Matrix Norms in Data Streams: Faster, Multi-Pass and Row-Order

Vladimir Braverman* Johns Hopkins University
Stephen R. Chestnut† ETH Zurich
Robert Krauthgamer‡ Weizmann Institute of Science
Yi Li§ Nanyang Technological University
David P. Woodruff¶ Carnegie Mellon University
Lin F. Yang‖ Princeton University

October 25, 2018

Abstract

A central problem in data streams is to characterize which functions of an underlying frequency vector can be approximated efficiently. Recently there has been considerable effort in extending this problem to that of estimating functions of a matrix that is presented as a data-stream. This setting generalizes classical problems to the analogous ones for matrices. For example, instead of estimating frequent-item counts, we now wish to estimate "frequent-direction" counts. A related example is to estimate norms, which now correspond to estimating a vector norm on the singular values of the matrix. Despite recent efforts, the current understanding for such matrix problems is considerably weaker than that for vector problems.

We study a number of aspects of estimating matrix norms in a stream that have not previously been considered: (1) multi-pass algorithms, (2) algorithms that see the underlying matrix one row at a time, and (3) time-efficient algorithms. Our multi-pass and row-order algorithms use less memory than what is provably required in the single-pass and entrywise-update models, and thus give separations between these models (in terms of memory). Moreover, all of our algorithms are considerably faster than previous ones. We also prove a number of lower bounds, and obtain for instance, a near-complete characterization of the memory required of row-order algorithms for estimating Schatten p-norms of sparse matrices.

*Email: vova@cs.jhu.edu. This material is based upon work supported by the NSF Grants IIS-1447639, EAGER CCF-1650041, and CAREER CCF-1652257

†Email: stephenc@ethz.ch.

‡Email: robert.krauthgamer@weizmann.ac.il. Work supported in part by the Israel Science Foundation grant #897/13.

§Email: leeyi@umich.edu.

¶Email: dwoodruf@cs.cmu.edu.

‖Email: lin.yang@princeton.edu. This material is based upon work supported by the NSF Grant IIS-1447639. Work was done while the author was in Johns Hopkins University.

1 Introduction

Modern datasets, from text documents and images to social graphs, are often represented as a large matrix $A \in \mathbb{R}^{m \times n}$. In many application domains, including database queries, data mining, network transactions and sensor networks (see e.g. [Lib13, WLL+16, HK15] for recent examples), the input matrix A is presented to the algorithm as a data stream, i.e., a sequence of items/updates that can take several forms. In the *entry-wise (or insertion-only) model*, each item specifies (i, j, A_{ij}) and provides the value of one entry, in arbitrary order (and the unspecified entries are set to 0). The *row-order model* is similar, except that the items follow the natural order (sorted with i as the primary key, and j as the secondary one). In the *turnstile model*, each stream item has the form (i, j, δ) and represents an update $A_{ij} \leftarrow A_{ij} + \delta$ for $\delta \in \mathbb{R}$ (after initializing A to the all-zeros matrix). These models capture different access patterns, but all three can represent sparse matrices quite efficiently, because zero entries are implicit. As usual, the key parameters of an algorithm in the data-stream model are its memory (also referred to as storage/space requirements) and its runtime (per update and to report its output).

Many properties of a matrix are directly related to its spectral characteristics, i.e., its singular values. For example, the number of non-zero singular values is just the matrix rank, which determines the degrees of freedom of a corresponding linear system; the maximum and minimum singular values of a matrix determine its condition number, which in turn determines the hardness of many problems, such as optimization problems; the leading singular values of a matrix determine how well a matrix can be represented by the principal components; and so forth. It is generally hard to compute directly the singular values of a matrix, especially in the streaming model, but luckily, the Schatten norms of the matrix can often be used as surrogates for its spectrum, see e.g. [ZWJ15, KV16, DNPS16, KO17]. Formally, the *Schatten p-norm* of a matrix $A \in \mathbb{R}^{m \times n}$ is defined, for every $p \geq 1$, as

$$\|A\|_{S_p} := \Big(\sum_{j \geq 1} \sigma_j^p \Big)^{1/p},$$

where $\sigma_1 \geq \cdots \geq \sigma_{\min(m,n)}$ are the singular values of A. This definition naturally extends to all $0 < p < 1$ although then it is not a norm, and also to $p = 0, \infty$ by taking the limit. This is a very important family of matrix norms, and includes as special cases the well-known trace/nuclear norm $\|A\|_* = \sum_{j \geq 1} \sigma_j = \|A\|_{S_1}$, the Frobenius norm $\|A\|_F = \big(\sum_{j \geq 1} \sigma_j^2 \big)^{1/2} = \|A\|_{S_2}$, and the spectral/operator norm $\|A\|_{op} = \sigma_1(A) = \|A\|_{S_\infty}$.

We study algorithms that approximate the Schatten p-norm of a matrix A presented in a data stream. While this problem has attracted significant attention lately [AN13, LNW14, LW16a, LW16b, LW17], our results address three new aspects. First, we design faster and more space-efficient *multi-pass* algorithms. Second, we consider the *row-order model*, which is a common access pattern for matrix data (see, e.g. [Lib13]). Third, we design algorithms with faster *update time and/or query time*. The above three aspects were not considered previously for matrix norms, and our work opens the door for further diversification of prevailing models (and thereby of current algorithms). In particular, our results can be applicable to classical scenarios, e.g., where data is stored on disk (or any media where a linear scan is much faster than random access), and potentially lead to performance improvements in other such domains. In the next few subsections, we present our contributions in more detail.

1.1 New Estimator for PSD Matrices (or Even p)

Our first results rely on a new method for estimating the Schatten p-norm $\|A\|_{S_p}$ of a positive semidefinite matrix (PSD) matrix $A \in \mathbb{R}^{n \times n}$ for integer $p \geq 2$. This method yields two new

Problem: Schatten p-norm of PSD A, integer $p \geq 2$ (or general A, even p)				
passes	space	update time	query time	
1	$\epsilon^{-2}n^{2-4/p}$	$\epsilon^{-2}n^{2-4/p}$	$\epsilon^{-2}n^{p-2}$	[LNW14]
1	$\epsilon^{-2}n^{2-4/p}$	ϵ^{-2}	$\epsilon^{-2}n^{(1-2/p)\omega}$	Theorems 3.3 and 3.8
$\lceil p/2 \rceil$	$\epsilon^{-2}n$	ϵ^{-2}	$\epsilon^{-2}n$	[Woo14, Theorem 6.1]
$\lceil p/2 \rceil$	$\epsilon^{-2}n^{1-1/(p-1)}$	ϵ^{-2}	$\epsilon^{-2}n^{(1-1/(p-1))}$	Theorems 3.6 and 3.8

Table 1: Streaming algorithms for $(1+\epsilon)$-approximation of the Schatten p-norm of $A \in \mathbb{R}^{n\times n}$. The bounds for storage/time omit $O_p(1)$ factors, and count space in words.

streaming algorithms in the turnstile model, which require, respectively, one pass and $\lceil p/2 \rceil$ passes over the input. Both algorithms are at least as good as the previous ones in all three standard performance measures of storage, update time, and query time; and each algorithm offers significant improvements in two out of these three. Our one-pass algorithm achieves update time $O(1)$ compared with the previous poly(n), and query time $O(n^{\omega(1-p/2)})$, where $\omega \leq 2.373$ is the matrix multiplication exponent [Le 14], compared with the previous n^{p-2}. And our multi-pass algorithm requires storage that is sublinear in n, compared with $O(n)$ previously. We note that if p is even, then the above results extend to arbitrary $A \in \mathbb{R}^{m\times n}$ (and not only PSD) by a standard argument. A detailed comparison of the bounds is given in Table 1, and the results themselves appear in Section 3.

Throughout the paper, a matrix is called *sparse* if it has at most $O(1)$ non-zero entries per row and per column. We write $\tilde{O}(f)$ as a shorthand for $O(f \cdot \log^{O(1)} f)$, and write $O_a(f)$ to indicate that the constant in O-notation depends on some parameter a.

Techniques Our technical innovation is an unbiased estimator of $\mathrm{Tr}(A^p)$ for a *symmetric* (and not only PSD) matrix $A \in \mathbb{R}^{n\times n}$. To see why this is useful, denote the eigenvalues of A by $\lambda_1 \geq \cdots \geq \lambda_n$, and observe that if A is PSD (or alternatively if p is even), then $\mathrm{Tr}(A^p) = \sum_i \lambda_i^p = \sum_i \sigma_i(A)^p = \|A\|_{S_p}^p$. Our estimator has the form

$$X \coloneqq \mathrm{Tr}(G_1 A G_2^T G_2 A G_3^T \cdots G_p A G_1^T), \tag{1}$$

where $G_i \in \mathbb{R}^{t\times n}$ are certain random matrices. This estimator X can be computed from the p bilinear sketches $\{G_i A G_{i+1}^T\}_{i\in[p]}$ by straightforward matrix multiplication, where $G_{p+1} \coloneqq G_1$ by convention. And if, say, $t = O(n^{1-2/p})$, then each bilinear sketch has dimension $O(t^2) = O(n^{2-4/p})$. These determine the streaming algorithm's storage requirement and query time, and, if the matrices $\{G_i\}_{i\in[p]}$ have sparse columns, the updates will be fast.

The main difficulty is to bound the estimator's variance, which highly depends on the choice of the matrices $\{G_i\}_{i\in[p]}$. The basics of this technique can be seen in the case $p = 4$, if the G_i's satisfy the following definition.

Definition 1.1. *A random matrix $S \in \mathbb{R}^{t\times n}$ is called an* (ϵ, δ, d)-Johnson-Lindenstrauss Transformation (JLT) *if for every $V \subseteq \mathbb{R}^n$ of cardinality $|V| \leq d$ it holds that*

$$\Pr\left[\forall x \in V,\ \|Sx\|_2^2 \in (1 \pm \epsilon)\|x\|_2^2\right] \geq 1 - \delta.$$

An (ϵ, δ, d)-JLT can be constructed with $t = O(\epsilon^{-2}\log(d/\delta))$ rows, which is optimal (see [KMN11] or [JW13]). While using independent $N(0, 1/t)$ Gaussians entries works, there is a construction with only $O(\epsilon^{-1}\log(1/\delta))$ non-zero entries per column [KN14].

The case $p = 4$ has a particularly short and simple analysis, whenever G_1 and G_2 are independent (ϵ, δ, n)-JLT matrices, which we can achieve with $t = O(\epsilon^{-2} \log n)$. The first idea is to "peel off" G_i from both sides, using that for any PSD matrix M, with high probability $\mathrm{Tr}(G_i M G_i^T) \in (1 \pm \epsilon)\,\mathrm{Tr}(M)$ (see Lemma 3.2 for a precise statement). A second idea is to use the identity $\mathrm{Tr}(BC) = \mathrm{Tr}(CB)$ to rewrite $\mathrm{Tr}(AA^T G_2^T G_2 AA^T) = \mathrm{Tr}(G_2 AA^T AA^T G_2^T)$. Now using the first idea once again, we are likely to arrive at an approximation to $\mathrm{Tr}(AA^T\ AA^T) = \|A\|_{S_4}$. The full details are given in Section 3.1.

The sketching method extends from $p = 4$ to any integer $p \geq 2$, but the simple analysis above breaks (because for $p > 4$ the "inside" matrix M is no longer PSD) and thus our analysis is much more involved. We first analyze G_i's with independent Gaussian entries, by a careful expansion of the fourth moment of X, which exploits certain cancellations occurring (only) for Gaussians. We then consider G_i's that are sampled from a particular sparse JLT due to [TZ04], and employ a symmetrization-and-decoupling argument to compare the variance of X in this case with that of Gaussian G_i's.

We make two technical remarks. First, proving $\mathbb{E}[X] = \mathrm{Tr}(A^p)$ is straightforward. Indeed, by the second idea above, we can rewrite $X = \mathrm{Tr}(G_1 AG_2^T G_2 AG_3^T \cdots G_p AG_1^T)$ as $X = \mathrm{Tr}(G_1^T G_1 AG_2^T G_2 A \cdots G_p^T G_p A)$. Now using $\mathbb{E}[G_i^T G_i] = I$ together with linearity of trace and of expectation, we obtain that $\mathbb{E}[X] = \mathrm{Tr}(A^p)$. Second, after setting $t = O(n^{1-2/p})$ (independent of ϵ), our bound on the variance is $O(\mathbb{E}[X]^2)$, which we can decrease in a standard way, taking $O(1/\epsilon^2)$ repetitions. See Sections 3.2 and 3.4 for details.

The multi-pass streaming algorithm is implemented slightly differently, in that $G_1 \in \mathbb{R}^{1\times n}$, i.e., has only one row. The other matrices $G_2, \ldots, G_p \in \mathbb{R}^{t\times n}$ are as before, although we now set $t = O(n^{1-1/(p-1)})$. Our estimator X can be computed in $\lceil p/2 \rceil$ passes with space only $2t$ as follows. In the first pass, compute vectors $X_L \leftarrow G_1 AG_2^T \in \mathbb{R}^{1\times t}$ and $X_R \leftarrow G_p^T AG_1 \in \mathbb{R}^{t\times 1}$, and then on the i-th pass update $X_L \leftarrow X_L G_i^T AG_{i+1}$ and $X_R \leftarrow G_{p-i+1} AG_{p-i+2}^T X_R$. Notice that the computation in each pass is linear in A. For even p, after completing $p/2$ passes, compute and output $X' = X_L X_R \in \mathbb{R}$ (and similarly for odd p). This X' is similar to the estimator X described above, except for the new dimensions of the G_i's. See Sections 3.3 and 3.4.

This multi-pass algorithm offers a very significant space savings over the one-pass algorithm. It is also a bit surprising because it is getting close to the corresponding vector norm, namely, ℓ_p-norm on $\mathbb{R}^n$, for which the optimal space for $O(p)$ passes is $\tilde{O}(n^{1-2/p})$ bits. In fact, for the vector norm, $O(p)$ passes do not significantly reduce the storage needed compared with one pass, which stands in sharp contrast to Schatten p-norms. As mentioned before, if p is even then the algorithms extends to arbitrary $A \in \mathbb{R}^{m\times n}$ by a standard argument.

1.2 Lower Bound for PSD Matrices

Recent work [LW16a] has improved the storage lower bound for estimating Schatten p-norms for non-integer values of p, by showing that $(1+\epsilon)$-approximation (in the one-pass entry-wise model) requires storage $n^{1-g(\epsilon)}$, for some function $g(\epsilon) \to 0$ as $\epsilon \to 0$, even for a sparse matrix. This contrasts with our algorithms for PSD matrices (from Section 1.1), where the exponent is independent of ϵ and bounded away from 1. However, the hard distribution used by [LW16a] is not over PSD matrices, leaving open the possibility that PSD matrices admit algorithms that use storage $O(n^c)$ for $c < 1$ independent of ϵ.

We close this gap in Section 4, by adapting the lower bound of [LW16a] to PSD matrices, to show, for every non-integer $p > 0$, a storage lower bound of $\Omega(n^{1-g'(\epsilon)})$ for some function $g'(\epsilon) \to 0$ as $\epsilon \to 0$ (again, in the one-pass entry-wise model and even for a sparse matrix). A key feature of our lower bounds for PSD matrices is that they hold in the model in which each entry of the matrix

Problem: Schatten p-norm of a sparse matrix in row-order stream			
	which $p > 0$	space	
Algorithms:	all p	$\tilde{O}(n)$	trivial (by sparsity), $\epsilon = 0$
	$p \equiv 0 \pmod 4$	$\tilde{O}_{p,\epsilon}(n^{1-4/p})$	Section 1.3
	$p \equiv 2 \pmod 4$	$\tilde{O}_{p,\epsilon}(n^{1-4/(p+2)})$	Theorem 6.1, $p \geq 6$
Lower Bounds:	$p \in 2\mathbb{Z}, p \geq 4$	$\Omega(n^{1-4/p})$	Theorem 5.4, for $\epsilon < \epsilon_0(p)$, even multi-pass
	$p \notin 2\mathbb{Z}$	$\Omega_t(n^{1-1/t})$	Theorem 5.3, for $\epsilon < \epsilon_0(t,p)$

Table 2: Bounds for $(1+\epsilon)$-approximation of the Schatten p-norm of a sparse matrix $A \in \mathbb{R}^{n\times n}$ in the one-pass row-order model. Space is counted in bits.

occurs exactly once in the stream. This models applications where the matrix resides in external memory and is being streamed through main memory; in such a model multiple updates to an entry may not appear. While it is possible to obtain lower bounds for PSD matrices by embedding the multiplayer Set-Disjointness lower bound [BJKS02] for vectors onto the diagonal of a matrix, to apply such lower bounds the diagonal entries need to be incremented repeatedly, that is, one such diagonal entry needs to be updated $n^{\Omega(1)}$ times. In contrast, in our lower bounds each matrix entry occurs exactly once in the stream, i.e., there are no updates to entries.

1.3 Results for Row-Order Model

For sparse matrices, estimating Schatten p-norms in the row-order model can be reduced to estimating Schatten $(p/2)$-norms in the turnstile model. Consider estimating $\|A\|_{S_p}^p$ for some sparse matrix A. The algorithm first forms $A^TA = \sum_i A_i^T A_i$ "on the fly", by reading each row A_i and immediately generating a stream of updates that corresponds to the non-zero entries in $A_i^T A_i$, and then it can just estimate the Schatten $(p/2)$-norm of that stream, because $\|A^TA\|_{S_{p/2}}^{p/2} = \|A\|_{S_p}^p$. Observe that each row A_i has only $O(1)$ non-zero entries, hence also $A_i^T A_i$ has only $O(1)$ non-zero entries, and the algorithm only needs $O(1)$ space to generate the updates to A^TA. Moreover, since A is sparse, also A^TA is sparse. It was shown in [LW16a] how to estimate the Schatten p-norm, for an even integer p, using $\tilde{O}_{p,\epsilon}(n^{1-2/p})$ bits of space, even in the turnstile model. For $p \in 4\mathbb{Z}$, the above yields an algorithm in the row-order model that uses $\tilde{O}_{p,\epsilon}(n^{1-4/p})$ bits of space for sparse matrices.

In Sections 5 and 6, we study the problem in the row-order model for all $p > 0$. When p is not an even integer, we prove that $(1+\epsilon)$-approximating the Schatten p-norm in the one-pass entry-wise model requires $\Omega_\epsilon(n^{1-g(\epsilon)})$ bits of space where $g(\epsilon) \to 0$ as $\epsilon \to 0$. This bound holds even for sparse matrices, in which case it is almost tight. When $p \geq 4$ is an even integer, we prove a lower bound of $\Omega_p(n^{1-4/p})$ bits of space, matching up to logarithmic factors the algorithm from above for $p \in 4\mathbb{Z}$. For the remaining case $p \equiv 2 \pmod 4$, we present an algorithm using $\tilde{O}_{p,\epsilon}(n^{1-4/(p+2)})$ space, leaving a slight polynomial gap from the lower bound of $\Omega_p(n^{1-4/p})$.

1.4 Previous Work

The aforementioned algorithm of [LNW14] uses a single sketching matrix G, for example, if A is PSD, then their sketch is $S = GAG^T$, where $G \in \mathbb{R}^{t\times n}$ is a Gaussian matrix. Its estimate for $\|A\|_{S_p}$ is produced by summing over all "cycles" $S_{i_1,i_2}S_{i_2,i_3}\cdots S_{i_p,i_1}$, where $i_1,\ldots,i_p \in [t]$ are distinct. Our sketch improves upon theirs in both update time and query time. The only other streaming

algorithm for Schatten p-norm that we are aware of is that of [LW16a, Theorem 7], which uses space $O(n^{1-\frac{2}{p}}\operatorname{poly}(\frac{1}{\epsilon}, \log n))$ but works only for matrices that have $O(1)$-entries per row and per column.

One possible approach to improve the update time would be to replace the Gaussian matrices in [LNW14] with a distribution over matrices that admit a fast multiplication algorithm. The analysis done in [LNW14] relies on the Gaussian entries (rotational invariance, in particular), so the replacement matrix should preserve the distribution of the sketch. Kapralov, Potluru, and Woodruff [KPW16] present just such a distribution on matrices $\tilde{G}$, where the multiplication $\tilde{G}A$ can be computed quickly and $\tilde{G}A$ is close to GA in total variation distance. Unfortunately, under the distribution of [KPW16], or any other with a similar guarantee on total variation distance, each coordinate update to A results in a dense rank-one update to the sketch, which means that the update time is not improved.

Several strong lower bounds are known for approximating Schatten p-norms and other matrix functions, both for the dimension of a sketch and for storage requirement (bits). Li, Nguyen and Woodruff [LNW14] prove that for $0 \leq p < 2$ every linear sketch that can approximate rank and Schatten p-norm must have dimension $\Omega(\sqrt{n})$ and every bilinear sketch must have dimension $\Omega(n^{1-\epsilon})$. Li and Woodruff [LW16b] show that every linear sketch for Schatten p-norm, $p \geq 2$, requires dimension $\Omega(n^{2-4/p})$. In [LW16a], they prove space complexity lower bounds that hold even when the input matrix is sparse. Specifically, they show that one-pass streaming algorithms which $(1 \pm \epsilon)$-approximate various functions of the singular values, including Schatten p-norms when p is not an even integer, require $\Omega(n^{1-g(\epsilon)})$ bits of space for some function $g(\epsilon) \to 0$ as $\epsilon \to 0$. Additional space lower bounds, e.g., for $p \in [1, 2)$, can be deduced from a general statement of [AKR15], see [LW16a, Table 1].

2 Notation and Preliminaries

The space bounds of sketching algorithms in the turnstile model are stated in terms of sketch dimension (number of entries). The number of bits required can be larger by a $\log nM$ factor, where M is the absolute ratio of the largest element in the matrix to the smallest. We call a matrix a *Gaussian matrix* if its entries are independent $N(0, 1)$ random variables. A matrix G of dimension $t \times n$ is a *column-normalized* Gaussian matrix if $G = G'/\sqrt{t}$, where G' is a Gaussian matrix. Now-standard techniques such as Nisan's Pseudo-random generator or k-wise independence can be used to derandomize Gaussian matrices for use in sketching algorithms. Column-normalized Gaussian matrices serve as JLTs. In particular, there exists a constant c such that if G be a $t \times n$ column-normalized Gaussian matrix with $t \geq \frac{c}{\epsilon^2} \log \frac{d}{\delta}$, then G is a (ϵ, δ, d)-JLT [IM98].

3 New Estimator for PSD Matrices (and Integer p)

The main result in this section is a new one-pass streaming algorithm for estimating the Schatten p-norm, for integer $p \geq 2$. When p is odd, it additionally requires that the input matrix is PSD. The first version of this algorithm, described in Section 3.2, has the same storage requirement of $\tilde{O}_p(n^{2-4/p}/\epsilon^2)$ bits as the previous algorithm of [LNW14] that uses cycle sums, but has simpler analysis and faster query time[1], which is roughly matrix multiplication time, n^ω, instead of n^p. Moreover, it is based on a new method that leads to a $\lceil p/2 \rceil$-pass algorithm with storage requirement

[1] In [KV16], Kong and Valiant independently improve the algorithm in [LNW14] to the same runtime as Theorem 3.3 in this paper by considering only "increasing cycles".

$\tilde{O}_p(n^{1-1/(p-1)}/\epsilon^2)$ bits, as described in Section 3.3. Previously, the algorithm in [Woo14, Theorem 6.1] has the same number of passes but larger storage requirement $O(n/\epsilon^2)$.[2] Finally, we improve the update time, as described in Section 3.4, by employing the sketching matrices G_i that are certain sparse matrices instead of Gaussians.

We start in Section 3.1 with the case $p = 4$, which is based on the same sketch but is significantly easier to analyse.

3.1 Schatten 4-Norm using JLT matrices

Theorem 3.1. *Let $G_1, G_2 \in \mathbb{R}^{t\times n}$ be independent $(\epsilon, \frac{\delta}{n}, 1)$-JLT matrices. Then for every $A \in \mathbb{R}^{n\times m}$,*

$$\Pr\left[\operatorname{Tr}(G_1AA^TG_2^TG_2AA^TG_1^T) \in (1\pm 2\epsilon)^2\|A\|_{S_4}^4\right] = 1-2\delta.$$

Thus, one can find a $(1\pm\epsilon)$-approximation to the Schatten-4 norm of a general matrix $A \in \mathbb{R}^{n\times m}$ using a linear sketch of dimension $O(\epsilon^{-2}n\log n)$.

Before proving the theorem, we remark that if each column of G_i has only s non-zero entries, it is easy to see that the update time of this linear sketch is $O(s)$, assuming any entry of G_1 and G_2 can be accessed in $O(1)$ time (in a streaming algorithm, the entries are usually computed from a small random seed in polylog(n) time). The query time is dominated by multiplying a matrix of size $t \times n$ with one of size $n \times t$, and thus take $O(t^\omega \cdot n/t) = \tilde{O}(n^\omega/\epsilon^{2(\omega-1)})$.

Now we prove Theorem 3.1, for which we need the following lemma.

Lemma 3.2. *Let $G \in \mathbb{R}^{t\times n}$ be an $(\epsilon, \delta/n, 1)$-JLT matrix. Then for every PSD matrix $A \in \mathbb{R}^{n\times n}$,*

$$\Pr\left[\operatorname{Tr}(GAG^T) \in (1\pm\epsilon)\operatorname{Tr}(A)\right] \geq 1-\delta.$$

Proof. By the Spectral Theorem, $A = U\Lambda U^T$, where Λ is a diagonal matrix and U is an orthonormal matrix. Then $G' = GU$ is a still $(\epsilon, \delta/n, 1)$-JLT. Thus

$$\operatorname{Tr}(GAG^T) = \operatorname{Tr}(G'\Lambda G'^T) = \operatorname{Tr}(\sqrt{\Lambda}G'^TG'\sqrt{\Lambda}) = \sum_{i=1}^n \lambda_i e_i^T G'^T G' e_i = \sum_{i=1}^n \lambda_i \|G'e_i\|_2^2.$$

By the JLT guarantee and a union bound, with probability at least $1-\delta$, for all $i \in [n]$ we have $\|G'e_i\|_2^2 \in [1-\epsilon, 1+\epsilon]$, in which case $\operatorname{Tr}(GAG^T) \in (1\pm\epsilon)\operatorname{Tr}(A)$. ☐

of Theorem 3.1. Apply Lemma 3.2 to the PSD matrix AA^TAA^T, to get that with probability at least $1-\delta$ (over the choice of G_2),

$$\operatorname{Tr}(G_2AA^TAA^TG_2^T) \in (1\pm 2\epsilon)\operatorname{Tr}(AA^TAA^T) = (1\pm 2\epsilon)\|A\|_{S_4},$$

where the left-hand side is equal to $\operatorname{Tr}(AA^TG_2^TG_2AA^T)$, by the identity $\operatorname{Tr}(MM^T) = \operatorname{Tr}(M^TM)$. Now suppose (by conditioning) that G_2 is already fixed, and apply the same lemma to the PSD matrix $AA^TG_2^TG_2AA^T$, to get that with probability at least $1-\delta$ (over the choice of G_1),

$$\operatorname{Tr}(G_1AA^TG_2^TG_2AA^TG_1^T) \in (1\pm 2\epsilon)\operatorname{Tr}(AA^TG_2^TG_2AA^T).$$

The proof follows by a union bound.

The linear sketch of A consists of the two matrices G_1A and G_2A, which suffices to estimate $\|A\|_{S_4}^4$ as above with $\delta = 1/8$. This sketch is linear and its dimension is $2tn$, where we can use say Gaussians to obtain $t = O(\epsilon^{-2}\log n)$. ☐

[2] We note that also in [Woo14, Theorem 6.1] it is required that p is even or that the input matrix is PSD, but this is erroneously omitted.

3.2 Schatten p-norm Using Gaussians

We now design a sketch for Schatten-p norm that uses column-normalized Gaussian matrices. We will later extend and refine it to improve the per-update processing time.

Theorem 3.3. *For every $0 < \epsilon < 1/2$ and integer $p \geq 2$, there is an algorithm that outputs at $(1 \pm \epsilon)$-approximation to the Schatten-p norm of a PSD matrix $A \in \mathbb{R}^{n \times n}$ using a randomized linear sketch of dimension $s = O_p(\epsilon^{-2} n^{2-4/p})$. The update time (for each entry in A) is $O(s)$ and the query time (for computing the estimate) is $O(\epsilon^{-2} n^{(1-2/p)\omega})$, where $\omega < 2.373$ is the matrix multiplication constant.*

If p is even, the above algorithm extends to a general matrix $A \in \mathbb{R}^{n \times m}$.

The first part of the theorem (for PSD matrices) follows directly from Proposition 3.4 below. The proposition is applicable to all symmetric matrices, but $\|A\|_{S_p}^p = \operatorname{Tr}(A^p)$ only for PSD matrices or even p. The linear sketch stores $G_i A G_{i+1}^T$ for $i = 1, \ldots, p$, where by convention $G_{p+1} = G_1$, repeated independently in parallel $O_p(1/\epsilon^2)$ times. Thus, the sketch has dimension $O_p(\epsilon^{-2} t^2)$. The estimator is obtained by computing the $O_p(1/\epsilon^2)$ independent copies of X and reporting their average. To analyze its accuracy, notice that a PSD matrix A satisfies $\mathbb{E}[X] = \operatorname{Tr}(A^p) = \|A\|_p^p$. Then setting $t = n^{1-2/p}$ gives $\operatorname{Var}(X) \leq O_p(\|A\|_{S_p})^{2p}$ and averaging multiple independent copies of X reduces the variance.

The second part (for general matrices), follows by using the same sketch for the symmetric matrix $B = \left(\begin{smallmatrix} 0 & A \\ A^T & 0 \end{smallmatrix}\right)$, because the nonzero singular values of B are those of A repeated twice and $\|B\|_{S_p}^p = 2\|A\|_{S_p}^p = 2 \operatorname{Tr}(A^p)$, where the last equality uses the assumption that p is even.

Because the correctness of the algorithm comes by bounding the variance of X, it is enough that the entries in each Gaussian matrix are four-wise independent, which is crucial for applications with limited storage like streaming.

Proposition 3.4. *For integer $p \geq 2$ and $t \geq 1$, let $G_1, \ldots, G_p$ be independent $t \times n$ column-normalized Gaussian matrices. Then for every symmetric matrix $A \in \mathbb{R}^{n \times n}$, the estimator $X = \operatorname{Tr}\left(G_1 A G_2^T G_2 A \ldots G_p^T \; G_p A G_1^T\right)$ satisfies*

$$\mathbb{E}[X] = \operatorname{Tr}(A^p) \qquad \text{and} \qquad \operatorname{Var}(X) = O_p\left(1 + \sum_{z=2}^{\lfloor \frac{p}{2} \rfloor + 1} \left(\frac{n^{1-\frac{2}{p}}}{t}\right)^z + \sum_{z=2}^{p} \left(\frac{n^{1-\frac{2}{z}}}{t}\right)^z\right) \|A\|_{S_p}^{2p}.$$

The full proof of this proposition appears in postponed to Section A. We outline the general idea here. It is standard that a Gaussian matrix is rotational invariant, i.e., G and GU are identically distributed for any orthogonal matrix U. Thus, by the Spectral Theorem, instead of considering symmetric matrix $A = U \Lambda U^T$, we can consider only its diagonalization Λ.

The proof of this proposition proceeds first by expanding X in terms of inner products of columns of the matrix G, i.e., $X = \sum_{i_1, i_2, \ldots, i_p \in [n]} \lambda_{i_1} \lambda_{i_2} \ldots \lambda_{i_p} \cdot \langle g_{i_1}^{(1)}, g_{i_2}^{(1)} \rangle \cdot \langle g_{i_2}^{(2)}, g_{i_3}^{(2)} \rangle \ldots \langle g_{i_p}^{(p)}, g_{i_1}^{(p)} \rangle$, where λ_i is the i-th eigenvalue of A and $g_{i_j}^{(j)}$ is the i_j-th column of G_j. We then expand $\mathbb{E}(X^2)$. The non-zero terms in $\mathbb{E}(X^2)$ are composed by only those terms of even powers in every eigenvalue. Computing the expectation of each term is straightforward because the entries of G are independent Gaussian random variables, but the crux of the proof is in bounding the sum of the terms. We introduce a collection of diagrams that aid in enumerating the terms according to their structure and computing the sum.

3.3 Multi-Pass Algorithm

The proof of Proposition 3.4 relies on the matrices G_i being Gaussians in two places. First, we assume that the matrix A is diagonal, and in general we need to consider G_iU instead of G_i. Second, the columns of these matrices have small variance/moments, as described in (7)-(8). We now generalize the proof to relax these requirements (e.g., to 4-wise independence) and obtain a multi-pass algorithm.

Lemma 3.5. *For integers $p \geq 2$ and $1 \leq t' \leq t$, let $G_1 \in \mathbb{R}^{t' \times n}$ and $G_2, \ldots, G_p \in \mathbb{R}^{t \times n}$ be independent column-normalized Gaussian matrices with 4-wise independent entries. The for every symmetric matrix $A \in \mathbb{R}^{n \times n}$, the estimator $X = \mathrm{Tr}\left(G_1 A G_2^T G_2 A \ldots G_p^T G_p A G_1^T\right)$ satisfies*

$$\mathbb{E}[X] = \mathrm{Tr}(A^p) \qquad \text{and} \qquad \mathrm{Var}(X) = O_p\left(1 + \sum_{z=2}^{\lfloor p/2 \rfloor} \frac{n^{z-1-2(z-1)/p}}{t' t^{z-1}} + \sum_{z=2}^{p} \frac{n^{z-2}}{t' t^{z-1}}\right) \|A\|_{S_p}^{2p}.$$

The proof of this lemma is postponed to B. It is a direct corollary of the proof of Proposition 3.4, except that t', the size of the first sketch matrix, is emphasized.

We can now use the above sketch to approximate the Schatten p-norm using $\tilde{O}(n^{1-1/(p-1)})$ bits of space with $\lceil p/2 \rceil$ passes over the input.

Theorem 3.6. *Let $p \geq 2$ be an even integer. There is a $\lceil p/2 \rceil$-pass streaming algorithm, that on input matrix $A \in \mathbb{R}^{n \times m}$ with $n \geq m$ given as a stream, outputs an estimate X such that with probability at least 0.9, $X \in (1 \pm \epsilon)\|A\|_{S_p}^p$ and uses $O_p(n^{1-1/(p-1)}/\epsilon^2)$ words of space. The above extends to all integers $p \geq 2$ if A is PSD.*

of Theorem 3.6. Without loss of generality we may assume that A is symmetric as argued in the proof of Theorem 3.3. We first describe a basic algorithm that produces an estimator for $\|A\|_{S_p}^p$ that is unbiased and has variance $O_p(\|A\|_{S_p}^{2p})$. We will later decrease the variance to $O(\epsilon^2\|A\|_{S_p}^{2p})$ using the standard technique of independent repetitions in parallel.

The basic algorithm uses a pseudo-random generator to produce a four-wise independent column-normalized Gaussian matrix. In fact, it samples p such matrices, namely, $G_1 \in \mathbb{R}^{1 \times n}$ and $G_2, \ldots, G_p \in \mathbb{R}^{t \times n}$ for $t = O(n^{1-1/(p-1)})$, where the p matrices are independent of each other. In the first pass, the algorithm computes $G_1 A G_2^T$ and $G_p A G_1^T$, and stores them in memory. Notice that these are linear sketches of A, each dimension t. In the second pass, the algorithm uses these results to compute $(G_1 A G_2^T) G_2 A G_3^T$ and $G_{p-1} A G_p^T (G_p A G_1^T)$ which are again linear sketches of the stream A (given the result of the first pass), each of dimension t. Continuing in this manner until pass number $\lceil p/2 \rceil$, the algorithm stores in memory the vectors $h = G_{\lfloor p/2 \rfloor} A G_{\lfloor p/2 \rfloor + 1}^T \cdot G_p A G_1^T$ and $h'^T = G_1 A G_2^T \cdots G_{\lfloor p/2 \rfloor - 1} A G_{\lfloor p/2 \rfloor}^T$, each of dimension t. Now compute $Y = h'^T h$. By Lemma 3.5 we have that $\mathbb{E}[Y] = \mathrm{Tr}(A^p) = \sum_i \lambda_i^p$ (where in the case that p is odd we use the assumption that A is PSD). Thus, Y is an unbiased estimator for $\|A\|_{S_p}^p$, and it remains to bound its variance. By Lemma 3.5,

$$\mathrm{Var}(Y) = O_p\left(\sum_{z=2}^{\lfloor p/2 \rfloor} \frac{n^{z-1-2\frac{z-1}{p}}}{n^{z-1-\frac{z-1}{p-1}}} + \sum_{z=2}^{p} \frac{n^{z-2}}{n^{z-1-\frac{z-1}{p-1}}}\right) \|A\|_{S_p}^{2p} = O_p(\|A\|_{S_p}^{2p}).$$

By repeating the basic algorithm $O_p(1/\epsilon^2)$ times in parallel and reporting the average of their estimates Y, we obtain estimator X for $\|A\|_{S_p}^p$ that is unbiased and has variance at most $\frac{1}{9}\epsilon^2\|A\|_{S_p}^{2p}$. The correctness of this estimator follows by Chebyshev's ineuqality. The basic algorithm is required to store $2p$ intermediate vectors of dimension t and random seeds for the p Gaussian matrices. By standard techniques, the length of the seeds is $O_p(\mathrm{polylog}\, n)$ bits. The final algorithm stores these for all the $O_p(1/\epsilon^2)$ repetitions, and Theorem 3.6 follows. ∎

3.4 Faster Update Time

Since Gaussian matrices are dense, a change to one coordinate of the input matrix A may lead to a change of every entry in the sketch. This means long update times for a streaming algorithm based on the sketch. In this section we extend our result for Gaussian sketching matrices to a distribution over $\{-1, 0, 1\}$ valued matrices with only one non-zero entry per column. The new sketch can be used to improve the update time of algorithms in the last two sections.

Definition 3.7 (Sparse ZD-sketch)**.** *Let $\mathcal{D}_{t,n}$ be the distribution over matrices $G := ZD \in \mathbb{R}^{t\times n}$, where $Z = (z_1, z_2, \ldots, z_n) \in \mathbb{R}^{t\times n}$ and $D = \mathrm{diag}(d_1, d_2, \ldots, d_n)$ are generated as follows. Let $h : [n] \to [t]$ be a 4-wise independent hash function, and set $Z_{i,j} = \mathbb{1}_{\{i=h(j)\}}$, i.e., in each z_j only the $h(j)$-th coordinate is set to 1, and all other coordinates are 0. The diagonal entries of D are four-wise independent uniform $\{-1, 1\}$ random variables, and D is independent from Z.*

Notice that each column of G has a single non-zero entry, which is actually a random sign, and the n columns are four-wise independent. This random matrix G is similar to the sketching matrix used in [TZ04] to speed up the update time when estimating the second frequency moment of a vector in $\mathbb{R}^n$. Also note that the ZD-sketch is a version of sparse JL matrices (see e.g., [KN14, DKS10]). In this paper we do not aim at optimizing the sparsity as we focus on approximating Schatten norms.

It is fairly easy to show that ZD-sketch works for approximating Schatten p-norm of matrices with all entries non-negative. The proof is presented in Section C. We now show that the conclusion of Theorem 3.3 and Theorem 3.6 still hold if we replace the Gaussian matrices in the sketch with independent samples from the sparse ZD-sketch. A major difficulty that arises in replacing the Gaussian matrix with the sparse ZD-sketch is the latter's lack of rotational invariance. To prove Theorem 3.3 we were able to expand X^2 in terms of the eigenvalues of A and compute the expectation term-by-term, but this is not possible for the sparse ZD-sketch. For example, let G be a Gaussian matrix, for any orthogonal matrix U, the matrix GU is again a Gaussian matrix with an identical distribution to G. This does not hold for sparse ZD-sketch. As a consequence, in the expansion of $\mathbb{E}(X^2)$ in the proof of Proposition 3.4, the non-zero terms would also include those monomials of odd powers of $\lambda_i(A)$. For example, for Schatten 3-norm, one cannot bound $\sum_{i_1,i_2,\ldots i_6\in[n]} \prod_{j=1}^{6} \lambda_{i_j}$ by $O(\|A\|_{S_3}^6)$. But this term appears in the expansion of $\mathbb{E}(X^2)$ of the Schatten 3-norm estimator if using the sparse ZD-sketch matrices.

To resolve this problem, we use a technique similar to the proof of Hanson-Wright Inequality in [RV13] to bound the variance of X. The proof is composed of three major steps. The first step is to decouple the dependent summands by injecting independence. The second step is to replace the independent random vectors with fully independent Gaussian vectors while preserving the variance. We can then apply our techniques for Gaussians to bound the variance of the final random variable. The case $p = 1$ is useful to illustrate the technique, even though Schatten 1-norm approximation can be easily accomplished in other ways. Let $G \in \mathbb{R}^{t\times n}$ be the sparse JLT matrix and let $A \in \mathbb{R}^{n\times n}$ be PSD. The sketch is GAG^T and

$$\mathrm{Tr}(GAG^T) - \mathrm{Tr}(A) = \sum_{i\neq j} a_{i,j}\langle g_i, g_j\rangle. \tag{2}$$

Since $i \neq j$, g_i and g_j are independent. However the summands are subtly dependent. We first decouple the summand by choosing $\delta_i \sim$Bernoulli(1/2), and write $\langle g_i, g_j\rangle = 4\,\mathbb{E}(\delta_i(1-\delta_j)\langle g_i, g_j\rangle)$. Let $V = \{i : \delta_i = 1\}$, then $\sum_{i\neq j} a_{i,j}\langle g_i, g_j\rangle = 4\,\mathbb{E}_\delta \sum_{i\in V, j\in\bar{V}} a_{i,j}\langle g_i, g_j\rangle$. Thus conditioning on δ and $\{g_j : j \in \bar{V}\}$, the set $\{\langle g_i, \sum_{j\in\bar{V}} a_{i,j} g_j\rangle : i \in V\}$ is a set of independent random variables. We can match these random variables with Gaussian random variables of the same variance, and thus

replace g_i with independent Gaussian vectors. The same process can be repeated for $g_j : j \in \bar{V}$, and replace every vector $g_i : i \in [n]$ by independent Gaussian vectors. This lets us apply similar techniques as used in the proof of Proposition 3.4 to bound the variance of the resulting random variable, and thus bound the variance of the original random variable $\mathrm{Tr}(GAG^T) - \mathrm{Tr}(A)$.

The analogue of (2) for the case of our general estimator, $X - \mathrm{Tr}(A^p)$, is much more complicated than the $p = 1$ case. We observe that these terms can be grouped as a sum of products of consecutive *walks*, i.e.,
$a_{i_1,i_2}a_{i_2,i_3}\dots a_{i_z,j_{z+1}}\langle g^{(z+1)}_{j_{z+1}}, g^{(z+1)}_{i_{z+1}}\rangle$ for some z. Notice that $\langle g^{(z')}_{j'}, g^{(z')}_{j'}\rangle = 1$ for any j' and z'. For each walk, we can apply similar idea to replace the g_i vectors with independent Gaussian vectors. Again, we apply similar techniques as used in the proof of Proposition 3.4 to bound the variance of each group. As a result, when replacing the Gaussian matrices by sparse JLT matrices, Lemma 3.5 still holds.

Using the sparse ZD-sketch, we are able to achieve the same space bound and query time as in Theorem 3.3 and Theorem 3.6. But our update time is improved to $O(1/\epsilon^2)$. We present the full statement of our theorem below. The full proof can be found in are presented in Section **??**.

Theorem 3.8. *For every $0 < \epsilon < 1/2$ and integer $p \geq 2$, there is a randomized one-pass streaming algorithm $\mathcal{A}$ with space requirement $O(n^{2-4/p}/\epsilon^2)$, that given as input a PSD matrix $A \in \mathbb{R}^{n\times n}$, outputs with high probability a $(1+\epsilon)$-approximation of $\|A\|^p_{S_p}$. The algorithm processes an update in time $O(1/\epsilon^2)$, and computes the output (after the updates) in time $O(n^{(1-2/p)\omega})/\epsilon^2$, where $\omega < 3$ is the matrix multiplication constant.*

There is similarly a randomized $\lceil p/2\rceil$-pass streaming algorithm $\mathcal{B}$ with space requirement $O(n^{1-1/(p-1)}/\epsilon^2)$, update time in a pass $O(1/\epsilon^2)$, and output time $O(n^{(1-2/p)}/\epsilon^2)$.

For even $p \geq 2$, both algorithms extend to general input $A \in \mathbb{R}^{n\times m}$ with $m \leq n$.

4 Lower Bound For PSD Matrices

Theorem 4.1. *Suppose that $p > 0$ and $X \in \mathbb{R}^{n\times n}$ is a PSD matrix given in the entry-wise streaming model.*

(a) When $p \in \mathbb{Z}$, there is $c = c(p) > 0$ such that every one-pass streaming algorithm that $(1+c)$-approximates $\|X\|_{S_p}$ with probability $2/3$ must use $\Omega_p(n^{1-2/p})$ bits of space for even p, and $\Omega_p(n^{1-2/(p-1)})$ bits of space for odd p.

(b) When $p \notin \mathbb{Z}$, for every integer $t \geq 2$, there is $c = c(p,t) > 0$ such that every one-pass streaming algorithm that $(1+c)$-approximates $\|X\|_{S_p}$ with probability $2/3$ must use $\Omega_{p,t}(n^{1-1/t})$ bits of space.

Proof. Let M be drawn from the hard input distribution for even integer p in [LW16a], which involves an integer parameter t but does not depend on the value of p. This M is drawn from one of two distributions with the properties that (i) for each even integer $r \geq 2t$, there exist a threshold L and a small constant η which both depend on r and t such that with high probability, $\|M\|^r_r \geq (1+\eta)L$ when M is drawn from one distribution and $\|M\|^r_r \leq (1-\eta)L$ when M is drawn from the other distribution; (ii) for any even integer $r < 2t$, there is no such gap in $\|M\|^r_r$ between the two distributions; (iii) distinguishing which distribution M is drawn from requires $\Omega_t(n^{1-1/t})$ bits of space for one-pass streaming algorithms, even in the insertion-only model.

It was also proved in [LW16a] that the maximum singular value of M is at most t. Then $A = \begin{pmatrix} tI_n & M \\ M^T & tI_n \end{pmatrix}$ is positive semidefinite since its eigenvalues are $t \pm \sigma_1(M)$, ..., $t \pm \sigma_n(M)$, all of

which are non-negative. We shall show that there is a constant-factor gap in $\|A\|_{S_p}^p$ when M is drawn from the two distributions, then the same lower bound in property (iii) above follows.

Consider two distributions over the PSD matrices of the above form, induced by the two distributions of M, respectively. Recall that if $p > 0$ is not an integer,

$$\forall |x| \le 1, \quad (1+x)^p = \sum_{k=0}^{\infty} \binom{p}{k} x^k.$$

Hence when $|\sigma| \le t$,

$$(t+\sigma)^p + (t-\sigma)^p = 2t^p \sum_{k \text{ even}} \binom{p}{k} \left(\frac{\sigma}{t}\right)^k.$$

Thus

$$\|A\|_{S_p}^p = 2 \sum_{k \text{ even}} \binom{p}{k} t^{p-k} \|M\|_k^k.$$

The existence of a gap in $\|A\|_{S_p}^p$ follows immediately from properties (i) and (ii) above. □

We remark that all lower bounds in Theorem 4.1 even hold for sparse matrices, since the hard instances are sparse. The lower bounds for non-integers p and even integers p are strengthenings of the same lower bounds in [LW16a], and are almost tight and tight up to polylogarithmic factors, respectively.

5 Row-Order Model: Lower Bounds

First we discuss lower bounds for estimating Schatten norms in the row-order model. Suppose that G is a graph with n nodes and $m = O(n)$ edges. Let $M \in \mathbb{R}^{m \times n}$ be the incidence matrix of G and $L \in \mathbb{R}^{n \times n}$ be the Laplacian matrix of G, then $L = M^T M$ and thus $\|M\|_{S_p}^p = \|L\|_{S_{p/2}}^{p/2}$. Similarly to the approach in [LW16a], we shall need a lower bound on distinguishing two families of graphs, while some matrix derived from the graph has different Schatten norms in the two cases. The lower bound on distinguishing graphs we shall use is due to Kogan and Krauthgamer [KK15] based on the Boolean Hypermatching Problem [VY11], defined as follows.

Proposition 5.1 ([KK15]). *Let $t \ge 2$ be an integer, and let G be an undirected 2-regular graph on n nodes consisting of either (a) vertex-disjoint $(2t+1)$-cycles or (b) vertex-disjoint $(4t+2)$-cycles. Every randomized one-pass insertion-only streaming algorithm that, with probability at least $2/3$, determines whether G is of type (a) or type (b) must use $\Omega_t(n^{1-1/t})$ bits of space.*

The next lemma shows that the Laplacian matrix has different Schatten p-norms between the two cases in the hard instance.

Lemma 5.2. *Suppose that $t \ge 2$ is an integer and $p > 0$ is not an integer. Let G be a graph as in Proposition 5.1, then the Schatten p-norm of the Laplacian matrix of G is different by a constant factor $c(t,p) \ne 1$ between the two types.*

Proof. Let $m = 4t+2$. To prove the lemma it suffices to show a gap in the Schatten-p norm of the Laplacian matrix between two $(m/2)$-cycles and one m-cycles. Let L_m denote the Laplacian matrix of an m-cycle. Since L_m is circulant, its eigenvalues (and thus singular values since L is PSD) are given by the following explicit expression:

$$\sigma_{m,j} = 2 - \omega_m^j - \omega_m^{j(m-1)}, \quad j = 0, \ldots, m-1$$

where

$$\omega_m = e^{2\pi i \frac{\pi}{m}}.$$

Thus

$$\|L_m\|_{S_p}^p = \begin{cases} 2\sum_{i=1}^{\lfloor m/2 \rfloor} \sigma_{m,j}^p, & m \text{ is odd;} \\ 4^p + 2\sum_{i=1}^{m/2-1} \sigma_{m,j}^p, & m \text{ is even.} \end{cases}$$

When m is an even integer, the eigenvalues of $L_{m/2}$ are eigenvalues of L_m, more specifically, $\sigma_{m/2,j} = \sigma_{m,2j}$. It follows that

$$2\|L_{m/2}\|_{S_p}^p - \|L_m\|_{S_p}^p = 2\sum_{j=1}^{\frac{n}{2}-1} (-1)^j \sigma_{m,j}^p - 4^p,$$

where we used the fact that $m = 4t+2$ in our setting and thus $m/2 = 2t+1$ is odd. Note that

$$\sigma_{m,j} = 2 - 2\cos\frac{2j\pi}{m} = 4\sin^2\frac{j\pi}{m},$$

we have that

$$2\|L_{m/2}\|_{S_p}^p - \|L_m\|_{S_p}^p = 2\cdot 4^p \sum_{j=1}^{\frac{n}{2}-1} (-1)^j \sin^{2p}\frac{j\pi}{m} - 4^p.$$

Our goal is therefore to show that

$$\sum_{j=1}^{\frac{n}{2}-1} (-1)^j \sin^{2p}\frac{j\pi}{m} \neq \frac{1}{2}.$$

Consider the Fourier cosine series expansion

$$\sin^{2p}\frac{j\pi}{m} = \frac{1}{2^{2p}}\frac{\Gamma(2p+1)}{\Gamma(p+1)^2} + \frac{1}{2^{2p-1}}\sum_{k=1}^{\infty} \frac{(-1)^k\Gamma(2p+1)}{\Gamma(p+k+1)(p-k+1)} \cos\left(2kj\frac{\pi}{m}\right),$$

where the Fourier coefficient can be obtained by using Binomial Theorem and Gauss Theorem for hypergeometric functions ${}_2F_1$ to evaluate the following integral (cf. Exercise 44 on p123 of [AAR99])

$$\int_{-\pi/2}^{\pi/2} (1-e^{2i\theta})^{p-k}(1-e^{-2i\theta})^{p+k} d\theta.$$

Next, observe that

$$\sum_{j=1}^{\frac{m}{2}-1} (-1)^j \cos\left(2kj\frac{\pi}{m}\right) = \begin{cases} 0, & k \text{ is even;} \\ \frac{m}{2}-1, & k \equiv \frac{m}{2} \pmod m; \\ -1, & \text{otherwise.} \end{cases} \tag{3}$$

The problem reduces to evaluate

$$S := \frac{1}{2^{2p-1}}\Bigg\{ \sum_{\substack{\text{odd } k \\ k \not\equiv \frac{m}{2}-1 (\text{mod } m)}} \gamma_p(k) - \left(\frac{m}{2}-1\right) \sum_{k \equiv \frac{m}{2}-1 (\text{mod } m)} \gamma_p(k) \Bigg\},$$

where

$$\gamma_p(k) = \frac{\Gamma(2p+1)}{\Gamma(p+k+1)\Gamma(p-k+1)}.$$

Observe that $\gamma_p(k) > 0$ for $k \le \lceil p \rceil$, $\gamma_p(\lceil p \rceil + 1) < 0$ and $\gamma_p(k)$ has alternating signs for $k \ge \lceil p \rceil + 1$.

When $\lceil p \rceil$ is even, it holds that $\gamma_p(k) < 0$ for $k \equiv m/2 - 1 (\mathrm{mod}\ m)$ and thus

$$S > \frac{1}{2^{2p-1}} \sum_{\text{odd } k} \gamma_p(k);$$

when $\lceil p \rceil$ is odd, it holds that $\gamma_p(k) > 0$ for $k \equiv m/2 - 1 (\mathrm{mod}\ m)$ and thus

$$S < \frac{1}{2^{2p-1}} \sum_{\text{odd } k} \gamma_p(k).$$

The result follows immediately once the following identity is established:

$$\frac{1}{2^{2p-1}} \sum_{\text{odd } k} \gamma_p(k) = \frac{1}{2^{2p-1}} \sum_{\text{odd } k} \frac{\Gamma(2p+1)}{\Gamma(p+k+1)\Gamma(p-k+1)} = \frac{1}{2}. \tag{4}$$

Consider the integral representation

$$\frac{\Gamma(2p+1)}{\Gamma(p+k+1)\Gamma(p-k+1)} = \frac{1}{2\pi i} \int_{-\infty}^{(0+)} t^{-(p-k)-1}(1-t)^{-(p+k)-1} dt,$$

where the contour integral goes from the upper edge of the negative real axis from $-\infty$ to 0, then goes clockwise around 0, and returns to $-\infty$ along the lower edge of the negative real axis. Summing under the integral yields that

$$\begin{aligned}
&\sum_{\text{odd } k} \gamma_p(k) \\
&= \frac{1}{2\pi i} \int_{-\infty}^{(0+)} \frac{1}{(1-2t)t^p(1-t)^p} dt \\
&= \frac{\sin(p\pi)}{\pi} {}_2F_1\left(\begin{matrix} 1, 1-p \\ 1+p \end{matrix} ; -1 \right) \frac{\Gamma(2p)\Gamma(1-p)}{\Gamma(1+p)} \\
&= \frac{\sin(p\pi)}{\pi} \cdot \frac{\Gamma(1+p)\Gamma(\frac{3}{2})}{\Gamma(2)\Gamma(\frac{1}{2}+p)} \cdot \frac{\Gamma(2p)\Gamma(1-p)}{\Gamma(1+p)} \\
&= \frac{\sin(p\pi)}{\pi} \cdot \frac{\sqrt{\pi}/2}{\Gamma(\frac{1}{2}+p)} \cdot \frac{2^{2p-1}}{\sqrt{\pi}} \Gamma(p)\Gamma\left(p+\frac{1}{2}\right)\Gamma(1-p) \\
&= 2^{2p-2},
\end{aligned}$$

where we used the integral representation of hypergeometric function ${}_2F_1$ (Equation (2.3.17) in [AAR99]) for the second equality, Kummer's identity ([AAR99, Corollary 3.1.2]) for the third equality, Legendre's duplication formula ([AAR99, Theorem 1.5.1]) for the fourth and Euler Reflection Formula ([AAR99, Theorem 1.2.1]) for the last equality. This establishes (4). □

The next theorem follows easily by combining Proposition 5.1 and Lemma 5.2.

Theorem 5.3. *Suppose that $t \ge 2$ is an integer and $p > 0$ is not an even integer. Every randomized streaming algorithm that with probability at least 2/3 estimates $\|A\|_{S_p}^p$ within factor $1+\epsilon$, for $\epsilon < \epsilon_0(t,p)$, when the input $A \in \mathbb{R}^{n \times n}$ is sparse and given in row-order model, must use $\Omega_t(n^{1-1/t})$ bits of space.*

Proof. Let A be the incidence matrix of G in Proposition 5.1. Since G has exactly n edges, the size of A is exactly $n \times n$. In the streaming model for G, each update describes an edge, which corresponds to a row of A. Thus a stream of G corresponds to a stream of A in row-order model. By Lemma 5.2, a Schatten-norm algorithm can distinguish the type of G, and the lower bound therefore follows from Proposition 5.1. □

The theorem above gives a nearly tight bound for estimating the Schatten p-norm for sparse matrices and $p \notin 2\mathbb{Z}$. For $p \in 2\mathbb{Z}$ we have the following theorem.

Theorem 5.4. *Suppose that $t \geq 2$ is an integer and $p \geq 4$ is an even integer. Every randomized streaming algorithm that estimates the Schatten p-norm of the input matrix up to a constant factor (depending on p) with probability $\geq 2/3$ in the row-order model must use $\Omega(n^{1-4/p})$ bits of space. This lower bound holds even for multi-pass algorithms.*

Proof. We reduce the problem to the communication complexity of multiparty Set-Disjointness [Gro09]. Suppose there are $k = 2n^{2/p}$ players. Each player is given a set in $\{1, \ldots, n\}$. Let A be an empty matrix and we shall show how to construct A according to the input of the Set-Disjointness problem. For each element j in each player's set, we add a row e_j to A, where e_j is the j-th row of the $n \times n$ identity matrix. With high probability the hard instance of multiparty Set-Disjointness has $m \leq n$ elements, and thus A will have m rows. By padding we may assume that A is $n \times n$, and is clearly given in the row-order model.

When the players' sets are disjoint, it is clear that all singular values of A are 1 and thus $\|A\| = m \leq n$. When the players' set have a common element, there is a singular value of $\sqrt{k}$ and hence $\|A\|_{S_p}^p \geq k^{p/2} = 2^{p/2}n$. Therefore $\|A\|_{S_p}^p$ is different by a constant factor in the two cases.

The communication complexity lower bound of unrestricted protocols for Set-Disjointness is $\Omega(n/k)$ bits, which implies that the streaming lower bound for estimating Schatten p-norm is $\Omega(n/k^2) = \Omega(n^{1-4/p})$ bits, even for multi-pass algorithms. □

As discussed in Introduction, Theorem 5.3 is asymptotically tight up to logarithmic factors for $p \in 4\mathbb{Z}$. For the remaining case $p \equiv 2 \pmod 4$, we present an algorithm using $\tilde{O}(n^{1-4/(p+2)})$ space in Section 6, leaving a slight polynomial gap from the lower bound of $\Omega(n^{1-4/p})$.

6 Row-Order Model: Algorithm For Even p

In this section, we present an algorithm which estimates the Schatten p-norm (where $p \equiv 2 \pmod 4$ is an integer) of $n \times n$ sparse matrices in row-order model using $\tilde{O}(n^{1-4/(p+2)})$ bits of space. The following algorithm is in a similar flavour to the algorithm in [LW16a], where the Precision Sampling structure was used to sample rows of a matrix proportionally to their row norms, and we shall omit such details in this section. Since we are reading A in row-order model, we can sample and obtain rows of A exactly with weighted reservoir sampling [ES06], but we shall use Precision Sampling to sample rows of A^TA.

Theorem 6.1. *Suppose that $p = 4k + 2$ for some integer $k \geq 1$ and $A \in \mathbb{R}^{n \times n}$ is a sparse matrix given in one-pass row-order model. Algorithm 1 returns Y such that $(1-\epsilon)\|A\|_{S_p}^p \leq Y \leq (1+\epsilon)\|A\|_{S_p}^p$ with probability $\geq 2/3$, using space $O_p(n^{1-\frac{4}{p+2}} \operatorname{poly}(1/\epsilon, \log n))$.*

Proof. The analysis is similar to [LW16a]. Let $B = A^TA$, $L = \|B\|_F^2$ and $Z = \|A\|_F^2$. For a matrix M we shall denote its i-row by M_i. We also denote by $\tilde{B}_i$ the approximation recovered by the algorithm to B_i. For notational convenience, we also define $K_1 = \cdots = K_k = K$ and $K_{k+1} = V$.

Algorithm 1 Algorithm for $p = 4k + 2$ and sparse matrices in row order model

Assume that matrix A has at most $O(1)$ non-zero entries per row and per column and is given in row order model and that $p = 4k + 2$ for some integer $k \geq 1$.

$T \leftarrow \Theta(\epsilon^{-2} n^{1-1/(k+1)})$

Maintain a sketch for estimating $\|A^T A\|_F^2$ and obtain an $(1+\epsilon)$-approximation L'

$Z \leftarrow \|A\|_F^2$ ▷ Can be computed exactly in row-order model

$K \leftarrow$ set of indices of rows of $A^T A$ with norm $\geq \sqrt{L'/(10T)}$ ▷ Count-Sketch, see [LW16a]

$V \leftarrow$ set of indices of rows of A with norm $\geq \sqrt{Z/(10T)}$ ▷ Maintaining $10T$ rows of largest norm

for $s = 1, \ldots, k$ **do**

  Sample T rows of $A^T A$ proportionally to row norm ▷ Precision sampling, see [LW16a]

  Obtain approximation to the sampled rows ▷ By-product of precision sampling, see [LW16a]

  $I_s \leftarrow$ the set of the indices of the sampled rows

  $I_s \leftarrow I_s \cup K$

end for

Sample T rows of A proportionally to row norm ▷ Reservoir sampling

$I_{k+1} \leftarrow$ the set of the indices of the sampled rows

$I_{k+1} \leftarrow I_{k+1} \cup V$

Return Y as defined in (5)

Next, we define for $s = 1, \ldots, k$

$$\tau_s(i) = \begin{cases} 1, & i \in K; \\ L/\|B_i\|_2^2, & i \in I_s \setminus K, \end{cases}$$

$$\tilde{\tau}_s(i) = \begin{cases} 1, & i \in K; \\ L'/\|\tilde{B}_i\|_2^2, & i \in I_s \setminus K \end{cases}$$

and

$$\tilde{\tau}_{k+1}(i) = \tau_{k+1}(i) = \begin{cases} 1, & i \in V; \\ Z/\|A_i\|_2^2, & i \in I_{k+1} \setminus V. \end{cases}$$

We further define

$$X(i_1, \ldots, i_t) = \prod_{j=1}^{k} \langle B_{i_j}, B_{i_{j+1}} \rangle \langle B_{i_{j+1}}, A_{i_{k+1}} \rangle \langle A_{i_{k+1}}, B_{i_1} \rangle \cdot \tau_1(i_1) \cdots \tau_{k+1}(i_{k+1}),$$

and

$$\tilde{X}(i_1, \ldots, i_t) = \prod_{j=1}^{k} \langle \tilde{B}_{i_j}, \tilde{B}_{i_{j+1}} \rangle \langle \tilde{B}_{i_{j+1}}, A_{i_{k+1}} \rangle \langle A_{i_{k+1}}, \tilde{B}_{i_1} \rangle \cdot \tilde{\tau}_1(i_1) \cdots \tilde{\tau}_{k+1}(i_{k+1}).$$

Since $B = A^T A$ is PSD, it holds that

$$\begin{aligned}
\|A\|_{S_p}^p &= \mathrm{Tr}(B^{p/2}) \\
&= \sum_{i_1} e_{i_1}^T B \cdot \underbrace{(B^T B) \cdots (B^T B)}_{k-1 \text{ times}} \cdot B \cdot B e_{i_1} \\
&= \sum_{i_1} B_{i_1} \cdot \underbrace{(B^T B) \cdots (B^T B)}_{k-1 \text{ times}} \cdot A^T A \cdot B_{i_1}^T \\
&= \sum_{i_1,\ldots,i_{k+1}} B_{i_1} (B_{i_2}^T B_{i_2}) \cdot (B_{i_k}^T B_{i_k}) (A_{i_{k+1}}^T A_{i_{k+1}}) B_{i_1}^T \\
&= \sum_{i_1,\ldots,i_{k+1}} \langle B_{i_j}, B_{i_{j+1}} \rangle \langle B_{i_{j+1}}, A_{i_{k+1}} \rangle \langle A_{i_{k+1}}, B_{i_1} \rangle.
\end{aligned}$$

Our estimator is

$$Y = \sum_{i \in I_1,\ldots,i_{k+1} \in I_{k+1}} \frac{1}{T^{\sigma(i_1,\ldots,i_{k+1})}} \tilde{X}(i_1,\ldots,i_{k+1}), \tag{5}$$

where

$$T^{\sigma(i_1,\ldots,i_{k+1})} = |\{1 \le s \le k+1 : i_s \notin K_s\}|$$

Following a similar analysis to that in [LW16a], we have that

$$\left| \mathbb{E} Y - \|A\|_{S_p}^p \right| \le \epsilon \|A\|_{S_p}^p,$$

where we have crucially used the fact that A and $A^T A$ are sparse matrices. The variance bound is similar, too. The covariance terms are sums over $i_1,\ldots,i_{k+1},i'_1,\ldots,i'_{k+1}$ and we split them into two kinds depending on whether $i_{k+1} = i'_{k+1}$. Eventually we shall have

$$\mathbb{E} Y^2 - (\mathbb{E} Y)^2 \lesssim \sum_{r=1}^{k} \frac{1}{T^r} \|B\|_F^{2r} \|A\|_{S_{2p-4r+4}}^{2p-4r+4} + \sum_{r=1}^{k+1} \frac{1}{T^r} \|B\|_F^{2(r-1)} \|A\|_F^2 \|A\|_{S_{2p-4r+2}}^{2p-4r+2} \lesssim \sum_{r=1}^{k+1} \frac{1}{T^r} n^{r-\frac{4r-2}{p}} \|A\|_{S_p}^{2p},$$

which implies that

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

if the constant C in $T = C n^{1-1/(k+1)}/\epsilon^2$ is large enough. □

References

[AAR99] G. E. Andrews, R. Askey, and R. Roy. *Special Functions*. Cambridge University Press, 1999.

[AKR15] A. Andoni, R. Krauthgamer, and I. Razenshteyn. Sketching and embedding are equivalent for norms. In *47th Annual ACM Symposium on Theory of Computing*, pages 479–488. ACM, 2015. doi:10.1145/2746539.2746552.

[AN13] A. Andoni and H. Nguyên. Eigenvalues of a matrix in the streaming model. In *24th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1729–1737. SIAM, 2013. doi:10.1137/1.9781611973105.124.

[BJKS02] Z. Bar-Yossef, T. S. Jayram, R. Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. In *43rd Annual IEEE Symposium on Foundations of Computer Science*, pages 209–218. IEEE, 2002. doi:10.1109/SFCS.2002.1181944.

[DKS10] A. Dasgupta, R. Kumar, and T. Sarlós. A sparse johnson: Lindenstrauss transform. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 341–350. ACM, 2010.

[DNPS16] E. Di Napoli, E. Polizzi, and Y. Saad. Efficient estimation of eigenvalue counts in an interval. *Numerical Linear Algebra with Applications*, 23(4):674–692, 2016.

[ES06] P. S. Efraimidis and P. G. Spirakis. Weighted random sampling with a reservoir. *Information Processing Letters*, 97(5):181 – 185, 2006.

[Gro09] A. Gronemeier. Asymptotically Optimal Lower Bounds on the NIH-Multi-Party Information Complexity of the AND-Function and Disjointness. In *26th International Symposium on Theoretical Aspects of Computer Science*, volume 3, pages 505–516, Dagstuhl, Germany, 2009. doi:10.4230/LIPIcs.STACS.2009.1846.

[HK15] H. Huang and S. P. Kasiviswanathan. Streaming anomaly detection using randomized matrix sketching. *Proc. VLDB Endow.*, 9(3):192–203, November 2015. doi:10.14778/2850583.2850593.

[IM98] P. Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *30th Annual ACM Symposium on Theory of Computing*, pages 604–613. ACM, 1998. doi:10.1145/276698.276876.

[JW13] T. S. Jayram and D. P. Woodruff. Optimal bounds for Johnson-Lindenstrauss transforms and streaming problems with subconstant error. *ACM Transactions on Algorithms*, 9(3):26, 2013. doi:10.1145/2483699.2483706.

[KK15] D. Kogan and R. Krauthgamer. Sketching cuts in graphs and hypergraphs. In *Conference on Innovations in Theoretical Computer Science*, pages 367–376. ACM, 2015. doi:10.1145/2688073.2688093.

[KMN11] D. Kane, R. Meka, and J. Nelson. Almost optimal explicit Johnson-Lindenstrauss families. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 628–639. Springer, 2011. doi:10.1007/978-3-642-22935-0_53.

[KN14] D. M. Kane and J. Nelson. Sparser Johnson-Lindenstrauss transforms. *Journal of the ACM*, 61(1):4, 2014. doi:10.1145/2559902.

[KO17] A. Khetan and S. Oh. Spectrum estimation from a few entries. *arXiv preprint arXiv:1703.06327*, 2017.

[KPW16] M. Kapralov, V. Potluru, and D. Woodruff. How to fake multiply by a gaussian matrix. In *International Conference on Machine Learning*, pages 2101–2110, 2016.

[KV16] W. Kong and G. Valiant. Spectrum estimation from samples. *arXiv preprint arXiv:1602.00061*, 2016.

[Le 14] F. Le Gall. Powers of tensors and fast matrix multiplication. In *39th International Symposium on Symbolic and Algebraic Computation*, pages 296–303. ACM, 2014. doi:10.1145/2608628.2608664.

[Lib13] E. Liberty. Simple and deterministic matrix sketching. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 581–588. ACM, 2013. doi:10.1145/2487575.2487623.

[LNW14] Y. Li, H. L. Nguyen, and D. P. Woodruff. On sketching matrix norms and the top singular vector. In *25th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA, pages 1562–1581. SIAM, 2014. doi:10.1137/1.9781611973402.114.

[LW16a] Y. Li and D. P. Woodruff. On approximating functions of the singular values in a stream. In *48th Annual ACM Symposium on Theory of Computing*, pages 726–739. ACM, 2016. doi:10.1145/2897518.2897581.

[LW16b] Y. Li and D. P. Woodruff. Tight bounds for sketching the operator norm, Schatten norms, and subspace embeddings. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, volume 60 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 39:1–39:11. Schloss Dagstuhl, 2016. doi:10.4230/LIPIcs.APPROX-RANDOM.2016.39.

[LW17] Y. Li and D. P. Woodruff. Embeddings of Schatten Norms with Applications to Data Streams. In *44th International Colloquium on Automata, Languages, and Programming (ICALP 2017)*, volume 80 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 60:1–60:14. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017. doi:10.4230/LIPIcs.ICALP.2017.60.

[RV13] M. Rudelson and R. Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electron. Commun. Probab*, 18(82):1–9, 2013. doi:10.1214/ECP.v18-2865.

[TZ04] M. Thorup and Y. Zhang. Tabulation based 4-universal hashing with applications to second moment estimation. In *SODA*, volume 4, pages 615–624, 2004.

[VY11] E. Verbin and W. Yu. The streaming complexity of cycle counting, sorting by reversals, and other problems. In *Proceedings of the 22nd ACM-SIAM SODA*, pages 11–25, 2011. doi:10.1137/1.9781611973082.2.

[WLL+16] Z. Wei, X. Liu, F. Li, S. Shang, X. Du, and J.-R. Wen. Matrix sketching over sliding windows. In *Proceedings of the 2016 International Conference on Management of Data*, SIGMOD '16, pages 1465–1480. ACM, 2016. doi:10.1145/2882903.2915228.

[Woo14] D. P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10:1–157, 2014. doi:10.1561/0400000060.

[ZWJ15] Y. Zhang, M. Wainwright, and M. Jordan. Distributed estimation of generalized matrix rank: Efficient algorithms and lower bounds. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 457–465, 2015.

A Proof of Proposition 3.4

Proof. Using the identity $\mathrm{Tr}(MM^T) = \mathrm{Tr}(M^TM)$ we have

$$X = \mathrm{Tr}\left(G_1AG_2^TG_2A\cdots G_p^TG_pA\cdot G_1^T\right) = \mathrm{Tr}\left(G_1^T\cdot G_1AG_2^TG_2A\cdots G_p^TG_pA\right).$$

By linearity of trace, expectation and matrix product, and by the fact that $\mathbb{E}[G_i^TG_i] = I_{n\times n}$ for all $i\in[p]$, we have

$$\begin{aligned}\mathbb{E}\,X &= \mathbb{E}\,\mathrm{Tr}\left(G_1^T\cdot G_1AG_2^TG_2A\cdots G_p^TG_pA\right)\\ &= \mathbb{E}\,\mathrm{Tr}\left(I_{n\times n}AG_2^TG_2A\cdots G_p^TG_pA\right)\\ &= \cdots = \mathrm{Tr}(A^p).\end{aligned}$$

It remains to bound the variance of X. Without loss of generality we can assume that A is a diagonal matrix $\mathrm{diag}(\lambda_1,\lambda_2,\ldots,\lambda_n)$, where $\lambda_1\geq\cdots\geq\lambda_n$. Indeed, in the case of a general symmetric A, we can write $A = U\Lambda U^T$ for an orthonormal matrix U and a diagonal matrix Λ. Then $G_iAG_{i+1}^T = (G_iU)\Lambda(G_{i+1}U)^T$, and the matrices $\{G_iU\}_{i\in[p]}$ have the same joint distribution as $\{G_i\}_{i\in[p]}$, hence $\mathrm{Var}(X)$ would not change if A is replaced with Λ.

Let us write $G_i = (g_1^{(i)}, g_2^{(i)}, \ldots, g_n^{(i)})$, where each $g_j^{(i)}\in\mathbb{R}^t$ is a column vector. It is easily verified that

$$X = \sum_{i_1,i_2,\ldots,i_p\in[n]}\lambda_{i_1}\lambda_{i_2}\cdots\lambda_{i_p}\langle g_{i_1}^{(1)}, g_{i_2}^{(1)}\rangle\langle g_{i_2}^{(2)}, g_{i_3}^{(2)}\rangle\cdots\langle g_{i_p}^{(p)}, g_{i_1}^{(p)}\rangle.$$

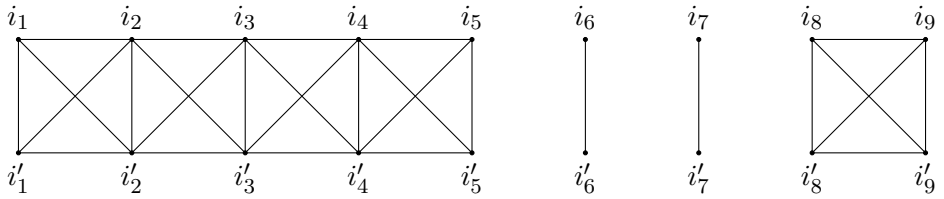

Figure 1: An example of a non-zero variance term ($p = 9$): $i_1 = \cdots = i_5 = i'_1 = i'_2 \cdots = i'_5$, $i_6 = i'_6$, $i_7 = i'_7$, $i_8 = i_9 = i'_8 = i'_9$ and i_1, i_6, i_7, i_8 are distinct. The term of the eigenvalues in the variance expression is $\lambda_{i_1}^{10}\lambda_{i_6}^2\lambda_{i_7}^2\lambda_{i_8}^4$.

Indeed, first write

$$(G_1^T G_1 A)_{i_1,i_2} = \sum_{k\in[t]} (G_1^T)_{i_1,k}(G_1)_{k,i_2}A_{i_2,i_2} = \langle g_{i_1}^{(1)}, g_{i_2}^{(1)}\rangle \lambda_{i_2},$$

and then expand the trace in
$X = \operatorname{Tr}\left(G_1^T G_1 A \cdot G_2^T G_2 A \cdots G_p^T G_p A\right)$ using all closed walks $(i_1, i_2, \ldots, i_p) \in [n]^p$.

It is not difficult to verify that for all $j \neq j' \in [n]$ and $i_1, i_2, i'_1, i'_2 \in [p]$,

$$\mathbb{E}[\langle g_{i_1}^{(j)}, g_{i_2}^{(j)}\rangle] = \mathbb{1}_{\{i_1=i_2\}}, \tag{6}$$

$$\mathbb{E}[\langle g_{i_1}^{(j)}, g_{i_2}^{(j)}\rangle \langle g_{i'_1}^{(j')}, g_{i'_2}^{(j')}\rangle] = \mathbb{1}_{\{i_1=i_2, i'_1=i'_2\}}. \tag{7}$$

$$\mathbb{E}[\langle g_{i_1}^{(j)}, g_{i_2}^{(j)}\rangle \langle g_{i'_1}^{(j)}, g_{i'_2}^{(j)}\rangle] = \mathbb{1}_{\{(i_1,i'_1)=(i_2,i'_2)\}} + \tfrac{1}{t}\mathbb{1}_{\{(i_1,i_2)=(i'_1,i'_2)\}} + \tfrac{1}{t}\mathbb{1}_{\{(i_1,i_2)=(i'_2,i'_1)\}}. \tag{8}$$

Notice that in the last equation, the events in the three indicators are not disjoint, and when $i_1 = i'_1 = i_2 = i'_2$ the righthand-side evaluates to $1 + 2/t$.

We proceed to bound $\operatorname{Var}(X) \leq \mathbb{E}[X^2]$. Denoting $I = (i_1, i_2, \ldots, i_p) \in [n]^p$ with the convention $i_{p+1} := i_1$, and similarly for I', we can write

$$X^2 = \left(\sum_I \prod_{j\in[p]} \lambda_{i_j} \langle g_{i_j}^{(j)}, g_{i_{j+1}}^{(j)}\rangle\right)^2 = \sum_{I,I'} \prod_{j\in[p]} \lambda_{i_j}\lambda_{i'_j} \langle g_{i_j}^{(j)}, g_{i_{j+1}}^{(j)}\rangle \langle g_{i'_j}^{(j)}, g_{i'_{j+1}}^{(j)}\rangle. \tag{9}$$

We can represent each term of X^2 (a fixed choice for I, I') by a diagram (see an example in Figure 1). Each node in the diagram represents an index i_j, and each square corresponds to a factor of the form $\langle g_{i_j}^{(j)}, g_{i_{j+1}}^{(j)}\rangle \langle g_{i'_j}^{(j)}, g_{i'_{j+1}}^{(j)}\rangle$. A line connecting two nodes represents that the respective indices are equal. Notice that for each square, if a vertical line exists, then both vertical lines must exist, otherwise the expectation of this square is zero, and it has no contribution to $\mathbb{E}[X^2]$. Thus, for a non-zero diagram, if it has at least one vertical line, then it actually has all possible vertical lines. Diagrams with no vertical lines can be non-zero diagrams only if they are made entirely by horizontal cross-lines and parallel lines, which we call the *trivial diagrams*, and they correspond to the coefficient of $\lambda_i^p \lambda_j^p$ for $i \neq j$. Each non-trivial diagram corresponds to an *integer partition* of p (i.e., a way of writing the integer p as the sum of positive integers, with the order of the summands/parts having no significance), but we should account also for permutations and cyclic shifts on the parts. Given an integer partition $[p_1, p_2, \ldots, p_z]$ of p, we write it as $(p_1^{(z_1)}, p_2^{(z_2)} \cdots p_{t'}^{(z_{t'})})$,

where $p_1 \geq p_2 \geq \cdots \geq p_{t'}$ are the distinct parts (or part sizes), and z_i counts how many parts are equal to p_i. Then the number of different diagrams for a given integer partition $[p_1, p_2, \ldots, p_z]$ is

$$C_{[p_1,p_2,\ldots,p_z]} = \frac{t'!p}{z_1!z_2!\cdots z_{t'}!}.$$

Observe that this number is upper bounded by a constant M_p determined only by p. Each integer partition of p corresponds to a monomial of the eigenvalues. A connected component in the diagram corresponding to a power of the eigenvalue, and this power is just the size of that component. For each connected component, the total number of indices is an even number because of the vertical lines. For a single square, the coefficient is given by Equations (7)-(8). Using the diagram representation, we can calculate

$$\begin{aligned} \mathbb{E}\left(\boxtimes \right) = 1 + \frac{2}{t}; \qquad \mathbb{E}\left(\overline{\quad} \right) = 1; \\ \mathbb{E}\left(\| \right) = \mathbb{E}\left(\times \right) = \frac{1}{t}. \end{aligned} \tag{10}$$

All other diagrams either do not exist in the expansion of X^2, or have a zero expectation. Diagrams corresponding to the same partition of p have the same coefficient. Since for each complete square there is a factor $1 + 2/t$, and for each incomplete square there is a factor $1/t$, the coefficient for a partition $[p_1, p_2, \ldots, p_z]$ with $z > 1$ parts is

$$Z_{[p_1,p_2,\ldots,p_z]} = \frac{1}{t^z}\left(1 + \frac{2}{t}\right)^{p-z}.$$

For non-trivial diagrams (i.e., have vertical lines) with $z > 1$ (i.e., excluding the completely connected graph) we collect all such terms as X_1 and bound their expectation by

$$\mathbb{E}[X_1] \leq \sum_{[p_1,p_2,\ldots,p_z]} \frac{M_p}{t^z} \sum_{i_1,i_2,\ldots,i_z \in [n]} \lambda_{i_1}^{2p_1}\lambda_{i_2}^{2p_2}\cdots\lambda_{i_z}^{2p_z}. \tag{11}$$

For non-trivial diagrams and $z = 1$, there cannot be any incomplete square, and we can compute the expression explicitly,

$$\begin{aligned} \mathbb{E}\left[\sum_{i\in[n]} \lambda_i^{2p} \prod_{j\in[p]} \langle g_i^{(j)}, g_i^{(j)}\rangle\langle g_i^{(j)}, g_i^{(j)}\rangle\right] &= \left(1 + \frac{2}{t}\right)^p \sum_{i=1}^n \lambda_i^{2p} \\ &= 3^p \|A\|_{S_{2p}}^{2p}. \end{aligned}$$

For trivial diagrams (no vertical lines), we collect the terms as X_2 and bound their expectation by

$$\begin{aligned} &\mathbb{E}[X_2] \\ &\leq \sum_{i\neq k\in[n]} \lambda_i^p\lambda_k^p \sum_{z=0}^p \binom{p}{z} \mathbb{E}\left(\times\right)^z \mathbb{E}\left(\overline{\quad}\right)^{p-z} \\ &\leq 2^p \sum_{i\neq k\in[n]} \lambda_i^p\lambda_k^p \\ &\leq M_p' \|A\|_{S_p}^{2p}, \end{aligned}$$

where M'_p is a constant that depends only on p.

We now turn to bounding $\mathbb{E}[X_1]$ using (11). For each integer partition $[p_1, p_2, \ldots, p_z]$ of p with $z > 1$ parts,

$$\sum_{i_1,i_2,\ldots,i_z\in[n]} \lambda_{i_1}^{2p_1}\lambda_{i_2}^{2p_2}\cdots\lambda_{i_z}^{2p_z} = \left(\sum_{i_1\in[n]}\lambda_{i_1}^{2p_1}\right)\cdots\left(\sum_{i_z\in[n]}\lambda_{i_z}^{2p_z}\right)$$
$$= \prod_{j=1}^{z}\|A\|_{S_{2p_j}}^{2p_j}.$$

Let z' be the number of parts with $2p_j \le p$. Clearly, $z' \ge z-1$, since at most one part can have $p_j \ge p/2$. Consider first the case $z' = z$. It is well-known (via an application of Hölder's inequality) that $\|x\|_q \le \|x\|_r \le n^{1/r-1/q}\|x\|_q$ holds for all $x \in \mathbb{R}^n$ and $1 \le r \le q$. This comparison of norms applies also to the Schatten norms of A (viewed as n-dimensional norms of the eigenvalues of A), proves that $\|A\|_{S_{2p_j}}^{2p_j} \le n^{1/(2p_j)-1/p}\|A\|_{S_p}$. We thus obtain

$$\prod_{j=1}^{z}\|A\|_{S_{2p_j}}^{2p_j} \le \prod_{j=1}^{z}\left(n^{1/(2p_j)-1/p}\|A\|_{S_p}\right)^{2p_j} = n^{z-2}\|A\|_{S_p}^{2p}. \tag{12}$$

In the case $z' = z-1$, there is a unique j^* such that $p_{j^*} > p/2$, and therefore $z \le (p-p_{j^*})+1 \le \lfloor p/2\rfloor + 1$. For $j \ne j^*$ we can use the comparison of Schatten norms as above, and for $j = j^*$ we simply use $\|A\|_{S_{2p_j}} \le \|A\|_{S_{2p}}$. We thus obtain

$$\begin{aligned}\prod_{j=1}^{z}\|A\|_{S_{2p_j}}^{2p_j} &\le \|A\|_{S_{2p}}^{2p_{j^*}}\prod_{j\ne j^*}\left(n^{1/(2p_j)-1/p}\|A\|_{S_p}\right)^{2p_j}\\ &\le n^{z-1-2(p-p_{j^*})/p}\|A\|_{S_p}^{2p}\\ &\le n^{z-2z/p}\|A\|_{S_p}^{2p}.\end{aligned} \tag{13}$$

where the last inequality follows by $1 + 2(p-p_{j^*})/p \ge 1 + 2(z-1)/p = 2z/p + (1-2/p)$. With also the $z = 1$ term considered, we have

$$\mathrm{Var}(X) \le M''_p\left(1 + \sum_{z=2}^{\lfloor p/2\rfloor+1}\left(\frac{n^{1-2/p}}{t}\right)^z + \sum_{z=2}^{p}\left(\frac{n^{1-2/z}}{t}\right)^z\right)\|A\|_{S_p}^{2p}.$$

where M''_p is a constant depends only on p. This completes the proof of Proposition 3.4. □

B Proof of Lemma 3.5

Proof. We first argue that it suffices to prove the corollary under the assumption that the entries of G_l are fully independent. Indeed, each of the terms we need to calculate is the expectation of a polynomial of total degree at most 4 in the random variables G_{ij}. For example, the factor contains G_l in a typical term of X^2 is $\langle g_{i_l}^{(l)} g_{j_l}^{(l)}\rangle\langle g_{i'_l}^{(l)} g_{j'_l}^{(l)}\rangle$ The expectation of such a polynomial when G_l's entries are 4-wise independent is exactly the same as when these entries are fully independent.

Assume henceforth that the entries of G_l are mutually independent. We repeat the proof of Proposition 3.4, except that when considering the square containing (i_1, i_2), we replace t with t' in

(8) and (10). In diagrams where this square is complete, the contribution to $\mathbb{E}[X^2]$, as given by (9), does not change. When this square is incomplete, we replace t by t' in subsequent calculations like (11) and (12). The proof is otherwise identical, but we kept the more precise bound obtained in (13). □

C A Simple Proof for Sparse Sketch of Matrices With Non-Negative Entries

Lemma C.1. *Let $G = (g_1, g_2, \ldots, g_n) \sim \mathcal{D}_{t,n}$, then the following conditions hold.*

1. *for each $i \in [n]$, $\mathbb{E}\langle g_i, g_i\rangle = 1$, $\mathbb{E}[\langle g_i, g_i\rangle^2] = 1$;*
2. *for each $i, j \in [n], i \neq j$, $\mathbb{E}\langle g_i, g_j\rangle = 0$, $\mathbb{E}[\langle g_i, g_j\rangle^2] = 1/t$;*
3. *for each $i, j, i', j' \in [n], \{i, j\} \neq \{i', j'\}, i \neq j, i' \neq j'$, $\mathbb{E}\langle g_i, g_j\rangle = \mathbb{E}[\langle g_i, g_j\rangle\langle g_{i'}, g_{j'}\rangle] = 0$;*

Proof. Property 1 follows immediately. For 2, $\mathbb{E}\langle g_i, g_j\rangle = 0$ and

$$\begin{aligned}\mathbb{E}\langle g_i, g_j\rangle^2 = \mathbb{E}\left(\sum_l g_{i,l}g_{j,l}\right)^2 &= \sum_{l,k}\mathbb{E}(g_{i,l}g_{j,l}g_{i,k}g_{j,k}) \\ &= \sum_{l=1}^{t} E(d_i^2 d_j^2 z_{i,l}z_{i,k}) \\ &= \sum_{l=1}^{t}\frac{1}{t^2} = \frac{1}{t}.\end{aligned}$$

For 3, we only need to consider the case when $\{i, j\}\cap\{i', j'\} \neq \emptyset$. Without loss of generality, assume $i = i'$, thus,

$$\begin{aligned}\mathbb{E}\langle g_i, g_j\rangle\langle g_i, g_{j'}\rangle = &\sum_l E(g_{i,l}g_{j,l}g_{i,l}, g_{j',l}) \\ &+ \sum_{l\neq k} E(g_{i,l}g_{j,l}g_{i,k}, g_{j',k}) = 0,\end{aligned}$$

where we use that $g_{i,l}g_{i,k} = 0$ when $l \neq k$. □

The following lemma is a simple case that the variance of a sparse sketch is smaller than the Gaussian sketch. We will show in the next section that the sparse sketch is superior to the Gaussian sketch for every symmetric matrix.

Lemma C.2. *Let $G_1 \sim \mathcal{D}_{t',n}$ and let $G_2, \ldots G_p$ be independent copies of $\mathcal{D}_{t,n}$, where $p \geq 2$ is an integer and c_1, c_2 are two absolute constants. Let A be a symmetric matrix with all entries non-negative and $1 \leq t' \leq t$. Let $X = \mathrm{Tr}\left(G_1 A G_2^T G_2 A G_3 \cdots G_p A G_1^T\right)$. Let X' be a random variable obtained by replacing G_i of X by a column normalized gassian matrix of the same size. Then,*

$$\mathbb{E}(X^2) \leq \mathbb{E}(X'^2).$$

Proof. Let $J = \{j_1, \dots j_p\} \in [n]^p$ and $I = \{i_1, \dots i_p\} \in [n]^p$. Define

$$X_{I,J} := a_{i_p,j_1} a_{i_1,j_2} \cdots a_{i_{p-1},j_p} \langle g_{j_1}^{(1)}, g_{i_1}^{(1)} \rangle \langle g_{j_2}^{(2)}, g_{i_2}^{(2)} \rangle \cdots \langle g_{j_p}^{(p)}, g_{i_p}^{(p)} \rangle.$$

We now expand X in a different form,

$$X = \sum_{I,J} X_{I,J}.$$

Thus,
$$X^2 = \sum_{I,J,I',J'} X_{I,J} X_{I',J'}.$$

Define $X'_{I,J}$ analogously by replacing g_i with Gaussian vectors. Since each $a_{i,j} \geq 0$, with Proposition 3.4 and Lemma C.1, we immediately have that $\mathbb{E}(X^2) \leq \mathbb{E}(X'^2)$. □

The preceding lemma leads to the following theorem.

Theorem C.3. *For every integer $p \geq 2$, there exists a randomized one-pass streaming algorithm $\mathcal{A}$ using space $O(n^{2-4/p}/\epsilon^2)$, and a $\lceil p/2 \rceil$-pass streaming algorithm $\mathcal{B}$ using space $O(n^{1-1/(p-1)}/\epsilon^2)$, given as input PSD matrix $A \in \mathbb{R}^{n \times n}$ with all entries non-negative, then the output of the algorithms $\mathcal{A}(A)$ and $\mathcal{B}(A)$ satisfy*

$$\Pr[\mathcal{A}(A) \in (1 \pm \epsilon)\|A\|_{S_p}^p] \geq 0.99;$$

and

$$\Pr[\mathcal{B}(A) \in (1 \pm \epsilon)\|A\|_{S_p}^p] \geq 0.99,$$

where the probability is over the randomness of the algorithms. Both algorithms require $O(1/\epsilon^2)$ time to process each update in a pass. After the updates, $\mathcal{A}$ requires time $O(n^{(1-2/p)\omega}/\epsilon^2)$ to compute its output and $\mathcal{B}$ requires time $O(n^{(1-2/p)}/\epsilon^2)$, where $\omega < 3$ is the matrix multiplication constant. For general input matrices A of size $n \times m$ for $m \leq n$, if A has all entries non-negative, the above claim holds for even integers $p \geq 2$.