Universal Matrix Sparsifiers and Fast Deterministic Algorithms for Linear Algebra

Rajarshi Bhattacharjee
University of Massachusetts Amherst, MA, USA

Gregory Dexter
Purdue University, West Lafayette, IN, USA

Cameron Musco
University of Massachusetts Amherst, MA, USA

Archan Ray
University of Massachusetts Amherst, MA, USA

Sushant Sachdeva
University of Toronto, Canada

David P. Woodruff
Carnegie Mellon University, Pittsburgh, PA, USA

Abstract

Let $\mathbf{S} \in \mathbb{R}^{n\times n}$ be any matrix satisfying $\|\mathbf{1} - \mathbf{S}\|_2 \leq \epsilon n$, where $\mathbf{1}$ is the all ones matrix and $\|\cdot\|_2$ is the spectral norm. It is well-known that there exists $\mathbf{S}$ with just $O(n/\epsilon^2)$ non-zero entries achieving this guarantee: we can let $\mathbf{S}$ be the scaled adjacency matrix of a Ramanujan expander graph. We show that, beyond giving a sparse approximation to the all ones matrix, $\mathbf{S}$ yields a *universal sparsifier* for any positive semidefinite (PSD) matrix. In particular, for any PSD $\mathbf{A} \in \mathbb{R}^{n\times n}$ which is normalized so that its entries are bounded in magnitude by 1, we show that $\|\mathbf{A} - \mathbf{A} \circ \mathbf{S}\|_2 \leq \epsilon n$, where $\circ$ denotes the entrywise (Hadamard) product. Our techniques also yield universal sparsifiers for non-PSD matrices. In this case, we show that if $\mathbf{S}$ satisfies $\|\mathbf{1} - \mathbf{S}\|_2 \leq \frac{\epsilon^2 n}{c \log^2(1/\epsilon)}$ for some sufficiently large constant c, then $\|\mathbf{A} - \mathbf{A} \circ \mathbf{S}\|_2 \leq \epsilon \cdot \max(n, \|\mathbf{A}\|_1)$, where $\|\mathbf{A}\|_1$ is the nuclear norm. Again letting $\mathbf{S}$ be a scaled Ramanujan graph adjacency matrix, this yields a sparsifier with $\widetilde{O}(n/\epsilon^4)$ entries. We prove that the above universal sparsification bounds for both PSD and non-PSD matrices are tight up to logarithmic factors.

Since $\mathbf{A} \circ \mathbf{S}$ can be constructed *deterministically* without reading all of $\mathbf{A}$, our result for PSD matrices derandomizes and improves upon established results for randomized matrix sparsification, which require sampling a random subset of $O(\frac{n \log n}{\epsilon^2})$ entries and only give an approximation to any fixed $\mathbf{A}$ with high probability. We further show that any randomized algorithm must read at least $\Omega(n/\epsilon^2)$ entries to spectrally approximate general $\mathbf{A}$ to error ϵn, thus proving that these existing randomized algorithms are optimal up to logarithmic factors. We leverage our deterministic sparsification results to give the first deterministic algorithms for several problems, including singular value and singular vector approximation and positive semidefiniteness testing, that run in faster than matrix multiplication time. This partially addresses a significant gap between randomized and deterministic algorithms for fast linear algebraic computation.

Finally, if $\mathbf{A} \in \{-1, 0, 1\}^{n\times n}$ is PSD, we show that a spectral approximation $\widetilde{\mathbf{A}}$ with $\|\mathbf{A} - \widetilde{\mathbf{A}}\|_2 \leq \epsilon n$ can be obtained by deterministically reading $\widetilde{O}(n/\epsilon)$ entries of $\mathbf{A}$. This improves the $1/\epsilon$ dependence on our result for general PSD matrices by a quadratic factor and is information-theoretically optimal up to a logarithmic factor.

2012 ACM Subject Classification Theory of computation → Sketching and sampling; Mathematics of computing → Computations on matrices; Theory of computation → Expander graphs and randomness extractors

Keywords and phrases sublinear algorithms, randomized linear algebra, spectral sparsification, expanders

Digital Object Identifier 10.4230/LIPIcs.ITCS.2024.13

15th Innovations in Theoretical Computer Science Conference (ITCS 2024).
Editor: Venkatesan Guruswami; Article No. 13; pp. 13:1–13:24
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Related Version *Full Version*: https://arxiv.org/abs/2305.05826 [14]

Funding RB, CM and AR was partially supported by an Adobe Research grant, along with NSF Grants 2046235, 1763618, and 1934846. AR was also partially supported by the Manning College of Information and Computer Sciences Dissertation Writing Fellowship. GD was partially supported by NSF AF 1814041, NSF FRG 1760353, and DOE-SC0022085. SS was supported by an NSERC Discovery Grant RGPIN-2018-06398 and a Sloan Research Fellowship. DW was supported by a Simons Investigator Award.

Acknowledgements We thank Christopher Musco for helpful conversations about this work.

1 Introduction

A common task in processing large matrices is *element-wise sparsification.* Given an $n \times n$ matrix $\mathbf{A}$, the goal is to choose a small subset S of coordinates in $[n] \times [n]$, where $[n] = \{1, 2, \ldots, n\}$, such that $\|\mathbf{A} - \mathbf{A} \circ \mathbf{S}\|_2$ is small, where $\circ$ denotes the Hadamard (entrywise) product and $\mathbf{S}$ is a sampling matrix which equals $\frac{n^2}{|S|}$ on the entries in S, but is 0 otherwise. As in previous work, we consider operator norm error, where for a matrix $\mathbf{B}$, $\|\mathbf{B}\|_2 \overset{\text{def}}{=} \sup_{\mathbf{x} \in \mathbb{R}^n} \frac{\|\mathbf{Bx}\|_2}{\|\mathbf{x}\|_2}$. Elementwise sparsification has been widely studied [2, 27, 1, 16] and has been used as a primitive in several applications, including low-rank approximation [2, 29], approximate eigenvector computation [8, 2], semi-definite programming [7, 26], and matrix completion [18, 19]. Without loss of generality, one can scale the entries of $\mathbf{A}$ so that the maximum entry is bounded by 1 in absolute value, and we refer to such matrices as having *bounded entries.* With this normalization, it will be convenient to consider the task of finding a small subset S with corresponding sampling matrix $\mathbf{S}$ such that, for a given error parameter $\epsilon \in (0, 1)$,

$$\|\mathbf{A} - \mathbf{A} \circ \mathbf{S}\|_2 \leq \epsilon \cdot n. \tag{1}$$

One can achieve the error guarantee in (1) for any bounded entry matrix $\mathbf{A}$ with high probability by uniformly sampling a set of $O\left(\frac{n \log n}{\epsilon^2}\right)$ entries of $\mathbf{A}$. See, e.g., Theorem 1 of [27].[1] However, there are no known lower bounds for this problem, even if we consider the harder task of *universal sparsification*, which requires finding a fixed subset S such that (1) *holds simultaneously for every bounded entry matrix* $\mathbf{A}$. The existence of such a fixed subset S corresponds to the existence of a *deterministic sublinear query algorithm* that constructs a spectral approximation to any $\mathbf{A}$ by forming $\mathbf{A} \circ \mathbf{S}$ (which requires reading just $|S|$ entries of $\mathbf{A}$). As we will see, such algorithms have applications to fast deterministic algorithms for linear algebraic computation. We ask:

What is the size of the smallest set S that achieves (1) *simultaneously for every bounded entry matrix* $\mathbf{A}$*?*

Previously, no bound on the size of S better than the trivial $O(n^2)$ was known.

1.1 Our Results

Our first result answers the above question for the class of *symmetric positive semidefinite* (PSD) matrices, i.e., those matrices $\mathbf{A}$ for which $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$ for all vectors $\mathbf{x} \in \mathbb{R}^n$, or equivalently, whose eigenvalues are all non-negative. PSD matrices arise e.g., as covariance

[1] To apply Thm. 1 of [27] we first rescale $\mathbf{A}$ so that all entries are $\leq 1/n$ in magnitude, and note that $\|\mathbf{A}\|_F^2 \leq 1$ in this case. Thus, they sample $s = O\left(\frac{n \log n}{\epsilon^2}\right)$ entries. Rescaling their error guarantee analogously gives (1).

Table 1 Summary of results. Algorithms in the first two rows output a spectral approximation that is sparse, as in (1). Our deterministic algorithms in this case follow directly from our universal sparsification results. Algorithms in the last two rows can output an approximation of any form.

Matrix Type	Approx. Type	Randomized Error	Randomized Sample Complexity	Deterministic Error	Deterministic Sample Complexity
$\|\mathbf{A}\|_\infty \leq 1$ $\mathbf{A}$ is PSD	sparse	ϵn	$\Theta(n/\epsilon^2)$ (Thms 1, 3)	ϵn	$\Theta(n/\epsilon^2)$ (Thms 1, 3)
$\|\mathbf{A}\|_\infty \leq 1$	sparse	ϵn	$O\left(\frac{n \log n}{\epsilon^2}\right)$ [2] $\Omega(n/\epsilon^2)$ (Thm 3)	$\epsilon \max(n, \|\mathbf{A}\|_1)$	$O\left(\frac{n \log^4(1/\epsilon)}{\epsilon^4}\right)$ (Thm 4) $\Omega(n/\epsilon^4)$ (Thm 7)
$\|\mathbf{A}\|_\infty \leq 1$ $\mathbf{A}$ is PSD	any	ϵn	$O\left(\frac{n \log(1/\epsilon)}{\epsilon}\right)$ [36]	ϵn	$O\left(\frac{n \log n}{\epsilon}\right)$ (Thm 10) for $\mathbf{A} \in \{-1, 0, 1\}^{n \times n}$ $\Omega(n/\epsilon)$ (Thm 11)
$\|\mathbf{A}\|_\infty \leq 1$	any	ϵn	$\Omega(n/\epsilon^2)$ (Thm 12)	$\epsilon \max(n, \|\mathbf{A}\|_1)$	$\Omega(n/\epsilon^2)$ (Thm 8)

matrices, kernel similarity matrices, and graph Laplacians, and significant work has considered efficient algorithms for approximating their properties [47, 23, 52, 21, 6, 33, 41]. We summarize our results below and give a comparison to previous work in Table 1.

We show that there exists a set S with $|S| = O(n/\epsilon^2)$ that achieves (1) simultaneously for all bounded entry PSD matrices. This improves the best known randomized bound of $O\left(\frac{n \log n}{\epsilon^2}\right)$ for algorithms which only succeed on a *fixed matrix* with high probability.

▶ **Theorem 1** (Universal Sparsifiers for PSD Matrices). *There exists a subset S of $s = O(n/\epsilon^2)$ entries of $[n] \times [n]$ such that, letting $\mathbf{S} \in \mathbb{R}^{n \times n}$ have $\mathbf{S}_{ij} = \frac{n^2}{s}$ for $(i,j) \in S$ and $\mathbf{S}_{ij} = 0$ otherwise, simultaneously for all PSD matrices $\mathbf{A} \in \mathbb{R}^{n \times n}$ with bounded entries, $\|\mathbf{A} - \mathbf{A} \circ \mathbf{S}\|_2 \leq \epsilon n$.*

Theorem 1 can be viewed as a significant strengthening of a classic spectral graph expander guarantee [32, 4]; indeed, letting $\mathbf{A} = \mathbf{1}$ be the all ones matrix, we have that if $\mathbf{A} \circ \mathbf{S}$ satisfies (1), then it matches the near-optimal spectral expansion of Ramanujan graphs, up to a constant factor.[2] In fact, we prove Theorem 1 by proving a more general claim: that any matrix $\mathbf{S}$ that sparsifies the all ones matrix also sparsifies every bounded entry PSD matrix. In particular, Theorem 1 follows as a direct corollary of:

▶ **Theorem 2** (Spectral Expanders are Universal Sparsifiers for PSD Matrices). *Let $\mathbf{1} \in \mathbb{R}^{n \times n}$ be the all ones matrix. Let $\mathbf{S} \in \mathbb{R}^{n \times n}$ be any matrix such that $\|\mathbf{1} - \mathbf{S}\|_2 \leq \epsilon n$. Then for any PSD matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ with bounded entries, $\|\mathbf{A} - \mathbf{A} \circ \mathbf{S}\|_2 \leq \epsilon n$.*

To prove Theorem 1 given Theorem 2, we let S be the edge set of a Ramanujan spectral expander graph with degree $d = O(1/\epsilon^2)$ and adjacency matrix $\mathbf{G}$. We have $s = nd = O(n/\epsilon^2)$ and $\mathbf{S} = \frac{n^2}{s}\mathbf{G} = \frac{n}{d}\mathbf{G}$. Thus, the top eigenvector of $\mathbf{S}$ is the all ones vector with eigenvalue $\lambda_1(\mathbf{S}) = \frac{n}{d}\lambda_1(\mathbf{G}) = n$. All other eigenvalues are bounded by $|\lambda_i(\mathbf{S})| = \frac{n}{d}|\lambda_i(\mathbf{G})| = O(\frac{n}{d} \cdot \sqrt{d}) = O(\epsilon n)$. Combined, after adjusting ϵ by a constant factor, this shows that $\|\mathbf{1} - \mathbf{S}\|_2 \leq \epsilon n$, as required.

[2] If we let $d = s/n = O(1/\epsilon^2)$ be the average number of entries per row of $\mathbf{S}$, and let $\mathbf{G} \in \{0,1\}^{n \times n}$ be the binary adjacency matrix of the graph with edges in S (i.e., $\mathbf{G} = \frac{s}{n^2} \cdot \mathbf{S}$), then by (1) applied with $\mathbf{A} = \mathbf{1}$, $\lambda_1(\mathbf{G}) = \frac{s}{n^2}\lambda_1(\mathbf{S}) = \frac{s}{n^2}\lambda_1(\mathbf{1} \circ \mathbf{S}) = \frac{s}{n^2}(\lambda_1(\mathbf{1}) \pm \epsilon n) = \Theta(s/n) = \Theta(d)$, while for $i > 1$, $|\lambda_i(\mathbf{G})| \leq \frac{s}{n^2}(|\lambda_i(\mathbf{1})| + \epsilon n) \leq \epsilon d = \Theta(\sqrt{d})$.

Sparsity Lower Bound. We show that Theorem 1 is tight. Even if we only seek to approximate the all ones matrix (rather than all bounded PSD matrices), $\mathbf{S}$ requires $\Omega(n/\epsilon^2)$ non-zero entries.

▶ **Theorem 3** (Sparsity Lower Bound – PSD Matrices). *Let $\mathbf{1} \in \mathbb{R}^{n\times n}$ be the all-ones matrix. Then, for any $\epsilon \in (0,1/2)$ with $\epsilon \geq c/\sqrt{n}$ for large enough constant c, any $\mathbf{S} \in \mathbb{R}^{n\times n}$ with $\|\mathbf{1}-\mathbf{S}\|_2 \leq \epsilon n$ must have $\Omega(n/\epsilon^2)$ nonzero entries.*

Theorem 3 resolves an open question of [16], which asked whether a spectral approximation must have $\Omega\left(\frac{\text{ns}(\mathbf{A})\,\text{sr}(\mathbf{A})}{\epsilon^2}\right)$ non-zeros, where $\text{sr}(\mathbf{A}) = \frac{\|\mathbf{A}\|_F^2}{\|\mathbf{A}\|_2^2}$ is the stable rank and $\text{ns}(\mathbf{A}) = \max_i \frac{\|\mathbf{a}_i\|_1^2}{\|\mathbf{a}_i\|_2^2}$, where $\mathbf{a}_i$ is the i^{th} row of $\mathbf{A}$, is the "numerical sparsity". They give a lower bound when $\text{sr}(\mathbf{A}) = \Theta(n)$ but ask if this can be extended to $\text{sr}(\mathbf{A}) = o(n)$. For the all ones matrix, $\text{sr}(\mathbf{A}) = 1$ and $\text{ns}(\mathbf{A}) = n$, and so Theorem 3 resolves this question. Further, by applying the theorem to a block diagonal matrix with r disjoint $k \times k$ blocks of all ones, we resolve the question for integer $\text{sr}(\mathbf{A})$ and $\text{ns}(\mathbf{A})$.

Non-PSD Matrices. The techniques used to prove Theorem 1 also give nearly tight universal sparsification bounds for general bounded entry (not necessarily PSD) matrices. In this case, we show that a subset S of $O\left(\frac{n\log^4(1/\epsilon)}{\epsilon^4}\right)$ entries suffices to achieve spectral norm error depending on the nuclear norm $\|\mathbf{A}\|_1$, which is the sum of singular values of $\mathbf{A}$.

▶ **Theorem 4** (Universal Sparsifiers for Non-PSD Matrices). *There exists a subset S of $s = O\left(\frac{n\log^4(1/\epsilon)}{\epsilon^4}\right)$ entries of $[n]\times[n]$ such that, letting $\mathbf{S} \in \mathbb{R}^{n\times n}$ have $\mathbf{S}_{ij} = \frac{n^2}{s}$ for $(i,j) \in S$ and $\mathbf{S}_{ij} = 0$ otherwise, simultaneously for all symmetric matrices $\mathbf{A} \in \mathbb{R}^{n\times n}$ with bounded entries,*

$$\|\mathbf{A} - \mathbf{A}\circ\mathbf{S}\|_2 \leq \epsilon \cdot \max(n, \|\mathbf{A}\|_1). \tag{2}$$

Note that for a bounded entry PSD matrix $\mathbf{A}$, we have $\|\mathbf{A}\|_1 = \text{tr}(\mathbf{A}) = O(n)$ and so for PSD matrices, (1) and (2) are equivalent.

▶ Remark 5. Although Theorem 4 is stated for symmetric $\mathbf{A}$, one can first symmetrize $\mathbf{A}$ by considering $\mathbf{B} = [\mathbf{0}, \mathbf{A}; \mathbf{A}^T, \mathbf{0}]$. Then for any $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{2n}$, we have $\mathbf{z}^T\mathbf{B}\mathbf{z} = 2\mathbf{x}^T\mathbf{A}\mathbf{y}$, and $\|\mathbf{B}\|_1 = 2\|\mathbf{A}\|_1$, so applying the above theorem to $\mathbf{B}$ gives us a spectral approximation to $\mathbf{A}$.

As with Theorem 1, Theorem 4 follows from a general claim which shows that sparsifying the all ones matrix to small enough error suffices to sparsify any bounded entry symmetric matrix.

▶ **Theorem 6** (Spectral Expanders are Universal Sparsifiers for Non-PSD Matrices). *Let $\mathbf{1} \in \mathbb{R}^{n\times n}$ be the all ones matrix. Let $\mathbf{S} \in \mathbb{R}^{n\times n}$ be any matrix such that $\|\mathbf{1}-\mathbf{S}\|_2 \leq \frac{\epsilon^2 n}{c\log^2(1/\epsilon)}$ for some large enough constant c. Then for any symmetric $\mathbf{A} \in \mathbb{R}^{n\times n}$ with bounded entries, $\|\mathbf{A} - \mathbf{A}\circ\mathbf{S}\|_2 \leq \epsilon \cdot \max(n, \|\mathbf{A}\|_1)$.*

Theorem 4 follows by applying Theorem 6 where $\mathbf{S}$ is taken to be the scaled adjacency matrix of a Ramanujan expander graph with degree $d = O(\log^4(1/\epsilon)/\epsilon^4)$.

Lower Bounds for Non-PSD Matrices. We can prove that Theorem 4 is tight up to logarithmic factors. Our lower bound holds even for the easier problem of top singular value (spectral norm) approximation and against a more general class of algorithms, which non-adaptively and deterministically query entries of the input matrix. The idea is simple:

since the entries read by the deterministic algorithm are fixed, we can construct two very different input instances on which the algorithm behaves identically: one which is the all ones matrix, and the other which is one only on the entries read by the algorithm and zero everywhere else. We show that if the number of entries s that are read is too small, the top singular value of the second instance is significantly smaller than the first (as compared to their Schatten-1 norms), which violates the desired error bound of $\epsilon \cdot \max(n, \|\mathbf{A}\|_1)$. To obtain a tight bound, we apply the above construction on just a subset of the rows of the matrix on which the algorithm does not read too many entries. Here, we critically use non-adaptivity, as this subset of rows can be fixed, independent of the input instance.

▶ **Theorem 7** (Non-Adaptive Query Lower Bound for Deterministic Spectral Approximation of Non-PSD Matrices)**.** *For any $\epsilon \in (1/n^{1/4}, 1/4)$, any deterministic algorithm that queries entries of a bounded entry matrix $\mathbf{A}$ non-adaptively and outputs $\widetilde{\sigma}_1$ satisfying $|\sigma_1(\mathbf{A}) - \widetilde{\sigma}_1| \leq \epsilon \cdot \max(n, \|\mathbf{A}\|_1)$ must read at least $\Omega\left(\frac{n}{\epsilon^4}\right)$ entries. Here $\sigma_1(\mathbf{A}) = \|\mathbf{A}\|_2$ is the largest singular value of $\mathbf{A}$.*

Observe that Theorem 4 implies the existence of a non-adaptive deterministic algorithm for the above problem with query complexity $O\left(\frac{n \log^4(1/\epsilon)}{\epsilon^4}\right)$, since $\mathbf{A} \circ \mathbf{S}$ can be computed with non-adaptive determinstic queries and since $\|\mathbf{A} - \mathbf{A} \circ \mathbf{S}\|_2 \leq \epsilon \cdot \max(n, \|\mathbf{A}\|_1)$ implies via Weyl's inequality that $|\sigma_1(\mathbf{A}) - \sigma_1(\mathbf{A} \circ \mathbf{S})| \leq \epsilon \cdot \max(n, \|\mathbf{A}\|_1)$. Also recall that while Theorem 4 is stated for bounded entry symmetric matrices, it applies to general (possibly asymmetric) matrices by Remark 5.

Note that Theorem 7 establishes a separation between universal sparsifiers and randomized sparsification since randomly sampling $O\left(\frac{n \log n}{\epsilon^2}\right)$ entries of any $\mathbf{A} \in \mathbb{R}^{n \times n}$ achieves error ϵn by [27]. In fact, for the problem of just approximating $\sigma_1(\mathbf{A})$ to error $\pm \epsilon n$, randomized algorithms using just $\text{poly}(\log n, 1/\epsilon)$ samples are known [10, 13]. Theorem 7 shows that universal sparsifiers for general bounded entry matrices (and in fact all non-adaptive deterministic algorithms for spectral approximation) require a worse $1/\epsilon^4$ dependence and error bound of $\epsilon \cdot \max(n, \|\mathbf{A}\|_1)$. We can extend Theorem 7 to apply to general deterministic algorithms that query $\mathbf{A}$ possibly *adaptively*, however the lower bound weakens to $\Omega(n/\epsilon^2)$. Understanding if this gap between adaptive and non-adaptive query deterministic algorithms is real is an interesting open question.

▶ **Theorem 8** (Adaptive Query Lower Bound for Deterministic Spectral Approximation of Non-PSD Matrices)**.** *For any $\epsilon \in (1/\sqrt{n}, 1/4)$, any deterministic algorithm that queries entries of a bounded entry matrix $\mathbf{A}$ (possibly adaptively) and outputs $\widetilde{\sigma}_1$ satisfying $|\sigma_1(\mathbf{A}) - \widetilde{\sigma}_1| \leq \epsilon \cdot \max(n, \|\mathbf{A}\|_1)$ must read at least $\Omega\left(\frac{n}{\epsilon^2}\right)$ entries. Here $\sigma_1(\mathbf{A}) = \|\mathbf{A}\|_2$ is the largest singular value of $\mathbf{A}$.*

1.1.1 Applications to Fast Deterministic Algorithms for Linear Algebra

Given sampling matrix $\mathbf{S}$ such that $\|\mathbf{A} - \mathbf{A} \circ \mathbf{S}\|_2 \leq \epsilon \cdot \max(n, \|\mathbf{A}\|_1)$, one can use $\mathbf{A} \circ \mathbf{S}$ to approximate various linear algebraic properties of $\mathbf{A}$. For example, by Weyl's inequality [49], the eigenvalues of $\mathbf{A} \circ \mathbf{S}$ approximate those of $\mathbf{A}$ up to additive error $\pm \epsilon \cdot \max(n, \|\mathbf{A}\|_1)$. Thus, our universal sparsification results (Theorems 1, 4) immediately give the first known deterministic algorithms for approximating the eigenspectrum of a bounded entry matrix up to small additive error with *sublinear entrywise query complexity.* Previously, only randomized sublinear query algorithms were known [50, 36, 37, 45, 10, 13].

Further, our results yield the first deterministic algorithms for several problems that run in $o(n^\omega)$ time, where $\omega \approx 2.373$ is the matrix multiplication exponent [3]. Consider e.g., approximating the top singular value $\sigma_1(\mathbf{A})$ (the spectral norm) of $\mathbf{A}$. A $(1 + \epsilon)$-relative error approximation can be computed in $\widetilde{O}(n^2/\sqrt{\epsilon})$ time with high probability using

$O(\log(n/\epsilon)/\sqrt{\epsilon})$ iterations of the Lanczos method with a random initialization [28, 35]. This can be further accelerated e.g., via randomized entrywise sparsification [27], allowing an additive $\pm\epsilon n$ approximation to $\sigma_1(\mathbf{A})$ to be computed in $\widetilde{O}(n/\operatorname{poly}(\epsilon))$ time. However, prior to our work, no fast deterministic algorithms were known, even with coarse approximation guarantees. The fastest approach was just to perform a full SVD of $\mathbf{A}$, requiring $\Theta(n^\omega)$ time. In fact, this gap between randomized and deterministic methods exists for many linear algebraic problems, and resolving it is a central open question.

By combining our universal sparsification results with a derandomized power method that uses a full subspace of vectors for initalization, and iteratively approximates the singular values of $\mathbf{A} \circ \mathbf{S}$ via "deflation" [40], we give to the best of our knowledge, the first $o(n^\omega)$ time deterministic algorithm for approximating *all singular values* of a bounded entry matrix $\mathbf{A}$ to small additive error.

▶ **Theorem 9** (Deterministic Singular Value Approximation). *Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a bounded entry symmetric matrix with singular values $\sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq \ldots \geq \sigma_n(\mathbf{A})$. Then there exists a deterministic algorithm that, given $\epsilon \in (0,1)$, reads $\widetilde{O}\left(\frac{n}{\epsilon^4}\right)$ entries of $\mathbf{A}$, runs in $\widetilde{O}\left(\frac{n^2}{\epsilon^8}\right)$ time, and returns singular value approximations $\widetilde{\sigma}_1(\mathbf{A}), \ldots, \widetilde{\sigma}_n(\mathbf{A})$ satisfying for all i,*

$$|\sigma_i(\mathbf{A}) - \widetilde{\sigma}_i(\mathbf{A})| \leq \epsilon \cdot \max(n, \|\mathbf{A}\|_1).$$

Further, for all $i \leq 1/\epsilon$, the algorithm returns a unit vector $\mathbf{z}_i$ such that $|\|\mathbf{A}\mathbf{z}_i\|_2 - \sigma_i(\mathbf{A})| \leq \epsilon \cdot \max(n, \|\mathbf{A}\|_1)$ and the returned vectors are all mutually orthogonal.

The runtime of Theorem 9 is stated assuming we have already constructed a deterministic sampling matrix $\mathbf{S}$ that samples $\tilde{O}(\frac{n}{\epsilon^4})$ entries of $\mathbf{A}$. It is well known that, if we set $\mathbf{S}$ to be the adjacency matrix of a Ramanujan expander graph, it can indeed be constructed deterministically in time $\widetilde{O}(n/\operatorname{poly}(\epsilon))$ [5]. This is lower order as compared to the runtime of $\widetilde{O}\left(\frac{n^2}{\epsilon^8}\right)$ unless $\epsilon < 1/n^c$ for some large enough constant c. Further, for a fixed input size n (and in fact, for a range of input sizes), $\mathbf{S}$ needs to be constructed only once. See Section 6 of [14] for further discussion.

Recall that if we further assume $\mathbf{A}$ to be PSD, then the error in Theorem 9 is bounded by $\epsilon \cdot \max(n, \|\mathbf{A}\|_1) \leq \epsilon n$. The sample complexity and runtime also improve by $\operatorname{poly}(1/\epsilon)$ factors due to the tighter universal sparisfier bound for PSD matrices given in Theorem 1. Also observe that while Theorem 9 gives additive error approximations to all of $\mathbf{A}$'s singular values, these approximations are only meaningful for singular values larger than $\epsilon \cdot \max(n, \|\mathbf{A}\|_1)$, of which there are at most $1/\epsilon$. Similar additive error guarantees have been given using randomized algorithms [13, 51]. Related bounds have also been studied in work on randomized methods for spectral density estimation [48, 30, 17].

We further leverage Theorem 9 to give the first $o(n^\omega)$ time deterministic algorithm for testing if a bounded entry matrix is either PSD or has at least one large negative eigenvalue $\leq -\epsilon \cdot \max(n, \|\mathbf{A}\|_1)$. Recent work has focused on optimal randomized methods for this problem [10, 38]. We also show that, under the assumption that $\sigma_1(\mathbf{A}) \geq \alpha \cdot \max(n, \|\mathbf{A}\|_1)$ for some $\alpha \in (0,1)$, one can deterministically compute $\widetilde{\sigma}_1(\mathbf{A})$ with $|\sigma_1(\mathbf{A}) - \widetilde{\sigma}_1(\mathbf{A})| \leq \epsilon \cdot \max(n, \|\mathbf{A}\|_1)$ in $\widetilde{O}\left(\frac{n^2 \log(1/\epsilon)}{\operatorname{poly}(\alpha)}\right)$ time. That is, one can compute a highly accurate approximation to the top singular value in roughly linear time in the input matrix size. Again, this is the first $o(n^\omega)$ time deterministic algorithm for this problem, and matches the runtime of the best known randomized methods for high accuracy top singular value computation, up to a $\operatorname{poly}(\log(n, 1/\alpha))$ factor.

Note that the details of our results on fast deterministic algorithms for linear algebra have been omitted due to a lack of space. They can be found in Section 6 of the full version [14].

1.1.2 Beyond Sparse Approximations

It is natural ask if it is possible to achieve better than $O(n/\epsilon^2)$ sample complexity for spectral approximation by using an algorithm that does not output a sparse approximation to $\mathbf{A}$, but can output more general data structures, allowing it to avoid the sparsity lower bound of Theorem 3. Theorems 7 and 8 already rule this out for both non-adaptive and adaptive deterministic algorithms for non-PSD matrices. However, it is known that this *is possible* with randomized algorithms for PSD matrices. For example, following [9], one can apply Theorem 3 of [36] with error parameter $\lambda = \epsilon n$. Observing that all ridge leverage score sampling probabilities are bounded by $1/\lambda$ (e.g., via Lemma 6 of the same paper and the bounded entry assumption), one can show that a Nyström approximation $\widetilde{\mathbf{A}}$ based on $\widetilde{O}(1/\epsilon)$ uniformly sampled columns satisfies $\|\mathbf{A} - \widetilde{\mathbf{A}}\|_2 \leq \epsilon n$ with high probability. Further $\widetilde{\mathbf{A}}$ can be constructed by reading just $\widetilde{O}(1/\epsilon)$ columns and thus $\widetilde{O}(n/\epsilon)$ entries of $\mathbf{A}$, giving a linear rather than quadratic dependence on $1/\epsilon$ as compared to Theorem 1. Unfortunately, derandomizing such a column-sampling-based approach seems difficult – any deterministic algorithm must read entries in $\Omega(n)$ columns of $\mathbf{A}$, as otherwise it will fail when $\mathbf{A}$ is entirely supported on the unread columns.

Nevertheless, in the special case where $\mathbf{A}$ is PSD and has entries in $\{-1, 0, 1\}^{n\times n}$, we show that a spectral approximation can be obtained by deterministically reading just $\widetilde{O}(n/\epsilon)$ entries of $\mathbf{A}$.

▶ **Theorem 10** (Deterministic Spectral Approximation of Binary Magnitude PSD Matrices). *Let $\mathbf{A} \in \{-1, 0, 1\}^{n\times n}$ be PSD. Then for any $\epsilon \in (0, 1)$, there exists a deterministic algorithm which reads $O\left(\frac{n \log n}{\epsilon}\right)$ entries of $\mathbf{A}$ and returns PSD $\widetilde{\mathbf{A}} \in \{-1, 0, 1\}^{n\times n}$ such that $\|\mathbf{A} - \widetilde{\mathbf{A}}\|_2 \leq \epsilon n$.*

Using Turán's theorem, we show that Theorem 10 is information theoretically optimal up to a $\log n$ factor, even for the potentially much easier problem of eigenvalue approximation:

▶ **Theorem 11** (Deterministic Spectral Approximation of Binary PSD Matrices – Lower Bound). *Let $\mathbf{A} \in \{0, 1\}^{n\times n}$ be PSD. Then for any $\epsilon \in (0, 1)$, any possibly adaptive deterministic algorithm which approximates all eigenvalues of $\mathbf{A}$ up to ϵn additive error must read $\Omega\left(\frac{n}{\epsilon}\right)$ entries of $\mathbf{A}$.*

An interesting open question is if $\widetilde{O}(n/\epsilon)$ sample complexity can be achieved for deterministic spectral approximation of *any bounded entry PSD matrix*, with a matrix $\widetilde{\mathbf{A}}$ of any form, matching what is known for randomized algorithms.

Finally, we show that the PSD assumption is critical to achieve $o(n/\epsilon^2)$ query complexity, for both randomized and deterministic algorithms. Without this assumption, $\Omega(n/\epsilon^2)$ queries are required, even to achieve (1) with constant probability for a single input $\mathbf{A}$. For bounded entry matrices, the randomized element-wise sparsification algorithms of [2, 27] read just $\widetilde{O}(n/\epsilon^2)$ entries of $\mathbf{A}$, and so our lower bounds are the first to show near-optimality of the number of entries read of these algorithms, which may be of independent interest.

▶ **Theorem 12** (Lower Bound for Randomized Spectral Approximation). *Any randomized algorithm that (possibly adaptively) reads entries of a binary matrix $\mathbf{A} \in \{0, 1\}^{n\times n}$ to construct a data structure $f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ that, for $\epsilon \in (0, 1)$ satisfies with probability at least 99/100,*

$$|f(\mathbf{x}, \mathbf{y}) - \mathbf{x}^T \mathbf{A}\mathbf{y}| \leq \epsilon n, \text{ for all unit } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n,$$

must read $\Omega\left(\frac{n}{\epsilon^2}\right)$ entries of $\mathbf{A}$ in the worst case, provided $\epsilon = \Omega\left(\frac{\log n}{\sqrt{n}}\right)$.

Note that details of our results on upper and lower bounds beyond sparse approximation have been omitted due to lack of space and can be found in Section 5 of the full version [14].

1.1.3 Relation to Spectral Graph Sparsification

We remark that while our work focuses on general bounded entry symmetric (PSD) matrices, when $\mathbf{A}$ is a PSD graph Laplacian matrix, it is possible to construct a sparsifier $\widetilde{\mathbf{A}}$ with just $\widetilde{O}(n/\epsilon^2)$ entries, that achieves a *relative error* spectral approximation guarantee that can be much stronger than the additive error guarantee of (1) [12]. In particular, one can achieve $(1-\epsilon)\mathbf{x}^T\mathbf{A}\mathbf{x} \leq \mathbf{x}^T\widetilde{\mathbf{A}}\mathbf{x} \leq (1+\epsilon)\mathbf{x}^T\mathbf{A}\mathbf{x}$ for all $\mathbf{x} \in \mathbb{R}^n$. To achieve such a guarantee however, it is not hard to see that the set of entries sampled in $\widetilde{\mathbf{A}}$ must depend on $\mathbf{A}$, and cannot be universal. Fast randomized algorithms for constructing $\widetilde{\mathbf{A}}$ have been studied extensively [43, 42, 12]. Recent work has also made progress towards developing fast deterministic algorithms [22].

2 Notation and Preliminaries

We start by defining notation used throughout. For any integer n, let $[n]$ denote the set $\{1, 2, \dots, n\}$.

Matrices and Vectors. Matrices are represented with bold uppercase literals, e.g., $\mathbf{A}$. Vectors are represented with bold lowercase literals, e.g., $\mathbf{x}$. $\mathbf{1}$ and $\mathbf{0}$ denote all ones (resp. all zeros) matrices or vectors. The identity matrix is denoted by $\mathbf{I}$. The size of these matrices vary based on their applications. For a vector $\mathbf{x}$, $\mathbf{x}(j)$ denotes its j^{th} entry. For a matrix $\mathbf{A}$, $\mathbf{A}_{ij}$ denotes the entry in the i^{th} row and j^{th} column. For a vector $\mathbf{x}$ (or matrix $\mathbf{A}$), $\mathbf{x}^T$ (resp. $\mathbf{A}^T$) denotes its transpose. For two matrices $\mathbf{A}, \mathbf{B}$ of the same size, $\mathbf{A} \circ \mathbf{B}$ denotes the entrywise (Hadamard) product.

Matrix Norms and Properties. For a vector $\mathbf{x}$, $\|\mathbf{x}\|_2$ denotes its Euclidean norm and $\|\mathbf{x}\|_1 = \sum_{i=1}^n |\mathbf{x}(i)|$ denotes its ℓ_1 norm. We denote the eigenvalues of a symmetric matrix $\mathbf{A}$ as $\lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq \ldots \geq \lambda_n(\mathbf{A})$ in decreasing order. A symmetric matrix is positive semidefinite (PSD) if $\lambda_i \geq 0$ for all $i \in [n]$. The singular values of a matrix $\mathbf{A}$ are denoted as $\sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq \ldots \geq \sigma_n(\mathbf{A}) \geq 0$ in decreasing order. We let $\|\mathbf{A}\|_2 = \max_x \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \sigma_1(\mathbf{A})$ denote the spectral norm, $\|\mathbf{A}\|_\infty$ denote the largest magnitude of an entry, $\|\mathbf{A}\|_F = (\sum_{i,j} \mathbf{A}_{ij}^2)^{1/2}$ denote the Frobenius norm, and $\|\mathbf{A}\|_1 = \sum_{i=1}^n \sigma_i(\mathbf{A})$ denote the Schatten-1 norm (also called the trace norm or nuclear norm).

Expander Graphs. Our universal sparsifier constructions are based on Ramanujan expander graphs [31, 32, 4], defined below.

▶ **Definition 13** (Ramanujan Expander Graphs [31]). *Let G be a connected d-regular, unweighted and undirected graph on n vertices. Let $\mathbf{G} \in \{0,1\}^{n \times n}$ be the adjacency matrix corresponding to G and λ_i be its i^{th} eigenvalue, such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. Then G is called a Ramanujan graph if:*

$$|\lambda_i| \leq 2\sqrt{d-1}, \quad \textit{for all } i > 1.$$

Equivalently, letting $\mathbf{1}$ be the $n \times n$ all ones matrix, $\|\mathbf{1} - \frac{n}{d}\mathbf{G}\|_2 \leq \frac{n}{d} \cdot 2\sqrt{d-1}$.

Efficient construction of Ramanujan graphs. Significant work has studied efficient constructions for Ramanujan Graphs [32, 31, 34, 24], and nearly linear time constructions, called *strongly explicit* constructions [5], have been proposed. E.g., by Proposition 1.1 of [5] we can reconstruct for any n and d, a graph on $n(1+o(1))$ vertices with second eigenvalue at most $(2+o(1))\sqrt{d}$ in $\tilde{O}(nd+\mathrm{poly}(d))$ time. Additionally, in our applications, the expander just needs to be constructed once and can then be used for any input of size n. In fact, a single expander can be used for a range of input sizes, by the argument below.

Though the size of the expander above is $(1+o(1))n$ instead of exactly n, and its expansion does not exactly hit the tight Ramanujan bound of $2\sqrt{d-1}$, we can let our input matrix $\mathbf{A}$ be the top $n \times n$ principal submatrix of a slightly larger $n(1+o(1)) \times n(1+o(1))$ matrix $\mathbf{A}'$ that is zero everywhere else. Then, the top $n \times n$ principal submatrix of a spectral approximation to $\mathbf{A}'$ is a spectral approximation to $\mathbf{A}$. Thus, obtaining a spectral approximation to $\mathbf{A}'$ via a Ramanujan sparsifier on $n(1+o(1))$ vertices suffices. Similarly, in our applications, we can adjust $d = 1/\mathrm{poly}(\epsilon)$ by at most a constant factor to account for the $(2+o(1))\sqrt{d}$ bound on the second eigenvalue, rather than the tight Ramanujan bound of $2\sqrt{d-1}$.

3 Universal Sparsifier Upper Bounds

We now prove our main results on universal sparsifiers for PSD matrices (Theorem 1) and general symmetric matrices (Theorem 4). Both theorems follow from general reductions which show that any sampling matrix $\mathbf{S}$ that sparsifies the all ones matrix to sufficient accuracy (i.e., is a sufficiently good spectral expander) yields a universal sparsifier. We prove the reduction for the PSD case in Section 3.1, and then extend it to the non-PSD case in Section 3.2.

3.1 Universal Sparsifiers for PSD Matrices

In the PSD case, we prove the following:

▶ **Theorem 2** (Spectral Expanders are Universal Sparsifiers for PSD Matrices). *Let $\mathbf{1} \in \mathbb{R}^{n\times n}$ be the all ones matrix. Let $\mathbf{S} \in \mathbb{R}^{n\times n}$ be any matrix such that $\|\mathbf{1}-\mathbf{S}\|_2 \le \epsilon n$. Then for any PSD matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ with bounded entries, $\|\mathbf{A} - \mathbf{A}\circ\mathbf{S}\|_2 \le \epsilon n$.*

Theorem 1 follows directly from Theorem 2 by letting $\mathbf{S}$ be the scaled adjacency matrix of a Ramanujan expander graph with degree $d = O(1/\epsilon^2)$ (Definition 13).

Proof of Theorem 2. To prove the theorem it suffices to show that for any $\mathbf{x} \in \mathbb{R}^n$ with $\|\mathbf{x}\|_2 = 1$, $|\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{x}^T(\mathbf{A}\circ\mathbf{S})\mathbf{x}| \le \epsilon n$. Let $\mathbf{v}_1,\ldots,\mathbf{v}_n$ and $\lambda_1,\ldots,\lambda_n$ be the eigenvectors and eigenvalues of $\mathbf{A}$ so that $\mathbf{A} = \sum_{i=1}^n \lambda_i \mathbf{v}_i\mathbf{v}_i^T$. Then we can expand out this error as:

$$\begin{aligned}|\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{x}^T(\mathbf{A}\circ\mathbf{S})\mathbf{x}| &= \left|\sum_{i=1}^n \lambda_i \cdot \mathbf{x}^T\mathbf{v}_i\mathbf{v}_i^T\mathbf{x} - \sum_{i=1}^n \lambda_i\mathbf{x}^T(\mathbf{v}_i\mathbf{v}_i^T\circ\mathbf{S})\mathbf{x})\right| \\ &= \left|\sum_{i=1}^n \lambda_i\mathbf{x}^T[\mathbf{v}_i\mathbf{v}_i^T\circ(\mathbf{1}-\mathbf{S})]\mathbf{x}\right|,\end{aligned}$$

where we use that for any matrix $\mathbf{M}$, $\mathbf{M}\circ\mathbf{1} = \mathbf{M}$. Now observe that if we let $\mathbf{D}_i \in \mathbb{R}^{n\times n}$ be a diagonal matrix with the entries of $\mathbf{v}_i$ on its diagonal, then we can write $\mathbf{v}_i\mathbf{v}_i^T\circ(\mathbf{1}-\mathbf{S}) =$

$\mathbf{D}_i(\mathbf{1}-\mathbf{S})\mathbf{D}_i$. Plugging this back in we have:

$$\begin{aligned}|\mathbf{x}^T\mathbf{A}\mathbf{x}-\mathbf{x}^T(\mathbf{A}\circ\mathbf{S})\mathbf{x}| &= \left|\sum_{i=1}^n \lambda_i \mathbf{x}^T\mathbf{D}_i(\mathbf{1}-\mathbf{S})\mathbf{D}_i\mathbf{x}\right| \\ &\leq \sum_{i=1}^n \lambda_i \|\mathbf{1}-\mathbf{S}\|_2\cdot \mathbf{x}^T\mathbf{D}_i^2\mathbf{x} \\ &\leq \epsilon n\cdot \sum_{i=1}^n \lambda_i \mathbf{x}^T\mathbf{D}_i^2\mathbf{x}. \end{aligned} \tag{3}$$

In the second line we use that $\mathbf{A}$ is PSD and thus λ_i is non-negative for all i. We also use that $|\mathbf{x}^T\mathbf{D}_i(\mathbf{1}-\mathbf{S})\mathbf{D}_i\mathbf{x}| \leq \|\mathbf{1}-\mathbf{S}\|_2\cdot\|\mathbf{D}_i\mathbf{x}\|_2^2 = \|\mathbf{1}-\mathbf{S}\|_2\cdot\mathbf{x}^T\mathbf{D}_i^2\mathbf{x}$. In the third line we bound $\|\mathbf{1}-\mathbf{S}\|_2 \leq \epsilon n$ using the assumption of the theorem statement.

Finally, writing $\mathbf{x}^T\mathbf{D}_i^2\mathbf{x} = \sum_{j=1}^n \mathbf{x}(j)^2\mathbf{v}_i(j)^2$ and plugging back into (3), we have:

$$\begin{aligned}|\mathbf{x}^T\mathbf{A}\mathbf{x}-\mathbf{x}^T(\mathbf{A}\circ\mathbf{S})\mathbf{x}| &\leq \epsilon n\cdot\sum_{i=1}^n\sum_{j=1}^n \lambda_i\mathbf{x}(j)^2\mathbf{v}_i(j)^2 \\ &= \epsilon n\cdot\sum_{j=1}^n \mathbf{x}(j)^2\sum_{i=1}^n\lambda_i\mathbf{v}_i(j)^2 \\ &= \epsilon n\cdot\sum_{j=1}^n\mathbf{x}(j)^2\mathbf{A}_{jj} \\ &\leq \epsilon n,\end{aligned}$$

where in the last step we use that $\mathbf{A}_{jj}\leq 1$ by our bounded entry assumption, and that $\mathbf{x}$ is a unit vector. This completes the theorem. ◀

▶ Remark 14. Note that we can state a potentially stronger bound for Theorem 2 by observing that in the second to last step, $|\mathbf{x}^T\mathbf{A}\mathbf{x}-\mathbf{x}^T(\mathbf{A}\circ\mathbf{S})\mathbf{x}| \leq \epsilon n\cdot\sum_{j=1}^n\mathbf{x}(j)^2\mathbf{A}_{jj} = \epsilon n\cdot\mathbf{x}^T\mathbf{D}\mathbf{x}$, where $\mathbf{D}$ is a diagonal matrix containing the diagonal elements of $\mathbf{A}$. That is, letting $\mathbf{M}\preceq\mathbf{N}$ denote that $\mathbf{N}-\mathbf{M}$ is PSD, we have the following spectral approximation bound: $-\epsilon n\cdot\mathbf{D}\preceq\mathbf{A}-(\mathbf{A}\circ\mathbf{S})\preceq\epsilon n\cdot\mathbf{D}$.

3.2 Universal Sparsifiers for Non-PSD Matrices

We next extend the above approach to give a similar reduction for universal sparsification of general symmetric matrices.

▶ **Theorem 6** (Spectral Expanders are Universal Sparsifiers for Non-PSD Matrices). *Let $\mathbf{1}\in\mathbb{R}^{n\times n}$ be the all ones matrix. Let $\mathbf{S}\in\mathbb{R}^{n\times n}$ be any matrix such that $\|\mathbf{1}-\mathbf{S}\|_2\leq\frac{\epsilon^2 n}{c\log^2(1/\epsilon)}$ for some large enough constant c. Then for any symmetric $\mathbf{A}\in\mathbb{R}^{n\times n}$ with bounded entries, $\|\mathbf{A}-\mathbf{A}\circ\mathbf{S}\|_2\leq\epsilon\cdot\max(n,\|\mathbf{A}\|_1)$.*

Observe that as compared to the PSD case (Theorem 2), here we require that $\mathbf{S}$ gives a stronger approximation to the all ones matrix. Theorem 4 follows directly by letting $\mathbf{S}$ be the scaled adjacency matrix of a Ramanujan expander graph with degree $d = O\left(\frac{\log^4(1/\epsilon)}{\epsilon^4}\right)$ (Definition 13).

Proof of Theorem 6. To prove the theorem, it suffices to show that for any $\mathbf{x} \in \mathbb{R}^n$ with $\|\mathbf{x}\|_2 = 1$, $|\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{x}^T(\mathbf{A} \circ \mathbf{S})\mathbf{x}| \leq \epsilon \cdot \max(n, \|\mathbf{A}\|_1)$. We will split the entries of $\mathbf{x}$ into level sets according to their magnitude. In particular, let $\bar{\epsilon} = \frac{\epsilon}{\log_2(1/\epsilon)}$ and let $\ell = \log_2(1/\bar{\epsilon})$. Then we can write $\mathbf{x} = \sum_{i=0}^{\ell+1} \mathbf{x}_i$, such that for any $t \in [n]$:

$$\mathbf{x}_i(t) = \begin{cases} \mathbf{x}(t) & \text{if } i \in \{1, 2, \ldots, \ell\} \text{ and } |\mathbf{x}(t)| \in \left(\frac{2^{i-1}}{\sqrt{n}}, \frac{2^i}{\sqrt{n}}\right] \\ \mathbf{x}(t) & \text{if } i = 0 \text{ and } |\mathbf{x}(t)| \in \left[0, \frac{1}{\sqrt{n}}\right] \\ \mathbf{x}(t) & \text{if } i = \ell + 1 \text{ and } |\mathbf{x}(t)| \in \left(\frac{1}{\bar{\epsilon}\sqrt{n}}, 1\right] \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Via triangle inequality, we have:

$$\left|\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{x}^T(\mathbf{A} \circ \mathbf{S})\mathbf{x}\right| \leq \sum_{i=0}^{\ell+1} \left|\mathbf{x}_i^T\mathbf{A}\mathbf{x} - \mathbf{x}_i^T(\mathbf{A} \circ \mathbf{S})\mathbf{x}\right|. \tag{5}$$

We will bound each term in (5) by $O(\bar{\epsilon} \cdot \max(n, \|\mathbf{A}\|_1)) = O\left(\frac{\epsilon \cdot \max(n, \|\mathbf{A}\|_1)}{\log(1/\epsilon)}\right) = O\left(\frac{\epsilon \cdot \max(n, \|\mathbf{A}\|_1)}{\ell+2}\right)$. Summing over all $\ell + 2$ terms we achieve a final bound of $O(\epsilon \cdot \max(n, \|\mathbf{A}\|_1))$. The theorem then follows by adjusting ϵ by a constant factor.

Fixing $i \in \{0, \ldots, \ell\}$, let $\mathbf{x}_L(t) = \mathbf{x}(t)$ if $|\mathbf{x}(t)| \leq \frac{1}{2^i \cdot \bar{\epsilon}\sqrt{n}}$ and $\mathbf{x}_L(t) = 0$ otherwise. Let $\mathbf{x}_H(t) = \mathbf{x}(t)$ if $|\mathbf{x}(t)| > \frac{1}{2^i \cdot \bar{\epsilon}\sqrt{n}}$ and $\mathbf{x}_H(t) = 0$ otherwise. In the edge case, for $i = \ell + 1$, let $\mathbf{x}_H = \mathbf{x}$ and $\mathbf{x}_L = \mathbf{0}$. Writing $\mathbf{x} = \mathbf{x}_H + \mathbf{x}_L$ and applying triangle inequality, we can bound:

$$\left|\mathbf{x}_i^T\mathbf{A}\mathbf{x} - \mathbf{x}_i^T(\mathbf{A} \circ \mathbf{S})\mathbf{x}\right| \leq \left|\mathbf{x}_i^T\mathbf{A}\mathbf{x}_L - \mathbf{x}_i^T(\mathbf{A} \circ \mathbf{S})\mathbf{x}_L\right| + \left|\mathbf{x}_i^T\mathbf{A}\mathbf{x}_H - \mathbf{x}_i^T(\mathbf{A} \circ \mathbf{S})\mathbf{x}_H\right|. \tag{6}$$

In the following, we separately bound the two terms in (6). Roughly, since $\mathbf{x}_L$ has relatively small entries and thus is relatively well spread, we will be able to show, using a similar approach to Theorem 2, that the first term is small since sparsification with $\mathbf{S}$ approximately preserves $\mathbf{x}_i^T\mathbf{A}\mathbf{x}_L$. On the otherhand, since $\mathbf{x}_H$ has relatively large entries and thus is relatively sparse, we can show that the second term is small since $\mathbf{x}_i^T\mathbf{A}\mathbf{x}_H$ is small (and $\mathbf{x}_i^T(\mathbf{A} \circ \mathbf{S})\mathbf{x}_H$ cannot be much larger).

Term 1: Well-Spread Vectors. We start by bounding $\left|\mathbf{x}_i^T\mathbf{A}\mathbf{x}_L - \mathbf{x}_i^T(\mathbf{A} \circ \mathbf{S})\mathbf{x}_L\right|$. Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ and $\lambda_1, \ldots, \lambda_n$ be the eigenvectors and eigenvalues of $\mathbf{A}$ so that $\mathbf{A} = \sum_{k=1}^n \lambda_k \mathbf{v}_k \mathbf{v}_k^T$. Then we write:

$$\begin{aligned} |\mathbf{x}_i^T\mathbf{A}\mathbf{x}_L - \mathbf{x}_i^T(\mathbf{A} \circ \mathbf{S})\mathbf{x}_L| &= \left|\sum_{k=1}^n \lambda_k \cdot \mathbf{x}_i^T\mathbf{v}_k\mathbf{v}_k^T\mathbf{x}_L - \sum_{k=1}^n \lambda_k \mathbf{x}_i^T(\mathbf{v}_k\mathbf{v}_k^T \circ \mathbf{S})\mathbf{x}_L)\right| \\ &\leq \sum_{k=1}^n |\lambda_k| \cdot \left|\mathbf{x}_i^T[\mathbf{v}_k\mathbf{v}_k^T \circ (\mathbf{1} - \mathbf{S})]\mathbf{x}_L\right|. \end{aligned}$$

As in the proof of Theorem 2, if we let $\mathbf{D}_k \in \mathbb{R}^{n \times n}$ be a diagonal matrix with the entries of $\mathbf{v}_k$ on its diagonal, then we have $\mathbf{v}_k\mathbf{v}_k^T \circ (\mathbf{1} - \mathbf{S}) = \mathbf{D}_k(\mathbf{1} - \mathbf{S})\mathbf{D}_k$. Further, $\mathbf{x}_i^T\mathbf{D}_k(\mathbf{1} - \mathbf{S})\mathbf{D}_k\mathbf{x}_L = \mathbf{v}_k^T\mathbf{D}_i(\mathbf{1} - \mathbf{S})\mathbf{D}_L\mathbf{v}_k$, where $\mathbf{D}_i, \mathbf{D}_L$ are diagonal with the entries of $\mathbf{x}_i$ and $\mathbf{x}_L$ on their diagonals.

Plugging this back in we have:

$$\begin{aligned}|\mathbf{x}_i^T\mathbf{A}\mathbf{x}_L-\mathbf{x}_i^T(\mathbf{A}\circ\mathbf{S})\mathbf{x}_L| &\leq \sum_{k=1}^{n}|\lambda_k|\cdot\left|\mathbf{v}_k^T\mathbf{D}_i(\mathbf{1}-\mathbf{S})\mathbf{D}_L\mathbf{v}_k^T\right|\\ &\leq \|\mathbf{1}-\mathbf{S}\|_2\cdot\sum_{k=1}^{n}|\lambda_k|\cdot\|\mathbf{v}_k^T\|_2\cdot\|\mathbf{D}_i\|_2\cdot\|\mathbf{D}_L\|_2\cdot\|\mathbf{v}_k\|_2\\ &\leq \|\mathbf{1}-\mathbf{S}\|_2\cdot\sum_{k=1}^{n}|\lambda_k|\cdot\|\mathbf{D}_i\|_2\cdot\|\mathbf{D}_L\|_2. \qquad (7)\end{aligned}$$

Finally, observe that by definition, for any $i\in\{0,\dots,\ell\}$, $\|\mathbf{D}_i\|_2=\max_{t\in[n]}|\mathbf{x}_i(t)|\leq\frac{2^i}{\sqrt{n}}$ and $\|\mathbf{D}_L\|_2=\max_{t\in[n]}|\mathbf{x}_L(t)|\leq\frac{1}{2^i\bar{\epsilon}\sqrt{n}}$. Thus, $\|\mathbf{D}_i\|_2\cdot\|\mathbf{D}_L\|_2\leq\frac{1}{\bar{\epsilon}n}$. For $i=\ell+1$, $\mathbf{x}_L=\mathbf{0}$ by definition and so, the bound $\|\mathbf{D}_i\|_2\cdot\|\mathbf{D}_L\|_2\leq\frac{1}{\bar{\epsilon}n}$ holds vacuously. Also, by the assumption of the theorem, $\|\mathbf{1}-\mathbf{S}\|_2\leq\frac{\epsilon^2 n}{c\log^2(1/\epsilon)}\leq\bar{\epsilon}^2 n$. Plugging back into (7),

$$|\mathbf{x}_i^T\mathbf{A}\mathbf{x}_L-\mathbf{x}_i^T(\mathbf{A}\circ\mathbf{S})\mathbf{x}_L|\leq\bar{\epsilon}^2 n\cdot\frac{1}{\bar{\epsilon}n}\cdot\|\mathbf{A}\|_1\leq\bar{\epsilon}\cdot\|\mathbf{A}\|_1, \qquad (8)$$

as required.

Term 2: Sparse Vectors. We next bound the second term of (6): $\left|\mathbf{x}_i^T\mathbf{A}\mathbf{x}_H-\mathbf{x}_i^T(\mathbf{A}\circ\mathbf{S})\mathbf{x}_H\right|$. We write $\mathbf{x}_i=\mathbf{x}_{i,P}+\mathbf{x}_{i,N}$, where $\mathbf{x}_{i,P}$ and $\mathbf{x}_{i,N}$ contain its positive and non-positive entries respectively. Similarly, write $\mathbf{x}_H=\mathbf{x}_{H,P}+\mathbf{x}_{H,N}$. We can then bound via triangle inequality:

$$\begin{aligned}|\mathbf{x}_i^T\mathbf{A}\mathbf{x}_H-\mathbf{x}_i^T(\mathbf{A}\circ\mathbf{S})\mathbf{x}_H| &\leq|\mathbf{x}_i^T\mathbf{A}\mathbf{x}_H|+|\mathbf{x}_i^T(\mathbf{A}\circ\mathbf{S})\mathbf{x}_H|\\ &\leq|\mathbf{x}_i^T\mathbf{A}\mathbf{x}_H|+|\mathbf{x}_{i,P}^T(\mathbf{A}\circ\mathbf{S})\mathbf{x}_{H,P}|+|\mathbf{x}_{i,P}^T(\mathbf{A}\circ\mathbf{S})\mathbf{x}_{H,N}|\\ &\quad+|\mathbf{x}_{i,N}^T(\mathbf{A}\circ\mathbf{S})\mathbf{x}_{H,P}|+|\mathbf{x}_{i,N}^T(\mathbf{A}\circ\mathbf{S})\mathbf{x}_{H,N}|. \qquad (9)\end{aligned}$$

We will bound each term in (9) by $O(\bar{\epsilon}n)$, giving that overall $|\mathbf{x}_i^T\mathbf{A}\mathbf{x}_H-\mathbf{x}_i^T(\mathbf{A}\circ\mathbf{S})\mathbf{x}_H|=O(\bar{\epsilon}n)$. We first observe that

$$|\mathbf{x}_i^T\mathbf{A}\mathbf{x}_H|\leq\|\mathbf{x}_i\|_1\cdot\|\mathbf{x}_H\|_1\cdot\|\mathbf{A}\|_\infty.$$

By assumption $\|\mathbf{A}\|_\infty\leq 1$. Further, for $i\in\{1,\dots,\ell+1\}$, since $|\mathbf{x}_i(t)|\geq\frac{2^{i-1}}{\sqrt{n}}$ for all t and since $\|\mathbf{x}_i\|_2\leq 1$, $\mathbf{x}_i$ has at most $\frac{n}{2^{2i-2}}$ non-zero entries. Thus, $\|\mathbf{x}_i\|_1\leq\frac{\sqrt{n}}{2^{i-1}}$. For $i=0$, this bound holds trivially since $\|\mathbf{x}_i\|_1\leq\sqrt{n}\cdot\|\mathbf{x}_i\|_2\leq\sqrt{n}\leq\frac{\sqrt{n}}{2^{i-1}}$.

Similarly, for $i\in\{0,\dots\ell\}$, since $\|\mathbf{x}_H\|_2\leq 1$ and since $|\mathbf{x}_H(t)|\geq\frac{1}{2^i\bar{\epsilon}\sqrt{n}}$ by definition, $\mathbf{x}_H$ has at most $2^{2i}\cdot\bar{\epsilon}^2 n$ non-zero entries. So $\|\mathbf{x}_H\|_1\leq 2^i\bar{\epsilon}\cdot\sqrt{n}$. For $i=\ell+1=\log_2(1/\bar{\epsilon})+1$ this bound holds trivially since $\|\mathbf{x}_H\|_1\leq\sqrt{n}\cdot\|\mathbf{x}_H\|_2\leq 2^i\bar{\epsilon}\cdot\sqrt{n}$. Putting these all together, we have

$$|\mathbf{x}_i^T\mathbf{A}\mathbf{x}_H|\leq\|\mathbf{x}_i\|_1\cdot\|\mathbf{x}_H\|_1\cdot\|\mathbf{A}\|_\infty\leq\frac{\sqrt{n}}{2^{i-1}}\cdot 2^i\bar{\epsilon}\cdot\sqrt{n}=2\bar{\epsilon}n. \qquad (10)$$

We next bound $|\mathbf{x}_{i,P}^T(\mathbf{A}\circ\mathbf{S})\mathbf{x}_{H,P}|$. Since $\mathbf{x}_{i,P}$ and $\mathbf{x}_{H,P}$ are both all positive vectors, and since by assumption $\|\mathbf{A}\|_\infty\leq 1$,

$$\begin{aligned}|\mathbf{x}_{i,P}^T(\mathbf{A}\circ\mathbf{S})\mathbf{x}_{H,P}| &\leq|\mathbf{x}_{i,P}^T\mathbf{S}\mathbf{x}_{H,P}|\\ &\leq|\mathbf{x}_{i,P}^T\mathbf{1}\mathbf{x}_{H,P}|+|\mathbf{x}_{i,P}^T(\mathbf{1}-\mathbf{S})\mathbf{x}_{H,P}|\\ &\leq\|\mathbf{x}_{i,P}\|_1\cdot\|\mathbf{x}_{H,P}\|_1\cdot\|\mathbf{1}\|_\infty+\|\mathbf{x}_{i,P}\|_2\cdot\|\mathbf{x}_{H,P}\|_2\cdot\|\mathbf{1}-\mathbf{S}\|_2. \qquad (11)\end{aligned}$$

Following the same argument used to show (10), $\|\mathbf{x}_{i,P}\|_1 \cdot \|\mathbf{x}_{H,P}\|_1 \cdot \|\mathbf{1}\|_\infty \leq 2\bar{\epsilon}n$. Further, since by the assumption of the theorem, $\|\mathbf{1} - \mathbf{S}\|_2 \leq \frac{\epsilon^2 n}{c\log^2(1/\epsilon)} \leq \bar{\epsilon}^2 n \leq \bar{\epsilon}n$, and since $\mathbf{x}_{i,P}$ and $\mathbf{x}_{H,P}$ both are at most unit norm, $\|\mathbf{x}_{i,P}\|_2 \cdot \|\mathbf{x}_{H,P}\|_2 \cdot \|\mathbf{1} - \mathbf{S}\|_2 \leq \bar{\epsilon}n$. Thus, plugging back into (11), we have $|\mathbf{x}_{i,P}^T(\mathbf{A} \circ \mathbf{S})\mathbf{x}_{H,P}| \leq 3\bar{\epsilon}n$. Identical bounds will hold for the remaining three terms of (9), since for each, the two vectors in the quadratic form have entries that either always match or never match on sign. Plugging these bounds and (10) back into (9), we obtain

$$|\mathbf{x}_i^T \mathbf{A}\mathbf{x}_H - \mathbf{x}_i^T(\mathbf{A} \circ \mathbf{S})\mathbf{x}_H| \leq 2\bar{\epsilon}n + 4 \cdot 3\bar{\epsilon}n \leq 14\bar{\epsilon}n, \tag{12}$$

which completes the required bound in the sparse case.

Concluding the Proof. We finally plug our bounds for the sparse case (12) and the well-spread case (8) into (6) to obtain:

$$\left|\mathbf{x}_i^T \mathbf{A}\mathbf{x} - \mathbf{x}_i^T(\mathbf{A} \circ \mathbf{S})\mathbf{x}\right| \leq 14\bar{\epsilon}n + \bar{\epsilon}\|\mathbf{A}\|_1 \leq 15\bar{\epsilon} \cdot \max(n, \|\mathbf{A}\|_1).$$

Plugging this bound into (5) gives that $|\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{x}^T(\mathbf{A} \circ \mathbf{S})\mathbf{x}| = O(\epsilon \cdot \max(n, \|\mathbf{A}\|_1))$, which completes the theorem after adjusting ϵ by a constant factor. ◀

4 Spectral Approximation Lower Bounds

We now show that our universal sparsifier upper bounds for both PSD matrices (Theorem 1) and non-PSD matrices (Theorem 4) are nearly tight. We also give $\Omega(n/\epsilon^2)$ query lower bounds against general deterministic (possibly adaptive) spectral approximation algorithms (Theorem 8) and general randomized spectral approximation algorithms (Theorem 12).

4.1 Sparsity Lower Bound for PSD Matrices

We first prove that every matrix which is an ϵn spectral approximation to the all-ones matrix must have $\Omega(\frac{n}{\epsilon^2})$ non-zero entries. This shows that Theorem 1 is optimal up to constant factors, even for algorithms that sparsify just a single bounded entry PSD matrix. The idea of the lower bound is simple: if a matrix $\mathbf{S}$ spectrally approximates the all-ones matrix, its entries must sum to $\Omega(n^2)$. Thus, if $\mathbf{S}$ has just s non-zero entries, it must have Frobenius norm at least $\Omega(n^2/\sqrt{s})$. Unless $s = \Omega(n/\epsilon^2)$, this Frobenius norm is too large for $\mathbf{S}$ to be a ϵn-spectral approximation of the all ones matrix (which has $\|\mathbf{1}\|_F = n$.)

▶ **Theorem 3** (Sparsity Lower Bound – PSD Matrices). *Let $\mathbf{1} \in \mathbb{R}^{n\times n}$ be the all-ones matrix. Then, for any $\epsilon \in (0, 1/2)$ with $\epsilon \geq c/\sqrt{n}$ for large enough constant c, any $\mathbf{S} \in \mathbb{R}^{n\times n}$ with $\|\mathbf{1} - \mathbf{S}\|_2 \leq \epsilon n$ must have $\Omega(n/\epsilon^2)$ nonzero entries.*

Proof. If we let $\mathbf{x} \in \mathbb{R}^n$ be the all ones vector, then since $\|\mathbf{x}\|_2^2 = n$, $\|\mathbf{1} - \mathbf{S}\|_2 \leq \epsilon n$ implies that $|\mathbf{x}^T\mathbf{S}\mathbf{x} - \mathbf{x}^T\mathbf{1}\mathbf{x}| = |\mathbf{x}^T\mathbf{S}\mathbf{x} - n^2| \leq \epsilon n^2$. This in turn implies:

$$\sum_{i,j\in[n]} |\mathbf{S}_{ij}| \geq \sum_{i,j\in[n]} \mathbf{S}_{ij} \geq n^2 - \epsilon n^2 \geq \frac{n^2}{2},$$

where the last inequality follows from the assumption that $\epsilon \leq \frac{1}{2}$. If $\mathbf{S}$ has s non-zero entries, since $\sum_{i,j\in[n]} |\mathbf{S}_{ij}|$ is the ℓ_1-norm of the entries, and $\|\mathbf{S}\|_F$ is the ℓ_2-norm of the entries, we conclude by $\ell_1 - \ell_2$ norm equivalence that $\|\mathbf{S}\|_F \geq \frac{1}{\sqrt{s}} \sum_{i,j\in[n]} |\mathbf{S}_{ij}| \geq \frac{n^2}{2\sqrt{s}}$. Therefore, by the triangle inequality,

$$\|\mathbf{1}-\mathbf{S}\|_F \geq \|\mathbf{S}\|_F - \|\mathbf{1}\|_F = \frac{n^2}{2\sqrt{s}} - n \geq \frac{n^2}{4\sqrt{s}},$$

as long as $s \leq \frac{n^2}{16}$ so that $\frac{n^2}{2\sqrt{s}} \geq 2n$. Now, using that $\|\mathbf{A}\|_F^2 = \sum_{i=1}^n \sigma_i^2(\mathbf{A}) \leq n \cdot \|\mathbf{A}\|_2^2$, we have:

$$\|\mathbf{1}-\mathbf{S}\|_F \geq \frac{n^2}{4\sqrt{s}} \implies \|\mathbf{1}-\mathbf{S}\|_F^2 \geq \frac{n^4}{16s} \implies \|\mathbf{1}-\mathbf{S}\|_2^2 \geq \frac{n^3}{16s} \implies \|\mathbf{1}-\mathbf{S}\|_2 \geq \frac{n^{3/2}}{4\sqrt{s}}.$$

By our assumption that $\|\mathbf{1}-\mathbf{S}\|_2 \leq \epsilon n$, this means that $s = \Omega\left(\frac{n}{\epsilon^2}\right)$, concluding the theorem. ◀

4.2 Lower Bounds for Deterministic Approximation of Non-PSD Matrices

We next show that our universal sparsification bound for non-PSD matrices (Theorem 4) is tight up to a $\log^4(1/\epsilon)$ factor. Our lower bound holds even for the easier problem of top singular value (spectral norm) approximation and against a more general class of algorithms, which non-adaptively and deterministically query entries of the input matrix. We show how to extend the lower bound to possibly adaptive deterministic algorithms in Theorem 8, but with a $1/\epsilon^2$ factor loss.

▶ **Theorem 7** (Non-Adaptive Query Lower Bound for Deterministic Spectral Approximation of Non-PSD Matrices). *For any $\epsilon \in (1/n^{1/4}, 1/4)$, any deterministic algorithm that queries entries of a bounded entry matrix $\mathbf{A}$ non-adaptively and outputs $\widetilde{\sigma}_1$ satisfying $|\sigma_1(\mathbf{A}) - \widetilde{\sigma}_1| \leq \epsilon \cdot \max(n, \|\mathbf{A}\|_1)$ must read at least $\Omega\left(\frac{n}{\epsilon^4}\right)$ entries. Here $\sigma_1(\mathbf{A}) = \|\mathbf{A}\|_2$ is the largest singular value of $\mathbf{A}$.*

Proof. Assume that we have a deterministic algorithm $\mathcal{A}$ that non-adaptively reads s entries of any bounded symmetric matrix $\mathbf{A}$ and outputs $\widetilde{\sigma}_1$ with $|\sigma_1(\mathbf{A}) - \widetilde{\sigma}_1| \leq \epsilon \max(n, \|\mathbf{A}\|_1)$. Assume for the sake of contradiction that $s \leq \frac{cn}{\epsilon^4}$ for some sufficiently small constant c.

Let $T \subset [n]$ be a set of $n^{3/2}/s^{1/2}$ rows on which $\mathcal{A}$ reads at most $\sqrt{ns}$ entries. Such a subset must exist since the average number of entries read in any set of $n^{3/2}/s^{1/2}$ rows is $\frac{n^{3/2}}{s^{1/2}} \cdot n \cdot \frac{s}{n^2} = \sqrt{ns}$.

Let $\mathbf{1}_T$ be the matrix which is 1 on all the rows in T and zero everywhere else. Let $\mathbf{S}_T$ be the matrix which matches $\mathbf{1}_T$ on all entries, except is 0 on any entry in the rows of T that is not read by the algorithm $\mathcal{A}$. Observe that $\mathcal{A}$ reads the same entries (all ones) and thus outputs the same approximation $\widetilde{\sigma}_1$ for $\mathbf{1}_T$ and $\mathbf{S}_T$. We now bound the allowed error of this approximation. $\mathbf{1}_T$ is rank-1 and so:

$$\|\mathbf{1}_T\|_1 = \sigma_1(\mathbf{1}_T) = \|\mathbf{1}_T\|_F = \frac{n^{5/4}}{s^{1/4}}.$$

We have $\sigma_1(\mathbf{S}_T) \leq \|\mathbf{S}_T\|_F \leq n^{1/4}s^{1/4}$ since $\mathbf{S}_T$ has just $\sqrt{ns}$ entries set to 1 – the entries that $\mathcal{A}$ reads in the rows of T. Using that $\mathbf{S}_T$ is supported only on the $n^{3/2}/s^{1/2}$ rows of T, $\mathbf{S}_T$ has rank at most $n^{3/2}/s^{1/2}$. Thus, we can bound:

$$\|\mathbf{S}_T\|_1 \leq \frac{n^{3/4}}{s^{1/4}} \cdot \|\mathbf{S}_T\|_F \leq n.$$

Now, since $|\sigma_1(\mathbf{1}_T) - \sigma_1(\mathbf{S}_T| \geq \left|\frac{n^{5/4}}{s^{1/4}} - n^{1/4}s^{1/4}\right|$, our algorithm incurs error on one of the two input instances at least

$$\frac{\left|\frac{n^{5/4}}{s^{1/4}} - n^{1/4}s^{1/4}\right|}{2} \geq \frac{n^{5/4}}{4s^{1/4}},$$

where the inequality follows since, by assumption, $s \leq \frac{cn}{\epsilon^4}$ for some small constant c and $\epsilon \geq \frac{1}{n^{1/4}}$, and thus $n^{1/4}s^{1/4} \leq \frac{n^{5/4}}{2s^{1/4}}$.

The above is a contradiction when $\epsilon < 1/4$ since the above error is at least $1/4 \cdot \|\mathbf{1}_T\|_1$ and further, for $s \leq \frac{cn}{\epsilon^4}$, the error it at least $\frac{\epsilon n}{4c^{1/4}} > \epsilon n \geq \epsilon \cdot \|\mathbf{S}_T\|_1$ if we set c small enough. Thus, on at least one of the two input instances the error exceeds $\epsilon \cdot \max(n, \|\mathbf{A}\|_1)$, yielding a contradiction. ◀

We can prove a variant on Theorem 7 when the algorithm is allowed to make adaptive queries. Here, our lower bound reduces to $\Omega(n/\epsilon^2)$, as we are not able to restrict our hard case to a small set of rows of the input matrix. Closing the gap here – either by giving a stronger lower bound in the adaptive case or giving an adaptive query deterministic algorithm that achieves $\widetilde{O}(n/\epsilon^2)$ query complexity is an interesting open question.

▶ **Theorem 8** (Adaptive Query Lower Bound for Deterministic Spectral Approximation of Non-PSD Matrices). *For any $\epsilon \in (1/\sqrt{n}, 1/4)$, any deterministic algorithm that queries entries of a bounded entry matrix $\mathbf{A}$ (possibly adaptively) and outputs $\widetilde{\sigma}_1$ satisfying $|\sigma_1(\mathbf{A}) - \widetilde{\sigma}_1| \leq \epsilon \cdot \max(n, \|\mathbf{A}\|_1)$ must read at least $\Omega\left(\frac{n}{\epsilon^2}\right)$ entries. Here $\sigma_1(\mathbf{A}) = \|\mathbf{A}\|_2$ is the largest singular value of $\mathbf{A}$.*

Proof. Assume that we have a deterministic algorithm $\mathcal{A}$ that reads at most s entries of any bounded entry matrix $\mathbf{A}$ and outputs $\widetilde{\sigma}_1$ with $|\sigma_1(\mathbf{A}) - \widetilde{\sigma}_1| \leq \epsilon \cdot \max(n, \|\mathbf{A}\|_1)$. Assume for the sake of contradiction that $s \leq \frac{cn}{\epsilon^2}$ for some sufficiently small constant c.

Let S be the set of entries that $\mathcal{A}$ reads when given the all ones matrix $\mathbf{1}$ as input. Let $\mathbf{S}$ be the matrix which is one on every entry in S and zero elsewhere. Observe that $\mathcal{A}$ reads the same entries and thus outputs the same approximation $\widetilde{\sigma}_1$ for $\mathbf{1}$ and $\mathbf{S}$. We now bound the allowed error of this approximation. $\mathbf{1}$ is rank-1 and has $\|\mathbf{1}\|_1 = \sigma_1(\mathbf{1}) = n$. We have $\sigma_1(\mathbf{S}) \leq \|\mathbf{S}\|_F = \sqrt{s}$ and can bound $\|\mathbf{S}\|_1 \leq \sqrt{n} \cdot \|\mathbf{S}\|_F \leq \sqrt{sn} \leq \frac{c^{1/2}n}{\epsilon}$, where the last bound follows from our assumption that $s \leq \frac{cn}{\epsilon^2}$. Now, since $|\sigma_1(\mathbf{1}) - \sigma_1(\mathbf{S}| \geq |n - \sqrt{s}|$, our algorithm incurs an error on one of the two input instances at least

$$\frac{|n - \sqrt{s}|}{2} \geq \frac{n}{4},$$

where the inequality holds since, by assumption, $s \leq \frac{cn}{\epsilon^2}$ for some small constant c and $\epsilon \geq \frac{1}{n^{1/2}}$, and thus $\sqrt{s} \leq \frac{n}{2}$.

The above is a contradiction when $\epsilon < 1/4$ since the error mentioned above is at least $1/2 \cdot \|\mathbf{1}\|_1 = 1/2 \cdot \max(n, \|\mathbf{1}\|_1)$. Furthermore, since we have bounded $\|\mathbf{S}\|_1 \leq \frac{c^{1/2}n}{\epsilon}$, the error is greater than $\epsilon \cdot \max(n, \|\mathbf{S}\|_1)$ when $c < 1$. Thus, on at least one of the two input instances the error exceeds $\epsilon \cdot \max(n, \|\mathbf{A}\|_1)$, yielding a contradiction. ◀

4.3 Lower Bound for Randomized Approximation of Non-PSD Matrices

Finally, we prove Theorem 12, which gives an $\Omega(n/\epsilon^2)$ query lower bound for spectral approximation of bounded entry matrices that holds even for randomized and adaptive algorithms that approximate $\mathbf{A}$ in an arbitrary manner (not necessarily via sparsification). In particular, we show the lower bound against algorithms that produce any data structure, $f(\cdot,\cdot)$, that satisfies $|f(\mathbf{x},\mathbf{y}) - \mathbf{x}^T\mathbf{A}\mathbf{y}| \leq \epsilon n$ for all unit norm $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$

To prove this lower bound, we let $\mathbf{A}$ be a random matrix where each row has binary entries that are either i.i.d. fair coin flips, or else are coin flips with bias $+\epsilon$. Each row is unbiased with probability 1/2 and biased with probability 1/2. We show that an ϵn-spectral approximation to $\mathbf{A}$ suffices to identify for at least a 9/10 fraction of the rows, whether or not they are biased – roughly since the approximation must preserve the fact that $\mathbf{x}^T\mathbf{A}\mathbf{x}$ is large when $\mathbf{x}$ is supported just on this biased set. Consider a communication problem in the blackboard model, in which each of n^2 players can access just a single entry of $\mathbf{A}$. It is known that if the n players corresponding to a single row of the matrix want to identify with good probability whether or not it is biased, they must communicate at least $\Omega(1/\epsilon^2)$ bits [44]. Further, via a direct sum type argument, we can show that for the n^2 players to identify the bias of a 9/10 fraction of the rows with good probability, $\Omega(n/\epsilon^2)$ bits must be communicated. I.e., at least $\Omega(n/\epsilon^2)$ of the players must read their input bits, yielding our query complexity lower bound.

Note that there may be other proof approaches here, based on a direct sum of n 2-player Gap-Hamming instances [11, 20, 15], but our argument is simple and already gives optimal bounds.

Section Roadmap. In Section 4.3.1, we formally define the problem of identifying the bias of a large fraction of $\mathbf{A}$'s rows as the (ϵ, n)-*distributed detection problem* (Definition 17). We prove a $\Omega(n/\epsilon^2)$ query lower bound for this problem in Lemma 20.. Then, in Section 4.3.2, we prove Theorem 12 by showing a reduction from the distributed detection problem to spectral approximation.

4.3.1 Distributed Detection Lower Bound

Here, we define additional concepts and notation specific to this section. For random variables X, Y and Z, let $H(X)$ denote the entropy, $H(X|Y)$ denote the conditional entropy, $I(X;Y)$ denote mutual information, and $I(X;Y|Z)$ denote conditional mutual information [25]. We also use some ideas from communication complexity. Namely, we work in the *blackboard model*, where T parties communicate by posting messages to a public blackboard with access to public randomness with unlimited rounds of adaptivity. Let $\Pi \in \{0,1\}^*$ denote the transcript of a protocol posted to the blackboard, and let $|\Pi|$ denote the length of a transcript. For a fixed protocol with fixed probability of success, we may assume, without loss of generality, that Π has a fixed pre-determined length (see Section 2.2 of [39]).

Our query complexity lower bound for spectral approximation will follow from a reduction from the following testing problem.

▶ **Definition 15** (ϵ-Distributed detection problem [44]). *For fixed distributions* $\mu_0 = \text{Bernoulli}(1/2)$ *and* $\mu_1 = \text{Bernoulli}(1/2+\epsilon)$*, with* $\epsilon \in [0, \frac{1}{2}]$*, let* $\mathbf{x} \in \{0,1\}^n$ *be a random vector such that* $\mathbf{x}_1, \ldots, \mathbf{x}_n$ *are sampled i.i.d. from* μ_V*, for* $V \in \{0,1\}$*. The distributed detection problem is the task of determining whether* $V = 0$ *or* $V = 1$*, given the values of* $\mathbf{x}$*.*

This decision problem can be naturally interpreted as a communication problem in the blackboard model, where n players each have a single private bit corresponding to whether a unique entry of $\mathbf{x}$ is zero or one. Prior work takes this view to lower bound the mutual information between the transcript of a protocol which correctly solves the ϵ-Distributed detection problem with constant advantage and the sampled bits in $\mathbf{x}$ in the case $V = 0$.

▶ **Theorem 16** (Theorem 6 in [44]). *Let Π be the transcript of a protocol that solves ϵ-distributed detection problem with probability $1-p$ for any fixed choice of $p \in [0, 0.5)$. Then*

$$I(X; \Pi | V = 0) = \Omega(\epsilon^{-2}).$$

Next, we define the (ϵ, n)-Distributed detection problem, which combines n length-n ϵ-Distributed detection problems into a single joint detection problem.

▶ **Definition 17** ((ϵ, n)-Distributed detection problem). *For $\mathbf{v} \in \{0,1\}^n$ distributed uniformly on the Hamming cube, generate the matrix $\mathbf{A} \in \{0,1\}^{n \times n}$ such that if $\mathbf{v}_i = 0$, then all entries in the i-th row of $\mathbf{A}$ are i.i.d. samples from* Bernoulli$(1/2)$. *Otherwise, let all entries in the i-th row be sampled from* Bernoulli$(1/2 + \epsilon)$. *The (ϵ, n)-Distributed detection problem is the task of recovering a vector $\hat{\mathbf{v}} \in \{0,1\}^n$ such that $\|\hat{\mathbf{v}} - \mathbf{v}\|_1 \leq \frac{n}{20}$.*

Again, this vector recovery problem has a natural interpretation as a communication problem, where n^2 players each hold a single bit of information that corresponds to whether a unique entry of $\mathbf{A}$ is zero or one. Throughout this section, we will view Definition 17 as a decision problem or communication problem as needed. Let Π be the transcript of a protocol that solves the (ϵ, n)-Distributed detection problem in the communication model introduced at the beginning of this section. Lower bounding the mutual information between the transcript, Π, and the private information held by the players (the entries of $\mathbf{A}$) lower bounds the length of the transcript via the following argument.

▶ **Lemma 18.** *If Π is the transcript of a protocol that solves the (ϵ, n)-Distributed detection problem and $I(\mathbf{A}; \Pi) \geq b$, then $|\Pi| \geq \frac{b}{\log 2}$.*

Proof. For random variables X, Y, $I(X;Y) \leq \min\{H(X), H(Y)\}$ [25]. If a random variable X has finite support, then $H(X) \leq \log|\mathrm{supp}(X)|$. Recall that $|\Pi|$ is the number of bits in the transcript and hence $\log 2^{|\Pi|} = \log 2 \cdot |\Pi| \geq H(\Pi) \geq I(\mathbf{A}; \Pi) \geq b$. Hence, we conclude the statement. ◀

Next, we lower bound the number of entries in the matrix $\mathbf{A}$ that must be observed to solve a variant of the (ϵ, n)-Distributed detection problem, where we aim to correctly decide each index of $\mathbf{v}$ with constant success probability, rather than constructing $\hat{\mathbf{v}}$ such that $\|\hat{\mathbf{v}} - \mathbf{v}\|_1 \leq \frac{n}{20}$. There are three sources of randomness in this problem: 1) the initial randomness in sampling $\mathbf{v}$, 2) the random variable $\mathbf{A}$ which depends on $\mathbf{v}$, and 3) the random transcript Π which depends on $\mathbf{A}$. When necessary to avoid confusion, we will be explicit regarding which of these variables are considered fixed and which are considered as random in probabilistic statements.

▶ **Lemma 19.** *Let $\mathbf{A}$ and $\mathbf{v}$ be distributed as in the (ϵ, n)-Distributed detection problem. Any protocol with transcript Π that constructs a vector $\hat{\mathbf{v}}$ such that $\mathbb{P}(\mathbf{v}_i = \hat{\mathbf{v}}_i) \geq \frac{9}{10}$, for all $i \in [n]$, (where the randomness is with respect to $\mathbf{A}$, $\mathbf{v}$, and Π), must observe $\Omega(\frac{n}{\epsilon^2})$ entries of $\mathbf{A}$ in the worst case.*

Proof. Consider $\mathbf{A}$ as n^2 players each holding a single bit of information corresponding to whether a unique entry of $\mathbf{A}$ is zero or one. Here, let $\Pi \in \{0,1\}^*$ be the transcript of a protocol that satisfies $\mathbb{P}(\mathbf{v}_i = \hat{\mathbf{v}}_i) \geq \frac{9}{10}$ for every $i \in [n]$. Note that any algorithm which solves this problem by querying k entries of $\mathbf{A}$ implies a protocol for the communication problem which communicates k bits, since the algorithm could be simulated by (possibly adaptively) posting each queried bit to the blackboard.

First, we lower bound the un-conditional mutual information between the private information in the i-th row of $\mathbf{A}$ and the transcript Π. Note that $\mathbf{v}_i$ is one or zero with equal probability, therefore,

$$I(\mathbf{A}_i;\Pi|\mathbf{v}_i) = \frac{1}{2}I(\mathbf{A}_i;\Pi|\mathbf{v}_i=0) + \frac{1}{2}I(\mathbf{A}_i;\Pi|\mathbf{v}_i=1) \geq \frac{1}{2}I(\mathbf{A}_i;\Pi|\mathbf{v}_i=0).$$

Next, we decompose the mutual information by its definition and the chain rule, then, we use the fact that $0 \leq H(\mathbf{v}_i) \leq 1$ and that conditioning can only decrease the entropy of a random variable.

$$\begin{aligned}&I(\mathbf{A}_i;\Pi) = I(\mathbf{A}_i;\Pi|\mathbf{v}_i) + H(\mathbf{v}_i|\mathbf{A}_i,\Pi) - H(\mathbf{v}_i|\mathbf{A}_i) - H(\mathbf{v}_i|\Pi) + H(\mathbf{v}_i)\\ &\Rightarrow I(\mathbf{A}_i;\Pi) \geq I(\mathbf{A}_i;\Pi|\mathbf{v}_i) - 2.\end{aligned}$$

Observe that determining $\hat{\mathbf{v}}_i$ with the guarantee $\mathbb{P}(\mathbf{v}_i = \hat{\mathbf{v}}_i) \geq \frac{9}{10}$ solves the ϵ-Distributed detection problem with probability at least $\frac{9}{10}$. Therefore, by Theorem 16 $I(\mathbf{A}_i;\Pi|\mathbf{v}_i=0) = \Omega(\epsilon^{-2})$, and so we can use the previous two equations to lower bound for the conditional mutual information:

$$I(\mathbf{A}_i;\Pi) \geq \frac{1}{2}I(\mathbf{A}_i;\Pi|\mathbf{v}_i=0) - 2 = \Omega(\frac{1}{\epsilon^2}). \tag{13}$$

Next, we lower bound the total mutual information between $\mathbf{A}$ and Π. Mirroring the argument of Lemma 1 in [44], we use that, for independent random variables, entropy is additive and conditional entropy is subadditive to lower bound the mutual information between $\mathbf{A}$ and Π:

$$\begin{aligned}I(\{\mathbf{A}_i\}_{i\in[n]};\Pi) &= H(\{\mathbf{A}_i\}_{i\in[n]}) - H(\{\mathbf{A}_i\}_{i\in[n]}|\Pi)\\ &\geq \sum_{i=1}^{n} H(\mathbf{A}_i) - H(\mathbf{A}_i,\Pi)\\ &= \sum_{i=1}^{n} I(\mathbf{A}_i;\Pi) = \Omega\left(\frac{n}{\epsilon^2}\right),\end{aligned}$$

where the last step follows from (13). By Lemma 18, $|\Pi| = \Omega(I(\{\mathbf{A}_i\}_{i\in[n]};\Pi)) = \Omega(\frac{n}{\epsilon^2})$. Therefore, every algorithm which samples entries of $\mathbf{A}$ to construct a vector $\hat{\mathbf{v}}$ satisfying $\mathbb{P}(\mathbf{v}_i = \hat{\mathbf{v}}_i) \geq \frac{9}{10}$ must observe at least $\Omega(\frac{n}{\epsilon^2})$ entries of $\mathbf{A}$. ◀

We now have the necessary results to prove a lower bound on the number of entries of $\mathbf{A}$ that must be observed to solve the (ϵ, n)-Distributed detection problem, which we do by reducing to the problem in Lemma 19.

▶ **Lemma 20.** *Any adaptive randomized algorithm which solves the (ϵ, n)-Distributed detection problem with probability at least $\frac{19}{20}$ (with respect to the randomness in $\mathbf{v}$, $\mathbf{A}$, and Π) must observe $\Omega(\frac{n}{\epsilon^2})$ entries of $\mathbf{A}$.*

Proof. Any algorithm that can produce a vector $\hat{\mathbf{v}}$ such that $\|\hat{\mathbf{v}} - \mathbf{v}\|_1 \leq \frac{n}{20}$ with probability at least $\frac{19}{20}$ could be used to guarantee that $\mathbb{P}(\hat{\mathbf{v}}_i = \mathbf{v}_i) \geq \frac{9}{10}$ by the following argument.

For any input matrix $\mathbf{A}$, create the matrix $\bar{\mathbf{A}}$ such that $\bar{\mathbf{A}}_i = \mathbf{A}_{\sigma(i)}$, where σ is a permutation sampled uniformly from the symmetric group of size n. Let $\mathbf{v}_\sigma$ be the vector which satisfies $[\mathbf{v}_\sigma]_i = \mathbf{v}_{\sigma(i)}$. Run the considered protocol on $\bar{\mathbf{A}}$ to recover $\mathbf{u}$ such that $\|\mathbf{u} - \mathbf{v}_\sigma\|_1 \leq \frac{n}{20}$ with probability $\frac{19}{20}$. Finally, let $\hat{\mathbf{v}}_i = \mathbf{u}_{\sigma^{-1}(i)}$ for all $i \in [n]$.

Then, $\mathbb{P}(\hat{\mathbf{v}}_i = \mathbf{v}_i) = \mathbb{P}(\|\mathbf{u}-\mathbf{v}_\sigma\|_1 \le \frac{n}{20}) \cdot \mathbb{P}(\mathbf{u}_{\sigma^{-1}(i)} - \mathbf{v}_{\sigma^{-1}(i)} | \|\mathbf{u}-\mathbf{v}_\sigma\|_1 \le \frac{n}{20}) = \frac{19}{20} \cdot \frac{19}{20} \ge \frac{9}{10}$. Since $\sigma(i)$ is uniformly distributed over $[n]$, $\mathbb{P}(\mathbf{u}_{\sigma^{-1}(i)} - \mathbf{v}_{\sigma^{-1}(i)} | \|\mathbf{u}-\mathbf{v}_\sigma\|_1 \le \frac{n}{20})$ is the number of correctly decided entries divided by n.

By Lemma 19, we conclude that solving the (ϵ, n)-Distributed detection with probability at least $\frac{19}{20}$ requires observing $\Omega(\frac{n}{\epsilon^2})$ entries of $\mathbf{A}$ in the worst case. ◀

4.3.2 Spectral Approximation Query Lower Bound

Next, we show that the (ϵ, n)-Distributed detection problem can be solved using a spectral approximation of $\mathbf{A}$, thereby implying a query complexity lower bound for spectral approximation.

▶ **Theorem 12** (Lower Bound for Randomized Spectral Approximation). *Any randomized algorithm that (possibly adaptively) reads entries of a binary matrix $\mathbf{A} \in \{0,1\}^{n\times n}$ to construct a data structure $f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ that, for $\epsilon \in (0,1)$ satisfies with probability at least 99/100,*

$$|f(\mathbf{x}, \mathbf{y}) - \mathbf{x}^T \mathbf{A}\mathbf{y}| \le \epsilon n, \text{ for all unit } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n,$$

must read $\Omega\left(\frac{n}{\epsilon^2}\right)$ entries of $\mathbf{A}$ in the worst case, provided $\epsilon = \Omega\left(\frac{\log n}{\sqrt{n}}\right)$.

Proof. We reduce solving the (ϵ, n)-Distributed detection problem to constructing a spectral approximation satisfying the guarantee in the above theorem statement. Throughout this proof, let ϵ_d be the parameter associated with the (ϵ, n)-Distributed detection problem and ϵ_s be the accuracy of the spectral approximation.

Let $\mathbf{A}$ be the matrix associated with the (ϵ_d, n)-Distributed detection problem. Suppose that by reading r entries of $\mathbf{A}$, we can create a data structure $f(\cdot,\cdot)$ satisfying:

$$|f(\mathbf{x}, \mathbf{y}) - \mathbf{x}^T \mathbf{A}\mathbf{y}| \le \epsilon_s n \|\mathbf{x}\|_2 \|\mathbf{y}\|_2. \text{ for all input } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

We show that for ϵ_s sufficiently small respect to ϵ_d, such a spectral approximation is sufficient to solve the (ϵ_d, n)-Distributed detection problem with no further queries to $\mathbf{A}$.

First, let B_i denote the number of ones in the i-th row of $\mathbf{A}$. Observe that $B_i \sim \text{Binomial}(n, 1/2)$ if $\mathbf{v}_i = 0$, and, $B_i \sim \text{Binomial}(n, 1/2 + \epsilon_d)$ if $\mathbf{v}_i = 1$. By Hoeffding's inequality [46],

$$P\left(|B_i - E[B_i]| > \sqrt{4n \log n}\right) \le 2 \exp\left(\frac{-4n \log n}{2n}\right) = \frac{2}{n^2}.$$

Therefore, by the union bound over all $i \in [n]$,

$$P\left(\max_{i\in S} |B_i - E[B_i]| > \sqrt{4n \log n}\right) \le \frac{2}{n}.$$

Let $S_1, S_2 \subset [n]$, such that $|S_1| = |S_2| = \frac{n}{2}$. Then with probability at least $1 - \frac{2}{n}$,

$$\begin{aligned}\left|\sum_{i\in S_1} B_i - \sum_{i\in S_2} B_i\right| &\le \left|E\left[\sum_{i\in S_1} B_i - \sum_{i\in S_2} B_i\right]\right| + \frac{2n}{2} \cdot \sqrt{4n \log n} \\ &= \left|E\left[\sum_{i\in S_1} B_i - \sum_{i\in S_2} B_i\right]\right| + O(n^{3/2} \log n). \qquad (14)\end{aligned}$$

For large enough n, we have with probability at least $\frac{99}{100}$ that there are at least $\frac{n}{4}$ biased rows, since the number of biased rows is distributed as Binomial$(n, 1/2)$. Define the set S^* such that $|S^*| = \frac{n}{2}$ and S^* contains a maximal number of biased rows. Define the set $\bar{S}$ as,

$$\bar{S} = \arg\max_{|S|=\frac{n}{2}} f(\mathbf{1}, \mathbf{1}_S).$$

We will show that $\bar{S}$ can contain at most $k \leq \frac{n}{80} + o(n)$ fewer biased rows than S^*. By guarantee of the spectral approximation and (14), (note $\mathbf{1}_S^T \mathbf{A} \mathbf{1}_S = \sum_{i,j \in S} \mathbf{A}_{ij}$),

$$\begin{aligned} f(\mathbf{1}, \mathbf{1}_{S^*}) - f(\mathbf{1}, \mathbf{1}_{\bar{S}}) &\geq \mathbf{1}^T \mathbf{A} \mathbf{1}_{S^*} - \mathbf{1}^T \mathbf{A} \mathbf{1}_{\bar{S}} - 2\epsilon_s n \|\mathbf{1}\|_2 \|\mathbf{1}_{S^*}\|_2 \\ &\geq E[\mathbf{1}^T \mathbf{A} \mathbf{1}_{S^*} - \mathbf{1}^T \mathbf{A} \mathbf{1}_{\bar{S}}] + O(n^{3/2} \log n) - 2\epsilon_s n^2. \end{aligned}$$

If $\bar{S}$ has k fewer biased rows than S^*, then, $E[\mathbf{1}^T \mathbf{A} \mathbf{1}_{S^*} - \mathbf{1}^T \mathbf{A} \mathbf{1}_{\bar{S}}] = \epsilon_d n k$. By the optimality of $\bar{S}$, $f(\mathbf{1}, \mathbf{1}_{S^*}) - f(\mathbf{1}, \mathbf{1}_{\bar{S}}) \leq 0$. Therefore, if $\epsilon_s = \frac{\epsilon_d}{160}$, then,

$$0 \geq f(\mathbf{1}, \mathbf{1}_{S^*}) - f(\mathbf{1}, \mathbf{1}_{\bar{S}}) \geq \epsilon_d n k - 2\epsilon_s n^2 + O(n^{3/2} \log n) = \epsilon_d n k - \frac{\epsilon_d n^2}{80} + O(n^{3/2} \log n).$$

Solving for k implies, $k \leq \frac{n}{80} + O(\frac{n^{1/2} \log n}{\epsilon_d})$. By assumption of the theorem statement, $\frac{1}{\epsilon_d} = o(\frac{\sqrt{n}}{\log n})$, therefore, $k \leq \frac{n}{80} + o(n)$.

Let b be the number of biased rows in $\mathbf{A}$. First, we consider the case where $b = \frac{n}{2}$ exactly. In this case, S^* contains $\frac{n}{2}$ biased rows, and hence, $\bar{S}$ contains at least $\frac{n}{2} - \frac{n}{80} - o(n) \geq \frac{n}{2} - \frac{n}{40}$ biased rows for large enough n. Therefore, deciding all rows in $\bar{S}$ (i.e., $\hat{\mathbf{v}}_i = 1$ for all $i \in \bar{S}$) are biased will correctly decide $(\frac{n}{2} - \frac{n}{40}) / \frac{n}{2} \geq \frac{19}{20}$ of the biased rows. By looking at the complement of S^* and $\bar{S}$, we conclude that $\frac{19}{20}$ of the unbiased rows are decided correctly as well. Hence, we can construct $\hat{\mathbf{v}}$ such that $\|\mathbf{v} - \hat{\mathbf{v}}\|_1 \leq \frac{n}{20}$ by the assignment $\hat{\mathbf{v}}_i = 1$ for $i \in \bar{S}$ and $\hat{\mathbf{v}}_i = 0$ otherwise.

The number of biased rows is distributed as a binomial random variable, i.e., $b \sim$ Binomial$(n, 1/2)$. Therefore, by Hoeffding's inequaliy, $\mathbb{P}(|b - \frac{n}{2}| > 10\sqrt{n}) < 0.01$. For large enough n, $10\sqrt{n} \leq 0.01 \cdot n$. Therefore, with probability at least $\frac{99}{100}$, we can reduce to the case $b = \frac{n}{2}$ by assuming that at most an additional $0.01 \cdot n$ rows are misclassified as biased or unbiased. Hence, with probability at least $\frac{99}{100}$, a spectral approximation of $\mathbf{A}$ with $\epsilon_s = \Theta(\epsilon_d)$ accuracy is sufficient to recover at least $\frac{9}{10}$ of the biased rows with probability at least $\frac{99}{100}$.

Correctly classifying $\frac{9}{10}$ of the rows as biased or unbiased requires observing $\Omega(\frac{n}{\epsilon_d^2})$ entries of $\mathbf{A}$ by Lemma 20. Since $\Omega(\frac{n}{\epsilon_d^2}) = \Omega(\frac{n}{\epsilon_s^2})$ entries of $\mathbf{A}$ must be observed to construct the spectral approximation used to solve the (ϵ, n)-Distributed detection problem, we conclude the lower bound of the theorem statement. ◀

Note that the assumption $\frac{1}{\epsilon} = o\left(\frac{\sqrt{n}}{\log n}\right)$ in the previous theorem is mild, since if this assumption does not hold, then we must read nearly all entries of the matrix anyways. We also show that our construction provides a lower bound even when the input is restricted to be symmetric.

▶ **Corollary 21.** *The lower bound in Theorem 12 applies when the input is restricted to **symmetric** binary matrices.*

Proof. While construction used in Theorem 12 is not symmetric, we can modify it to give the same lower bound for **symmetric** input. Let $\mathbf{A}$ be defined as in the proof of Theorem 12, and let $\bar{\mathbf{A}}$ be the Hermitian dilation of $\mathbf{A}$, i.e.,

$$\bar{\mathbf{A}} = \begin{bmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{A}^T & \mathbf{0} \end{bmatrix}.$$

Any query $\mathbf{x}^T\mathbf{A}\mathbf{y}$ can be simulated by the query $\bar{\mathbf{x}}^T\bar{\mathbf{A}}\bar{\mathbf{y}}$, where $\bar{\mathbf{x}} = [\mathbf{x}, \mathbf{0}]^T$ and $\bar{\mathbf{y}} = [\mathbf{0}, \mathbf{y}]^T$. The size of $\mathbf{A}$ is n and the size of $\bar{\mathbf{A}}$ is $2n$, hence, if $\bar{f}$ is a spectral approximation of $\bar{\mathbf{A}}$ such that,

$$|\bar{f}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) - \bar{\mathbf{x}^T}\bar{\mathbf{A}}\bar{\mathbf{y}}| \leq \left(\frac{\epsilon_s}{2}\right)(2n)\|\bar{\mathbf{x}}\|_2\|\bar{\mathbf{y}}\|_2, \text{ for all } \bar{\mathbf{x}}, \bar{\mathbf{y}} \in \mathbb{R}^n,$$

then we can simulate $f(\cdot, \cdot)$ such that,

$$|f(\mathbf{x}, \mathbf{y}) - \mathbf{x}^T\mathbf{A}\mathbf{y}| \leq \epsilon_s n\|\mathbf{x}\|_2\|\mathbf{y}\|_2, \text{ for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

Therefore, the proof of Theorem 12 applies to the Hermitian dilation $\bar{\mathbf{A}}$ after adjusting for a constant factor in the accuracy parameter ϵ_s. ◀

5 Conclusion

Our work has shown that it is possible to deterministically construct an entrywise sampling matrix $\mathbf{S}$ (a *universal sparsifier*) with just $\widetilde{O}(n/\epsilon^c)$ non-zero entries such that for any bounded entry $\mathbf{A}$ with $\|\mathbf{A}\|_\infty \leq 1$, $\|\mathbf{A} - \mathbf{A} \circ \mathbf{S}\|_2 \leq \epsilon \cdot \max(n, \|\mathbf{A}\|_1)$. We show how to achieve sparsity $O(n/\epsilon^2)$ when $\mathbf{A}$ is restricted to be PSD (Theorem 1) and $\widetilde{O}(n/\epsilon^4)$ when $\mathbf{A}$ is general (Theorem 4), and prove that both these bounds are tight up to logarithmic factors (Theorems 3 and 7). Further, our proofs are based on simple reductions, which show that any $\mathbf{S}$ that spectrally approximates the all ones matrix to sufficient accuracy (i.e., is a sufficiently good spectral expander) yields a universal sparsifier.

In Section 5 of the full version of the paper [14], we also apply our universal sparsification bounds to give the first $o(n^\omega)$ time deterministic algorithms for several core linear algebraic problems, including singular value/vector approximation and positive semidefiniteness testing. Additionally, when $\mathbf{A}$ is restricted to be PSD and have entries in $\{-1, 0, 1\}$, in Section 6 of the full version, we show how to give achieve improved deterministic query complexity of $\widetilde{O}(n/\epsilon)$ to construct a general spectral approximation $\widetilde{\mathbf{A}}$, which may not be sparse (Theorem 10). We again show that this bound is tight up a to a logarithmic factor (Theorem 11)

Our work leaves several open questions:

1. An interesting question is if $\widetilde{O}(n/\epsilon)$ sample complexity can be achieved for deterministic spectral approximation of any bounded entry PSD matrix if the sparsity of the output is not restricted, thereby generalizing the upper bound for PSD $\{-1, 0, 1\}$-matrices proven in Theorem 10. This query complexity is known for randomized algorithms based on column sampling [36], however it is not currently known how to derandomize such results.
2. It would also be interesting to close the $\widetilde{O}(1/\epsilon^2)$ factor gap between our universal sparsification upper bound of $\widetilde{O}(n/\epsilon^4)$ queries for achieving $\epsilon \cdot \max(n, \|\mathbf{A}\|_1)$ spectral approximation error for non-PSD matrices (Theorem 4) and our $\Omega(n/\epsilon^2)$ query lower bound for general deterministic algorithms that make possibly adaptive queries to $\mathbf{A}$ (Theorem 8). By Theorem 7, our universal sparsification bound is tight up to log factors for algorithms that make *non-adaptive* deterministic queries to $\mathbf{A}$. It is unknown if adaptive queries can be used to give improved algorithms.
3. Finally, it would be interesting to understand if our deterministic algorithms for spectrum approximation can be improved. For example, can one compute an ϵn additive error or a $(1 + \epsilon)$ relative error approximation to the top singular value $\|\mathbf{A}\|_2$ for bounded entry $\mathbf{A}$ and constant ϵ in $o(n^\omega)$ time deterministically? Are there fundamental barriers that make doing so difficult?

References

1 Dimitris Achlioptas, Zohar Shay Karnin, and Edo Liberty. Near-optimal entrywise sampling for data matrices. In *Advances in Neural Information Processing Systems 26 (NeurIPS)*, 2013.

2 Dimitris Achlioptas and Frank McSherry. Fast computation of low rank matrix approximations. In *Proceedings of the 39th Annual ACM Symposium on Theory of Computing (STOC)*, 2007.

3 Josh Alman and Virginia Vassilevska Williams. A refined laser method and faster matrix multiplication. In *Proceedings of the 32nd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2021.

4 Noga Alon. Eigenvalues and expanders. *Combinatorica*, 1986.

5 Noga Alon. Explicit expanders of every degree and size. *Combinatorica*, 2021.

6 Alexandr Andoni, Jiecao Chen, Robert Krauthgamer, Bo Qin, David P Woodruff, and Qin Zhang. On sketching quadratic forms. In *Proceedings of the 7th Conference on Innovations in Theoretical Computer Science (ITCS)*, 2016.

7 Sanjeev Arora, Elad Hazan, and Satyen Kale. Fast algorithms for approximate semidefinite programming using the multiplicative weights update method. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2005.

8 Sanjeev Arora, Elad Hazan, and Satyen Kale. A fast random sampling algorithm for sparsifying matrices. In *Proceedings of the 10th International Workshop on Randomization and Computation (RANDOM)*, 2006.

9 Haim Avron, Michael Kapralov, Cameron Musco, Christopher Musco, Ameya Velingker, and Amir Zandieh. Random Fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

10 Ainesh Bakshi, Nadiia Chepurko, and Rajesh Jayaram. Testing positive semi-definiteness via random submatrices. In *Proceedings of the 61st Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2020.

11 Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *Journal of Computer and System Sciences*, 2004.

12 Joshua Batson, Daniel A Spielman, Nikhil Srivastava, and Shang-Hua Teng. Spectral sparsification of graphs: theory and algorithms. *Communications of the ACM*, 2013.

13 Rajarshi Bhattacharjee, Gregory Dexter, Petros Drineas, Cameron Musco, and Archan Ray. Sublinear time eigenvalue approximation via random sampling. *Proceedings of the 50th International Colloquium on Automata, Languages and Programming (ICALP)*, 2023.

14 Rajarshi Bhattacharjee, Gregory Dexter, Cameron Musco, Archan Ray, Sushant Sachdeva, and David P Woodruff. Universal matrix sparsifiers and fast deterministic algorithms for linear algebra, 2023. arXiv:2305.05826.

15 Mark Braverman, Ankit Garg, Denis Pankratov, and Omri Weinstein. Information lower bounds via self-reducibility. In *Proceedings of the 8th International Computer Science Symposium in Russia (CSR)*, 2013.

16 Vladimir Braverman, Robert Krauthgamer, Aditya R Krishnan, and Shay Sapir. Near-optimal entrywise sampling of numerically sparse matrices. In *Proceedings of the 34th Annual Conference on Computational Learning Theory (COLT)*, 2021.

17 Vladimir Braverman, Aditya Krishnan, and Christopher Musco. Sublinear time spectral density estimation. In *Proceedings of the 54th Annual ACM Symposium on Theory of Computing (STOC)*, 2022.

18 Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 2012.

19 Emmanuel J. Candès and Terence Tao. The power of convex relaxation: near-optimal matrix completion. *IEEE Transations on Information Theory*, 2010.

20 Amit Chakrabarti and Oded Regev. An optimal lower bound on the communication complexity of gap-Hamming-distance. *SIAM Journal on Computing*, 2012.

21 Françoise Chatelin. *Spectral approximation of linear operators*. SIAM, 2011.

22 Julia Chuzhoy, Yu Gao, Jason Li, Danupon Nanongkai, Richard Peng, and Thatchaphol Saranurak. A deterministic algorithm for balanced cut with applications to dynamic connectivity, flows, and beyond. In *Proceedings of the 61st Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2020.

23 Kenneth L Clarkson and David P Woodruff. Low-rank PSD approximation in input-sparsity time. In *Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2017.

24 Michael B Cohen. Ramanujan graphs in polynomial time. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2016.

25 Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

26 Alexandre d'Aspremont. Subsampling algorithms for semidefinite programming. *Stochastic Systems*, 2011.

27 Petros Drineas and Anastasios Zouzias. A note on element-wise matrix sparsification via a matrix-valued Bernstein inequality. *Information Processing Letters*, 2011.

28 Jacek Kuczyński and Henryk Woźniakowski. Estimating the largest eigenvalue by the power and Lanczos algorithms with a random start. *SIAM Journal on Matrix Analysis and Applications*, 1992.

29 Abhisek Kundu. *Element-wise matrix sparsification and reconstruction*. PhD thesis, Rensselaer Polytechnic Institute, USA, 2015.

30 Lin Lin, Yousef Saad, and Chao Yang. Approximating spectral densities of large matrices. *SIAM Review*, 2016.

31 Alexander Lubotzky, Ralph Phillips, and Peter Sarnak. Ramanujan graphs. *Combinatorica*, 1988.

32 Grigorii Aleksandrovich Margulis. Explicit constructions of concentrators. *Problemy Peredachi Informatsii*, 1973.

33 Raphael A Meyer, Cameron Musco, Christopher Musco, and David P Woodruff. Hutch++: Optimal stochastic trace estimation. In *Symposium on Simplicity in Algorithms (SOSA)*. SIAM, 2021.

34 Moshe Morgenstern. Existence and explicit constructions of $q+1$ regular Ramanujan graphs for every prime power q. *Journal of Combinatorial Theory, Series B*, 1994.

35 Cameron Musco and Christopher Musco. Randomized block Krylov methods for stronger and faster approximate singular value decomposition. *Advances in Neural Information Processing Systems 28 (NeurIPS)*, 2015.

36 Cameron Musco and Christopher Musco. Recursive sampling for the Nyström method. *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

37 Cameron Musco and David P. Woodruff. Sublinear time low-rank approximation of positive semidefinite matrices. In *Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2017.

38 Deanna Needell, William Swartworth, and David P. Woodruff. Testing positive semidefiniteness using linear measurements. *Proceedings of the 63rd Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2022.

39 Tim Roughgarden. Communication complexity (for algorithm designers). *Foundations and Trends in Theoretical Computer Science*, 2016.

40 Yousef Saad. *Numerical methods for large eigenvalue problems: Revised edition*. SIAM, 2011.

41 Florian Schäfer, Matthias Katzfuss, and Houman Owhadi. Sparse Cholesky factorization by Kullback–Leibler minimization. *SIAM Journal on Scientific Computing*, 2021.

42 Daniel A Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing (STOC)*, 2008.

43 Daniel A. Spielman and Shang-Hua Teng. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing (STOC)*, 2004.

44 Vaidehi Srinivas, David P Woodruff, Ziyu Xu, and Samson Zhou. Memory bounds for the experts problem. *Proceedings of the 54th Annual ACM Symposium on Theory of Computing (STOC)*, 2022.

45 JA Tropp, A Yurtsever, M Udell, and V Cevher. Randomized single-view algorithms for low-rank matrix approximation. acm report 2017-01, caltech, pasadena, jan. 2017, 2017.

46 Roman Vershynin. *High-dimensional probability: An introduction with applications in data science.* Cambridge University Press, 2018.

47 Shusen Wang, Luo Luo, and Zhihua Zhang. SPSD matrix approximation vis column selection: Theories, algorithms, and extensions. *The Journal of Machine Learning Research*, 2016.

48 Alexander Weisse, Gerhard Wellein, Andreas Alvermann, and Holger Fehske. The kernel polynomial method. *Reviews of modern physics*, 2006.

49 Hermann Weyl. The asymptotic distribution law of the eigenvalues of linear partial differential equations (with an application to the theory of cavity radiation). *Mathematical Annals*, 1912.

50 Christopher Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. *Advances in Neural Information Processing Systems 13 (NeurIPS)*, 2000.

51 David Woodruff and William Swartworth. Optimal eigenvalue approximation via sketching. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing (STOC)*, 2023.

52 Jianlin Xia and Ming Gu. Robust approximate Cholesky factorization of rank-structured symmetric positive definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 2010.