Input-Sparsity Low Rank Approximation in Schatten Norm

Yi Li[*1] David P. Woodruff[*2]

Abstract

We give the first input-sparsity time algorithms for the rank-k low rank approximation problem in every Schatten norm. Specifically, for a given $m \times n$ ($m \geq n$) matrix A, our algorithm computes $Y \in \mathbb{R}^{m \times k}$, $Z \in \mathbb{R}^{n \times k}$, which, with high probability, satisfy $\|A - YZ^T\|_p \leq (1+\varepsilon)\|A - A_k\|_p$, where $\|M\|_p = \left(\sum_{i=1}^n \sigma_i(M)^p\right)^{1/p}$ is the Schatten p-norm of a matrix M with singular values $\sigma_1(M), \ldots, \sigma_n(M)$, and where A_k is the best rank-k approximation to A. Our algorithm runs in time $\tilde{O}(\mathrm{nnz}(A) + mn^{\alpha_p} \operatorname{poly}(k/\varepsilon))$, where $\alpha_p = 0$ for $p \in [1, 2)$ and $\alpha_p = (\omega - 1)(1 - 2/p)$ for $p > 2$ and $\omega \approx 2.374$ is the exponent of matrix multiplication. For the important case of $p = 1$, which corresponds to the more "robust" nuclear norm, we obtain $\tilde{O}(\mathrm{nnz}(A) + m \cdot \operatorname{poly}(k/\epsilon))$ time, which was previously only known for the Frobenius norm ($p = 2$). Moreover, since $\alpha_p < \omega - 1$ for every p, our algorithm has a better dependence on n than that in the singular value decomposition for every p. Crucial to our analysis is the use of dimensionality reduction for Ky-Fan p-norms.

1. Introduction

A common task in processing or analyzing large-scale datasets is to approximate a large matrix $A \in \mathbb{R}^{m \times n}$ ($m \geq n$) with a low-rank matrix. Often this is done with respect to the Frobenius norm, that is, the objective function is to minimize the error $\|A - X\|_F$ over all rank-k matrices $X \in \mathbb{R}^{m \times n}$ for a rank parameter k. It is well-known that the optimal solution is $A_k = P_L A = AP_R$, where P_L is the orthogonal projection onto the top k left singular vectors of A, and P_R is the orthogonal projection onto the top k right singular vectors of A. Typically this is found via the singular value decomposition (SVD) of A, which is an expensive operation.

For large matrices A this is too slow, so we instead allow for randomized approximation algorithms in the hope of achieving a much faster running time. Formally, given an approximation parameter $\varepsilon > 0$, we would like to find a rank-k matrix X for which $\|A - X\|_F \leq (1 + \varepsilon)\|A - A_k\|_F$ with large probability. For this relaxed problem, a number of efficient methods are known, which are based on dimensionality reduction techniques such as random projections, importance sampling, and other sketching methods, with running times[1,2] $\tilde{O}(\mathrm{nnz}(A) + m \operatorname{poly}(k/\varepsilon))$, where $\mathrm{nnz}(A)$ denotes the number of non-zero entries of A. This is significantly faster than the SVD, which takes $\tilde{\Theta}(mn^{\omega-1})$ time, where ω is the exponent of matrix multiplication. See (Woodruff, 2014) for a survey.

In this work, we consider approximation error with respect to general matrix norms, i.e., to the Schatten p-norm. The Schatten p-norm, denoted by $\|\cdot\|_p$, is defined to be the ℓ_p-norm of the singular values of the matrix. Below is the formal definition of the problem.

Definition 1.1 (Low-rank Approximation)**.** Let $p \geq 1$. Given a matrix $A \in \mathbb{R}^{m \times n}$, find a rank-$k$ matrix $\hat{X} \in \mathbb{R}^{m \times n}$ for which

$$\left\|A - \hat{X}\right\|_p \leq (1+\varepsilon) \min_{X:\operatorname{rank}(X)=k} \|A - X\|_p . \quad (1)$$

It is a well-known fact (Mirsky's Theorem) that the optimal solution for general Schatten norms coincides with the optimal rank-k matrix A_k for the Frobenius norm, given by the SVD. However, approximate solutions for the Frobenius norm loss function may give horrible approximations for other Schatten p-norms.

Of particular importance is the Schatten 1-norm, also called the nuclear norm or the trace norm, which is the sum of the singular values of a matrix. It is typically considered to be more robust than the Frobenius norm (Schatten 2-norm) and has been used in robust PCA applications (see, e.g., (Xu et al., 2010; Candès et al., 2011; Yi et al., 2016)).

[*] Equal contribution [1] School of Physical and Mathematical Sciences, Nanyang Technological University [2] Department of Computer Science, Carnegie Mellon University. Correspondence to: Yi Li <yili@ntu.edu.sg>, David Woodruff <dwoodruf@andrew.cmu.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020. Copyright 2020 by the author(s).

[1] We use the notation $\tilde{O}(f)$ to hide the polylogarithmic factors in $O(f \operatorname{poly}(\log f))$.

[2] Since outputting X takes $O(mn)$ time, these algorithms usually output X in factored form, where each factor has rank k.

For example, suppose the top singular value of an $n \times n$ matrix A is 1, the next $2k$ singular values are $1/\sqrt{k}$, and the remaining singular values are 0. A Frobenius norm rank-k approximation could just choose the top singular direction and pay a cost of $\sqrt{2k \cdot 1/k} = \sqrt{2}$. Since the Frobenius norm of the bottom $n-k$ singular values is $(k+1) \cdot \frac{1}{k}$, this is a $\sqrt{2}$-approximation. On the other hand, if a Schatten 1-norm rank-k approximation algorithm were to only output the top singular direction, it would pay a cost of $2k \cdot 1/\sqrt{k} = 2\sqrt{k}$. The bottom $n-k$ singular values have Schatten 1-norm $(k+1) \cdot \frac{1}{\sqrt{k}}$. Consequently, the approximation factor would be $2(1-o(1))$, and one can show if we insisted on a $\sqrt{2}$-approximation or better, a Schatten 1-norm algorithm would need to capture a constant fraction of the top k directions, and thus capture more of the underlying data than a Frobenius norm solution.

Consider another example where the top k singular values are all 1s and the $(k+i)$-th singular value is $1/i$. When $k = o(\log n)$, capturing only the top singular direction gives a $(1+o(1))$-approximation for the Schatten 1-norm but a $\Theta(\sqrt{k})$-approximation for the Frobenius norm. This example, together with the preceding one, shows that the Schatten norm is a genuinely a different error metric.

Surprisingly, no algorithms for low-rank approximation in the Schatten p-norm were known to run in time $\tilde{O}(\mathrm{nnz}(A) + m \operatorname{poly}(k/\varepsilon))$ prior to this work, except for the special case of $p = 2$. We note that the case of $p = 2$ has special geometric structure that is not shared by other Schatten p-norms. Indeed, a common technique for the $p = 2$ setting is to first find a $\operatorname{poly}(k/\epsilon)$-dimensional subspace V containing a rank-k subspace inside of it which is a $(1+\epsilon)$-approximate subspace to project the rows of A on. Then, by the Pythagorean theorem, one can first project the rows of A onto V, and then find the best rank-k subspace of the projected points inside of V. For other Schatten p-norms, the Pythagorean theorem does not hold, and it is not hard to construct counterexamples to this procedure for $p \neq 2$.

To summarize, the SVD runs in time $\Theta(mn^{\omega-1})$, which is much slower than $\mathrm{nnz}(A) \le mn$. It is also not clear how to adapt existing fast Frobenius-norm algorithms to generate $(1+\varepsilon)$-factor approximations with respect to other Schatten p-norms.

Our Contributions In this paper we obtain the first provably efficient algorithms for the rank-k $(1+\varepsilon)$-approximation problem with respect to the Schatten p-norm for every $p \ge 1$.

Theorem 1.1 (informal, combination of Theorems 3.6 and 4.4)**.** *Suppose that $m \ge n$ and $A \in \mathbb{R}^{m \times n}$. There is a randomized algorithm which outputs two matrices $Y \in \mathbb{R}^{m \times k}$ and $Z \in \mathbb{R}^{n \times k}$ for which $\hat{X} = YZ^T$ satisfies (1) with probability at least 0.9. The algorithm runs in time $O(\mathrm{nnz}(A)\log n) + \tilde{O}(mn^{\alpha_p} \operatorname{poly}(k/\varepsilon))$, where*

$$\alpha_p = \begin{cases} 0, & 1 \le p \le 2; \\ (\omega-1)(1-\frac{2}{p}), & p > 2, \end{cases}$$

and the hidden constants depend only on p.

In the particular case of $p = 1$, and more generally for all $p \in [1,2]$, our algorithm achieves a running time of $O(\mathrm{nnz}(A)\log n + m \operatorname{poly}(k/\varepsilon))$, which was previously known to be possible for $p = 2$ only. When $p > 2$, the running time begins to depend polynomially on m and n but the dependence remains $o(mn^{\omega})$ for all larger p. Thus, even for larger values of p, when k is subpolynomial in n, our algorithm runs substantially faster than the SVD. Empirical evaluations are also conducted to demonstrate our improved algorithm when $p = 1$ in Section 5.

It was shown by Musco & Woodruff (2017) that computing a constant-factor low-rank approximation to A^TA, given only A, requires $\Omega(\mathrm{nnz}(A) \cdot k)$ time. Given that the squared singular values of A are the singular values of A^TA, it is natural to suspect that obtaining a constant-factor low rank approximation to the Schatten 4-norm low-rank approximation would therefore require $\Omega(\mathrm{nnz}(A) \cdot k)$ time. Surprisingly, we show this is not the case, and obtain an $\tilde{O}(\mathrm{nnz}(A) + mn^{(\omega-1)/2} \operatorname{poly}(k/\varepsilon))$ time algorithm.

In addition, we generalize the error metric from matrix norms to a wide family of general loss functions, see Section 6 for details. Thus, we considerably broaden the class of loss functions for which input sparsity time algorithms were previously known for.

Technical Overview. We illustrate our ideas for $p = 1$. Our goal is to find an orthogonal projection $\hat{Q}'$ for which $\left\|A(I-\hat{Q}')\right\|_1 \le (1+O(\varepsilon))\left\|A-A_k\right\|_1$. The crucial idea in the analysis is to split $\|\cdot\|_1$ into a head part $\|\cdot\|_{(r)}$, which, known as the Ky-Fan norm, equals the sum of the top r singular values, and a tail part $\|\cdot\|_{(-r)}$ (this is just a notation—the tail part is not a norm), which equals the sum of all the remaining singular values. Observe that for $r \ge k/\varepsilon$ it holds that $\left\|A(I-\hat{Q}')\right\|_{(-r)} \le \|A\|_{(-r)} \le \left\|A-A_k\right\|_{(-r)} + \varepsilon\left\|A-A_k\right\|_1$ for any rank-k orthogonal projection $\hat{Q}'$ and it thus suffices to find $\hat{Q}'$ for which $\left\|A(I-\hat{Q}')\right\|_{(r)} \le (1+\varepsilon)\left\|A-A_k\right\|_{(r)}$. To do this, we sketch $A(I-Q)$ on the left by a projection-cost preserving matrix S by Cohen et al. (2017) such that $\left\|SA(I-Q)\right\|_{(r)} = (1\pm\varepsilon)\left\|A(I-Q)\right\|_{(r)} \pm \varepsilon\|A-A_k\|_1$ for all rank-k projections Q. Then we solve $\min_Q \left\|SA(I-Q)\right\|_{(r)}$ over all rank-k projections Q and obtain a $(1+\varepsilon)$-approximate projection $\hat{Q}'$, which, intuitively, is close to the best projection P_R for $\min_Q \left\|A(I-Q)\right\|_{(r)}$, and can be shown to satisfy the desired property above.

The last step is to approximate $A\hat{Q}'$, which could be expensive if done trivially, so we reformulate it as a regression problem $\min_Y \left\|A - YZ^T\right\|_1$ over $Y \in \mathbb{R}^{n\times k}$, where Z is an $n \times k$ matrix whose columns form an orthonormal basis of the target space of the projection $\hat{Q}'$. This latter idea has been applied successfully for Frobenius-norm low-rank approximation (see, e.g., (Clarkson & Woodruff, 2017)). Here we need to argue that the solution to the Frobenius-norm regression $\min_Y \left\|A - YZ^T\right\|_F$ problem gives a good solution to the Schatten 1-norm regression problem. Finally we output Y and Z.

2. Preliminaries

Notation For an $m \times n$ matrix A, let $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_s(A)$ denote its singular values, where $s = \min\{m, n\}$. The Schatten p-norm ($p \geq 1$) of A is defined to be $\|A\|_p := \sum_{i=1}^s (\sigma_i(A)^p)^{1/p}$ and the singular (p, r)-norm ($r \leq s$) to be $\|A\|_{(p,r)} = \sum_{i=1}^r (\sigma_i(A)^p)^{1/p}$. It is clear that $\|A\|_p = \|A\|_{(p,s)}$. When $p = 2$, the Schatten p-norm coincides with the Frobenius norm and we shall use the notation $\|\cdot\|_F$ in preference to $\|\cdot\|_2$.

Suppose that A has the singular value decomposition $A = U\Sigma V^T$, where Σ is a diagonal matrix of the singular values. For $k \leq \min\{m, n\}$, let Σ_k denote the diagonal matrix for the largest k singular values only, i.e., $\Sigma_k = \operatorname{diag}\{\sigma_1(A), \ldots, \sigma_k(A), 0, \ldots, 0\}$. We define $A_k = U\Sigma_k V^T$. The famous Mirsky's theorem states that A_k is the best rank-k approximation to A for any rotationally invariant matrix norm.

For a subspace $E \subseteq \mathbb{R}^n$, we define P_E to be an $n \times \dim(E)$ matrix whose columns form an orthonormal basis of E.

Toolkit There has been extensive research on randomized numerical linear algebra in recent years. Below are several existing results upon which our algorithm will be built.

Definition 2.1 (Sparse Embedding Matrix)**.** Let $\varepsilon > 0$ be an error parameter. The (n, ε)-sparse embedding matrix R of dimension $n \times r$ is constructed as follows, where r is to be specified later. Let $h : [n] \to [r]$ be a random function and $\sigma : [n] \to \{-1, 1\}$ be a random function. The matrix R has only n nonzero entries: $R_{i,h(i)} = \sigma(i)$ for all $i \in [n]$. The value of r is chosen to be $r = \Theta(1/\varepsilon^2)$ such that $\Pr_R\{\left\|A^T RR^T B - A^T B\right\|_F^2 \leq \varepsilon^2 \|A\|_F^2 \|B\|_F^2\} \geq 0.99$ for all A with orthonormal columns. This is indeed possible by (Clarkson & Woodruff, 2017; Meng & Mahoney, 2013; Nelson & Nguyen, 2013).

It is clear that, for a matrix A with n columns and an (n, ε)-sparse embedding matrix R, the matrix product AR can be computed in $O(\operatorname{nnz}(A))$ time.

Lemma 2.1 (Thin SVD (Demmel et al., 2007))**.** *Suppose that $A \in \mathbb{R}^{m\times n}$ with $m \geq n$ and the (thin) singular value decomposition is $A = U\Sigma V^T$, where $U \in \mathbb{R}^{m\times n}$ and $\Sigma, V \in \mathbb{R}^{n\times n}$. Computing the full thin SVD takes time $\tilde{O}(mn^{\omega-1})$.*

Lemma 2.2 (Multiplicative Spectral Approximation (Cohen et al., 2015b))**.** *Suppose that $A \in \mathbb{R}^{m\times n}$ ($n \leq m \leq \operatorname{poly}(n)$) has rank r. There exists a sampling matrix R of $O(\varepsilon^{-2} r \log r)$ rows such that $(1 - \varepsilon)A^T A \preceq (RA)^T(RA) \preceq (1 + \varepsilon)A^T A$ with probability at least 0.9 and R can be computed in $O(\operatorname{nnz}(A) \log n + n^\omega \log^2 n + n^2\varepsilon^{-2})$ time, where θ is an arbitrary constant in $(0, 1]$.*

Lemma 2.3 (Additive-Multiplicative Spectral Approximation (Cohen et al., 2017; Musco, 2018))**.** *Suppose that $A \in \mathbb{R}^{m\times n}$ with $m \geq n$, error parameters $\varepsilon \geq \eta \geq 1/\operatorname{poly}(n)$. Let $K = k + \varepsilon/\eta$. There exists a randomized algorithm which runs in $O(\operatorname{nnz}(A) \log n) + \tilde{O}(mK^{\omega-1})$ time and outputs a matrix C of $t = \Theta(\varepsilon^{-2} K \log K)$ columns, which are rescaled column samples of A without replacement, such that with probability at least 0.99,*

$$\begin{aligned}(1 - \varepsilon)AA^T - \eta \|A - A_k\|_F^2 I &\preceq CC^T \\ &\preceq (1 + \varepsilon)AA^T + \eta \|A - A_k\|_F^2 I. \quad (2)\end{aligned}$$

We also need an elementary inequality.

Lemma 2.4. *Suppose that $p \geq 1$ and $\varepsilon \in (0, 1]$. Let $C_{p,\varepsilon} = p(1 + 1/\varepsilon)^{p-1}$. It holds for $x \in [\varepsilon, 1]$ that $(1 + x)^p \leq 1 + C_{p,\varepsilon}x^p$ and that $(1 - x)^p \geq 1 - C_{p,\varepsilon}x^p$.*

Proof. It is easy to see that for $x \in [\varepsilon, 1]$, $(1 + x)^p \leq 1 + \frac{(1+\varepsilon)^p - 1}{\varepsilon^p} x^p$ and $(1 - x)^p \geq 1 - \frac{1-(1+\varepsilon)^p}{\varepsilon^p} x^p$. Then note that $p\left(1 + \frac{1}{\varepsilon}\right)^{p-1} \geq \frac{(1+\varepsilon)^p - 1}{\varepsilon^p} \geq \frac{1-(1+\varepsilon)^p}{\varepsilon^p}$. □

3. Case $p < 2$

The algorithm is presented in Algorithm 1. In this section we shall prove the correctness and analyze the running time for a constant $p \in [1, 2)$. Throughout this section we set $r = k/\varepsilon$.

Lemma 3.1. *Suppose that $p \in (0, 2)$ and S satisfies (3). It then holds for all rank-k orthogonal projections Q that*

$$\begin{aligned}&(1 - \varepsilon) \|A(I - Q)\|_{(p,r)}^p - r\eta_1^{\frac{p}{2}} \|A - A_k\|_p^p \\ &\leq \|SA(I - Q)\|_{(p,r)}^p \\ &\leq (1 + \varepsilon) \|A(I - Q)\|_{(p,r)}^p + r\eta_1^{\frac{p}{2}} \|A - A_k\|_p^p .\end{aligned}$$

Proof. Since S satisfies (3), it holds for any rank-k orthogonal projection Q that

$$\begin{aligned}&(1 - \varepsilon)(I - Q)A^T A(I - Q) - \eta_1 \|A - A_k\|_F^2 I \\ &\preceq (I - Q)A^T S^T SA(I - Q)\end{aligned}$$

Algorithm 1 Outline of the algorithm for finding a low-rank approximation

```
1: if $p < 2$ then
2:     $\eta_1 \leftarrow O((\varepsilon^2/k)^{2/p})$, $\eta_2 \leftarrow O(\varepsilon^2/k^{2/p-1})$
3: else
4:     $\eta_1 \leftarrow O(\varepsilon^{1+2/p}/k^{2/p}n^{1-2/p})$, $\eta_2 \leftarrow O(\varepsilon^2/n^{1-2/p})$
5: end if
6: Use Lemma 2.3 to obtain a sampling matrix $S$ of $s$ rows such that
       $(1-\varepsilon)A^TA - \eta_1 \|A - A_k\|_F^2 I$
       $\preceq A^TS^TSA$                                        (3)
       $\preceq (1+\varepsilon)A^TA + \eta_1 \|A - A_k\|_F^2 I.$
7: $T \leftarrow$ subspace embedding matrix for $s$-dimensional subspaces with error $O(\varepsilon)$
8: $W' \leftarrow$ projection onto the top $k$ left singular vectors of $SAT$
9: $Z \leftarrow$ matrix whose columns are an orthonormal basis of the row space of $W'SA$
10: $R \leftarrow (n, \Theta(\sqrt{\eta_2/k}))$-sparse embedding matrix
11: $\hat{Y} \leftarrow ARP_{\text{rowspace}(Z^TR)}$
12: return $\hat{Y}, Z$
```

$$\preceq (1+\varepsilon)(I-Q)A^TA(I-Q) + \eta_1 \|A - A_k\|_F^2 I.$$

The following relationship between singular values of $SA(I-Q)$ and $A(I-Q)$ is an immediate corollary via the max-min characterization of singular values (cf., e.g., Lemma 7.2 of (Li et al., 2019))

$$\begin{aligned} &(1-\varepsilon)\sigma_i^2(A(I-Q)) - \eta_1 \|A - A_k\|_F^2 \\ &\le \sigma_i^2(SA(I-Q)) \\ &\le (1+\varepsilon)\sigma_i^2(A(I-Q)) + \eta_1 \|A - A_k\|_F^2 \end{aligned} \tag{4}$$

Since $p < 2$ and thus $\|\cdot\|_F \le \|\cdot\|_p$, we have from (4) that

$$\begin{aligned} &(1-\varepsilon)\sigma_i^p(A(I-Q)) - \eta_1^{\frac{p}{2}} \|A - A_k\|_p^p \\ &\le \sigma_i^p(SA(I-Q)) \\ &\le (1+\varepsilon)\sigma_i^p(A(I-Q)) + \eta_1^{\frac{p}{2}} \|A - A_k\|_p^p . \end{aligned}$$

Passing to the (p,r)-singular norm yields the desired result. □

Lemma 3.2. *When $p \in (0,2)$ is a constant and $\varepsilon \in (0,1/2]$, let $\hat{Q}' = ZZ^T$ be the projection onto the column space of Z, where Z is as defined in Line 9 of Algorithm 1. With probability at least 0.99, it holds that*

$$\left\|SA(I-\hat{Q}')\right\|_{(p,r)} \le (1+\varepsilon) \min_Q \|SA(I-Q)\|_{(p,r)} , \tag{5}$$

where the minimization on the right-hand side is over all rank-k orthogonal projections Q.

Proof. Observe that

$$\min_Q \|SA - SAQ\|_{(p,r)} = \min_W \|SA - WSA\|_{(p,r)} ,$$

where the minimizations are over all rank-k orthogonal projections Q and all rank-k orthogonal projections W, and the equality is achieved when Q is the projection onto the right top k singular vectors of SA and W the left top k singular vectors.

Since T is an oblivious subspace embedding matrix and preserves all singular values of $(I-W)SA$ up to a factor of $(1 \pm \varepsilon)$, we have

$$\min_W \|SAT - WSAT\|_{(p,r)} = (1 \pm \varepsilon) \min_W \|SA - WSA\|_{(p,r)} .$$

The minimization on the left-hand side above is easy to solve: the minimizer W' is exactly the projection onto the top k singular vectors of SAT, as computed in Line 8 of Algorithm 1. Since $\hat{Q}'$ is the projection onto the row space of $W'SA$, it holds that the row space of $SA\hat{Q}'$ is the closest space to that of SA in the row space of $W'SA$. Hence

$$\left\|SA - SA\hat{Q}'\right\|_{(p,r)} \le \|SA - W'SA\|_{(p,r)} .$$

The claimed result (5) then follows from

$$\begin{aligned} \|SA - W'SA\|_{(p,r)} &\le \frac{1}{1-\varepsilon} \|SAT - W'SAT\|_{(p,r)} \\ &= \frac{1}{1-\varepsilon} \min_W \|SAT - WSAT\|_{(p,r)} \\ &\le \frac{1+\varepsilon}{1-\varepsilon} \min_W \|SA - WSA\|_{(p,r)} \\ &\le (1+4\varepsilon) \min_Q \|SA - SAQ\|_{(p,r)} \end{aligned}$$

and rescaling ε. □

Lemma 3.3. *Let $\varepsilon \in (0,1/2]$. Suppose that $\hat{Q}'$ satisfies (5). Then it holds that*

$$\begin{aligned} \left\|A(I-\hat{Q}')\right\|_{(p,r)}^p \le (1+c_1\varepsilon) \|A - A_k\|_{(p,r)}^p \\ + c_2 k \eta_1^{p/2} \|A - A_k\|_p^p . \end{aligned}$$

for some absolute constants $c_1, c_2 > 0$.

Proof. Let $\hat{Q} = \arg\min_Q \|SA(I-Q)\|_{(p,r)}$, where the minimization is over all rank-k projections Q. Let Q^* be the orthogonal projection onto the top k right singular vectors of A. It follows that

$$\begin{aligned} &\left\|A(I-\hat{Q}')\right\|_{(p,r)}^p \\ &\le \frac{1}{1-\varepsilon} \left\|SA(I-\hat{Q}')\right\|_{(p,r)}^p + \frac{1}{1-\varepsilon} k \eta_1^{\frac{p}{2}} \|A - A_k\|_p^p \end{aligned}$$

$$
\begin{aligned}
&\le \frac{(1+\varepsilon)^p}{1-\varepsilon}\left\|SA(I-\hat{Q})\right\|_{(p,r)}^p + \frac{1}{1-\varepsilon}k\eta_1^{\frac{p}{2}}\left\|A-A_k\right\|_p^p \\
&\le \frac{(1+\varepsilon)^p}{1-\varepsilon}\left\|SA(I-Q^*)\right\|_{(p,r)}^p + \frac{1}{1-\varepsilon}k\eta_1^{\frac{p}{2}}\left\|A-A_k\right\|_p^p \\
&\le \frac{(1+\varepsilon)^p}{1-\varepsilon}\Big((1+\varepsilon)\left\|A(I-Q^*)\right\|_{(p,r)}^p + k\eta_1^{\frac{p}{2}}\left\|A-A_k\right\|_p^p\Big) \\
&\qquad + \frac{1}{1-\varepsilon}k\eta_1^{\frac{p}{2}}\left\|A-A_k\right\|_p^p \\
&= \frac{(1+\varepsilon)^{p+1}}{1-\varepsilon}\left\|A-A_k\right\|_{(p,r)}^p + \frac{(1+\varepsilon)^p+1}{1-\varepsilon}k\eta_1^{\frac{p}{2}}\left\|A-A_k\right\|_p^p,
\end{aligned}
$$

where the first inequality follows from Lemma 3.1, the second inequality Lemma 3.2, the third inequality follows from the optimality of $\hat{Q}$ and the fourth inequality again from Lemma 3.1. □

The next lemma is an immediate corollary of the preceding lemma.

Lemma 3.4. *Let $\varepsilon \in (0, 1/2]$. Suppose that $\hat{Q}'$ satisfies (5). Then it holds for some absolute constants $c_1, c_2 > 0$ that*

$$\left\|A(I-\hat{Q}')\right\|_p^p \le (1+c_1\varepsilon)\left\|A-A_k\right\|_p^p$$

whenever $\eta_1 \le (\varepsilon^2/(c_2 k))^{2/p}$.

Proof. Again let $\hat{Q} = \arg\min_Q \|SA(I-Q)\|_{(p,r)}$, where the minimization is over all rank-k projections Q. Observe that

$$
\begin{aligned}
&\left\|A(I-\hat{Q}')\right\|_p^p \\
&= \left\|A(I-\hat{Q}')\right\|_{(p,r)}^p + \sum_{i\ge r+1}\sigma_i^p(A(I-\hat{Q}')) \\
&\le (1+c_1\varepsilon)\left\|A-A_k\right\|_{(p,r)}^p + c_2 r\eta_1^{p/2}\left\|A-A_k\right\|_p^p \\
&\qquad + \sum_{i\ge r+1}\sigma_i^p(A) \\
&\le (1+c_1\varepsilon)\left\|A-A_k\right\|_p^p + c_2 r\eta_1^{p/2}\left\|A-A_k\right\|_p^p \\
&\qquad + \sum_{i=r+1}^{r+k+1}\sigma_i^p(A) \\
&\le (1+c_1\varepsilon)\left\|A-A_k\right\|_p^p + c_2 r\eta_1^{p/2}\left\|A-A_k\right\|_p^p \\
&\qquad + \frac{k}{r}\left\|A-A_k\right\|_p^p \\
&\le (1+(c_1+1)\varepsilon)\left\|A-A_k\right\|_p^p + c_2 r\eta_1^{p/2}\left\|A-A_k\right\|_p^p
\end{aligned}
$$

where we used the preceding lemma (Lemma 3.3) in the first inequality and $r = k/\varepsilon$ in the last inequality. The claimed result holds when $\eta \le (\varepsilon/(c_2 r))^{2/p}$. □

So far we have found a rank-k orthogonal projection $\hat{Q}' = ZZ^T$ such that

$$\left\|A - A\hat{Q}'\right\|_p \le (1+c_1\varepsilon)\left\|A-A_k\right\|_p$$

for some absolute constant c_1. However, it is not clear how to compute the matrix product $A\hat{Q}'$ efficiently. Hence we consider the regression problem

$$\min_{Y:\operatorname{rank}(Y)=k}\left\|A - YZ^T\right\|_p.$$

It is clear that the minimizer is $Y = AZ$, which satisfies that $\left\|A - YZ^T\right\|_p = \left\|A - A\hat{Q}'\right\|_p$, since the rowspace of YZ^T is a k-dimensional subspace of the rowspace of Z^T and thus it is exactly the rowspace of Q. The next lemma shows that $\hat{Y}$ is an approximation to Y.

Lemma 3.5. *When $1 \le p < 2$ is a constant, the matrix $\hat{Y}$ defined in Line 11 of Algorithm 1 satisfies with probability at least 0.99 that*

$$\left\|A - \hat{Y}Z^T\right\|_p \le (1+c\varepsilon)\min_{Y:\operatorname{rank}(Y)=k}\left\|A - YZ^T\right\|_p,$$

for some absolute constant $c > 0$, whenever $\eta_2 \le \varepsilon^2/(2k)^{2/p-1}$.

Proof. First, it is clear that the optimal solution to $\min_Y \left\|A - YZ^T\right\|_p$ is $Y = AZ$, where the minimization is over all rank-k $n \times k$ matrices Y.

Note that

$$\hat{Y} = ARP_{\text{rowspace}(Z^T R)}$$

is the minimizer to the Frobenius-norm minimization problem $\min_Y \|(A - YZ^T)R\|_F$. Since R is a sparse embedding matrix of error $\Theta(\sqrt{\eta_2/k})$, one can show that (see, e.g., Lemma 7.8 of (Clarkson & Woodruff, 2017)) with probability at least 0.99,

$$\left\|AZ - \hat{Y}\right\|_F \le \sqrt{\eta_2}\left\|A - AZZ^T\right\|_F.$$

It follows that

$$
\begin{aligned}
&\left\|A - \hat{Y}Z^T\right\|_p \\
&\le \left\|A - AZZ^T\right\|_p + \left\|\hat{Y}Z^T - AZZ^T\right\|_p \\
&\le (1+c_1\varepsilon)\left\|A - AZZ^T\right\|_p + \left\|\hat{Y} - AZ\right\|_p \\
&\le (1+c_1\varepsilon)\left\|A - AZZ^T\right\|_p + (2k)^{\frac{1}{p}-\frac{1}{2}}\left\|\hat{Y} - AZ\right\|_F \\
&\le (1+c_1\varepsilon)\left\|A - AZZ^T\right\|_p + (2k)^{\frac{1}{p}-\frac{1}{2}}\sqrt{\eta_2}\left\|A - AZZ^T\right\|_F \\
&= (1+c_1\varepsilon)\left\|A - AZZ^T\right\|_p + (2k)^{\frac{1}{p}-\frac{1}{2}}\sqrt{\eta_2}\left\|A - AZZ^T\right\|_p \\
&\le (1+(c_1+1)\varepsilon)\left\|A - AZZ^T\right\|_p
\end{aligned}
$$

provided that $\eta_2 \le \varepsilon^2/(2k)^{2/p-1}$. □

Remark 1. The preceding lemma (Lemma 3.5) may be of independent interest, as it solves Schatten p-norm regression efficiently, which has not been discussed in the literature in the context of dimensionality reduction before.

In summary, we conclude the section with our main theorem.

Theorem 3.6. *Let $p \in [1, 2)$. Suppose that $A \in \mathbb{R}^{m\times n}$ with $m \geq n$. There is a randomized algorithm which outputs $Y \in \mathbb{R}^{m\times k}$ and $Z \in \mathbb{R}^{n\times k}$ such that $\hat{X} = YZ^T$ satisfies (1) with probability at least 0.97. The algorithm runs in time $O(\operatorname{nnz}(A)\log n) + \tilde{O}_p(mk^{2(\omega-1)/p}/\varepsilon^{(4/p-1)(\omega-1)} + k^{2\omega/p}/\varepsilon^{(4/p-1)(2\omega+2)})$.*

Proof. The correctness of the output is clear from the preceding lemmata by rescaling ε. We discuss the runtime below.

We first examine the runtime to obtain Z. For η_1 in Lemma 3.4 we have $k\eta_1 \leq \varepsilon$ and thus $K = \Theta(\varepsilon/\eta_1)$. Applying Lemma 2.3, we have $s = \tilde{O}(K/\varepsilon^2)$ and obtaining the matrix S takes time $O(\operatorname{nnz}(A)\log n)+\tilde{O}(mK^{\omega-1})$. By Lemma 2.2, we can obtain $(SA)T$ in time $O(\operatorname{nnz}(SA)) + \tilde{O}(s^\omega/\varepsilon^2)$ for a matrix T of $\tilde{O}(s/\varepsilon^2)$ columns and thus the subsequent SVD of SAT, by Lemma 2.1, takes $\tilde{O}(s^\omega/\varepsilon^2)$. These three steps take time $\tilde{O}(mK^{\omega-1}) + O(\operatorname{nnz}(SA)) + \tilde{O}(s^\omega/\varepsilon^2) = O(\operatorname{nnz}(A)) + \tilde{O}(mK^2 + K^\omega/\varepsilon^{2\omega+2})$, where we used the fact that S samples the rows of A without replacement and so $\operatorname{nnz}(SA) \leq \operatorname{nnz}(A)$. Calculating the row span of $W'SA$, which is a k-by-n matrix, takes $O(nk^{\omega-1})$ time. The total runtime is $O(\operatorname{nnz}(A)\log n) + \tilde{O}_p(mK^{\omega-1} + K^\omega/\varepsilon^{2\omega+2})$. Plugging in $K = \varepsilon/\eta_1 = \Theta(k^{2/p}/\varepsilon^{4/p-1})$ yields the runtime $O(\operatorname{nnz}(A)\log n) + \tilde{O}(mk^{2(\omega-1)/p}/\varepsilon^{(4/p-1)(\omega-1)} + k^{2\omega/p}/\varepsilon^{(4/p-1)(2\omega+2)})$.

Next we examine the runtime to obtain $\hat{Y}$. Since R has $t = \Theta(k/\eta_2)$ rows and AR can be computed in $O(\operatorname{nnz}(A))$ time, Z^TR can be computed in $O(nk)$ time, the row space of Z^TR (which is a $k \times t$ matrix) in $O(k^{\omega-1}t) = \tilde{O}(k^\omega/\eta_2)$ time, and the final matrix product $(AR)P_{\text{rowspace}(Z^TR)}$ in $O(mt^{\omega-1}) = \tilde{O}(mk^{\omega-1}/\eta_2)$ time. Overall, computing $\hat{Y}$ takes time $O(\operatorname{nnz}(A)) + \tilde{O}(mk^{\omega-1}/\eta_2) = O(\operatorname{nnz}(A)) + \tilde{O}(mk^{\omega+2/p-2}/\varepsilon^2)$.

The overall runtime follows immediately. □

4. Case $p > 2$

The algorithm remains the same in Algorithm 1. In this section we shall prove the correctness and analyse the runtime for constant $p > 2$. The outline of the proof is the same and we shall only highlight the differences. Again we let $r = k/\varepsilon$.

In place of Lemma 3.1, we now have:

Lemma 4.1. *Suppose that $p > 2$ and S satisfies (3). It then holds for all rank-k orthogonal projection Q that*

$$\begin{aligned}&(1-K_p\varepsilon)\left\|A(I-Q)\right\|_{(p,r)}^p - C_{p/2,\varepsilon} r\eta_1^{p/2}\left\|A-A_k\right\|_F^p\\ &\leq \left\|SA(I-Q)\right\|_{(p,r)}^p\\ &\leq (1+K_p\varepsilon)\left\|A(I-Q)\right\|_{(p,r)}^p + C_{p/2,\varepsilon} r\eta_1^{p/2}\left\|A-A_k\right\|_F^p,\end{aligned}$$

where $K_p \geq 1$ is some constant that depends only on p.

Proof. We now have two cases based on (4).

- When $\sigma_i(A(I-Q))^2 \geq (1/\varepsilon)\eta_1\|A-A_k\|_F^2$, we have
$$\begin{aligned}(1-O_p(\varepsilon))\sigma_i^p(A(I-Q)) &\leq \sigma_i^p(SA(I-Q))\\ &\leq (1+O_p(\varepsilon))\sigma_i^p(A(I-Q))\end{aligned}$$
- When $\sigma_i(A(I-Q))^2 < (1/\varepsilon)\eta_1\|A-A_k\|_F^2$, we have from Lemma 2.4 that
$$\begin{aligned}&(1-\varepsilon)\sigma_i^p(A(I-Q)) - C_{p/2,\varepsilon}\eta_1^{p/2}\left\|A-A_k\right\|_F^p\\ &\leq \sigma_i^p(SA(I-Q))\\ &\leq (1+\varepsilon)\sigma_i^p(A(I-Q)) + C_{p/2,\varepsilon}\eta_1^{p/2}\left\|A-A_k\right\|_F^p\end{aligned}$$

The claimed result follows in the same manner as in the proof of Lemma 3.1. □

The analogy of Lemma 3.4 is the following, where we apply Hölder's inequality that $\|A-A_k\|_F \leq n^{1/2-1/p}\|A-A_k\|_p$.

Lemma 4.2. *Let $\varepsilon \in (0, 1/2]$. Suppose that $\hat{Q}'$ satisfies (5). Then it holds for some constants $c_p, c_p' > 0$ which depend only on p that*

$$\left\|A(I-\hat{Q}')\right\|_p^p \leq (1+c_p\varepsilon)\left\|A-A_k\right\|_p^p,$$

whenever $\eta_1 \leq c_p'\varepsilon^{1+2/p}/(k^{2/p}n^{1-2/p})$.

Proof. Similarly we have

$$\begin{aligned}\left\|A(I-\hat{Q}')\right\|_p^p &\leq (1+c_p\varepsilon)\left\|A-A_k\right\|_p^p\\ &\qquad + rC_{p/2,\varepsilon}\eta_1^{p/2}n^{\frac{p}{2}-1}\left\|A-A_k\right\|_p^p.\end{aligned}$$

The conclusion follows when

$$C_{p/2,\varepsilon}\eta_1^{p/2}n^{p/2-1} \leq \frac{\varepsilon}{r} = \frac{\varepsilon^2}{k},$$

that is,

$$\left(\frac{p}{2}\left(1+\frac{1}{\varepsilon}\right)^{\frac{p}{2}-1}\right)\eta_1^{\frac{p}{2}}n^{\frac{p}{2}-1} \leq \frac{\varepsilon^2}{k}.$$ □

The analogy of Lemma 3.5 is the following.

Lemma 4.3. *When $p > 2$ is a constant, the matrix $\hat{Y}$ defined in Line 11 of Algorithm 1 satisfies with probability at least 0.9 that*

$$\left\|A-\hat{Y}Z^T\right\|_p \leq (1+c_p\varepsilon)\min_{Y:\operatorname{rank}(Y)=k}\left\|A-YZ^T\right\|_p,$$

for some constant that depends only on p, whenever $\eta_2 \leq \varepsilon^2/n^{1-2/p}$.

Proof. The proof is similar to that of Lemma 3.5 except that we have instead in the last part of the argument that

$$
\begin{aligned}
&\left\|A-\hat{Y}Z\right\|_p \\
&\leq (1+c_p\varepsilon)\left\|A-AZZ^T\right\|_p+\left\|\hat{Y}-AZ\right\|_p \\
&\leq (1+c_p\varepsilon)\left\|A-AZZ^T\right\|_p+\left\|\hat{Y}-AZ\right\|_F \\
&\leq (1+c_p\varepsilon)\left\|A-AZZ^T\right\|_p+\sqrt{\eta_2}\left\|A-AZZ^T\right\|_F \\
&\leq (1+c_p\varepsilon)\left\|A-AZZ^T\right\|_p+\sqrt{\eta_2}n^{\frac{1}{2}-\frac{1}{p}}\left\|A-AZZ^T\right\|_p
\end{aligned}
$$

and we would need $\eta_2 \leq \varepsilon^2/n^{1-\frac{2}{p}}$. □

In summary, we have the following main theorem.

Theorem 4.4. *Let $p > 2$ be a constant. Suppose that $A \in \mathbb{R}^{m\times n}$ ($m \geq n$). There is a randomized algorithm which outputs $Y \in \mathbb{R}^{m\times k}$ and $Z \in \mathbb{R}^{n\times k}$ such that $\hat{X} = YZ^T$ satisfies (1) with probability at least 0.97. The algorithm runs in time $O(\operatorname{nnz}(A)\log n) + \tilde{O}_p(n^{\omega(1-2/p)}k^{2\omega/p}/\varepsilon^{2\omega/p+2} + mn^{(\omega-1)(1-2/p)}(k/\varepsilon)^{2(\omega-1)/p})$.*

Proof. The correctness follows from the previous lemmata as in the proof of Theorem 3.6. Below we discuss the running time.

First we examine the time required to obtain Z. It is easy to verify that $k\eta_1 \leq \varepsilon$ and so $K = \Theta(\varepsilon/\eta_1)$. Similar to the analysis in Theorem 3.6, we have the total runtime $O(\operatorname{nnz}(A)\log n) + \tilde{O}_p(mK^{\omega-1} + K^{\omega}/\varepsilon^{2\omega+2}))$. Note that $K = \Theta(\varepsilon/\eta_1) = \Theta(n^{1-2/p}(k/\varepsilon)^{2/p})$, so the runtime becomes $O(\operatorname{nnz}(A)\log n) + \tilde{O}_p(mn^{(\omega-1)(1-2/p)}(k/\varepsilon)^{2(\omega-1)/p} + n^{\omega(1-2/p)}k^{2\omega/p}/\varepsilon^{2\omega/p+2})$.

Next we examine the time required to obtain $\hat{Y}$. Again similarly the runtime is $O(\operatorname{nnz}(A)) + \tilde{O}(mk^{\omega-1}/\eta_2) = O(\operatorname{nnz}(A)) + \tilde{O}(mn^{1-2/p}k^{\omega-1}/\varepsilon^2)$.

Combining the two runtimes above yields the overall runtime. □

5. Experiments

The contribution of our work is primarily theoretical: an input sparsity time algorithm for low-rank approximation for any Schatten p-norm. In this section, nevertheless, we give an empirical verification of the advantage of our algorithm on both synthetic and real-world data. We focus on the most important case of the nuclear norm, i.e., $p = 1$.

In addition to the solution provided by our algorithm, we also consider a natural candidate for a low-rank approximation algorithm, which is the solution in Frobenius norm, that is, a rank-k matrix X for which $\|A - X\|_F \leq (1+\varepsilon)\|A - A_k\|_F$. This problem admits a simple solution as follows. Take S to be a Count-Sketch matrix and let Z be an $n \times k$ matrix whose columns form an orthonormal basis of the top-k right singular vectors of SA. Then $X = AZZ^T$ is a Frobenius-norm solution with high probability (Cohen et al., 2015a).

Table 1: Performance of our algorithm on synthetic data compared with the approximate Frobenius-norm solution and the SVD.

	$k = 5$	$k = 10$	$k = 20$
median of ε_1	0.00372	0.00377	0.00486
median of ε_2	0.00412	0.00485	0.00637
median runtime of $\|\cdot\|_1$ algorithm	0.067s	0.196s	0.428s
median runtime of $\|\cdot\|_F$ algorithm	0.044s	0.073s	0.191s
median runtime of SVD	5.788s		

We shall compare the quality (i.e., approximation ratio measured in Schatten 1-norm) of both solutions and the running times[3].

Synthetic Data. We adopt a simpler version of Algorithm 1 by taking S to be a COUNT-SKETCH matrix of target dimension k^2 and both R and T to be identity matrices of appropriate dimension. For the Frobenius-norm solution, we also take S to be a COUNT-SKETCH matrix of target dimension k^2. We choose $n = 3000$ and generate a random $n \times n$ matrix A of independent entries, each of which is uniform in $[0, 1]$ with probability 0.05 and 0 with probability 0.95. Since the regime of interest is $k \ll n$, we vary k among $\{5, 10, 20\}$. For each value of k, we run our algorithm 50 times and record the relative approximation error $\varepsilon_1 = \|A - YZ^T\|_1/\|A - A_k\|_1 - 1$ with the running time and the relative approximation error of the Frobenius-norm solution $\varepsilon_2 = \|A - X\|_1/\|A - A_k\|_1 - 1$ with the running time. The same matrix A is used for all tests. In Table 1 we report the median of ε_1, the median of ε_2, the median running time of both algorithms, among 50 independent runs for each k, and the median running time of a full SVD of A among 10 runs.

We can observe that our algorithm achieves a good (relative) approximation error, which is less than 0.005 in all such cases of k. Our algorithm also outperforms the approximate Frobenius-norm solution by 10%–30% in terms of approximation error. Our algorithm also runs about 13-fold faster than a regular SVD.

[3] All tests are run under MATLAB 2019b on a machine of Intel Core i7-6550U CPU@2.20GHz with 2 cores.

Table 2: Performance of our algorithm on KOS data compared with the approximate Frobenius-norm solution and the SVD.

	$k = 5$	$k = 10$	$k = 20$
median of ε_1	0.0149	0.0145	0.0132
median of ε_2	0.0183	0.0216	0.0259
median runtime of $\|\cdot\|_1$ algorithm	0.155s	0.204s	0.323s
median runtime of $\|\cdot\|_F$ algorithm	0.113s	0.154s	0.242s
median runtime of SVD	4.999s		

KOS data. For real-world data, we use a word frequency dataset, named KOS, from UC Irvine.[4] The matrix represents word frequencies in blogs and has dimension 3430×6906 with 353160 non-zero entries. Again we report the median relative approximation error and the median running time of our algorithm and those of the Frobenius-norm algorithm among 50 independent runs for each value of $k \in \{5, 10, 20\}$. The results are shown in Table 2.

Our algorithm achieves a good approximation error, less than 0.015, and surpasses the approximate Frobenius-norm solution for all such values of k. The gap between two solutions in the approximation error widens as k increases. When $k = 10$, our algorithm outperforms the approximate Frobenius-norm by 30%; when $k = 20$, this increases to almost 50%. Our algorithm, although stably slower than the Frobenius norm algorithm by 30%–40%, still displays a 14.5-fold speed-up compared with the regular SVD.

6. Generalization

More generally, one can ask to solve the problem of low-rank approximation with respect to some function Φ on the matrix singular values, i.e.,

$$\min_{X:\operatorname{rank}(X)=k} \Phi(A - X) \tag{6}$$

Here we consider $\Phi(A) = \sum_i \phi(\sigma_i(A))$ for some increasing function $\phi : [0, \infty) \to [0, \infty)$. It is clear that Φ is rotationally invariant and that $\Phi(A) \geq \Phi(B)$ if $\sigma_i(A) \geq \sigma_i(B)$ for all i. These two properties allow us to conclude that A_k remains an optimal solution for such general Φ.

We further assume that ϕ satisfies the following conditions.

(a) there exists $\alpha > 0$ such that $\phi((1 + \varepsilon)x) \leq (1 + \alpha\varepsilon)\phi(x)$ and $\phi((1 - \varepsilon)x) \geq (1 - \alpha\varepsilon)\phi(x)$ for all sufficiently small ε.

(b) it holds that for each sufficiently small ε, $K^1_{\phi,\varepsilon} = \sup_{x>0} \sup_{y\in[\varepsilon x,x]} (\phi(x + y) - \phi(x))/\phi(y) < \infty$ and $K^2_{\phi,\varepsilon} = \sup_{x>0} \sup_{y\in[\varepsilon x,x]} (\phi(x) - \phi(x - y))/\phi(y) < \infty$.

(c) it holds that for each sufficiently small ε, $L_{\phi,\varepsilon} = \sup_{x>0} \phi(\varepsilon x)/\phi(x) < \infty$.

(d) there exists $\gamma > 0$ such that $\phi(x + y) \leq \gamma(\phi(x) + \phi(y))$.

When the function ϕ is clear from the text, we also abbreviate $K^i_{\phi,\varepsilon}$ and $L_{\phi,\varepsilon}$ as K^i_ε and L_ε, respectively. Let $K_\varepsilon = \max\{K^1_\varepsilon, K^2_\varepsilon\}$.

It follows from a similar argument to Lemma 4.1 and Conditions (a)–(c) that

$$\begin{aligned}
&(1 - \alpha\varepsilon)\phi(\sigma_i(A(I - Q))) - L_{\sqrt{\eta_1}} K_\varepsilon \phi(\|A - A_k\|_F) \\
&\leq \phi(\sigma_i(SA(I - Q))) \\
&\leq (1 + \alpha\varepsilon)\phi(\sigma_i(A(I - Q))) + L_{\sqrt{\eta_1}} K_\varepsilon \phi(\|A - A_k\|_F)
\end{aligned}$$

Note Condition (c) implies that $\phi\big(\sqrt{\sum_i x_i^2}\big) \leq \phi\big(\sum_i x_i\big) \leq \gamma \sum_i \phi(x_i)$, which further implies that $\phi(\|A - A_k\|_F) \leq \gamma\Phi(A - A_k)$. Therefore

$$\begin{aligned}
&(1 - \alpha\varepsilon)\phi(\sigma_i(A(I - Q))) - \gamma L_{\sqrt{\eta_1}} K_\varepsilon \Phi(A - A_k) \\
&\leq \phi(\sigma_i(SA(I - Q))) \\
&\leq (1 + \alpha\varepsilon)\phi(\sigma_i(A(I - Q))) + \gamma L_{\sqrt{\eta_1}} K_\varepsilon \Phi(A - A_k)
\end{aligned}$$

Analogously to the singular (p, r)-norm, we define $\Phi_r(A) = \sum_{i=1}^r \phi(\sigma_i(A))$. It is easy to verify that the argument of Lemmata 3.2 to 3.4 will go through with minimal changes, yielding that

$$\begin{aligned}
\Phi_k(A(I - \hat{Q}')) \leq (1 + c_1\varepsilon)\Phi_k(A - A_k) \\
+ c_2 r \Phi(A - A_k)
\end{aligned}$$

for some constants $c_1, c_2 > 0$ that depend on $\alpha, \gamma, K_\varepsilon, L_\varepsilon$. When $\eta_1 \leq c_3(\varepsilon/r)^{1/\alpha}$ we have

$$\Phi(A(I - \hat{Q}')) \leq (1 + c_4\varepsilon)\Phi_k(A - A_k).$$

We can then output AZ and Z in time $O(\operatorname{nnz}(A) \cdot k + nk)$. Performing a similar analysis on the running time as before, we arrive at the following theorem.

Theorem 6.1. *Suppose that $\phi : [0, \infty) \to [0, \infty)$ is increasing and satisfies Conditions (a)–(d) and that $K_\varepsilon = \operatorname{poly}(1/\varepsilon)$ and $L_\varepsilon = \operatorname{poly}(1/\varepsilon)$. Let $A \in \mathbb{R}^{n\times n}$. There is a randomized algorithm which outputs matrices $Y, Z \in \mathbb{R}^{n\times k}$ such that $X = YZ^T$ satisfies (6) with probability at least 0.98. The algorithm runs in time $O(\operatorname{nnz}(A)(k + \log n)) + \tilde{O}(n \operatorname{poly}(k/\varepsilon))$, where the hidden constants depend on α, γ and the polynomial exponents for K_ε and L_ε.*

We remark that a few common loss functions satisfy our conditions for ϕ. These include the Tukey p-norm loss function $\phi(x) = x^p \cdot \mathbf{1}_{\{x\leq\tau\}} + \tau^p \cdot \mathbf{1}_{\{x>\tau\}}$, the ℓ_1-ℓ_2 loss function $\phi(x) = 2\sqrt{1 + x^2/2} - 1$ and the Huber loss function $\phi(x) = x^2/2 \cdot \mathbf{1}_{\{x\leq\tau\}} + \tau(x - \tau/2) \cdot \mathbf{1}_{\{x>\tau\}}$.

[4] https://archive.ics.uci.edu/ml/datasets/Bag+of+Words

Acknowledgements

The authors would like to thank the anonymous reviewers for helpful comments. Y. Li was supported in part by Singapore Ministry of Education (AcRF) Tier 2 grant MOE2018-T2-1-013. D. P. Woodruff was supported in part by Office of Naval Research (ONR) grant N00014-18-1-2562. Part of this work was done during a visit of D. P. Woodruff to Nanyang Technological University, funded by the aforementioned Singapore Ministry of Education grant.

References

Candès, E. J., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *J. ACM*, 58(3), June 2011.

Clarkson, K. L. and Woodruff, D. P. Low-rank approximation and regression in input sparsity time. *J. ACM*, 63(6), January 2017. ISSN 0004-5411.

Cohen, M. B., Elder, S., Musco, C., Musco, C., and Persu, M. Dimensionality reduction for k-means clustering and low rank approximation. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '15, pp. 163–172, New York, NY, USA, 2015a. Association for Computing Machinery.

Cohen, M. B., Lee, Y. T., Musco, C., Musco, C., Peng, R., and Sidford, A. Uniform sampling for matrix approximation. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, ITCS '15, pp. 181–190, 2015b.

Cohen, M. B., Musco, C., and Musco, C. Input sparsity time low-rank approximation via ridge leverage score sampling. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA'17, pp. 1758–1777, USA, 2017. Society for Industrial and Applied Mathematics.

Demmel, J., Dumitriu, I., and Holtz, O. Fast linear algebra is stable. *Numer. Math.*, 108(1):59–91, October 2007. ISSN 0029-599X.

Li, Y., Nguyen, H. L., and Woodruff, D. P. On approximating matrix norms in data streams. *SIAM Journal on Computing*, 48(6):1643–1697, 2019.

Meng, X. and Mahoney, M. W. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*, pp. 91–100, 2013.

Musco, C. *Faster linear algebra for data analysis and machine learning*. PhD thesis, MIT, 2018.

Musco, C. and Woodruff, D. P. Is input sparsity time possible for kernel low-rank approximation? In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4438–4448, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Nelson, J. and Nguyen, H. L. OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 117–126, Oct 2013.

Woodruff, D. P. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1–2):1–157, 2014.

Xu, H., Caramanis, C., and Sanghavi, S. Robust PCA via outlier pursuit. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems 23*, pp. 2496–2504. Curran Associates, Inc., 2010.

Yi, X., Park, D., Chen, Y., and Caramanis, C. Fast algorithms for robust PCA via gradient descent. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 4159–4167, Red Hook, NY, USA, 2016. Curran Associates Inc.