A Note on Element-wise Matrix Sparsification via a Matrix-valued Bernstein Inequality

Petros Drineas* Anastasios Zouzias†

June 4, 2018

Abstract

Given a matrix $A \in \mathbb{R}^{n \times n}$, we present a simple, element-wise sparsification algorithm that zeroes out all sufficiently small elements of A and then retains some of the remaining elements with probabilities proportional to the square of their magnitudes. We analyze the approximation accuracy of the proposed algorithm using a recent, elegant non-commutative Bernstein inequality, and compare our bounds with all existing (to the best of our knowledge) element-wise matrix sparsification algorithms.

1 Introduction

Element-wise matrix sparsification was pioneered by Achlioptas and McSherry [AM01, AM07], who described sampling-based algorithms to select a small number of elements from an input matrix $A \in \mathbb{R}^{n \times n}$ in order to construct a sparse sketch $\widetilde{A} \in \mathbb{R}^{n \times n}$, which is close to A in the operator norm. Such sketches were used in approximate eigenvector computations [AM01, AHK06, AM07], semi-definite programming solvers [AHK05, d'A09], and matrix completion problems [CR09, CT10]. Motivated by their work, we present a simple matrix sparsification algorithm that achieves the best known upper bounds for element-wise matrix sparsification.

Our main algorithm (Algorithm 1) zeroes out "small" elements of A and randomly samples the remaining elements of A with respect to a probability distribution that favors "larger" entries. In Algorithm 1, we let $e_1, e_2, \ldots, e_n \in \mathbb{R}^n$ denote the standard basis vectors for $\mathbb{R}^n$ (see Section 3.1 for more notation). Our sampling procedure selects s entries from A (note that $\widehat{A}$ from the description of Algorithm 1 is simply A, but with elements less than or equal to $\epsilon/(2n)$ zeroed out) in s independent, identically distributed (i.i.d.) trials with replacement. In each trial, elements of A are retained with probability proportional to their squared magnitude. Note that the same element of A could be selected multiple times and that $\widetilde{A}$ contains at most s non-zero entries. Theorem 1 is our main quality-of-approximation result for Algorithm 1 and achieves sparsity bounds proportional to $\|A\|_F^2$.

*Department of Computer Science, Rensselaer Polytechnic Institute, drinep@cs.rpi.edu
†Department of Computer Science, University of Toronto, zouzias@cs.toronto.edu

1: **Input:** $A \in \mathbb{R}^{n \times n}$, accuracy parameter $\epsilon > 0$.
2: **Let** $\widehat{A} = A$ and **zero-out** all entries of $\widehat{A}$ that are smaller (in absolute value) than $\epsilon/2n$.
3: **Set** s as in Eqn. (1).
4: **For** $t = 1 \ldots s$ (i.i.d. trials with replacement) **randomly sample** indices (i_t, j_t) (entries of $\widehat{A}$), with

$$\mathbb{P}\left((i_t, j_t) = (i, j)\right) \; = \; p_{ij}, \quad \text{where } p_{ij} := \widehat{A}_{ij}^2 / \left\|\widehat{A}\right\|_{\mathrm{F}}^2 \text{ for all } (i,j) \in [n] \times [n].$$

5: **Output:**

$$\widetilde{A} = \frac{1}{s} \sum_{t=1}^{s} \frac{\widehat{A}_{i_t j_t}}{p_{i_t j_t}} e_{i_t} e_{j_t}^T \in \mathbb{R}^{n \times n}.$$

Algorithm 1: Matrix Sparsification Algorithm

Theorem 1 *Let $A \in \mathbb{R}^{n \times n}$ be any matrix, let $\epsilon > 0$ be an accuracy parameter, and let $\widetilde{A}$ be the sparse sketch of A constructed via Algorithm 1. If*

$$s = \frac{28 n \ln\left(\sqrt{2} n\right)}{\epsilon^2} \|A\|_{\mathrm{F}}^2 , \tag{1}$$

then, with probability at least $1 - n^{-1}$,

$$\left\|A - \widetilde{A}\right\|_2 \leq \epsilon.$$

$\widetilde{A}$ has at most s non-zero entries and the construction of $\widetilde{A}$ can be implemented in one pass over the input matrix A (see Section 3.2).

We conclude this section with Corollary 1, which is a re-statement of Theorem 1 involving the *stable rank* of A, denoted by $\mathbf{sr}(A)$ (recall that the stable rank of any matrix A is defined as the ratio $\mathbf{sr}(A) := \|A\|_{\mathrm{F}}^2 / \|A\|_2^2$, which is upper bounded by the rank of A). The corollary guarantees relative error approximations for matrices of – say – constant stable rank, such as the ones that arise in [Rec09, CT10].

Corollary 1 *Let $A \in \mathbb{R}^{n \times n}$ be any matrix and let $\varepsilon > 0$ be an accuracy parameter. Let $\widetilde{A}$ be the sparse sketch of A constructed via Algorithm 1 (with $\epsilon = \varepsilon \|A\|_2$). If $s = 28 n \mathbf{sr}(A) \ln\left(\sqrt{2} n\right) / \varepsilon^2$, then, with probability at least $1 - n^{-1}$,*

$$\left\|A - \widetilde{A}\right\|_2 \leq \varepsilon \|A\|_2 .$$

It is worth noting that the sampling algorithm implied by Corollary 1 can not be implemented in one pass, since we would need a priori knowledge of the spectral norm of A in order to implement Step 2 of Algorithm 1.

2 Related Work

In this section (as well as in Table 1), we present a head-to-head comparison of our result with all existing (to the best of our knowledge) bounds on matrix sparsification. In [AM01, AM07] the authors presented a sampling method that requires in *expectation* $16n\|A\|_F^2/\epsilon^2 + 8^4 n\log^4 n$ non-zero entries in $\widetilde{A}$ in order to achieve an accuracy guarantee ϵ with a failure probability of at most $e^{-19\log^4 n}$. Compared with our result, their bound holds only when $\epsilon > 4\sqrt{n}\cdot\max_{i,j}|A_{ij}|$ and, in this range, our bounds are superior when $\|A\|_F^2/(\max_{i,j}|A_{ij}|)^2 = o(n\log^3 n)$. It is worth mentioning that the constant involved in [AM01, AM07] is two orders of magnitude larger than ours and, more importantly, that the results of [AM01, AM07] hold only when $n \geq 700\cdot 10^6$.

In [GT09], the authors study the $\|\cdot\|_{\infty\to 2}$ and $\|\cdot\|_{\infty\to 1}$ norms in the matrix sparsification context and they also present a sampling scheme analogous to ours. They achieve (in expectation) a sparsity bound of $Rn\|A\|_F^2\max_{i,j}|A_{ij}|/\epsilon^2$ when $\epsilon \geq \sqrt{nR}\max_{i,j}|A_{ij}|$; here $R = \max_{ij}|A_{ij}|/\min_{A_{ij}\neq 0}|A_{ij}|$. Thus, our results are superior (in the above range of ϵ) when $R\cdot\max_{i,j}|A_{ij}| = \omega(\log n)$.

It is harder to compare our method to the work of [AHK06], which depends on the $\sum_{i,j=1}^n |A_{ij}|$. The latter quantity is, in general, upper bounded only by $n\|A\|_F$, in which case the sampling complexity of [AHK06] is much worse, namely $O(n^{3/2}\|A\|_F^2/\epsilon)$. Finally, the recent bounds on matrix sparsification via the non-commutative Khintchine's inequality in [NDT09] are inferior compared to ours in terms of sparsity guarantees by at least $O(\ln^2(n/\ln^2 n))$. However, we should mention that the bounds of [NDT09] can be extended to multi-dimensional matrices (tensors), whereas our result does not generalize to this setting; see [NDT10] for details.

Comparison with Prior Results

Sparsity of $\widetilde{A}$		**Failure Probability**	**Citation**	**Comments**
$16n\|A\|_F^2/\epsilon^2 + 8^4 n\log^4 n$	Expected	$e^{-19\log^4 n}$	[AM07]	$\epsilon > 4\sqrt{n}\cdot b$ $n \geq 700\cdot 10^6$
$R\cdot b\cdot n\|A\|_F^2/\epsilon^2$	Expected	$e^{-\Omega(R\cdot n)}$	[GT09]	$\epsilon > c_1\sqrt{n\cdot R}\cdot b,\ n\geq 1$
$c_2 n\log^2(\frac{n}{\log^2 n})\log n\|A\|_F^2/\epsilon^2$	Expected	$1/n$	[NDT09]	$\epsilon > 0,\ n\geq 300,$ $c_2 \leq 45^2$
$c_3 n\log^3 n\|A\|_F^2/\epsilon^2$	Expected	$1/n$	[NDT10]	$\epsilon > 0,\ n\geq 300$ Extends to tensors
$c_4\sqrt{n}\sum_{ij}|A_{ij}|/\epsilon$	Exact	$e^{-\Omega(n)}$	[AHK06]	$\epsilon > 0,\ n\geq 1$
$28n\ln\left(\sqrt{2}n\right)\|A\|_F^2/\epsilon^2$	Exact	$1/n$	Theorem 1	$\epsilon > 0,\ n\geq 1$

Table 1: Summary of prior work in matrix sparsification results. Given a matrix $A\in\mathbb{R}^{n\times n}$ and an accuracy parameter $\epsilon > 0$, we seek a sparse $\widetilde{A}\in\mathbb{R}^{n\times n}$ such that $\left\|A-\widetilde{A}\right\|_2 \leq \epsilon$. The first column indicates the number of non-zero entries in $\widetilde{A}$, whereas the second column indicates whether this number is exact or simply holds in expectation. In terms of notation, we let b denote the $\max_{i,j}|A_{ij}|$ and R denote the $\max_{ij}|A_{ij}|/\min_{A_{ij}\neq 0}|A_{ij}|$. Finally, c_1, c_2, c_3, c_4 denote unspecified constants.

3 Background

3.1 Notation

We let $[n]$ denote the set $\{1, 2, \ldots, n\}$. We will use the notation $\mathbb{P}(\cdot)$ to denote the probability of the event in the parentheses and $\mathbb{E}(X)$ to denote the expectation of a random variable X. When X is a matrix, $\mathbb{E}(X)$ denotes the element-wise expectation of each entry of X. For a matrix $X \in \mathbb{R}^{n\times n}$, $X^{(j)}$ will denote the j-th column of X as a column vector and, similarly, $X_{(i)}$ will denote the i-th row of X as a row vector (for any i or j in $[n]$). The Frobenius norm $\|X\|_{\mathrm{F}}$ of the matrix X is defined as $\|X\|_{\mathrm{F}}^2 = \sum_{i,j=1}^{n} X_{ij}^2$, and the spectral norm $\|X\|_2$ of the matrix X is defined as $\|X\|_2 = \max_{\|y\|_2=1} \|Xy\|_2$. For two symmetric matrices X, Y we say that $Y \succeq X$ if and only if $Y - X$ is a positive semi-definite matrix. Finally, $\mathbf{I}_n$ denotes the identity matrix of size n and $\ln x$ denotes the natural logarithm of x.

3.2 Implementing the Sampling in one Pass over the Input Matrix

We now discuss the implementation of Algorithm 1 in one pass over the input matrix A. Towards that end, we will leverage (a slightly modified version of) Algorithm SELECT (p. 137 of [DKM06]). We note that Step 3 essentially operates on $\widehat{A}$. Clearly, in a single pass over the data we can run

1: **Input:** A_{ij} for all $(i, j) \in [n] \times [n]$, arbitrarily ordered and $\epsilon > 0$.
2: $N = 0$.
3: **For all** $(i, j) \in [n] \times [n]$ **such that** $A_{ij}^2 > \frac{\epsilon^2}{4n^2}$

- $N = N + A_{ij}^2$.
- **Set** $(I, J) = (i, j)$ and $S = A_{ij}$ **with probability** $\frac{A_{ij}^2}{N}$.

4: **Output:** Return (I, J), S and N.

Algorithm 2: One-pass SELECT algorithm

in parallel s copies of the SELECT Algorithm (using a total of $O(s)$ memory) to effectively return s independent samples from $\widehat{A}$. Lemma 1 (page 136 of [DKM06], note that the sequence of the A_{ij}^2's is all-positive) guarantees that each of the s copies of SELECT returns a sample satisfying:

$$\mathbb{P}((i_t, j_t) = (i, j)) = \frac{\widehat{A}_{ij}^2}{\sum_{i,j=1}^{n} \widehat{A}_{ij}^2} = \frac{\widehat{A}_{ij}^2}{\left\|\widehat{A}\right\|_{\mathrm{F}}^2}, \quad \text{for all } t = 1, \ldots, s.$$

Finally, in the parlance of Step 5 of Algorithm 1, (i_t, j_t) is set to (I, J) and $p_{i_t j_t}$ is set to S^2/N for all $t \in [s]$.

4 Proof of Theorem 1

The proof of Theorem 1 will combine Lemmas 1 and 4 in order to bound $\left\|A - \widetilde{A}\right\|_2$ as follows:

$$\left\|A - \widetilde{A}\right\|_2 = \left\|A - \widehat{A} + \widehat{A} - \widetilde{A}\right\|_2 \leq \left\|A - \widehat{A}\right\|_2 + \left\|\widehat{A} - \widetilde{A}\right\|_2 \leq \epsilon/2 + \epsilon/2 = \epsilon.$$

The failure probability of Theorem 1 emerges from Lemma 4, which fails with probability at most n^{-1} for the choice of s in Eqn. (1). The proof of Lemma 4 will involve an elegant matrix-valued Bernstein bound proven in [Rec09]. See also [Gro09] or [Tro10, Theorem 2.10] for similar bounds.

4.1 Bounding $\left\|A - \widehat{A}\right\|_2$

Lemma 1 *Using the notation of Algorithm 1,* $\left\|A - \widehat{A}\right\|_2 \leq \epsilon/2$.

Proof: Recall that the entries of $\widehat{A}$ are either equal to the corresponding entries of A or they are set to zero if the corresponding entry of A is (in absolute value) smaller than $\epsilon/(2n)$. Thus,

$$\left\|A - \widehat{A}\right\|_2^2 \leq \left\|A - \widehat{A}\right\|_{\mathrm{F}}^2 = \sum_{i,j=1}^{n} \left(A - \widehat{A}\right)_{ij}^2 \leq \sum_{i,j=1}^{n} \frac{\epsilon^2}{4n^2} \leq \frac{\epsilon^2}{4}.$$

$\diamond$

4.2 Bounding $\left\|\widehat{A} - \widetilde{A}\right\|_2$

In order to prove our main result in this section (Lemma 4) we will leverage a powerful matrix-valued Bernstein bound originally proven in [Rec09] (Theorem 3.2). We restate this theorem, slightly rephrased to better suit our notation.

Theorem 2 [THEOREM 3.2 OF [REC09]] *Let $M_1, M_2, \ldots, M_s$ be independent, zero-mean random matrices in $\mathbb{R}^{n\times n}$. Suppose* $\max_{t\in[s]} \left\{\left\|\mathbb{E}\left(M_t M_t^T\right)\right\|_2, \left\|\mathbb{E}\left(M_t^T M_t\right)\right\|_2\right\} \leq \rho^2$ *and* $\|M_t\|_2 \leq \gamma$ *for all* $t \in [s]$. *Then, for any* $\tau > 0$,

$$\left\|\frac{1}{s}\sum_{t=1}^{s} M_t\right\|_2 \leq \tau$$

holds, subject to a failure probability of at most

$$2n \exp\left(-\frac{s\tau^2/2}{\rho^2 + \gamma\tau/3}\right).$$

In order to apply the above theorem, using the notation of Algorithm 1, we set $M_t = \frac{\widehat{A}_{i_t j_t}}{p_{i_t j_t}} e_{i_t} e_{j_t}^T - \widehat{A}$ for all $t \in [s]$ to obtain

$$\frac{1}{s}\sum_{t=1}^{s} M_t = \frac{1}{s}\sum_{t=1}^{s} \left[\frac{\widehat{A}_{i_t j_t}}{p_{i_t j_t}} e_{i_t} e_{j_t}^T - \widehat{A}\right] = \widetilde{A} - \widehat{A}. \tag{2}$$

Let $\mathbf{0}_n$ denote the all-zeros matrix of size n. It is easy to argue that $\mathbb{E}\left(M_t\right) = \mathbf{0}_n$ for all $t \in [s]$. Indeed, if we consider that $\sum_{i,j=1}^{n} p_{ij} = 1$ and $\widehat{A} = \sum_{i,j=1}^{n} \widehat{A}_{ij} e_i e_j^T$ we obtain

$$\mathbb{E}\left(M_t\right) = \sum_{i,j=1}^{n} p_{ij} \left(\frac{\widehat{A}_{ij}}{p_{ij}} e_i e_j^T - \widehat{A}\right) = \sum_{i,j=1}^{n} \widehat{A}_{ij} e_i e_j^T - \sum_{i,j=1}^{n} p_{ij}\widehat{A} = \mathbf{0}_n.$$

Our next lemma bounds $\|M_t\|_2$ for all $t \in [s]$.

Lemma 2 *Using our notation,* $\|M_t\|_2 \leq 4n\epsilon^{-1}\left\|\widehat{A}\right\|_{\mathrm{F}}^2$ *for all* $t \in [s]$.

Proof: First, using the definition of M_t and the fact that $p_{i_tj_t} = \widehat{A}^2_{i_tj_t}/\left\|\widehat{A}\right\|_{\mathrm{F}}^2$,

$$\|M_t\|_2 = \left\|\frac{\widehat{A}_{i_tj_t}}{p_{i_tj_t}}e_{i_t}e_{j_t}^T - \widehat{A}\right\|_2 \leq \frac{\left\|\widehat{A}\right\|_{\mathrm{F}}^2}{\left|\widehat{A}_{i_tj_t}\right|} + \left\|\widehat{A}\right\|_2 \leq \frac{2n\left\|\widehat{A}\right\|_{\mathrm{F}}^2}{\epsilon} + \left\|\widehat{A}\right\|_{\mathrm{F}}.$$

The last inequality follows since all entries of $\widehat{A}$ are at least $\epsilon/(2n)$ and the fact that $\left\|\widehat{A}\right\|_2 \leq \left\|\widehat{A}\right\|_{\mathrm{F}}$. We can now assume that

$$\left\|\widehat{A}\right\|_{\mathrm{F}} \leq \frac{2n\left\|\widehat{A}\right\|_{\mathrm{F}}^2}{\epsilon} \tag{3}$$

to conclude the proof of the lemma. To justify our assumption in Eqn. (3), we note that if it is violated, then it must be the case that $\left\|\widehat{A}\right\|_{\mathrm{F}} < \epsilon/(2n)$. If that were true, then all entries of $\widehat{A}$ would be equal to zero. (Recall that all entries of $\widehat{A}$ are either zero or, in absolute value, larger than $\epsilon/(2n)$.) Also, if $\widehat{A}$ were identically zero, then *(i)* $\widetilde{A}$ would also be identically zero and, *(ii)* all entries of A would be at most $\epsilon/(2n)$. Thus,

$$\left\|A - \widetilde{A}\right\|_2 = \|A\|_2 \leq \|A\|_{\mathrm{F}} \leq \sqrt{n^2\frac{\epsilon^2}{4n^2}} = \frac{\epsilon}{2}.$$

Thus, if the assumption of Eqn. (3) is not satisfied, the resulting all-zeros $\widetilde{A}$ still satisfies Theorem 1. $\diamond$

Our next step towards applying Theorem 2 involves bounding the spectral norm of the expectation of $M_tM_t^T$. The spectral norm of the expectation of $M_t^TM_t$ admits a similar analysis and the same bound and is omitted.

Lemma 3 *Using our notation,* $\left\|\mathbb{E}\left(M_tM_t^T\right)\right\|_2 \leq n\left\|\widehat{A}\right\|_{\mathrm{F}}^2$ *for any* $t \in [s]$.

Proof: We start by evaluating $\mathbb{E}\left(M_tM_t^T\right)$; recall that $p_{ij} = \widehat{A}^2_{ij}/\left\|\widehat{A}\right\|_{\mathrm{F}}^2$:

$$\begin{aligned}
\mathbb{E}\left(M_tM_t^T\right) &= \mathbb{E}\left(\left(\frac{\widehat{A}_{i_tj_t}}{p_{i_tj_t}}e_{i_t}e_{j_t}^T - \widehat{A}\right)\left(\frac{\widehat{A}_{i_tj_t}}{p_{i_tj_t}}e_{j_t}e_{i_t}^T - \widehat{A}^T\right)\right) \\
&= \sum_{i,j=1}^n p_{ij}\left(\frac{\widehat{A}_{ij}}{p_{ij}}e_ie_j^T - \widehat{A}\right)\left(\frac{\widehat{A}_{ij}}{p_{ij}}e_je_i^T - \widehat{A}^T\right) \\
&= \sum_{i,j=1}^n \left(\frac{\widehat{A}_{ij}^2}{p_{ij}}e_ie_i^T - \widehat{A}_{ij}\widehat{A}e_je_i^T - \widehat{A}_{ij}e_ie_j^T\widehat{A}^T + p_{ij}\widehat{A}\widehat{A}^T\right) \\
&= \left\|\widehat{A}\right\|_{\mathrm{F}}^2\sum_{i=1}^n m_i\cdot e_ie_i^T - \sum_{j=1}^n \widehat{A}e_j\sum_{i=1}^n \widehat{A}_{ij}e_i^T - \sum_{j=1}^n\left(\sum_{i=1}^n \widehat{A}_{ij}e_i\right)\left(\widehat{A}e_j\right)^T + \sum_{i,j=1}^n p_{ij}\widehat{A}\widehat{A}^T,
\end{aligned}$$

where m_i is the number of non-zeroes of the i-th row of $\widehat{A}$. We now simplify the above result using a few simple observations: $\sum_{i,j=1}^{n} p_{ij} = 1$, $\widehat{A}e_j = \widehat{A}^{(j)}$, $\sum_{i=1}^{n} \widehat{A}_{ij} e_i = \widehat{A}^{(j)}$, and $\sum_{j=1}^{n} \widehat{A}^{(j)} \left(\widehat{A}^{(j)}\right)^T = \widehat{A}\widehat{A}^T$. Thus, we get

$$\begin{aligned}
\mathbb{E}\left(M_t M_t^T\right) &= \left\|\widehat{A}\right\|_{\mathrm{F}}^2 \sum_{i=1}^{n} m_i \cdot e_i e_i^T - \sum_{j=1}^{n} \widehat{A}^{(j)} \left(\widehat{A}^{(j)}\right)^T - \sum_{j=1}^{n} \widehat{A}^{(j)} \left(\widehat{A}^{(j)}\right)^T + \widehat{A}\widehat{A}^T \\
&= \left\|\widehat{A}\right\|_{\mathrm{F}}^2 \sum_{i=1}^{n} m_i \cdot e_i e_i^T - \widehat{A}\widehat{A}^T.
\end{aligned}$$

Since $0 \leq m_i \leq n$ and using Weyl's inequality (Theorem 4.3.1 of [HJ90]), which states that by adding a positive semi-definite matrix to a symmetric matrix all its eigenvalues will increase, we get that

$$-\widehat{A}\widehat{A}^T \preceq \mathbb{E}\left(M_t M_t^T\right) \preceq n \left\|\widehat{A}\right\|_{\mathrm{F}}^2 \mathbf{I}_n.$$

Consequently $\left\|\mathbb{E}\left(M_t M_t^T\right)\right\|_2 = \max\left\{\left\|\widehat{A}\right\|_2^2, n\left\|\widehat{A}\right\|_{\mathrm{F}}^2\right\} = n\left\|\widehat{A}\right\|_{\mathrm{F}}^2$.

$\diamond$

We can now apply Theorem 2 on Eqn. (2) with $\tau = \epsilon/2$, $\gamma = 4n\epsilon^{-1}\left\|\widehat{A}\right\|_{\mathrm{F}}^2$ (Lemma 2), and $\rho^2 = n\left\|\widehat{A}\right\|_{\mathrm{F}}^2$ (Lemma 3) . Thus, we get that $\left\|\widehat{A} - \widetilde{A}\right\|_2 \leq \epsilon/2$ holds, subject to a failure probability of at most

$$2n \exp\left(-\frac{\epsilon^2 s/8}{(1 + 4/6)\, n \left\|\widehat{A}\right\|_{\mathrm{F}}^2}\right).$$

Bounding the failure probability by δ and solving for s, we get that

$$s \geq \frac{14}{\epsilon^2} n \left\|\widehat{A}\right\|_{\mathrm{F}}^2 \ln\left(\frac{2n}{\delta}\right).$$

Using $\left\|\widehat{A}\right\|_{\mathrm{F}} \leq \|A\|_{\mathrm{F}}$ (by construction) concludes the proof of the following lemma, which is the main result of this section.

Lemma 4 *Using the notation of Algorithm 1, if $s \geq 14n\epsilon^{-2}\|A\|_{\mathrm{F}}^2 \ln(2n/\delta)$, then, with probability at least $1 - \delta$,*

$$\left\|\widehat{A} - \widetilde{A}\right\|_2 \leq \epsilon/2.$$

5 Acknowledgments

We would like to thank the anonymous reviewers for numerous comments that significantly improved the presentation of our work. This research has been supported by the National Science Foundation through NSF CCF 1016501, NSF DMS 1008983, and NSF CCF 545538 awards to Petros Drineas.

References

[AHK05] S. Arora, E. Hazan, and S. Kale. Fast Algorithms for Approximate Semidefinite Programming using the Multiplicative Weights Update Method. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 339–348, 2005. (Cited on page 1)

[AHK06] S. Arora, E. Hazan, and S. Kale. A Fast Random Sampling Algorithm for Sparsifying Matrices. In *Proceedings of the International Workshop on Randomization and Approximation Techniques (RANDOM)*, pages 272–279, 2006. (Cited on pages 1 and 3)

[AM01] D. Achlioptas and F. McSherry. Fast Computation of Low Rank Matrix Approximations. In *Proceedings of the Symposium on Theory of Computing (STOC)*, pages 611–618, 2001. (Cited on pages 1, 2 and 3)

[AM07] D. Achlioptas and F. McSherry. Fast Computation of Low Rank Matrix Approximations. *Journal of the ACM*, 54(2), 2007. (Cited on pages 1, 2 and 3)

[CR09] E. J. Candès and B. Recht. Exact Matrix Completion via Convex Optimization. *Foundations of Computational Mathematics*, 9(3):717–772, 2009. (Cited on page 1)

[CT10] E. J. Candès and T. Tao. The Power of Convex Relaxation: Near-optimal Matrix Completion. *IEEE Trans. Inf. Theor.*, 56(5):2053–2080, 2010. (Cited on pages 1 and 2)

[d'A09] A. d'Aspremont. Subsampling Algorithms for Semidefinite Programming. Available at arxiv:0803.1990v5, November 2009. (Cited on page 1)

[DKM06] P. Drineas, R. Kannan, and M. W. Mahoney. Fast Monte Carlo Algorithms for Matrices I: Approximating Matrix Multiplication. *SIAM J. Comput.*, 36(1):132–157, 2006. (Cited on page 4)

[Gro09] D. Gross. Recovering Low-rank Matrices from Few Coefficients in any Basis. Available at arxiv:0910.1879, December 2009. (Cited on page 5)

[GT09] A. Gittens and J. A. Tropp. Error Bounds for Random Matrix Approximation Schemes. Available at arxiv:0911.4108, November 2009. (Cited on page 3)

[HJ90] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990. (Cited on page 7)

[NDT09] N. H. Nguyen, P. Drineas, and T. D. Tran. Matrix Sparsification via the Khintchine Inequality. Manuscript, 2009. (Cited on page 3)

[NDT10] N. H. Nguyen, P. Drineas, and T. D. Tran. Tensor Sparsification via a Bound on the Spectral Norm of Random Tensors. Available at arxiv:1005.4732, May 2010. (Cited on page 3)

[Rec09] B. Recht. A Simpler Approach to Matrix Completion. Available at arxiv:0910.0651, October 2009. (Cited on pages 2 and 5)

[Tro10] J. A. Tropp. User-Friendly Tail Bounds for Sums of Random Matrices. Available at arxiv:1004.4389, April 2010. (Cited on page 5)