A Note on Randomized Element-wise Matrix Sparsification

Abhisek Kundu * Petros Drineas †

Abstract

Given a matrix $A \in \mathbb{R}^{m \times n}$, we present a randomized algorithm that sparsifies $A$ by retaining some of its elements by sampling them according to a distribution that depends on both the square and the absolute value of the entries. We combine the ideas of [4, 1] and provide an elementary proof of the approximation accuracy of our algorithm following [4] without the truncation step.

1 Introduction

Element-wise matrix sparsification was pioneered in [2, 3] and was later improved in [4, 1]. More specifically, the original work of [2, 3] sampled entries from a matrix with probabilities depending on the square of an entry for “large” entries and on the absolute value of an entry for “small” entries. [4] proposed to zero out the small entries and then used sampling with respect to the squares of the remaining entries in order to sparsify the matrix; an elegant proof was possible via a matrix-Bernstein inequality. Very recently, [1] argued that the zeroing out step could be avoided by sampling with respect to the absolute values of the matrix entries. Theorem 1 combines the ideas of [4, 1] to provide an elementary proof that bypasses the zeroing out step. More specifically, we avoid zeroing out the small elements of the input matrix by constructing a sampling probability distribution that depends on both the absolute values as well as the squares of the entries of the input matrix.

2 Our Result

We present our main algorithm (Algorithm 1) and the related Theorem 1, which is our main quality-of-approximation result for Algorithm 1.

2.1 Notation

We use bold capital letters (e.g., $X$) to denote matrices and bold lowercase letters (e.g., $x$) to denote column vectors. Let $[n]$ denote the set $\{1, 2, ..., n\}$. We use $E(X)$ to denote the expectation of a random variable $X$; when $X$ is a random matrix, $E(X)$ denotes the element-wise expectation of each entry of $X$. For a matrix $X \in \mathbb{R}^{m \times n}$, the Frobenius norm $\|X\|_F$ is defined as $\|X\|_F^2 = \sum_{i,j=1}^{m,n} X_{ij}^2$, and the spectral norm $\|X\|_2$ is defined as $\|X\|_2 = \max\|y\|_2 \leq 1 \|Xy\|_2$. For symmetric matrices $A, B$ we say that $B \succeq A$ if and only if $B - A$ is a positive semi-definite matrix. $I_n$ denotes the $n \times n$ identity matrix and $\ln x$ denotes the natural logarithm of $x$. Finally, we use $e_i$ to denote standard basis vectors whose dimensionalities will be clear from the context.

*Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY, kundua2@rpi.edu.
†Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY, drinep@rpi.edu.
2.2 Algorithm

Our main algorithm (Algorithm 1) randomly samples (in independent, identically distributed trials) \( s \) elements of a given matrix \( X \) according to a probability distribution \( \{p_{ij}\}_{i,j=1}^{m,n} \) over the elements of \( X \).

**Algorithm 1** Element-wise Matrix Sparsification Algorithm

1. **Input:** \( X \in \mathbb{R}^{m \times n} \), \( \{p_{ij}\}_{i,j=1}^{m,n} \) such that \( p_{ij} \geq 0 \) (for all \( i, j \)) and \( \sum_{i,j=1}^{m,n} p_{ij} = 1 \), integer \( s > 0 \).
2. **For** \( t = 1 \ldots s \) (i.i.d. trials with replacement) **randomly sample** pairs of indices \((i_t, j_t) \in \{1 \ldots m\} \times \{1 \ldots n\}\) with \( \mathbb{P}[ (i_t, j_t) = (i, j)] = p_{ij} \).
3. **Output:** set of sampled pairs of indices \( \Omega = \{(i_t, j_t), t = 1 \ldots s\} \).
4. **Sampling operator:** \( S_\Omega : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{m \times n} \) with \( S_\Omega(X) = \frac{1}{s} \sum_{t=1}^{s} X_{i_t j_t} e_i e_j^T \).

**Theorem 1** Let \( X \in \mathbb{R}^{m \times n} \) and let \( \epsilon > 0 \) be an accuracy parameter. Let \( S_\Omega : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{m \times n} \) be the sampling operator of the element-wise sampling algorithm (Algorithm 1) and assume that the sampling probabilities \( \{p_{ij}\}_{i,j=1}^{m,n} \) satisfy

\[
p_{ij} \geq \frac{\beta}{2} \left( \frac{X_{ij}^2}{\|X\|_F^2} + \frac{|X_{ij}|}{\sum_{i,j=1}^{m,n} |X_{ij}|} \right)
\]

for all \( i, j \) and some \( \beta \in (0, 1] \). Then, with probability at least \( 1 - \delta \),

\[
\|S_\Omega(X) - X\|_2 \leq \epsilon,
\]

if either (i) \( \epsilon \leq \|X\|_F \) and \( s \geq \frac{6 \max\{m,n\} \ln((m+n)/\delta)}{\beta \epsilon^2} \frac{\|X\|_F^2}{\|X\|_F} \),

or (ii) \( \epsilon > \|X\|_F \) and \( s \geq \frac{6 \max\{m,n\} \ln((m+n)/\delta)}{\beta \epsilon} \|X\|_F \).

We now restate the above bound in terms of the stable rank of the input matrix. Recall that the stable rank is defined as \( \text{sr}(X) := \frac{\|X\|_F^2}{\|X\|_2^2} \) and is upper bounded by the rank of \( X \).

**Corollary 1** Let \( X \in \mathbb{R}^{m \times n} \), let \( \epsilon > 0 \) be an accuracy parameter such that \( \text{sr}(X) \geq \epsilon^2 \), and let \( S_\Omega(X) \) be the sparse sketch of \( X \) constructed via Algorithm 1 with the \( p_{ij} \)'s satisfying the bounds of eqn. (1). If

\[
s \geq \frac{6 \max\{m,n\} \ln((m+n)/\delta)}{\beta \epsilon^2} \text{sr}(X),
\]

then, with probability at least \( 1 - \delta \),

\[
\|X - S_\Omega(X)\|_2 \leq \epsilon \|X\|_2.
\]

3 Proof of Theorem

In this section we provide a proof of Theorem following the lines of [4]. First, we rephrase the non-commutative matrix-valued Bernstein bound theorem of [5] using our notation.
Lemma 2. Using our notation, let $M_1, M_2, \ldots, M_s$ be independent, zero-mean random matrices in $\mathbb{R}^{m \times n}$. Suppose $\max_{t \in [s]} \{ \| \mathbb{E}(M_t M_t^T) \|_2, \| \mathbb{E}(M_t^T M_t) \|_2 \} \leq \rho^2$ and $\| M_t \|_2 \leq \gamma$ for all $t \in [s]$. Then, for any $\epsilon > 0$,

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

holds, subject to a failure probability of at most

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

For all $t \in [s]$ we define the matrix $M_t \in \mathbb{R}^{m \times n}$ as follows:

$$M_t = \frac{X_{ij}}{p_{ij}} e_i e_j^T - X. \quad (2)$$

It now follows that

$$\frac{1}{s} \sum_{t=1}^{s} M_t = \frac{1}{s} \sum_{t=1}^{s} \left[ \frac{X_{ij}}{p_{ij}} e_i e_j^T - X \right] = S_\Omega(X) - X.$$ 

Let $0_{m \times n}$ denote the $m \times n$ all-zeros matrix and note that $X = \sum_{i,j=1}^{m,n} X_{ij} e_i e_j^T$. The following derivation is immediate (for all $t \in [s]$):

$$\mathbb{E}(M_t) = \mathbb{E}(S_\Omega(X)) - X = \sum_{i,j=1}^{m,n} p_{ij} \frac{X_{ij}}{p_{ij}} e_i e_j^T - X = 0_{m \times n}.$$ 

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

Lemma 1. Using our notation, $\| M_t \|_2 \leq \frac{3\sqrt{mn}}{\beta} \| X \|_F$ for all $t \in [s]$.

Proof: Notice that sampling according to the element-wise probabilities of eqn. (1) satisfies

$$p_{ij} \geq \frac{\beta}{2} \frac{|X_{ij}|}{\sum_{i,j=1}^{m,n} |X_{ij}|}.$$ 

We can use the above inequality to get

$$\| M_t \|_2 = \left\| \frac{X_{ij}}{p_{ij}} e_i e_j^T - X \right\|_2 \leq \frac{2}{\beta} \sum_{i=1}^{m} \sum_{j=1}^{n} |X_{ij}| \leq \frac{3\sqrt{mn}}{\beta} \| X \|_F.$$ 

In the above we used $\beta \leq 1$, $\| X \|_2 \leq \| X \|_F$, and (from the Cauchy-Schwarz inequality)

$$\sum_{i,j=1}^{m,n} |X_{ij}| \leq \sqrt{\sum_{i,j=1}^{m,n} X_{ij}^2} = \sqrt{mn} \| X \|_F.$$ 

Thus, we get a new bound for Lemma 2 of [4], bypassing the need for a truncation step.

Next we bound the spectral norm of the expectation of $M_t M_t^T$. The spectral norm of the expectation of $M_t M_t^T$ can be bounded using a similar analysis.

Lemma 2. Using our notation, $\| \mathbb{E}(M_t M_t^T) \|_2 \leq \frac{2n}{\beta} \| X \|_F^2$ for all $t \in [s]$. 


Proof: Recall that $X = \sum_{i,j=1}^{m,n} X_{ij} e_i e_j^T$ and $\sum_{i,j=1}^{m,n} p_{ij} = 1$ to derive

$$
\mathbb{E}[M_t M_t^T] = E \left( \begin{pmatrix} X_{ij} e_i e_j^T - X \end{pmatrix} \begin{pmatrix} X_{ij} e_i e_j^T - X^T \end{pmatrix} \right)
= \sum_{i,j=1}^{m,n} p_{ij} \begin{pmatrix} X_{ij} e_i e_j^T - X \end{pmatrix} \begin{pmatrix} X_{ij} e_i e_j^T - X^T \end{pmatrix}
= \sum_{i,j=1}^{m,n} \left( \frac{X_{ij}^2}{p_{ij}} e_i e_j^T e_i e_j^T \right) - \left( \sum_{i,j=1}^{m,n} X_{ij} e_i e_j^T \right) X^T - X \left( \sum_{i,j=1}^{m,n} X_{ij} e_i e_j^T \right) + \sum_{i,j=1}^{m,n} p_{ij} XX^T
= \sum_{i,j=1}^{m,n} \left( \frac{X_{ij}^2}{p_{ij}} e_i e_j^T \right) - XX^T.
$$

Notice that sampling according to the element-wise sampling probabilities of eqn. (1) satisfies $p_{ij} \geq \frac{\beta}{\|X\|_F}$ and so we get

$$
\mathbb{E}[M_t M_t^T] = \sum_{i,j=1}^{m,n} \left( \frac{X_{ij}^2}{p_{ij}} e_i e_j^T \right) - XX^T \leq \frac{2\|X\|_F^2}{\beta} \sum_{i,j=1}^{m,n} e_i e_j^T - XX^T = \frac{2n\|X\|_F^2}{\beta} I_m - XX^T.
$$

Using Weyl’s inequality we get

$$
\|\mathbb{E}[M_t M_t^T]\|_2 \leq \max \left\{ \|XX^T\|_2, \frac{2n\|X\|_F^2}{\beta} \|I_m\|_2 \right\} = \frac{2n}{\beta} \|X\|_F^2.
$$

We can now apply Theorem 2 with $\rho^2 = \frac{2n}{\beta} \|X\|_F^2$ and $\gamma = 3\sqrt{mn}$ to conclude that $\|S_\Omega(X) - X\|_2 \leq \epsilon$ holds subject to a failure probability at most

$$(m + n) \exp \left( \frac{-s\beta e^2}{4n \|X\|_F^2 + 2\sqrt{mn} \|X\|_F^2} \right).$$

Setting the failure probability equal to $\delta$, we conclude that it suffices to set $s$ as follows:

$$
s \geq \frac{1}{\beta e^2} (4n \|X\|_F^2 + 2\sqrt{mn} \|X\|_F) \ln \left( \frac{m+n}{\delta} \right).
$$

We now consider two cases. First, if $\epsilon \leq \|X\|_F$,

$$
4n \|X\|_F^2 + 2\sqrt{mn} \|X\|_F \leq \max\{m,n\} (4 \|X\|_F^2 + 2\epsilon \|X\|_F) \leq 6 \max\{m,n\} \|X\|_F^2,
$$

which immediately proves the first case of Theorem 1. Similarly, if $\epsilon > \|X\|_F$,

$$
4n \|X\|_F^2 + 2\sqrt{mn} \|X\|_F \leq 6 \max\{m,n\} \|X\|_F
$$

and the second case of Theorem 1 follows.
References

[1] D. Achlioptas, Z. Karnin, and E. Liberty. Matrix entry-wise sampling: Simple is best. In *Neural Information Processing Systems*, 2013.

[2] D. Achlioptas and F. McSherry. Fast computation of low rank matrix approximations. In *Proceedings of Symposium on the Theory of Computing*, pages 611–618, 2001.

[3] D. Achlioptas and F. McSherry. Fast computation of low-rank matrix approximations. *Journal of the ACM*, page 54(2):9, 2007.

[4] P. Drineas and A. Zouzias. A note on element-wise matrix sparsification via a matrix-valued Bernstein inequality. In *Information Processing Letters*, pages 385–389, 111(8), 2011.

[5] B. Recht. A simpler approach to matrix completion. In *The Journal of Machine Learning Research*, pages 3413–3430, 12, 2011.