.3}$$

where

$$I_p = \int_0^4 x^{\frac{p}{2}} \cdot \frac{\sqrt{(4-x)x}}{2\pi x}dx \le 2^p. \tag{3.4}$$

Hence $\|G\|_p \le 1.1 \cdot I_p^{1/p} n^{1/2+1/p} \le 1.1 \cdot 2 \cdot n^{1/2+1/p}$ with high probability. By the triangle inequality,

$$\begin{aligned}
\|G+X\|_p \ge \|X\|_p - \|G\|_p &\ge (4.9-2.2)n^{1/p+1/2}\\
&\ge 1.2 \cdot 2.2n^{1/p+1/2} \ge 1.2\|G\|_p
\end{aligned}$$

with high probability. □

3.2 General $p > 0$ The following theorem is a generalization of a result of Jiang [26]. Following his notation, we let Z_n denote the distribution of the upper-left $r_n \times s_n$ block of an orthonormal matrix chosen uniformly at random from O_n and $G_n \sim \frac{1}{\sqrt{n}}\mathcal{G}(r_n, s_n)$.

Theorem 3.3. *If $r_n s_n = o(n)$ and $s_n \le r_n \le n^{d/(d+2)}$ as $n \to \infty$ for some integer $d \ge 1$, then $\lim_{n\to\infty} d_{TV}(Z_n, G_n) = 0$.*

Jiang's original result restricts r_n and s_n to $r_n = o(\sqrt{n})$ and $s_n = o(\sqrt{n})$. We follow the general notation and outline in his paper, making a few modifications to remove this restriction. We postpone the proof to Appendix A.

The lower bound on bilinear sketches follows from Theorem 3.3, with the hard pair of rotationally invariant distributions being a Gaussian random matrix versus a random orthonormal matrix. The proof follows from Theorem 3.3 and is postponed to Appendix B.

Theorem 3.4. *Let $A \in \mathbb{R}^{n\times n}$ and $p > 0$. Suppose that an algorithm takes a bilinear sketch SAT ($S \in \mathbb{R}^{r\times n}$ and $T \in \mathbb{R}^{n\times s}$) and computes Y with $(1-c_p)\|A\|_p^p \le Y \le (1+c_p)\|A\|_p^p$ for any $A \in \mathbb{R}^{n\times n}$ with probability at least $\frac{3}{4}$, where $c_p = \frac{|I_p-1|}{2(I_p+1)}$. Then $rs = \Omega(n^{1-\eta})$ for any constant $\eta > 0$.*

3.3 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 3.5. *Let $U, V \sim \mathcal{G}(n,d)$, $G \sim \gamma\mathcal{G}(n,n)$ 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*

$$d_{TV}(\mathcal{L}_1(S), \mathcal{L}_2(S)) \le C|S|\left(n^{-2} + dc^d\right), \tag{3.5}$$

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

Proof. (sketch, see Appendix D 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 $(1 \pm 1/\mathrm{poly}(n))$.

The idea is to start with the row vectors $U_1, \dots, U_n$ of U and $V_1, \dots, V_n$ 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 $U_1', \dots, U_n'$ and $V_1', \dots, V_n'$ 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 for which $\|U_i\|_2 = (1 \pm 1/\text{poly}(n))\|U_i'\|_2$ and $\|V_i\|_2 = (1 \pm 1/\text{poly}(n))\|V_i'\|_2$ 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 $(U_1', \dots, U_n', V_1', \dots, V_n')$ is the same as that evaluated of $(U_1, \dots, U_n, V_1, \dots, V_n)$ up to a $(1 \pm 1/\text{poly}(n))$-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 i-th 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 $(i,j) \in Q$ 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 $d = n/2$ and so $\text{rank}(UV^T) \leq n/2$ while $\text{rank}(G) = n$ with probability 1. Note that both distributions are rotationally invariant, and so the lower bound on bilinear sketches follows immediately.

Theorem 3.6. *Let $A \in \mathbb{R}^{n \times n}$. Suppose that an algorithm takes a bilinear sketch SAT ($S \in \mathbb{R}^{r \times n}$ and $T \in \mathbb{R}^{n \times s}$) and computes Y with $(1 - c_p)\,\text{rank}(A) \leq Y \leq (1 + c_p)\,\text{rank}(A)$ for any $A \in \mathbb{R}^{n \times n}$ with probability at least $3/4$, where $c_p \in (0, 1/3)$ is a constant. Then $rs = \Omega(n^2)$.*

As an aside, given that w.h.p. over $A \sim \mathcal{L}_2$ in Theorem 3.5, 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 non-adaptive 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 an algorithm of Krauthgamer and Sasson is optimal [29].

4 Bilinear Sketch Algorithms

By the Johnson-Lindenstrauss Transform, or in fact, any subspace embedding with near-optimal dimension (which can lead to better time complexity [44, 12, 37, 39]), we can reduce the problem of general matrices to square matrices (see Appendix C for details), and henceforth we shall assume the input matrices are square.

We present a sketching algorithm to compute a $(1+\epsilon)$-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 (arbitrary number of positive and negative additive updates to entries given in an arbitrary order). The algorithm works for arbitrary A when $p \geq 4$ is an even integer.

Algorithm 1 The sketching algorithm for even $p \geq 4$.

```
Input: $n$, $\epsilon > 0$, even integer $p \geq 4$ and $A \in \mathbb{R}^{n \times n}$.
$N \leftarrow \Omega(\epsilon^{-2})$
Let $\{G_i\}$ and $\{H_i\}$ be independent $n^{1-2/p} \times n$ matrices with i.i.d. $N(0,1)$ entries
Maintain each $G_i A H_i^T$, $i = 1, \dots, N$
Compute $Z$ as defined in (4.6)
return $Z$
```

Suppose that $p = 2q$. We define a cycle σ to be an ordered pair of a sequence of length q: $\sigma = ((i_1, \dots, i_q), (j_1, \dots, j_q))$ 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 $i_{k+1} = i_1$. Let C denote the set of cycles. We define

$$Z = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|C|} \sum_{\sigma \in C} (G_i A H_i^T)_\sigma \tag{4.6}$$

for even p.

Theorem 4.1. *With probability $\geq 3/4$, the output Z returned by Algorithm 1 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 $r \cdot s = O_p(\epsilon^{-2} n^{2-4/p})$.*

See Appendix E for the proof. A similar algorithm works for odd p and PSD matrices A. See Appendix F for details.

5 Lower Bounds for General Linear Sketches

5.1 Lower bound for $p > 2$

Let $G \sim \mathcal{G}(n,n)$, $u, v \sim N(0, I_n)$ be independent. Define two distributions: $\mathcal{L}_1 = \mathcal{G}(n,n)$, while $\mathcal{L}_2$ is the distribution induced by $G + 5n^{-\frac{1}{2}+\frac{1}{p}} uv^T$.

A unit linear sketch can be described using an $n \times n$ matrix L. Applying this linear sketch to an $n \times n$ matrix Q results in $\langle L, Q \rangle$. More generally, consider k orthonormal linear sketches (which as was argued earlier, is equivalent to *any* k linear forms, since, given a sketch, one can left-multiply it by an arbitrary matrix to change its row space) corresponding to $\{L_i\}_{i=1}^k$ with $\text{tr}(L_i^T L_j) = \delta_{ij}$.

Let $X \sim \mathcal{L}_2$ and $Z_i = \langle L_i, X \rangle$. We define a distribution $\mathcal{D}_{n,k}$ on $\mathbb{R}^k$ to be the distribution of $(Z_1, \ldots, Z_k)$.

THEOREM 5.1. *Suppose that $p > 2$. For all sufficiently large n, whenever $k \leq cn^{\frac{3}{2}(1-\frac{2}{p})}$, it holds that $d_{TV}(N(0, I_k), \mathcal{D}_{n,k}) \leq 0.24$, where $c > 0$ is an absolute constant.*

Proof. It is clear that $\mathcal{D}_{n,k}$ is a Gaussian mixture with shifted mean

$$\begin{aligned} X_{u,v} &= 5n^{-1/2+1/p}(u^T L_1 v, u^T L_2 v, \ldots, u^T L_k v)^T \\ &=: 5n^{-1/2+1/p} Y_{u,v}. \end{aligned}$$

Without loss of generality we may assume that $k \geq 16$. Consider the event $\mathcal{E}_{u,v} = \{\|Y_{u,v}\|_2 \leq 4\sqrt{k}\}$. Since $\mathbb{E}\|Y_{u,v}\|_2^2 = k$, it follows from Markov's inequality that $\Pr_{u,v}\{\mathcal{E}_{u,v}\} \geq 15/16$. Let $\widetilde{\mathcal{D}}_{n,k}$ be the marginal distribution of $\mathcal{D}_{n,k}$ conditioned on $\mathcal{E}_{u,v}$. Then

$$d_{TV}(\widetilde{\mathcal{D}}_{n,k}, \mathcal{D}_{n,k}) \leq 1/8$$

and it suffices to bound $d_{TV}(N(0, I_n), \widetilde{\mathcal{D}}_{n,k})$. Resorting to χ^2 divergence by invoking Proposition 2.1 and Proposition 2.2, we have that

$$d_{TV}(N(0, I_n), \widetilde{\mathcal{D}}_{n,k}) \leq \sqrt{\mathbb{E}e^{\langle X_{u,v}, X_{u',v'}\rangle} - 1},$$

where $u, v, u', v' \sim N(0, I_n)$ conditioned on $\mathcal{E}_{u,v}$ and $\mathcal{E}_{u',v'}$. We first calculate that

$$\begin{aligned} \langle X_{u,v}, X_{u',v'}\rangle &= c_p^2 n^{-1+\frac{2}{p}} \sum_{a,b,c,d=1}^{n} \sum_i (L^i)_{ab}(L^i)_{cd} u_a u'_b v_c v'_d \\ &=: D \sum_{a,b,c,d} A_{a,b,c,d} u_a u'_b v_c v'_d, \end{aligned}$$

where $D = c_p^2 n^{-1+\frac{2}{p}}$ and $A_{a,b,c,d}$ is an array of order 4 such that

$$A_{a,b,c,d} = \sum_{i=1}^{k} L^i_{ab} L^i_{cd}.$$

We shall compute the partition norms of $A_{a,b,c,d}$ as needed in Latała's tail bound.

Partition of size 1. The only possible partition is $\{1, 2, 3, 4\}$. We have

$$\begin{aligned} \|A\|_{\{1,2,3,4\}} &= \left(\sum_{a,b,c,d} \left(\sum_{i=1}^{k} L^i_{a,b} L^i_{c,d} \right)^2 \right)^{1/2} \\ &= \left(\sum_{a,b,c,d} \sum_{i,j=1}^{k} L^i_{a,b} L^i_{c,d} L^j_{a,b} L^j_{c,d} \right)^{1/2} \\ &= \left(\sum_{a,b,c,d} \sum_{i=1}^{k} (L^i_{a,b})^2 (L^i_{c,d})^2 \right)^{1/2} \\ &= \sqrt{k} \end{aligned}$$

Partition of size 2 and 3. The norms are automatically upper-bounded by $\|A\|_{\{1,2,3,4\}} = \sqrt{k}$.

Partition of size 4. The only partition is $\{1\}, \{2\}, \{3\}, \{4\}$. We have

$$\begin{aligned} \|A\|_{\{1\},\{2\},\{3\},\{4\}} &= \sup_{u,v,u',v' \in \mathbb{S}^{n-1}} \sum_{i=1}^{k} u^T L^i v u'^T L^i v' \\ &\leq \sup_{u,v,u',v'} \frac{1}{2} \left(\sum_{i=1}^{k} \langle uv^T, L^i\rangle^2 + \langle u'v'^T, L^i\rangle^2 \right) \leq 1 \end{aligned}$$

The last inequality follows the fact that uv^T is a unit vector in $\mathbb{R}^{n^2}$ and L^i's are orthonormal vectors in $\mathbb{R}^{n^2}$.

Latała's inequality (Theorem 2.2) states that for $t \in [\sqrt{k}, k^2]$,

$$\begin{aligned} &\Pr\left\{ \left| \textstyle\sum_{a,b,c,d} A_{a,b,c,d} u_a u'_b v_c v'_d \right| > t \right\} \\ &\leq C_1 \exp\left(-c \min\left\{ \frac{t}{\sqrt{k}}, \frac{t^2}{k}, \frac{t^{\frac{2}{3}}}{k^{\frac{1}{3}}}, \sqrt{t} \right\} \right) \\ &\leq C_1 \exp\left(-c \cdot \frac{t^{\frac{2}{3}}}{k^{\frac{1}{3}}} \right) \end{aligned}$$

with no conditions imposed on u, v, u', v'. It follows that

$$\begin{aligned} &\Pr\left\{ |\langle Y_{u,v}, Y_{u',v'}\rangle| > t \middle| \mathcal{E}_{u,v}\mathcal{E}_{u',v'} \right\} \\ &\leq \frac{\Pr\{|\langle Y_{u,v}, Y_{u',v'}| > t\}}{\Pr\{\mathcal{E}_{u',v'}\}\Pr\{\mathcal{E}_{u,v}\}} \\ &\leq C_2 \exp\left(-c \cdot \frac{t^{\frac{2}{3}}}{k^{\frac{1}{3}}} \right), \quad t \in [\sqrt{k}, k^2]. \end{aligned}$$

Note that conditioned on $\mathcal{E}_{u,v}$ and $\mathcal{E}_{u',v'}$,

$$|\langle Y_{u,v}, Y_{u',v'}\rangle| \leq \|Y_{u,v}\|_2 \|Y_{u',v'}\|_2 \leq 16k.$$

Let $\epsilon = 1/8$, then for $t \in [k^{1/2+\epsilon}, 16k]$, it holds that

$$tD - \frac{ct^{2/3}}{k^{1/3}} \leq -\frac{ct^{2/3}}{2k^{1/3}},$$

provided that $k \leq c' n^{\frac{3}{2}(1-\frac{2}{p})}$. Integrating the tail bound gives that

$$\begin{aligned} &\mathbb{E}e^{X_{u,v}, X_{u',v'}} \\ &= 1 + \int_0^{16k} e^{tD} \Pr\{|\langle Y_{u,v}, Y_{u',v'}\rangle| > t\} dt \\ &= 1 + D \int_0^{k^{1/2+\epsilon}} e^{tD} \Pr\{|\langle Y_{u,v}, Y_{u',v'}\rangle| > t\} dt \end{aligned}$$

$$
\begin{aligned}
&+ D\int_{k^{1/2+\epsilon}}^{16k} e^{tD}\Pr\{|\langle Y_{u,v}, Y_{u',v'}\rangle| > t\}dt \\
&\leq 1 + D\int_0^{k^{1/2+\epsilon}} e^{tD}dt + C_2 D\int_{k^{1/2+\epsilon}}^{16k} e^{tD - ct^{2/3}/k^{1/3}}dt \\
&\leq e^{Dk^{1/2+\epsilon}} + C_2 D\int_{k^{1/2+\epsilon}}^{16k} e^{-\frac{ct^{2/3}}{2k^{1/3}}}dt \\
&\leq \exp(Dk^{1/2+\epsilon}) + 16C_2 kD\cdot\exp\left(-\frac{ck^{2\epsilon/3}}{2}\right) \\
&\leq \exp\left(\frac{c_p^2}{n^{(\frac{1}{4}-\frac{3}{2}\epsilon)(1-\frac{2}{p})}}\right) \\
&\qquad + 16C_2 c' c_p^2 n^{\frac{1}{2}(1-\frac{2}{p})}\exp\left(-\frac{cc' n^{\epsilon(1-\frac{2}{p})}}{2}\right) \\
&\leq 1.01
\end{aligned}
$$

when n is large enough. It follows immediately that $d_{TV}(N(0,I_n), \widetilde{\mathcal{D}}_{n,k}) \leq 1/10$ and thus

$$
d_{TV}(N(0,I_n), \mathcal{D}_{n,k}) \leq 1/10 + 1/8 < 0.24.
$$

□

The lower bound on the number of linear sketches follows immediately as a corollary.

THEOREM 5.2. *Let $X \in \mathbb{R}^{n\times n}$ and $p > 2$. Suppose that an algorithm takes k linear sketches of X and computes Y with $(1-c_p)\|X\|_p^p \leq Y \leq (1+c_p)\|X\|_p^p$ for any $X \in \mathbb{R}^{n\times n}$ with probability at least $3/4$, where $c_p = (1.2^p - 1)/(1.2^p + 1)$. Then $k = \Omega(n^{(3/2)(1-2/p)})$.*

5.2 General $p \geq 0$

Consider a random matrix (G, GM), where $G \sim \mathcal{G}(n, n/2)$ and $M \sim O_{n/2}$.

A unit linear sketch can be described using an $n\times n$ matrix $L = (A, B)$, where $A, B \in \mathbb{R}^{n\times(n/2)}$ such that $\|A\|_F^2 + \|B\|_F^2 = 1$. Applying this linear sketch to an $n\times n$ matrix $Q = (Q_1, Q_2)$ (where $Q_1, Q_2 \in \mathbb{R}^{n\times(n/2)}$) results in $\langle A, Q_1\rangle + \langle B, Q_2\rangle$.

More generally, consider k orthonormal linear sketches (which as was argued earlier, is equivalent to *any* k linear forms, since, given a sketch, one can left-multiply it by an arbitrary matrix to change its row space) corresponding to $\{L_i\}_{i=1}^k$ with $\operatorname{tr}(L_i^T L_j) = \delta_{ij}$.

Now we define a probability distribution $\mathcal{D}_{n,k}$ on $\mathbb{R}^k$. For each i write L_i as $L_i = (A^{(i)}, B^{(i)})$. Then by orthonormality, $\langle A^{(i)}, A^{(j)}\rangle + \langle B^{(i)}, B^{(j)}\rangle = \delta_{i,j}$. Define $Z_i = \langle A^{(i)}, G\rangle + \langle B^{(i)}, GM\rangle$ and $\mathcal{D}_{n,k}$ to be the distribution of $(Z_1, \ldots, Z_k)$.

THEOREM 5.3. *Let $\mathcal{D}_{n,k}$ be defined as above and $\zeta \in (0,1)$. Then for $k \leq (\zeta/3)^{3/2}\sqrt{n}$ it holds that*

$$
d_{TV}(\mathcal{D}_{n,k}, N(0, I_k)) \leq \zeta,
$$

where $N(0, I_k)$ is the standard k-dimensional Gaussian distribution.

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 $A^{(i,\sigma)}$, $B^{(i,\sigma)}$ and M, and $g \sim \mathcal{G}(n^2/2, 1)$. Assume that Φ has full row rank (we shall justify this assumption below). Fix Φ (by fixing M). Then $\Phi g \sim N(0, \Phi\Phi^T)$. It is known that ([27, Lemma 22])

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

where $\lambda_1, \ldots, \lambda_k$ are the eigenvalues values of $\Phi\Phi^T$. Write $\Phi\Phi^T = I + P$. Define an event $E = \left\{M : \|P\|_F^2 \leq \frac{12}{\zeta}\cdot\frac{k^2}{n}\right\}$. When E happens, the eigenvalues of P are bounded by $\sqrt{\frac{12}{\zeta}}\cdot\frac{k}{\sqrt{n}} \leq 2/3$. Let $\mu_1, \ldots, \mu_k$ be the eigenvalues of P, then $\lambda_i = 1 + \mu_i$ and $|\mu_i| \leq 2/3$. Hence

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

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

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

as advertised.

Now we show that $\mathbb{E}P_{ij}^2 \leq 4/n$ for all i,j. Suppose that $M = (m_{ij})$. Notice that the r-th row of Φ is

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

Hence by a straightforward calculation, the inner product of r-th and s-th row is

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

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

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

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

$$\begin{aligned} u_{jk}^2 \leq 2 \Big\{ & \Big(\sum\nolimits_i |A_{ik}^{(r)}|^2 \Big) \Big(\sum\nolimits_i |B_{ij}^{(s)}|^2 \Big) \\ & + \Big(\sum\nolimits_i |A_{ik}^{(s)}|^2 \Big) \Big(\sum\nolimits_i |B_{ij}^{(r)}|^2 \Big) \Big\} \end{aligned}$$

and thus

$$\begin{aligned} \|U\|_F^2 &\leq 2 \sum\nolimits_{j,k} \Big\{ \Big(\sum\nolimits_i |A_{ik}^{(r)}|^2 \Big) \Big(\sum\nolimits_i |B_{ij}^{(s)}|^2 \Big) + \\ & \qquad \Big(\sum\nolimits_i |A_{ik}^{(s)}|^2 \Big) \Big(\sum\nolimits_i |B_{ij}^{(r)}|^2 \Big) \Big\} \\ &\leq 2 \Big(\|A^{(r)}\|_F^2 \|B^{(s)}\|_F^2 + \|A^{(s)}\|_F^2 \|B^{(r)}\|_F^2 \Big) \leq 2. \end{aligned}$$

We conclude that

$$\begin{aligned} \mathbb{E}[P_{rs}^2] &= \sum\nolimits_{j,k} u_{jk}^2 \mathbb{E}[m_{kj}^2] + \sum\nolimits_{(j,k)\neq(i,\ell)} u_{jk} u_{i\ell} \mathbb{E}[m_{kj} m_{i\ell}] \\ &= \frac{2}{n} \sum\nolimits_{j,k} u_{jk}^2 = \frac{2\|U\|_F^2}{n} \leq \frac{4}{n}. \end{aligned}$$

This completes the proof. □

Without loss of generality, we can normalize our matrix by a factor of $1/\sqrt{n}$. Let $X \sim \frac{1}{\sqrt{n}}\mathcal{G}(n,n)$ and $Y = (G, GM)$, where $G \sim \mathcal{G}(n, n/2)$ and $M \sim O_{n/2}$. It is not difficult to see that $\|X\|_p^p$ and $\|Y\|_p^p$ differs by a constant with high probability. The following theorem is now an immediate corollary of Theorem 5.3. The proof is postponed to Appendix H.

Theorem 5.4. *Let $X \in \mathbb{R}^{n\times n}$ and $p \geq 0$. Suppose that an algorithm takes k linear sketches of X and computes Y with*

- *(when $p > 0$) $(1 - c_p)\|X\|_p^p \leq Y \leq (1 + c_p)\|X\|_p^p$, or*
- *(when $p = 0$) $(1 - c_p)\|X\|_0 \leq Y \leq (1 + c_p)\|X\|_0$*

for any $X \in \mathbb{R}^{n\times n}$ with probability at least $\frac{3}{4}$, where c_p is some constant depends only on p. Then $k = \Omega(\sqrt{n})$.

References

[1] N. Alon, Y. Matias, and M. Szegedy. The Space Complexity of Approximating the Frequency Moments. *J. Comput. Syst. Sci.*, 58(1):137–147, 1999.

[2] A. Andoni. Nearest neighbor search in high-dimensional spaces. In *the workshop: Barriers in Computational Complexity II*, 2010. http://www.mit.edu/ andoni/nns-barriers.pdf.

[3] A. Andoni, R. Krauthgamer, and K. Onak. Streaming algorithms from precision sampling. In *FOCS*, pages 363–372, 2011.

[4] A. Andoni and H. L. Nguyen. Eigenvalues of a matrix in the streaming model. In *SODA*, 2013.

[5] A. Andoni, H. L. Nguyen, Y. Polyansnkiy, and Y. Yu. Tight lower bound for linear sketches of moments. In *ICALP*, 2013.

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

[7] L. Bhuvanagiri, S. Ganguly, D. Kesh, and C. Saha. Simpler algorithm for estimating frequency moments of data streams. In *SODA*, pages 708–713, 2006.

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

[9] A. Chakrabarti, S. Khot, and X. Sun. Near-optimal lower bounds on the multi-party communication complexity of set disjointness. In *CCC*, 2003.

[10] 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)*, pages 693–703, 2002.

[11] H. Y. Cheung, L. C. Lau, and K. M. Leung. Graph connectivities, network coding, and expander graphs. In *FOCS*, pages 190–199, 2011.

[12] K. L. Clarkson and D. P. Woodruff. Low rank approximation and regression in input sparsity time. In *STOC*, pages 81–90, 2013.

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

[14] M. S. Crouch and A. McGregor. Periodicity and cyclic shifts via linear sketches. In *APPROX-RANDOM*, pages 158–170, 2011.

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

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

[17] P. Flajolet and G. N. Martin. Probabilistic counting algorithms for data base applications. *J. Comput. Syst. Sci.*, 31(2):182–209, 1985.

[18] F. Götze and A. Tikhomirov. Rate of convergence in probability to the marchenko-pastur law. *Bernoulli*, 10(3):pp. 503–548, 2004.

[19] S. Guha, P. Indyk, and A. McGregor. Sketching information divergences. *Machine Learning*, 72(1-2):5–19, 2008.

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

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

[22] P. Indyk. Sketching, streaming and sublinear-space algorithms, 2007. Graduate course notes available at `http://stellar.mit.edu/S/course/6/fa07/6.895/`.

[23] P. Indyk and A. McGregor. Declaring independence via the sketching of sketches. In *SODA*, pages 737–745, 2008.

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

[25] Y. Ingster and I. A. Suslina. *Nonparametric Goodness-of-Fit Testing Under Gaussian Models.* Springer, 1st edition, 2002.

[26] T. Jiang. How many entries of a typical orthogonal matrix can be approximated by independent normals? *The Annals of Probability*, 34(4):pp. 1497–1529, 2006.

[27] A. T. Kalai, A. Moitra, and G. Valiant. Efficiently learning mixtures of two gaussians. In *Proceedings of the 42nd STOC*, pages 553–562, 2010.

[28] 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, pages 745–754, 2011.

[29] R. Krauthgamer and O. Sasson. Property testing of data dimensionality. In *SODA*, pages 18–27, 2003.

[30] R. Latała. Estimates of moments and tails of Gaussian chaoses. *Ann. Probab.*, 34(6):2315–2331, 2006.

[31] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1302–1338, 2000.

[32] J. Lehec. Moments of the Gaussian chaos. In *Séminaire de Probabilités XLIII*, volume 2006 of *Lecture Notes in Math.*, pages 327–340. Springer, Berlin, 2011.

[33] C. Li and G. Miklau. Measuring the achievable error of query sets under differential privacy. *CoRR*, abs/1202.3399, 2012.

[34] Y. Li and D. P. Woodruff. A tight lower bound for high frequency moment estimation with small error. In *RANDOM*, 2013.

[35] M. W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011.

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

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

[38] S. Muthukrishnan. Data Streams: Algorithms and Applications. *Foundations and Trends in Theoretical Computer Science*, 1(2):117–236, 2005.

[39] J. Nelson and H. L. Nguyen. OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings. *CoRR*, abs/1211.1002, 2012.

[40] E. Price and D. P. Woodruff. Applications of the Shannon-Hartley theorem to data streams and sparse recovery. In *ISIT*, pages 2446–2450, 2012.

[41] M. Rudelson and R. Vershynin. Hanson-Wright inequality and sub-gaussian concentration. Manuscript. http://arxiv.org/abs/1306.2872v2.

[42] M. Rudelson and R. Vershynin. The Littlewood-Offord problem and invertibility of random matrices. *Advances in Mathematics*, 218(2):600 – 633, 2008.

[43] M. E. Saks and X. Sun. Space lower bounds for distance approximation in the data stream model. In *STOC*, pages 360–369, 2002.

[44] T. Sarlós. Improved approximation algorithms for large matrices via random projections. In *FOCS*, pages 143–152, 2006.

[45] A. B. Tsybakov. *Introduction to Nonparametric Estimation.* Springer, 1st edition, 2008.

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

Appendices

A Proof of Theorem 3.3

Proof. First we strengthen Lemma 2.6 in [26] and in this proof we shall follow Jiang's notation that the we read the upper-left $p \times q$ block despite the usage of r_n and s_n in the statement of the theorem.

It holds that $f_s'(s,t) = -2/(1-2s-t) = -2 + O(n^{-1/(d+2)})$ and $f_t'(s,t) = -1 + O(n^{1/(d+2)})$. The bounds for second-order derivatives of $f(s,t)$ still hold. Thus

$$B_n = n^2 \iint \ln(1-2s-t)dsdt + \frac{3kq}{2n} + O\left(\frac{1}{n^{1/(d+2)}}\right)$$
$$= n^2 \iint \ln(1-2s-t)dsdt + o(1).$$

Now, we expand the Taylor series into r terms,

$$\left|\ln(1+s+t) - \left\{(s+t) - \frac{(s+t)^2}{2} + \cdots + (-1)^{d+1}\frac{(s+t)^d}{d}\right\}\right| \leq \frac{(s+t)^{d+1}}{d+1}.$$

So

$$\int_0^v \int_0^u \ln(1+s+t)dsdt$$
$$= \sum_{k=1}^{d} \frac{(-1)^{k+1}}{k(k+1)(k+2)}((u+v)^{k+2} - u^{k+2} - v^{k+2}) + O((u+v)^{d+3})$$

as $n \to \infty$. Substituting $u = -(p+2)/n$ and $v = -q/n$ back into two integrals in (2.5), (2.7) becomes

$$\frac{n^2}{2}\int_0^v \int_0^u \ln(1+s+t)dsdt$$

$$= -\frac{1}{2}\sum_{k=1}^{d}\frac{1}{k(k+1)(k+2)}\frac{(p+q+2)^{k+2}-(p+2)^{k+2}-q^{k+2}}{n^k} + O\left(\frac{(p+q+2)^{d+3}}{n^{d+1}}\right)$$

and

$$\frac{n^2}{2}\int_0^v\int_0^{-2/n}\ln(1+s+t)dsdt = -\frac{1}{2}\sum_{k=1}^{d}\frac{1}{k(k+1)(k+2)}\frac{(q+2)^{k+2}-2^{k+2}-q^{k+2}}{n^k} + O\left(\frac{q^{d+3}}{n^{d+1}}\right)$$

Hence

$$B_n = -\frac{1}{2}\sum_{k=1}^{d}\frac{1}{k(k+1)(k+2)}\left(\sum_{i=1}^{k+1}\binom{k+2}{i}\frac{p^i q^{k+2-i}}{n^k}\right)+o(1).$$

Notice that for $i \geq 2$, it holds that $p^i q^{k+2-i}/n^k \leq p^k q^2/n^k \to 0$ as $n \to \infty$. It follows that

$$B_n = -\frac{1}{2}\sum_{k=1}^{d}\frac{p^{k+1}q}{k(k+1)n^k} + o(1).$$

This is our desired form of Lemma 2.6.

Now, following the proof of Lemma 2.7, we expand into more terms

$$\ln\left(1-\frac{x}{n}\right) = -\frac{x}{n}-\frac{x^2}{n^2}-\cdots-\frac{x^{d+1}}{(d+1)n^{d+1}}-\frac{x^{d+2}}{d+2}\cdot\frac{1}{(\xi-n)^{d+2}}.$$

so

$$f(x) = \frac{p+q+1}{2n}x-\frac{n-p-q-1}{4n^2}x-\cdots-\frac{n-p-q-1}{2(d+1)n^{d+1}}x^{d+1} + g_n(x)\frac{x^{d+2}}{n^{d+1}},$$

where

$$g_n(x) = -\frac{n^{d+1}(n-p-q-1)}{2(d+2)(\xi-n)^{d+2}}.$$

It is trivial to see that

$$\sup_{0\leq x\leq\alpha n}|g_n(x)| \leq \frac{1}{(1-\alpha)^{d+2}}.$$

The eigenvalues are bounded by $O((\sqrt{p}+\sqrt{q})^2) = O(n^{1-1/(d+2)})$. Define Ω accordingly. Then

$$\sum_{i=1}^{q} f(\lambda_i) = \frac{p+q+1}{2n}\operatorname{tr}(X^TX) - \sum_{j=2}^{d+1}\frac{n-p-q-1}{2jn^j}\operatorname{tr}((X^TX)^j) + \tilde{g}_n\frac{\operatorname{tr}((X^TX)^{d+2})}{n^{d+2}},$$

where $g_n \in [0,2)$ for n sufficiently large.

We want to show that

$$B_n + \sum_{i=1}^{q} f(\lambda_i) \to 0 \tag{A.1}$$

in probability as $n \to \infty$. First, we look at the last term. It is clear from

$$\operatorname{tr}((X_n^TX_n)^m) = \sum_{i=1}^{q}\sum_{j\in[q]^{m-1}}\sum_{k\in[p]^m} x_{k_1,i}x_{k_1,j_1}x_{k_2,j_1}x_{k_2,j_2} \cdots x_{k_{m-1},j_{m-2}}x_{k_{m-1},j_{m-1}}x_{k_m,j_{m-1}}x_{k_m,i}$$

that $\mathbb{E}\operatorname{tr}((X_n^TX_n)^m) = O(p_n^m q_n)$, where the constant in $O(\cdot)$ notation depends on m but not p and q. For any $\epsilon > 0$,

$$\Pr\left\{\operatorname{tr}((X_n^TX_n)^{r+2} > \epsilon n^{r+2}\right\} \leq \frac{\mathbb{E}[(\operatorname{tr}(X_n^TX_n)^{r+2})^2]}{\epsilon^2 n^{2r+4}} \leq \frac{\mathbb{E}[q\operatorname{tr}(X_n^TX_n)^{2r+4}]}{\epsilon^2 n^{2r+4}} = O\left(\frac{p^{2r+4}q^2}{\epsilon^2 n^{2r+4}}\right) \to 0$$

as $n \to \infty$. So the last term goes to 0 in probability. For the other terms, fix $\epsilon > 0$, we see that

$$\Pr\left\{|\operatorname{tr}((X_n^TX_n)^m) - \mathbb{E}\operatorname{tr}((X_n^TX_n)^m)| \geq \epsilon n^m\right\} \leq \frac{\mathbb{E}[(\operatorname{tr}((X_n^TX_n)^m)^2]}{\epsilon^2 n^{2m}} \leq \frac{\mathbb{E}[q(\operatorname{tr}((X_n^TX_n)^{2m})]}{\epsilon^2 n^{2m}} = O\left(\frac{p^{2m}q^2}{\epsilon^2 n^{2m}}\right) \to 0$$

hence

$$\sum_{i=1}^{q} f(\lambda_i) \to \frac{p+q+1}{2n}\mathbb{E}\operatorname{tr}(X^TX) - \sum_{j=2}^{d+1}\frac{n-p-q-1}{2jn^j}\mathbb{E}\operatorname{tr}((X^TX)^j)$$

in probability. It suffices to show that

$$B_n+\left(\frac{p+q+1}{2n}\mathbb{E}\operatorname{tr}(X^TX)-\sum_{j=2}^{d+1}\frac{n-p-q-1}{2jn^j}\mathbb{E}\operatorname{tr}((X^TX)^j)\right)$$

goes to 0. Rearrange the terms in the bracket as

$$\sum_{j=1}^{d}\left(\frac{p+q+1}{jn^j}\mathbb{E}(\operatorname{tr}(X^TX)^j) - \frac{1}{(j+1)n^j}\mathbb{E}(\operatorname{tr}(X^TX)^{j+1})\right) + \frac{p+q+1}{(d+1)n^{d+1}}\mathbb{E}(\operatorname{tr}(X^TX)^{j+1})$$

The last term goes to 0 as $\mathbb{E}(\operatorname{tr}(X^TX)^{j+1}) = O(p^{j+1}q)$ and $p^{d+1}/n^d \to 0$. Hence it suffices to show that for each k

$$-\frac{p^{k+1}q}{k(k+1)n^k} + \frac{p+q+1}{kn^k}\mathbb{E}(\operatorname{tr}(X^TX)^k) - \frac{1}{(k+1)n^k}\mathbb{E}(\operatorname{tr}(X^TX)^{k+1}) \to 0,$$

or

$$\frac{-p^{k+1}q+(k+1)(p+q+1)\mathbb{E}(\operatorname{tr}(X^TX)^k)-k\mathbb{E}(\operatorname{tr}(X^TX)^{k+1})}{n^k} \to 0,$$

which can be verified easily using that $\mathbb{E}\operatorname{tr}((X_n^TX_n)^m) = p_n^m q_n + o(n^m)$. Therefore (A.1) holds and the rest of the argument in Jiang's paper follows.

B Proof of Theorem 3.4

Proof. Choose A from $\mathcal{D}_0$ with probability 1/2 and from O_n with probability 1/2. Let $b \in \{0,1\}$ indicate which distribution A is chosen from, that is, $b = 0$ when A is chosen from $\mathcal{D}_0$ and $b = 1$ otherwise. Note that both $\mathcal{D}_0$ and $O(n)$ are rotationally invariant, so without loss of generality, we can assume that $S = \begin{pmatrix} I_r & 0 \end{pmatrix}$ and $T = \begin{pmatrix} I_s \\ 0 \end{pmatrix}$. Hence $SAT \sim Z_n$ when $b = 0$ and $SAT \sim G_n$ when $b = 1$. Notice that the singular values of an othorgonal matrix are all 1s. Then with probability $1 - o(1)$ we have that

$$(1-c_p)((I_p+o(1))n \le Y \le (1+c_p)((I_p+o(1))n, \quad b=0$$
$$(1-c_p)n \le Y \le (1+c_p)n, \quad b=1$$

so we can recover b from Y with probability $1 - o(1)$, provided that $c_p \le \frac{|I_p-1|}{2(I_p+1)}$.

Consider the event E that the algorithm's output indicates $b = 1$. Then as in the proof of Theorem 5.4,

$$d_{TV}(D_{n,k}, N(0, I_k)) \ge \frac{1}{2} + o(1).$$

On the other hand, by Theorem 3.3, $d_{TV}(D_{n,k}, N(0, I_k)) = o(1)$ when $rs \le n^{1-\eta}$ for some $\eta > 0$. This contradiction implies that $rs = \Omega(n^{1-\eta})$ for any $\eta > 0$.

C Reduction to Square Matrices for the upper bound

Suppose that $A \in \mathbb{R}^{n\times d}$ $(n > d)$. When $n = O(d/\epsilon^2)$, let $\tilde{A} = (A, 0)$ 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 $O(d/\epsilon^2)$ rows while roughly maintaining the singular values as follows. Call Φ a (d,δ)-subspace embedding matrix if with probability $\ge 1-\delta$ it holds that

$$(1-\epsilon)\|x\| \le \|\Phi x\| \le (1+\epsilon)\|x\|$$

for all x in a fixed d-dimensional subspace. In [44], it is proved that

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

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

In fact we can use more modern subspace embeddings [44, 12, 37, 39] 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 for the subspace embedding transform on singular values, which follows from the min-max principle for singular values.

Lemma C.2. *Let Φ be a (d,δ)-subspace embedding matrix. Then, with probability $\ge 1-\delta$, it holds that $(1-\epsilon)\sigma_i(\Phi A) \le \sigma_i(A) \le (1+\epsilon)\sigma_i(\Phi A)$ for all $1 \le i \le d$.*

Proof. [of Lemma C.2] The min-max principle for singular values says that

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

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 C.1 that with probability $\ge 1-\delta$,

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

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

Let $\tilde{A} = (\Phi A, 0)$ with zero padding so that $\tilde{A}$ is a square matrix of dimension $O(d/\epsilon^2)$. Then by the preceding lemma, with probability $\ge 1-\delta$, $\|\tilde{A}\|_p = \|\Phi A\|_p$ 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.

D Proof of Theorem 3.5

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 new notation, let us denote the marginal distribution of $\mathcal{L}_2$ with fixed G by $\mathcal{L}_2(G)$. Then

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

where $F(\delta) = \{G \in \mathbb{R}^{n\times n} : |G_{p_i,q_i}| \le \delta, \ \forall i = 1, \dots, k\}$ 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 PDF of a Gaussian random variable and that $k \le n^2$, it follows from a union bound that

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

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

$$\begin{aligned}
& d_{TV}(\mathcal{L}_1, \mathcal{L}_2(G)) \qquad \text{(D.4)} \\
&= \sup_{A\subset\mathcal{M}(\mathbb{R}^k)} \left|\Pr_{\mathcal{L}_1}(A) - \Pr_{\mathcal{L}_2(G)}(A)\right| \\
&= \sup_{A\subset\mathcal{M}(\mathbb{R}^k)} \left|\Pr_{\mathcal{L}_1}(A) - \Pr_{\mathcal{L}_1}(A - \xi)\right| \\
&\le \sup_{A\subset\mathcal{M}(\mathbb{R}^k)} \sum_{i=1}^{k} \left|\Pr_{\mathcal{L}_1}(A - \xi^{(i-1)}) - \Pr_{\mathcal{L}_1}(A - \xi^{(i)})\right| \qquad \text{(D.5)}
\end{aligned}$$

where $\mathcal{M}(\mathbb{R}^k)$ denotes the canonical Borel algebra on $\mathbb{R}^k$.

To bound (D.5), 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 $T_t : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d}$ (we defer the construction to the end of the proof) for which the following three properties hold:

1. $((T_tU)(T_tV)^T)_{pq} = (UV)_{pq} + t$ and for all $(p', q') \in S \setminus \{(p, q)\}$, we have $((T_tU)(T_tV)^T)_{p'q'} = (UV^T)_{p'q'}$.
2. $\|U - T_tU\|_F \le t'$, $\|V - T_tV\|_F \le t'$ with probability $1 - O(1/n^2 + dc^d)$, 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. $T_{-t} \cdot T_t = \text{id}$.

Property 3 implies that T_t is bijective. 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 that

$$\begin{aligned}
p(U) &= \frac{1}{(2\pi)^{nd/2}} \exp\left(-\frac{\|U\|_F^2}{2}\right) \qquad \text{(D.6)} \\
p(V) &= \frac{1}{(2\pi)^{nd/2}} \exp\left(-\frac{\|V\|_F^2}{2}\right)
\end{aligned}$$

we have that

$$\begin{aligned}
& \left|\Pr_{\mathcal{L}_1}(A) - \Pr_{\mathcal{L}_1}(A + te_i)\right| \\
&= \left| \int\limits_{E(x)} p(U)p(V)dUdV - \int\limits_{E(x-te_i)} p(U)p(V)dUdV \right| \\
&\le \left| \int\limits_{E_{\text{good}}(A)} p(U)p(V)dUdV - \int\limits_{E_{\text{good}}(A)} p(T_tU)p(T_tV)dUdV \right| \\
&\qquad + \int_{E_{\text{bad}}(A)} p(U)p(V) + p(T_tU)p(T_tV)dUdV \\
&\le \int_{E_{\text{good}}(A)} |p(U) - p(T_tU)|p(V)dUdV \\
&\qquad + \int\limits_{E_{\text{good}}(A)} p(T_tU)|p(V) - p(T_tV)|dUdV \\
&\qquad + O\left(\frac{1}{n^2} + dc^d\right). \qquad \text{(D.7)}
\end{aligned}$$

Using (D.6),

$$|p(U) - p(T_tU)| = p(U) \cdot \left|1 - \exp\left(\frac{\|U\|_F^2 - \|T_tU\|_F^2}{2}\right)\right|. \tag{D.8}$$

Notice that $\|U\|_F^2 \sim \chi^2(nd)$, where $\chi^2(nd)$ denotes the χ-squared distribution with nd degrees of freedom, and so by a tail bound for the χ^2-distribution [31, Lemma 1], $\|U\|_F^2 \le 6nd - t'$ (recall that $t' = 1/n^4$) with probability at least $1 - e^{-nd} \ge 1 - n^{-3}$. 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

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

Using $|1 - e^{|x|}| \le 2|x|$ for $|x| < 1$ and combining with (D.8), we have

$$|p(U) - p(T_tU)| \le p(U) \cdot 12n^{-3}. \tag{D.9}$$

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

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

It then follows that

$$\int_{E_{\text{good}}(A)} |p(U) - p(T_tU)| p(V) dU dV = O\left(\frac{1}{n^3}\right), \tag{D.10}$$

$$\int_{E_{\text{good}}(A)} p(T_tU) |p(V) - p(T_tV)| dU dV = O\left(\frac{1}{n^3}\right) \tag{D.11}$$

Plugging (D.10) and (D.11) into (D.7) yields that

$$\left|\Pr_{\mathcal{L}_1}(A) - \Pr_{\mathcal{L}_1}(A + te_i)\right| = O\left(\frac{1}{n^2} + dc^d\right), \tag{D.12}$$

which, combined with (D.5), (D.2), and (D.3), finally leads to

$$\begin{aligned} d_{TV}(\mathcal{L}_1, \mathcal{L}_2) &\le d_{TV}(\mathcal{L}_1, \mathcal{L}_2(G)) + 2\Pr\{F(\delta)^c\} \\ &= O\left(k(n^{-2} + dc^d)\right). \end{aligned}$$

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

Case 1a. Suppose that the p-th row contains $s \le d$ 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 p-th row.

Then we have (U_i denotes the i-th row of U and V_i the i-th 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

$$(U_i' - U_i) \begin{pmatrix} V_{q_1} & V_{q_2} & \cdots & V_{q_d} \end{pmatrix} = \Delta x \tag{D.13}$$

With probability 1, the matrix $\tilde{V} := (V_{q_1}, V_{q_2}, \dots, V_{q_d})$ has rank d. Hence we can solve $U_i' - U_i$ uniquely, and

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

Using the tail bound on the minimum singular value given in [42], we have $\Pr\{\sigma_d(\tilde{v}) \le \epsilon/\sqrt{d}\} \le C\epsilon + c^d$, where $C > 0$ and $0 < c < 1$ are absolute constants. Choosing $\epsilon = n^{-4}$ and recalling that $d \le n$, we see that with probability at least $1 - Cn^{-4} - c^d$, it holds that $\sigma_d(\tilde{V}) \ge 1/n^{9/2}$ and thus

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

This proves **property 2** of this case. To show **property 3** in this case, we can show something stronger, which is $T_{-t} \circ T_t = \mathrm{id}$, namely, this step is invertible. Indeed, if we replace Δx with $-\Delta x$ in (D.13) the solution $U_i' - U_i$ will be the opposite sign as well.

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

Case 2. Suppose that there are more than d entries read in both the p-th row and the q-th column. Define

$$J = \{i \in [n] : i\text{-th column has} \le d \text{ entries contained in } S\}$$
$$C_r = \{i \in [n] : (r, i) \in S\}, \quad R_c = \{i \in [n] : (i, c) \in S\}$$

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

Note that $|J^c| \le d$ since the total number of entries in S is at most d^2. Take the columns in $J^c \cap C_p$, and notice that $q \in J^c \cap C_p$. As in Case 1, 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 $q' \in (J^c \cap C_p) \setminus \{q\}$, since $|(J^c \cap C_p) \setminus \{q\}| \le d - 1$.

However, this new choice of U' possibly causes $(U'V^T)_{p,i} \neq (UV^T)_{p,i}$ for $i \in C_p \cap J$. For each $i \in C_p \cap J$, 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 1 applied to each i in $C_p \cap J$ 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 1) 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 $\|V_i - V_i'\| \le n^{-5}$. 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**. 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 $i \in J^c \cap C_p = \{q_1,\dots,q_d\}$, we have

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

Adding the two equation 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 $(V_{q_1}, V_{q_2}, \dots, V_{q_d})$ 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 $i \in C_p \cap J$. Suppose $R_i = \{p_1,\dots,p_d\} \ni p$. Similarly to the above we have that

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

and

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

Adding 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

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

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 $i \in C_p \cap J^c$, we conclude that $V'' = V$.

E Proof of Theorem 4.1

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

Proof. Let $X = U\Sigma V$ be the SVD of X. Let G and H be random matrices with i.i.d. $N(0,1)$ entries. By rotational invariance, GXH^T is identically distributed as $G\Sigma H^T$. Let $k = n^{1-2/p}$ and $\tilde{X}$ be the upper-left $k \times k$ block of $G\Lambda H^T =: \tilde{X}$. It is clear that

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

Define

$$Y = \frac{1}{|C|}\sum_{\sigma \in C} \tilde{X}_\sigma.$$

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

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

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 $\mathbb{E}Y = \|X\|_p^p$. Now we compute $\mathbb{E}Y^2$.

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

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

$$\begin{aligned}\mathbb{E}(\tilde{X}_\sigma \tilde{X}_\tau) = \sum_{\substack{\ell_1,\dots,\ell_q\\ \ell_1',\dots,\ell_q'\\ m_1,\dots,m_q\\ m_1',\dots,m_q'}} &\left(\prod\nolimits_{i=1}^q \sigma_{\ell_i}\sigma_{m_i}\sigma_{\ell_i'}\sigma_{m_i'}\right)\cdot\\ &\mathbb{E}\left\{\prod\nolimits_{s=1}^q G_{\ell_s,i_s} G_{m_s,i_{s+1}} G_{\ell_s',i_s'} G_{m_s',i_{s+1}'}\right\}\cdot\\ &\qquad \mathbb{E}\left\{\prod\nolimits_{s=1}^q H_{\ell_s,j_s} H_{m_s,j_s} H_{\ell_s',j_s'} H_{m_s',j_s'}\right\}.\end{aligned}$$

It is not difficult to see that (see Appendix G for details)

$$\mathbb{E}(\tilde{X}_\sigma \tilde{X}_\tau) \lesssim_{m,p}$$
$$\begin{cases} \|A\|_2^{2(2q-2m_1-1)} \|A\|_{2(m_2+1)}^{2(m_2+1)} \|A\|_4^{4(m_1-m_2)}, & m_2 \leq m_1 \leq q-1 \\ \|A\|_2^{2(2q-2m_2-1)} \|A\|_{2(m_1+1)}^{2(m_1+1)} \|A\|_4^{4(m_2-m_1)}, & m_1 \leq m_2 \leq q-1 \\ \|A\|_4^{4(q-m_2-1)} \|A\|_{2(m_2+1)}^{4(m_2+1)}, & m_1 = q, m_2 \leq q-1 \\ \|A\|_4^{4(q-m_1-1)} \|A\|_{2(m_1+1)}^{4(m_1+1)}, & m_2 = q, m_1 \leq q-1 \\ \|A\|_{2q}^{4q}, & m_1 = m_2 = q \end{cases}$$

By Hölder's inequality, $\|A\|_2^2 \leq n^{1-\frac{2}{p}} \|A\|_p^2$ and

$$\|A\|_{2(m+1)}^{2(m+1)} \leq n^{1-\frac{2(m+1)}{p}} \|A\|_p^{2(m+1)}$$

Thus,

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

There are $O_q(k^{q+m_1} k^{q+m_2}) = O_p(k^{p+m_1+m_2})$ pairs of (m_1, m_2)-disjoint cycles and $|C| = \Theta(k^p)$,

$$\begin{aligned} \frac{1}{|C|^2} \sum_{\substack{\sigma,\tau \in C \\ |\sigma\Delta\tau| = (m_1,m_2)}} \mathbb{E}(A_\sigma A_\tau) &\leq C_{m_1+m_2,p} \frac{n^{p-m_1-m_2-2}}{k^{p-m_1-m_2}} \|A\|_p^{2p} \\ &\leq C_{m_1+m_2,p} \|A\|_p^{2p}, \quad m_1, m_2 \leq q-1, \end{aligned}$$

$$\begin{aligned} \frac{1}{|C|^2} \sum_{\substack{\sigma,\tau \in C \\ |\sigma\Delta\tau| = (m_1,m_2)}} \mathbb{E}(A_\sigma A_\tau) &\leq C_{m_1+m_2,p} \frac{n^{q-m_1-1}}{k^{q-m_1}} \|A\|_p^{2p} \\ &\leq C_{m_1,p} \|A\|_p^{2p}, \quad \frac{p}{2} \leq m_2 = q, m_1 \leq q-1, \end{aligned}$$

$$\begin{aligned} \frac{1}{|C|^2} \sum_{\substack{\sigma,\tau \in C \\ |\sigma\Delta\tau| = (m_1,m_2)}} \mathbb{E}(A_\sigma A_\tau) &\leq C_{m_1+m_2,p} \frac{n^{p-2m_2-2}}{k^{q-m_2}} \|A\|_p^{2p} \\ &\leq C_{m_1,p} \|A\|_p^{2p}, \quad \frac{p}{2} \leq m_1 = q, m_2 \leq q-1, \end{aligned}$$

$$\frac{1}{|C|^2} \sum_{\substack{\sigma,\tau \in C \\ |\sigma\Delta\tau| = (m_1,m_2)}} \mathbb{E}(A_\sigma A_\tau) \leq \|A\|_p^{2p}, \quad m_1 + m_2 = p.$$

Therefore $\mathbb{E}Y^2 \leq C_p \|A\|_p^{2p}$ 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

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

and

$$Var(X) = \frac{Var(Y)}{N} \leq \frac{\mathbb{E}Y^2}{N} = \frac{1}{4} \epsilon^2 \|A\|_p^{2p}.$$

Finally, by Chebyshev inequality,

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

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

Algorithm 2 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 $G_i X G_i^T$, $i = 1, \ldots, N$
Compute $Z$ as defined in (F.14)
return $Z$
```

F Algorithms for Odd p

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 C denote the set of cycles. Call two cycles $\sigma, \tau \in C$ k-disjoint if $|\sigma \Delta \tau| = 2k$, where σ and τ are viewed as multisets.

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

Proof. Since X is symmetric it can be written as $X = O \Lambda O^T$, where Λ is a diagonal matrix and O an orthogonal matrix. Let G be a random matrix with i.i.d. $N(0,1)$ entries. By rotational invariance, GXG^T is identically distributed as $G \Lambda G^T$. Let $k = n^{1-2/p}$ and

$\tilde{X}$ be the upper-left $k \times k$ block of $G\Lambda G^T := \tilde{X}$. It is clear that

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

Define

$$Y = \frac{1}{C} \sum_{\sigma \in C} \tilde{X}_\sigma.$$

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

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

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

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

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

$$\mathbb{E}(\tilde{X}_\sigma \tilde{X}_\tau) \lesssim_{m,p} \|A\|_2^{2(p-m-1)} \|A\|_{2(m+1)}^{2(m+1)}$$

By Hölder's inequality, $\|A\|_2^2 \leq n^{1-\frac{2}{p}} \|A\|_p^2$ and

$$\|A\|_{2(m+1)}^{2(m+1)} \leq n^{1-\frac{2(m+1)}{p}} \|A\|_p^{2(m+1)}, \quad 2(m+1) \leq p$$

Thus,

$$\mathbb{E}(\tilde{X}_\sigma \tilde{X}_\tau) \lesssim_{m,p} \begin{cases} n^{p-m-2} \|A\|_p^{2p}, & 2(m+1) \leq p \\ k^{p-m-1} \|A\|_p^{2p}, & \text{otherwise.} \end{cases}$$

And when σ, τ are $(p-1)$- and p-disjoint,

$$\mathbb{E}(A_\sigma A_\tau) \leq \|A\|_p^{2p}.$$

There are $O_p(k^{p+m})$ pairs of m-disjoint cycles and $|C| = \Theta(k^p)$,

$$\frac{1}{|C|^2} \sum_{\substack{\sigma,\tau \in C \\ |\sigma\Delta\tau| = 2m}} \mathbb{E}(A_\sigma A_\tau) \leq C_{m,p} \frac{n^{p-m-2}}{k^{p-m}} \|A\|_p^{2p}$$

$$\leq C_{m,p} \|A\|_p^{2p}, \quad m \leq \frac{p}{2} - 1,$$

$$\frac{1}{|C|^2} \sum_{\substack{\sigma,\tau \in C \\ |\sigma\Delta\tau| = 2m}} \mathbb{E}(A_\sigma A_\tau) \leq C_{m,p} \frac{k^{p-m-1}}{k^{p-m}} \|A\|_p^{2p}$$

$$\leq C_{m,p} \|A\|_p^{2p}, \quad \frac{p}{2} - 1 < m \leq p-2,$$

$$\frac{1}{|C|^2} \sum_{\substack{\sigma,\tau \in C \\ |\sigma\Delta\tau| = 2m}} \mathbb{E}(A_\sigma A_\tau) \leq \frac{1}{k} \|A\|_p^{2p}, \quad m = p-1$$

$$\frac{1}{|C|^2} \sum_{\substack{\sigma,\tau \in C \\ |\sigma\Delta\tau| = 2m}} \mathbb{E}(A_\sigma A_\tau) \leq \|A\|_p^{2p}, \quad m = p.$$

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

$$Z = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|C|} \sum_{\sigma \in C} (G_i X G_i^T)_\sigma =: \frac{1}{N} \sum_{i=1}^{N} Y_i, \tag{F.14}$$

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

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

and

$$Var(X) = \frac{Var(Y)}{N} \leq \frac{\mathbb{E}Y^2}{N} = \frac{1}{4} \epsilon^2 \|A\|_p^{2p}.$$

Finally, by Chebyshev inequality,

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

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

G Omitted Details in the Proof of Theorem F.1

Suppose that $\sigma = (i_1, \ldots, i_p)$ and $\tau = (j_1, \ldots, j_p)$ are m-disjoint. Then

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

For the expectation to be non-zero, we must have each appeared entry repeated an even number of times. Hence, if some i_s appears only once among the index $\{i_s\}$ and $\{j_s\}$, it must hold that $\ell_s = \ell_{s+1}$. Thus for each of the summation term the indices $\{\ell_s\}$ breaks into a few blocks, in each of which all ℓ_s takes the same value, the same holds for $\{\ell_s'\}$. We also need to piece the blocks of $\{\ell_s\}$ which 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 can be made on m and p by taking the maximum constant over all possible block configurations. Notice that in a configuration, all r_w's are even, $\sum r_w = 2p$ and $w \le p - m$. Note that

$$\|A\|_r^r \|A\|_s^s \le \|A\|_{r-1}^{r-1} \|A\|_{s+1}^{s+1}, \quad r > s$$
$$\|A\|_{r+s}^{r+s} \le \|A\|_r^r \|A\|_s^s,$$

it is easy to see that the worst case of configuration is $w = p - m$ and $(r_1, \dots, r_w) = (2(m+1), 2 \dots, 2)$, giving the bound

$$C\|A\|_2^{2(p-m-1)} \|A\|_{2(m+1)}^{2(m+1)}.$$

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

H Proof of Theorem 5.4

Proof. First we show that $\|X\|_p^p$ and $\|Y\|_p^p$ differ by a constant factor. We have calculated $\|X\|_p^p = (I_p + o(1))n$ in (3.3). Similarly, it holds that

$$\|Y\|_p^p = (J_p + o(1))\, n,$$

where
(H.15)
$$J_p = 2^{\frac{p}{2}} \int_a^b x^{\frac{p}{2}} \frac{\sqrt{(b-x)(x-a)}}{\pi x} dx, \quad a, b = \frac{3}{2} \mp \sqrt{2}.$$

Extend the definition of I_p and J_p to $p = 0$ by $I_p = 1$ and $J_p = 1/2$. This agrees with $\|X\|_0 = \operatorname{rank}(X)$. Now it suffices to show that $I_p \neq J_p$ when $p \neq 2$. Indeed, let us consider the general integral

$$\int_{(1-\sqrt{\beta})^2}^{(1+\sqrt{\beta})^2} x^\gamma \frac{\sqrt{((1+\sqrt{\beta})^2 - x)(x - (1-\sqrt{\beta})^2)}}{2\pi\beta x} dx$$

Change the variable $y = (x - (1+\beta))/\sqrt{\beta}$, the integral above becomes

$$\frac{1}{2\pi} \int_{-2}^{2} (\sqrt{\beta} y + (1+\beta))^{\gamma-1} \sqrt{4 - y^2} dy.$$

Hence

$$J_p - I_p = \frac{1}{2\pi} \int_{-2}^{2} f_p(x) \sqrt{4 - x^2} dx,$$

where

$$f_p(x) = (\sqrt{2}x + 3)^{\frac{p}{2}-1} - (x+2)^{\frac{p}{2}-1}.$$

One can verify that $f_p < 0$ on $(-2, 2)$ when $0 < p < 2$, $f_p > 0$ on $(-2, 2)$ when $p > 2$ and $f_p = 0$ when $p = 2$. It is easy to compute that $I_2 = 1$. The case $p = 0$ is trivial. Therefore, $I_p > J_p$ for $p \in [0, 2)$, $I_p < J_p$ for $p \in (2, \infty)$, and $I_p = J_p = 1$ when $p = 2$.

Denote by $\mathcal{D}_0$ the distribution of X and by $\mathcal{D}_1$ that of Y. Suppose that the k linear forms are $a_1, \dots, a_k$. Choose X from $\mathcal{D}_0$ with probability $1/2$ and from $\mathcal{D}_1$ with probability $1/2$. Let $b \in \{0, 1\}$ indicate that $X \sim \mathcal{D}_b$. Without loss of generality, we can assume that $a_1, \dots, a_k$ are orthonormal. Hence $(a_1, \dots, a_k) \sim N(0, I_k)$ when $b = 0$ and $(a_1, \dots, a_k) \sim D_{n,k}$ when $b = 1$. Then with probability $3/4 - o(1)$ we have that

$$(1 - c_p)((I_p + o(1))n \le Y \le (1 + c_p)((I_p + o(1))n, \quad b = 0$$
$$(1 - c_p)((J_p + o(1))n \le Y \le (1 + c_p)((J_p + o(1))n, \quad b = 1$$

so we can recover b from Y with probability $1 - o(1)$, provided that $c_p \le \frac{|I_p - J_p|}{2(I_p + J_p)}$. Consider the event E that the algorithm's output indicates $b = 1$. Then $\Pr(E|X \sim D_0) \le 1/4 + o(1)$ while $\Pr(E|X \sim D_1) \ge 3/4 - o(1)$. By definition of total variation distance, $d_{TV}(D_{n,k}, N(0, I_k))$ is at least $|\Pr(E|X \sim D_0) - \Pr(E|X \sim D_1)|$, which is at least $\frac{1}{2} + o(1)$. On the other hand, by Theorem 5.3, $d_{TV}(D_{n,k}, N(0, I_k)) \le \frac{1}{4}$ when $k \le c\sqrt{n}$ for some $c > 0$. We meet a contradiction. Therefore it must hold that $k > c\sqrt{n}$.