Optimal Analysis of Subset-Selection Based $\ell_p$ Low-Rank Approximation

Chen Dan  
Carnegie Mellon University  
cdan@cs.cmu.edu  

Hong Wang*  
Princeton University  
hongwang@mit.edu  

Hongyang Zhang*  
Toyota Technological Institute at Chicago  
honyanz@ttic.edu  

Yuchen Zhou*  
University of Wisconsin, Madison  
yuchenzhou@stat.wisc.edu  

Pradeep Ravikumar  
Carnegie Mellon University  
pradeepr@cs.cmu.edu

Abstract

We study the low rank approximation problem of any given matrix $A$ over $\mathbb{R}^{n \times m}$ and $\mathbb{C}^{n \times m}$ in entry-wise $\ell_p$ loss, that is, finding a rank-$k$ matrix $X$ such that $\|A - X\|_p$ is minimized. Unlike the traditional $\ell_2$ setting, this particular variant is NP-Hard. We show that the algorithm of column subset selection, which was an algorithmic foundation of many existing algorithms, enjoys approximation ratio $(k + 1)^{1/p}$ for $1 \leq p \leq 2$ and $(k + 1)^{1-1/p}$ for $p > 2$. This improves upon the previous $O(k + 1)$ bound for $p \geq 2$ [1]. We complement our analysis with lower bounds; these bounds match our upper bounds up to constant $1$ when $p \geq 2$. At the core of our techniques is an application of Riesz-Thorin interpolation theorem from harmonic analysis, which might be of independent interest to other algorithmic designs and analysis more broadly.

As a consequence of our analysis, we provide better approximation guarantees for several other algorithms with various time complexity. For example, to make the algorithm of column subset selection computationally efficient, we analyze a polynomial time bi-criteria algorithm which selects $O(k \log m)$ columns. We show that this algorithm has an approximation ratio of $O((k + 1)^{1/p})$ for $1 \leq p \leq 2$ and $O((k + 1)^{1-1/p})$ for $p \geq 2$. This improves over the best-known bound with an $O(k + 1)$ approximation ratio. Our bi-criteria algorithm also implies an exact-rank method in polynomial time with a slightly larger approximation ratio.

1 Introduction

Low rank approximation has wide applications in compressed sensing, numerical linear algebra, machine learning, and many other domains. In compressed sensing, low rank approximation serves as an indispensable building block for data compression. In numerical linear algebra and machine learning, low rank approximation is the foundation of many data processing algorithms, such as PCA. Given a data matrix $A \in \mathbb{F}^{n \times m}$, low rank approximation aims at finding a low-rank matrix $X$ such that $\|A - X\|_p$ is minimized. Unlike the traditional $\ell_2$ setting, this particular variant is NP-Hard. We show that the algorithm of column subset selection, which was an algorithmic foundation of many existing algorithms, enjoys approximation ratio $(k + 1)^{1/p}$ for $1 \leq p \leq 2$ and $(k + 1)^{1-1/p}$ for $p \geq 2$. This improves upon the previous $O(k + 1)$ bound for $p \geq 2$ [1]. We complement our analysis with lower bounds; these bounds match our upper bounds up to constant $1$ when $p \geq 2$. At the core of our techniques is an application of Riesz-Thorin interpolation theorem from harmonic analysis, which might be of independent interest to other algorithmic designs and analysis more broadly.

As a consequence of our analysis, we provide better approximation guarantees for several other algorithms with various time complexity. For example, to make the algorithm of column subset selection computationally efficient, we analyze a polynomial time bi-criteria algorithm which selects $O(k \log m)$ columns. We show that this algorithm has an approximation ratio of $O((k + 1)^{1/p})$ for $1 \leq p \leq 2$ and $O((k + 1)^{1-1/p})$ for $p \geq 2$. This improves over the best-known bound with an $O(k + 1)$ approximation ratio. Our bi-criteria algorithm also implies an exact-rank method in polynomial time with a slightly larger approximation ratio.

1 Introduction

Low rank approximation has wide applications in compressed sensing, numerical linear algebra, machine learning, and many other domains. In compressed sensing, low rank approximation serves as an indispensable building block for data compression. In numerical linear algebra and machine learning, low rank approximation is the foundation of many data processing algorithms, such as PCA. Given a data matrix $A \in \mathbb{F}^{n \times m}$, low rank approximation aims at finding a low-rank matrix $X$ such that $\|A - X\|_p$ is minimized. Unlike the traditional $\ell_2$ setting, this particular variant is NP-Hard. We show that the algorithm of column subset selection, which was an algorithmic foundation of many existing algorithms, enjoys approximation ratio $(k + 1)^{1/p}$ for $1 \leq p \leq 2$ and $(k + 1)^{1-1/p}$ for $p \geq 2$. This improves upon the previous $O(k + 1)$ bound for $p \geq 2$ [1]. We complement our analysis with lower bounds; these bounds match our upper bounds up to constant $1$ when $p \geq 2$. At the core of our techniques is an application of Riesz-Thorin interpolation theorem from harmonic analysis, which might be of independent interest to other algorithmic designs and analysis more broadly.

As a consequence of our analysis, we provide better approximation guarantees for several other algorithms with various time complexity. For example, to make the algorithm of column subset selection computationally efficient, we analyze a polynomial time bi-criteria algorithm which selects $O(k \log m)$ columns. We show that this algorithm has an approximation ratio of $O((k + 1)^{1/p})$ for $1 \leq p \leq 2$ and $O((k + 1)^{1-1/p})$ for $p \geq 2$. This improves over the best-known bound with an $O(k + 1)$ approximation ratio. Our bi-criteria algorithm also implies an exact-rank method in polynomial time with a slightly larger approximation ratio.

1 Introduction

Low rank approximation has wide applications in compressed sensing, numerical linear algebra, machine learning, and many other domains. In compressed sensing, low rank approximation serves as an indispensable building block for data compression. In numerical linear algebra and machine learning, low rank approximation is the foundation of many data processing algorithms, such as PCA. Given a data matrix $A \in \mathbb{F}^{n \times m}$, low rank approximation aims at finding a low-rank matrix $X$ such that $\|A - X\|_p$ is minimized. Unlike the traditional $\ell_2$ setting, this particular variant is NP-Hard. We show that the algorithm of column subset selection, which was an algorithmic foundation of many existing algorithms, enjoys approximation ratio $(k + 1)^{1/p}$ for $1 \leq p \leq 2$ and $(k + 1)^{1-1/p}$ for $p \geq 2$. This improves upon the previous $O(k + 1)$ bound for $p \geq 2$ [1]. We complement our analysis with lower bounds; these bounds match our upper bounds up to constant $1$ when $p \geq 2$. At the core of our techniques is an application of Riesz-Thorin interpolation theorem from harmonic analysis, which might be of independent interest to other algorithmic designs and analysis more broadly.

As a consequence of our analysis, we provide better approximation guarantees for several other algorithms with various time complexity. For example, to make the algorithm of column subset selection computationally efficient, we analyze a polynomial time bi-criteria algorithm which selects $O(k \log m)$ columns. We show that this algorithm has an approximation ratio of $O((k + 1)^{1/p})$ for $1 \leq p \leq 2$ and $O((k + 1)^{1-1/p})$ for $p \geq 2$. This improves over the best-known bound with an $O(k + 1)$ approximation ratio. Our bi-criteria algorithm also implies an exact-rank method in polynomial time with a slightly larger approximation ratio.

1 Introduction

Low rank approximation has wide applications in compressed sensing, numerical linear algebra, machine learning, and many other domains. In compressed sensing, low rank approximation serves as an indispensable building block for data compression. In numerical linear algebra and machine learning, low rank approximation is the foundation of many data processing algorithms, such as PCA. Given a data matrix $A \in \mathbb{F}^{n \times m}$, low rank approximation aims at finding a low-rank matrix $X$ such that $\|A - X\|_p$ is minimized. Unlike the traditional $\ell_2$ setting, this particular variant is NP-Hard. We show that the algorithm of column subset selection, which was an algorithmic foundation of many existing algorithms, enjoys approximation ratio $(k + 1)^{1/p}$ for $1 \leq p \leq 2$ and $(k + 1)^{1-1/p}$ for $p \geq 2$. This improves upon the previous $O(k + 1)$ bound for $p \geq 2$ [1]. We complement our analysis with lower bounds; these bounds match our upper bounds up to constant $1$ when $p \geq 2$. At the core of our techniques is an application of Riesz-Thorin interpolation theorem from harmonic analysis, which might be of independent interest to other algorithmic designs and analysis more broadly.

As a consequence of our analysis, we provide better approximation guarantees for several other algorithms with various time complexity. For example, to make the algorithm of column subset selection computationally efficient, we analyze a polynomial time bi-criteria algorithm which selects $O(k \log m)$ columns. We show that this algorithm has an approximation ratio of $O((k + 1)^{1/p})$ for $1 \leq p \leq 2$ and $O((k + 1)^{1-1/p})$ for $p \geq 2$. This improves over the best-known bound with an $O(k + 1)$ approximation ratio. Our bi-criteria algorithm also implies an exact-rank method in polynomial time with a slightly larger approximation ratio.
\( X \in \mathbb{F}^{n \times m} \) such that

\[
\text{OPT} = \min_{X : \text{rank}(X) \leq k} \|X - A\|.
\]

Here the field \( \mathbb{F} \) can be both \( \mathbb{R} \) and \( \mathbb{C} \). The focus of this work is on the case when \( \| \cdot \| \) is the entry-wise \( \ell_p \) norm, and we are interested in an estimate \( \hat{X} \) with a tight approximation ratio \( \alpha \) so that we have the guarantee:

\[
\|\hat{X} - A\| \leq \alpha \cdot \text{OPT},
\]

As noted earlier, such low-rank approximation is a fundamental workhorse of machine learning. The key reason to focus on approximations with respect to general \( \ell_p \) norms, in contrast to the typical \( \ell_2 \) norm, is that these general \( \ell_p \) norms are better able to capture a broader range of realistic noise in complex datasets. For example, it is well-known that the \( \ell_1 \) norm is more robust to the sparse outlier \( [2, 4] \). So the \( \ell_1 \) low-rank approximation problem is a robust version of the classic PCA which uses the \( \ell_2 \) norm and has received tremendous attentions in machine learning, computer vision and data mining \([5, 6, 7]\). While it has better empirical and statistical properties, the key bottleneck to solving the problem in (1) is computational, and is known to be NP-hard in general. For example, the \( \ell_1 \) low-rank approximation is NP-hard to solve exactly even when \( k = 1 \) \([8]\), and is even hard to approximate with large error under the Exponential Time Hypothesis \([9]\). \([10]\) proved the NP-hardness of the problem when \( p = \infty \). A recent work \([11]\) proves that the problem has no constant factor approximation algorithm running in time \( O(2^k) \) for a constant \( \delta > 0 \), assuming the correctness of Small Set Expansion Hypothesis and Exponential Time Hypothesis. The authors also proposed a PTAS (Polynomial Time Approximation Scheme) with \( (1 + \varepsilon) \) approximation ratio when \( 0 < \varepsilon < 2 \). However, the running time is as large as \( O(n^{\text{poly}(k/\varepsilon)}) \).

Many other efforts have been devoted to designing approximation algorithms in order to alleviate the computational issues of \( \ell_p \) low-rank approximation. One promising approach is to apply subgradient descent based methods or alternating minimization \([12]\). Unfortunately, the loss surface of problem (1) suffers from spurious local minima and saddle points \([13]\), which might be arbitrarily worse than OPT. Therefore, they may not work well for the low-rank approximation problem as these local searching algorithms may easily get stuck at bad stationary points without any guarantee.

Instead, we consider another line of research—the heuristic algorithm of column subset selection (CSS). Here, the algorithm proceeds by choosing the best \( k \) columns of \( A \) as an estimation of column space of \( X \) and then solving an \( \ell_p \) linear regression problem in order to obtain the optimal row space of \( X \). See Algorithm 1 for the detailed procedure. Although the vanilla form of the subset selection based algorithm also has an exponential time complexity in terms of the rank \( k \), it can be slightly modified to polynomial time bi-criteria algorithms which selects more than \( k \) columns \([1]\). Most importantly, these algorithms are easy to implement and runs fast with nice empirical performance. Thus, subset selection based algorithms might seem to effectively alleviate the computational issues of problem (1). The caveat however is that CSS might seem like a simple heuristic, with potentially a very large worst-case approximation ratio \( \alpha \).

In this paper, we show that CSS yields surprisingly reasonable approximation ratios, which we also show to be tight by providing corresponding lower bounds, thus providing a strong theoretical backing for the empirical observations underlying CSS.

Due in part to its importance, there has been a burgeoning set of recent analyses of column subset selection. In the traditional low rank approximation problem with Frobenious norm error \( (p = 2) \) case in our setting, \([14]\) showed that CSS achieves \( \sqrt{k + 1} \) approximation ratio. The authors also showed that the \( \sqrt{k + 1} \) bound is tight (both upper and lower bounds can be recovered by our analysis). \([15, 18]\) improved the running time of CSS with different sampling schemes while preserving similar approximation bounds. The CSS algorithm and its variants are also applied and analyzed under various different settings. For instance, \([19]\) and \([20]\) studied the CUR decomposition with the Frobenius norm. \([21]\) studied the CSS problem under the missing-data case. With \( \ell_1 \) error, \([22]\) studied CSS for non-negative matrices in \( \ell_1 \) error. \([23]\) gave tight approximation bounds for CSS under finite-field binary matrix setting. Furthermore, \([9]\) considered the low rank tensor approximation with the Frobenius norm.

Despite a large amount of work on the subset-selection algorithm and the \( \ell_p \) low rank approximation problem, many fundamental questions remain unresolved. Probably one of the most important open questions is: what is the tight approximation ratio \( \alpha \) for the subset-selection algorithm in the \( \ell_p \) low rank approximation problem, up to a constant factor? In \([1]\), the approximation ratio is shown to be
upper bounded by $(k + 1)$ and lower bounded by $O(k^{1-\frac{2}{p}})$ when $p > 2$. This problem becomes even more challenging when one requires the approximation ratio to be tight up to factor 1, as little was known about a direct tool to achieve this goal in general. In this work, we improve both upper and lower bounds in [1] to optimal when $p > 2$. Note that our bounds are still applicable and improve over [1] when $1 < p < 2$, but there is an $O(k^{\frac{2}{p} - 1})$ gap between the upper and lower bounds.

1.1 Our Results

The best-known approximation ratio of subset selection based algorithms for $\ell_p$, low-rank approximation is $O(k + 1)$ [1]. In this work, we give an improved analysis of this algorithm. In particular, we show that the Column Subset Selection in Algorithm 1 is a $c_{p,k}$-approximation, where

$$c_{p,k} = \begin{cases} (k + 1)^{\frac{2}{p}}, & 1 \leq p \leq 2, \\ (k + 1)^{1 - \frac{1}{p}}, & p \geq 2. \end{cases}$$

This improves over Theorem 4 in [1] which proved that the algorithm is an $O(k + 1)$-approximation, for all $p \geq 1$. Below, we state our main theorem formally:

**Theorem 1.1** (Upper bound). The subset selection algorithm in Algorithm 1 is a $c_{p,k}$-approximation.

Our proof of Theorem 1.1 is built upon novel techniques of Riesz-Thorin interpolation theorem. In particular, with the proof of special cases for $p = 1, 2, \infty$, we are able to interpolate the approximation ratio of all intermediate $p$'s. Our techniques might be of independent interest to other $\ell_p$ norm or Schatten-$p$ norm related problem more broadly. See Section 1.2 for more discussions.

We also complement our positive result of subset selection algorithm with a negative result. Surprisingly, our upper bound matches our lower bound exactly up to constant 1 for $p \geq 2$. Below, we state our negative results formally:

**Theorem 1.2** (Lower bound). There exist infinitely many different values of $k$, such that approximation ratio of any $k$-subset-selection based algorithm is at least $(k + 1)^{1 - \frac{1}{p}}$ for $\ell_p$, rank-$k$ approximation.

Note that our lower bound strictly improves the $(k + 1)^{1 - \frac{2}{p}}$ bound in [1]. The main idea of the proof can be found in Section 1.2 and we put the whole proof of Theorem 1.2 in Appendix 3.

One drawback of Algorithm 1 is that the running time scales exponentially with the rank $k$. However, it serves as an algorithmic foundation of many existing computationally efficient algorithms. For example, a bi-criteria variant of this algorithm runs in polynomial time, only requiring the rank parameter to be a little over-parameterized. Our new analysis can be applied to this algorithm as well. Below, we state our result informally:

**Theorem 1.3** (Informal statement of Theorem 1.1). There is a bi-criteria algorithm which runs in $\text{poly}(m, n, k)$ time and selects $O(k \log m)$ columns of $A$. The algorithm is an $O(c_{p,k})$-approximation algorithm.

Our next result is a computationally-efficient, exact-rank algorithm with slightly larger approximation ratio. Below, we state our result informally:

**Theorem 1.4** (Informal statement of Theorem 1.2). There is an algorithm which solves problem (1) and runs in $\text{poly}(m, n)$ time with an $O(c_{p,k}^2 k \log m)$-approximation ratio, provided that $k = O\left(\frac{\log n}{\log \log n}\right)$.

1.2 Our Techniques

In this section, we give a detailed discussion about our techniques in the proofs. We start with the analysis of approximation ratio of column subset selection algorithm.

**Remark** Throughout this paper, we state the theorems for real matrices. The results can be naturally generalized for complex matrices as well.

**Notations:** We denote by $A \in \mathbb{R}^{n \times m}$ the input matrix, and $A_i$ the $i$-th column of $A$. $A^* = UV$ is the optimal rank-$k$ approximation, where $U \in \mathbb{R}^{n \times k}$, $V \in \mathbb{R}^{k \times m}$. $\Delta_i = A_i - UV_i$ is the error.
Algorithm 1 A $c_{p,k}$ approximation to problem (1) by column subset selection.
1: **Input:** Data matrix $A$ and rank parameter $k$.
2: **Output:** $X \in \mathbb{R}^{n \times m}$ such that $\text{rank}(X) \leq k$ and $\|X - A\|_p \leq c_{p,k} \cdot \text{OPT}$.
3: for $I \in \{S \subseteq [m]; |S| = k\}$ do
   4:    $U \leftarrow A_I$.
   5:    Run $\ell_p$ linear regression over $V$ that minimizes the loss $\|A - UV\|_p$.
   6:    Let $X_I = UV$.
7: end for
8: return $X_I$ which minimizes $\|A - X_I\|_p$ for $I \in \{S \subseteq [m]; |S| = k\}$.

vector on the $i$-th column, and $\Delta_i$ the $l$-th element of vector $\Delta_i$. For any $X \in \mathbb{R}^{n \times k}$, define the error of projecting $A$ onto $X$ by $\text{Err}(X) = \min_{Y \in \mathbb{R}^{k \times m}} \|A - XY\|_p$. Let $J = (j_1, \cdots, j_k) \in [m]^k$ be a subset of $[m]$ with cardinality $k$. We denote $X_J$ as the following column subset in matrix $X$: $X_J = [X_{j_1}, X_{j_2}, \cdots, X_{j_k}]$. Similarly, we denote by $X_{dJ}$ the following column subset in the $d$-th dimension of matrix $X$: $X_{dJ} = [X_{dj_1}, X_{dj_2}, \cdots, X_{dj_k}]$. Denote by $J^*$ the column subset which gives smallest approximation error, i.e., $J^* = \arg\min_{J \in [m]^k} \text{Err}(A_J)$.

**Analysis in Previous Work:** In order to show that the column subset selection algorithm gives an $\alpha$-approximation, we need to prove that

$$\text{Err}(A_{J^*}) \leq \alpha \|\Delta\|_p. \quad (2)$$

Directly bounding $\text{Err}(A_{J^*})$ is prohibitive. In [1], the authors proved an upper bound of $(k + 1)$ in two steps. First, the authors constructed a specific $S \in [m]^k$, and upper bounded $\text{Err}(A_{J^*})$ by $\text{Err}(A_S)$. Their construction is as follows: $S$ is defined as the minimizer of

$$S = \arg\min_{J \in [m]^k} \frac{|\det(V_J)|}{\prod_{j \in J} \|\Delta_j\|_p}.$$ 

In the second step, [1] upper bounded $\text{Err}(A_S)$ by considering the approximation error on each column $A_i$, and upper bounded the $\ell_p$ distance from $A_i$ to the subspace spanned by $A_S$ using triangle inequality of $\ell_p$ distance. They showed that the distance is at most $(k + 1)$ times of $\|\Delta_i\|_p$, uniformly for all columns $i \in [m]$. Therefore, the approximation ratio is bounded by $(k + 1)$.

**Weighted Average:** In the first step, we use a so-called weighted average technique, inspired by the approach in [23]. Instead of using the error of one specific column subset as an upper bound, we use a weighted average over all possible column subsets, i.e.,

$$\text{Err}^p(A_{J^*}) \leq \sum_{J \in [m]^k} w_J \text{Err}^p(A_J),$$

where the weight $w_J$’s are carefully chosen for each column subset $J$. This weighted average technique captures more information from all possible column subsets, rather than only from one specific subset, and leads to a tighter bound.

**Riesz-Thorin Interpolation Theorem:** In the second step, unlike [1] which simply used triangle inequality to prove the upper bound, our technique leads to more refined analysis of upper bounds for the approximation error for each subset $J \in [m]^k$. With the technique of weighted average in the first step, proving a technical inequality (Lemma 2.2) concerning the determinants suffices to complete the analysis of approximation ratio. In the proof of this lemma, we introduce several powerful tools from harmonic analysis, the theory of interpolating linear operators. Riesz-Thorin theorem is a classical result in interpolation theory that gives bounds for $L^p$ to $L^q$ operator norm. In general, it is easier to prove estimates within spaces like $L^2$, $L^1$ and $L^\infty$. Interpolation theory enables us to generalize results in those spaces to some $L^p$ and $L^q$ spaces with an explicit operator norm. By the Riesz-Thorin interpolation theorem, we are able to prove the lemma by just checking the special cases $p = 1, 2, \infty$, and then interpolate the inequality for all the intermediate value of $p$’s.

**Lower Bounds:** We now discuss the techniques in proving the lower bounds. Our proof is a generalization of [14], which shows that for the special case $p = 2, \sqrt{k + 1}$ is the best possible
approximation ratio. Their proof for the lower bound is constructive: they constructed a \((k + 1) \times (k + 1)\) matrix \(A\), such that using any \(k\)-subset leads to a sub-optimal solution by a factor no less than \((1 - \varepsilon)\sqrt{k + 1}\). However, since \(\ell_p\) norm is not rotationally-invariant in general, it is tricky to generalize their analysis to other values of \(p\)'s. To resolve the problem, we use a specialized version of their construction, the perturbed Hadamard matrices (see Section 3 for details), as they have nice symmetricity and are much easier to analyze. We give an example of special case \(k = 3\) for better intuition:

\[
A = \begin{pmatrix}
\varepsilon & \varepsilon & \varepsilon & \varepsilon \\
1 & 1 & -1 & -1 \\
1 & -1 & 1 & -1 \\
1 & -1 & -1 & 1
\end{pmatrix}.
\]

Here \(\varepsilon\) is a positive constant close to 0. We note that \(A\) is very close to a rank-3 matrix: if we replace the first row by four zeros, then it becomes rank-3. Thus, the optimal rank-3 approximation error is at most \((4\varepsilon^p)^{1/p} = 4^{1/p}\varepsilon\). Now we consider the column subset selection algorithm. For example, we use the first three columns \(A_1, A_2, A_3\) to approximate the whole matrix — the error only comes from the fourth column. We can show that when \(\varepsilon\) is small, the projection of \(A_4\) to \(\text{span}(A_1, A_2, A_3)\) is very close to

\[
-A_1 - A_2 - A_3 = (-3\varepsilon, -1, -1, 1)^T.
\]

Therefore, the column subset selection algorithm achieve about \(4\varepsilon\) error on this matrix, which is a \(4^{1 - \frac{1}{p}}\) factor from being optimal. The similar construction works for any integer \(k = 2^r - 1, r \in \mathbb{Z}^+\), where the lower bound is replaced by \((k + 1)^{1 - \frac{1}{p}}\), also matches with our upper bound exactly when \(p \geq 2\).

## 2 Analysis of Approximation Ratio

In this section, we will prove Theorem 1.1. Recall that our goal is to bound \(\text{Err}(A_{J^*})\). We first introduce two useful lemmas. Lemma 2.1 gives an upper bound on approximation error by choosing a single arbitrary column subset \(A_J\). Lemma 2.2 is our main technical lemma.

**Lemma 2.1.** If \(J\) satisfies \(\det(V_J) \neq 0\), the approximation error of \(A_J\) is upper bounded by

\[
\text{Err}^p(A_J) \leq \|\Delta - \Delta J V_J^{-1} V\|_p.
\]

**Lemma 2.2.** Let \(S = \{s_{ij}\} \in \mathbb{C}^{k \times m}\) be a complex matrix, \(r = (r_1, \ldots, r_m)\) be \(m\)-dimensional complex vector, and \(T = \begin{bmatrix} r \\ S \end{bmatrix} \in \mathbb{C}^{(k+1) \times m}\), we have

\[
\sum_{J \in [m]^{k+1}} |\det(T_J)|^p \leq C_{p,k} \sum_{l=1}^m |r_l|^p \sum_{J \in [m]^k} |\det(S_J)|^p,
\]

where

\[
C_{p,k} = \epsilon_{p,k}^p = \begin{cases} 
(k + 1), & 1 \leq p \leq 2, \\
(k + 1)^{p-1}, & p \geq 2.
\end{cases}
\]

We first show that Theorem 1.1 has a clean proof using the two lemmas, as stated below.

**Proof.** of Theorem 1.1 We can WLOG assume that \(\text{rank}(V) = k\). In fact, if \(\text{rank}(A) < k\), then of course \(\text{Err}^p(A_{J^*}) = 0\) and there is nothing to prove. Otherwise if \(\text{rank}(A) \geq k\), then by the definition of \(V\), we know that \(\text{rank}(V) = k\).

We will upper bound the approximation error of the best column by a weighted average of \(\text{Err}^p(A_J)\). In other words, we are going to choose a set of non-negative weights \(w_J\) such that \(\sum_{J \in [m]^k} w_J = 1\), and upper bound \(\text{Err}^p(A_{J^*})\) by

\[
\text{Err}^p(A_{J^*}) = \min_{J \in [m]^k} \text{Err}^p(A_J) \leq \sum_{J \in [m]^k} w_J \text{Err}^p(A_J).
\]
In the following analysis, our choice of \(w_J\) will be
\[
w_J = \frac{|\det(V_J)|^p}{\sum_{L \in [m]^k} |\det(V_L)|^p}.
\]

Since \(\text{rank}(V) = k\), \(w_J\) are well-defined. We first prove
\[
|\det(V_J)|^p \text{Err}^p(A_{J'}) \leq \sum_{d=1}^n \sum_{i=1}^m \left| \det \left( \frac{\Delta_{dL}}{V_L} \right) \right|^p.
\]
where we denote \(L = (i, J) = (i, j_1, \ldots, j_k) \in [m]^{k+1}\).

In fact, when \(\det(V_J) = 0\), of course LHS of (3) = 0 ≤ RHS. When \(\det(V_J) \neq 0\), we know that \(V_J\) is invertible. By Lemma 2.1
\[
|\det(V_J)|^p \text{Err}^p(A_{J'}) \leq |\det(V_J)|^p \|\Delta - \Delta_J V_J^{-1} V_J\|_p
\]
\[
= \|\det(V_J) (\Delta - \Delta_J V_J^{-1})\|_p
\]
\[
= \sum_{i=1}^m \|\det(V_J) (\Delta_i - \Delta_J V_J^{-1} V_J)\|_p
\]
\[
= \sum_{d=1}^n \sum_{i=1}^m \left| \det \left( \frac{\Delta_{di} - \Delta_{dJ} V_J}{V_J} \right) \right|^p
\]
\[
= \sum_{d=1}^n \sum_{i=1}^m \left| \det \left( \frac{\Delta_{di}}{V_i} - \frac{\Delta_{dJ}}{V_J} \right) \right|^p.
\]
The second last equality follows from the Schur’s determinant identity. Therefore (3) holds, and
\[
\text{Err}^p(A_{J'}) \leq \sum_{J \in [m]^{k+1}} \frac{|\det(V_J)|^p}{\sum_{J \in [m]^{k+1}} |\det(V_J)|^p} \text{Err}^p(A_J)
\]
\[
\leq \sum_{d=1}^n \left( \frac{1}{\sum_{J \in [m]^{k+1}} |\det(V_J)|^p} \sum_{J \in [m]^{k+1}} \sum_{i=1}^m \left| \det \left( \frac{\Delta_{dL}}{V_L} \right) \right|^p \right)
\]
\[
= \sum_{d=1}^n \left( \frac{1}{\sum_{J \in [m]^{k+1}} |\det(V_J)|^p} \sum_{L \in [m]^{k+1}} \left| \det \left( \frac{\Delta_{dL}}{V_L} \right) \right|^p \right).
\]

By Lemma 2.2
\[
\sum_{J \in [m]^{k+1}} |\det(V_J)|^p \sum_{L \in [m]^{k+1}} \left| \det \left( \frac{\Delta_{dL}}{V_L} \right) \right|^p \leq C_{p,k} \sum_{j=1}^m |\Delta_{dj}|^p.
\]
Therefore,
\[
\text{Err}^p(A_{J'}) \leq \sum_{d=1}^n \left( C_{p,k} \sum_{j=1}^m |\Delta_{dj}|^p \right) = C_{p,k} \sum_{j=1}^m \|\Delta_j\|_p^p = C_{p,k} \text{OPT}^p,
\]
which means
\[
\text{Err}(A_{J'}) \leq C_{p,k}^{1/p} \text{OPT} = c_{p,k} \text{OPT}.
\]

Therefore, we only need to prove the two lemmas. Lemma 2.1 is relatively easy to prove.
Proof. of Lemma 2.1 Recall that by definition of $\Delta_j$, $A_i = UV_i + \Delta_i$,

$$\text{Err}^p(A_j) = \min_{Y \in \mathbb{R}^{k \times m}} \|A - A_j Y\|^p_p$$

$$\leq \|A - A_j V_j^{-1}V\|^p_p$$

$$= \|(UV + \Delta) - (UV_j + \Delta_j)V_j^{-1}V\|^p_p$$

$$= \|\Delta - \Delta_j V_j^{-1}V\|^p_p.$$  \(\square\)

The main difficulty in our analysis comes from Lemma 2.2. The proof is based on Riesz-Thorin interpolation theorem from harmonic analysis. Although the technical details in verifying a key inequality (4) are quite complicated, the remaining part which connects Lemma 2.2 to the Riesz-Thorin interpolation theorem is not difficult to understand. Below we give a proof to Lemma 2.2 without verifying (4), and leave the complete proof of (4) in the appendix.

Proof. of Lemma 2.2 We first state a simplified version of the Riesz-Thorin interpolation theorem, which is the most convenient-to-use version for our proof. The general version can be found in the Appendix.

Lemma 2.3. [Simplified version of Riesz-Thorin] Let $\Lambda : \mathbb{C}^{n_1} \times \mathbb{C}^{n_2} \rightarrow \mathbb{C}^{n_0}$ be a multi-linear operator, such that

$$\|\Lambda(a, b)\|_{p_0} \leq M_{p_0} \|a\|_{p_0} \|b\|_{p_0},$$

$$\|\Lambda(a, b)\|_{p_1} \leq M_{p_1} \|a\|_{p_1} \|b\|_{p_1},$$

for all $a \in \mathbb{C}^{n_1}, b \in \mathbb{C}^{n_2}$, then

$$\|\Lambda(a, b)\|_{p_\theta} \leq M_{p_0}^{1-\theta} M_{p_1}^\theta \|a\|_{p_0} \|b\|_{p_1},$$

holds for all $\theta \in [0, 1]$, where

$$\frac{1}{p_\theta} := \frac{1 - \theta}{p_0} + \frac{\theta}{p_1}.$$  

Riesz-Thorin theorem is a classical result in interpolation theory that gives bounds for $L^p$ to $L^q$ operator norm. In general, it is easier to prove estimates within spaces like $L^2$, $L^1$ and $L^\infty$. Interpolation theory enables us to generalize results in those spaces to some $L^p$ and $L^q$ spaces in between with an explicit operator norm. In our application, the $X_i$ is a set of $n_i$ elements and $V_i$ is $\mathbb{C}^{n_i}$, the space of functions on $n_i$ elements.

Now we prove Lemma 2.2 In fact, by symmetry, Lemma 2.2 is equivalent to

$$(k + 1)! \sum_{I \in \binom{[m]}{k+1}} |\det(T_I)|^p \leq k! C_{p,k} \sum_{i=1}^m |r_i|^p \sum_{J \in \binom{[m]}{k}} |\det(S_J)|^p.$$  

Here, $\binom{[m]}{k} = \{(i_1, \cdots, i_k) | 1 \leq i_1 < i_2 < \cdots < i_k \leq m\}$.

Taking $\frac{1}{p}$-th power on both sides, we have the following equivalent form

$$\left( \sum_{I \in \binom{[m]}{k+1}} |\det(T_I)|^p \right)^{\frac{1}{p}} \leq \frac{c_{p,k}}{(k + 1)^{\frac{1}{p}}} \left( \sum_{i=1}^m |r_i|^p \right)^{\frac{1}{p}} \left( \sum_{J \in \binom{[m]}{k}} |\det(S_J)|^p \right)^{\frac{1}{p}}.$$  

By Laplace expansion on the first row of $\det(T_I)$, we have for every $I = (i_1, \cdots, i_{k+1}) \in \binom{[m]}{k+1}$

$$\det(T_I) = \sum_{t=1}^{k+1} (-1)^{t+1} r_{i_t} \det(S_{I_{-t}}).$$

Here, $I_{-t} = (i_1, \cdots, i_{t-1}, i_{t+1}, \cdots, i_{k+1}) \in \binom{[m]}{k}$.  

$$\sum_{I \in \binom{[m]}{k+1}} |\det(T_I)|^p \leq \frac{1}{(k+1)^{\frac{1}{p}}} \left( \sum_{i=1}^m |r_i|^p \right)^{\frac{1}{p}} \left( \sum_{J \in \binom{[m]}{k}} |\det(S_J)|^p \right)^{\frac{1}{p}}.$$  

Here, $I_{-t} = (i_1, \cdots, i_{t-1}, i_{t+1}, \cdots, i_{k+1}) \in \binom{[m]}{k}$.  

7
This motivates us to define the following multilinear map \( \Lambda : \mathbb{C}[^{m}] \times \mathbb{C}[^{n}] \to \mathbb{C}[^{m+1}] \): for all \( \{a_t\}_{t \in \{m\}} \in \mathbb{C}[^{m}], \{b_j\}_{j \in \{n\}} \in \mathbb{C}[^{n}], \) and index set \( I = (i_1, \cdots, i_{k+1}) \in \{m\}_{k+1} \), \( [\Lambda(a, b)]_I \) is defined as

\[
[\Lambda(a, b)]_I = \sum_{t=1}^{k+1} (-1)^{t+1} a_t b_{I-t}.
\]

Now, by letting \( a_t = r_t, b_J = \det(S_J) \), the inequality can be written as

\[
\|\Lambda(a, b)\|_p \leq \frac{c_{p,k}}{(k+1)^\theta} \|a\|_p \|b\|_p = \max \left( 1, (k+1)^{1-\frac{\theta}{p}} \right) \|a\|_p \|b\|_p.
\]  

(4)

Let \( M_p = \max \left( 1, (k+1)^{1-\frac{\theta}{p}} \right) \), this is exactly the same form as Riesz-Thorin Theorem! We will first prove the easier cases \( p = 1, p = 2 \) and \( p = \infty \). For general values of \( p \), we denote

\[
\frac{1}{p} = \frac{1 - \theta}{p_0} + \frac{\theta}{p_1}.
\]

here, when \( p \in [1, 2) \), we choose \( p_0 = 1, p_1 = 2 \); when \( p \in [2, +\infty) \), we choose \( p_0 = 2, p_1 = +\infty \). We can easily verify that \( M_p = M_{p_0}^{1-\theta} M_{p_1}^\theta \). By using the Riesz-Thorin interpolation theorem, we only need to show (4) holds for \( p = 1, 2, \infty \).

We leave the complete proof of (4) in the appendix. \( \square \)

3 Lower Bounds

In this section, we give a proof sketch of Theorem 1.2. The proof is constructive: we prove the theorem by showing for all \( \varepsilon > 0 \), we can construct a matrix \( A(\varepsilon) \), such that selecting every \( k \) columns of \( A(\varepsilon) \) leads to an approximation ratio at least \( (k+1)^{1-\frac{\theta}{1+o(1)}} \). Then, the theorem follows by letting \( \varepsilon \to 0^+ \). Our choice of \( A(\varepsilon) \) is a perturbation of Hadamard matrices.

Throughout the proof, we assume that \( k = 2^r - 1 \), for some \( r \in \mathbb{Z}^+ \), and \( \varepsilon > 0 \) is an arbitrarily small constant. We consider the well known Hadamard matrix of order \( (k+1) = 2^r \), defined below:

\[
H^{(1)} = 1,
\]

\[
H^{(2)} = \begin{pmatrix}
H^{(2^l-1)} & H^{(2^l-1)} \\
H^{(2^l-1)} & -H^{(2^l-1)}
\end{pmatrix}, l \geq 1.
\]

Now we can define \( A(\varepsilon) \), the construction of lower bound instance: it is a perturbation of \( H \) by replacing all the entries on the first row by \( \varepsilon \), i.e.,

\[
A(\varepsilon)_{ij} = \varepsilon \quad \text{when} \quad i = 1,
\]

\[
H_{ij} \quad \text{when} \quad i \neq 1.
\]

We can see that \( A(\varepsilon) \) is close to a rank-\( k \) matrix. In fact, \( A(0) \) has rank at most \( k \). Therefore, we can upper bound OPT by

\[
\text{OPT} \leq \|A(\varepsilon) - A(0)\|_p = ((k+1)\varepsilon)^{1/p} = (k+1)^{1/r} \varepsilon.
\]  

(5)

The remaining work is to give a lower bound on the approximation error using any \( k \) columns. For simplicity of notations, we use \( A \) as shorthand for \( A(\varepsilon) \) when it’s clear from context. Say we are using all \( (k+1) \) columns except the \( j \)-th, i.e. the column subset is \( A[k+1]_{-j} \). Obviously, we achieve zero error on all the columns other than the \( j \)-th. Therefore, the approximation error is essentially the \( \ell_p \) distance from \( A_j \) to \( \text{span}(A[k+1]_{-j}) \). We can show that the \( \ell_p \) projection from \( A_j \) to \( \text{span}(A[k+1]_{-j}) \) is very close to \( \sum_{i \neq j} (-A_i) \), in other words,

\[
\text{Err}(A[k+1]_{-j}) = (1 - o(1))\|A_j - \sum_{i \neq j} (-A_i)\|_p = (1 - o(1))(k+1) \varepsilon
\]  

(6)

The theorem follows by combining (5) and (6). The complete proof can be found in the appendix.
4 Analysis of Efficient Algorithms

One drawback of the column subset selection algorithm is its time complexity - it requires $O(m^k \text{poly}(n))$ time, which is not desirable since it’s exponential in $k$. However, several more efficient algorithms \[1\] are designed based on it. Our tighter analysis on Algorithm \[1\] implies better approximation guarantees on these algorithms as well. Our improved bounds can be stated as follows:

**Theorem 4.1.** Algorithm \[2\] which runs in $\text{poly}(m,n,k)$ time and selects $k \log m$ columns, is a bi-criteria $O(c_{p,k}) = O((k + 1)^\max(1/p, 1−1/p))$ approximation algorithm.

**Theorem 4.2.** Algorithm \[3\] which runs in $\text{poly}(m,n)$ time as long as $k = O(\log n)$ is an $O(c_{p,n}^3 k \log m) = O(k^\max(1+3/p,4−3/p) \log m)$ approximation algorithm.

These results improve the previous $O(k)$ and $O(k^4 \log(n))$ bounds respectively. We include the proof and the description of Algorithm \[2\] and Algorithm \[3\] in Appendix for completeness.

References


A Riesz-Thorin Interpolation Theorem

Lemma A.1 (Riesz-Thorin interpolation theorem, see Lemma 8.5 in [32]). Let \((X_i, \mathcal{M}_i, \mu_i), i = 0, 1, 2, \cdots, n\) be measure spaces. Let \(V_i\) represent the complex vector space of simple functions on \(X_i\). Suppose that \(\Lambda : V_1 \times V_2 \times \cdots \times V_n \to V_0\), is a multi-linear operator of types \(p_0\) and \(p_1\) where \(p_0, p_1 \in [1, \infty]\), with constants \(M_0\) and \(M_1\), respectively. i.e.,
\[
\|\Lambda(f_1, f_2, \cdots, f_n)\|_{p_i} \leq M_{p_i} \|f_1\|_{p_i} \|f_2\|_{p_i} \cdots \|f_n\|_{p_i}.
\]
for \(i = 0, 1.\) Let \(\theta \in [0, 1]\) and define
\[
\frac{1}{p_\theta} := \frac{1 - \theta}{p_0} + \frac{\theta}{p_1}.
\]
Then, \(\Lambda\) is of type \(p_\theta\) with constant \(M_{p_\theta} := M_1^{1-\theta} M_0^\theta\), that is,
\[
\|\Lambda(f_1, f_2, \cdots, f_n)\|_{p_\theta} \leq M_{p_\theta} \|f_1\|_{p_\theta} \|f_2\|_{p_\theta} \cdots \|f_n\|_{p_\theta}.
\]
Lemma A.2.3 is a direct corollary of this theorem.

B Lower Bounds

In this section, we will prove Theorem 1.2. The proof is constructive: we prove the theorem by showing for all \(\varepsilon > 0\), we can construct a matrix \(A(\varepsilon)\), such that selecting every \(k\) columns of \(A(\varepsilon)\) leads to an approximation ratio at least \(\frac{k+1}{k+1}\frac{1}{1+k\varepsilon^q}\). Then, the theorem follows by letting \(\varepsilon \to 0^+\).

Our choice of \(A(\varepsilon)\) is a perturbation of Hadamard matrices, defined below. Throughout the proof of Theorem 1.2 we assume that \(k = 2^r - 1\), for some \(r \in \mathbb{Z}^+\), and \(\varepsilon > 0\) is an arbitrarily small constant.

Proof. of Theorem 1.2 We consider the well known Hadamard matrix of order \((k+1) = 2^r\), defined below:
\[
H^{(1)} = 1,
\]
\[
H^{(2^l)} = \begin{pmatrix}     H^{(2^{l-1})} & H^{(2^{l-1})} \\     H^{(2^{l-1})} & -H^{(2^{l-1})} \end{pmatrix}, \quad l \geq 1.
\]
The Hadamard matrix has the following properties: (we will use \(H\) to represent \(H^{(2^r)}\) when it’s clear from context)

- \(H_{di} = 1\) or \(H_{di} = -1\).
- All entries on the first row are ones, i.e. \(H_{1j} = 1\).
- The columns of \(H\) are pairwisely orthogonal, i.e.
\[
\sum_{d=1}^{k+1} H_{di} H_{dj} = 0
\]
holds when \(i \neq j\).

Now we can define \(A(\varepsilon)\): it is a perturbation of \(H\) by replacing all the entries on the first row by \(\varepsilon\), i.e.,
\[
A(\varepsilon)_{ij} = \varepsilon \quad \text{when} \quad i = 1, \quad H_{ij} \quad \text{when} \quad i \neq 1.
\]
We can see that $A(\varepsilon)$ is close to a rank-$k$ matrix. In fact, $A(0)$ has rank at most $k$. Also, $A(0)$ is an $2^r \times 2^r$, or equivalently, $(k + 1) \times (k + 1)$ matrix, and it has all zeros on the first row. Therefore, we can upper bound OPT by

$$\text{OPT} \leq \|A(\varepsilon) - A(0)\|_p = ((k + 1)\varepsilon^p)^{1/p} = (k + 1)^{1/p} \varepsilon.$$  

The remaining work is to give a lower bound on the approximation error using any $k$ columns. For simplicity of notations, we use $A$ as shorthand for $A(\varepsilon)$ when it’s clear from context. Say we are using all $(k + 1)$ columns except the $j$-th, i.e. the column subset is $A_{[k+1]-\{j\}}$. Obviously, we achieve zero error on all the columns other than the $j$-th. Therefore, the approximation error is essentially the $\ell_p$ distance from $A_j$ to $\text{span}(A_{[k+1]-\{j\}})$. Thus,

$$\text{Err}(A_{[k+1]-\{j\}}) = \inf_{x_i \in \mathbb{R}} \|A_j - \sum_{i \neq j} x_i A_i\|_p$$

$$= \inf_{x_i \in \mathbb{R}} \left( \sum_{d=1}^{k+1} \left| A_{dj} - \sum_{i \neq j} x_i A_{di} \right|^p \right)^{1/p}$$

$$= \inf_{x_i \in \mathbb{R}} \left( \varepsilon^p \left| 1 - \sum_{i \neq j} x_i \right|^p + \sum_{d=2}^{k+1} \left| H_{dj} - \sum_{i \neq j} x_i H_{di} \right|^p \right)^{1/p}.$$  

By Hölder’s inequality,

$$\left( \varepsilon^p \left| 1 - \sum_{i \neq j} x_i \right|^p + \sum_{d=2}^{k+1} \left| H_{dj} - \sum_{i \neq j} x_i H_{di} \right|^p \right)^{1/p} \geq \left( 1 - \sum_{i \neq j} x_i \right) + \sum_{d=2}^{k+1} H_{dj} \left( H_{dj} - \sum_{i \neq j} x_i H_{di} \right)$$

where $\frac{1}{p} + \frac{1}{q} = 1$.

We can actually show that $RHS = k + 1$.

Using the fact that $H_{1i} = H_{1j} = 1$ and $\sum_{d=1}^{k+1} H_{di} H_{dj} = 0,$

$$RHS = \left( 1 - \sum_{i \neq j} x_i \right) + \sum_{d=2}^{k+1} H_{dj} \left( H_{dj} - \sum_{i \neq j} x_i H_{di} \right)$$

$$= \left( 1 - \sum_{i \neq j} H_{1i} H_{1j} x_i \right) + \sum_{d=2}^{k+1} \left( 1 - \sum_{i \neq j} x_i H_{di} H_{dj} \right)$$

$$= \sum_{d=1}^{k+1} \left( 1 - \sum_{i \neq j} x_i H_{di} H_{dj} \right)$$

$$= (k + 1) - \sum_{i \neq j} x_i \left( \sum_{d=1}^{k+1} H_{di} H_{dj} \right)$$

$$= k + 1.$$
Now we can finally bound the approximation error

$$\text{Err}(A_{[k+1]-(j)}) = \inf_{x_i \in \mathbb{R}} \left( \varepsilon^p \left| 1 - \sum_{i \neq j} x_i \right|^p + \sum_{d=2}^{k+1} H_d - \sum_{i \neq j} x_i H_d \right)^{1/p}$$

$$\geq \frac{k+1}{(\varepsilon^{-q} + 1)^{1/q}}$$

Thus,

$$\text{Err}(A_{[k+1]-(j)}) \geq \frac{(k+1)^{1-\frac{1}{p}}}{(1 + k\varepsilon)^{1/q}}.$$ 

Note that this bound can be arbitrarily close to $(k+1)^{1-\frac{1}{p}}$ when $\varepsilon$ is small enough, thus we complete the proof. □

C Proof of Equation (4)

Now we are going to prove (4). First, we need to extend the definition of $b_J$ for all $J = (j_1, \ldots, j_k) \in [m]^k$. This definition is similar to the property of determinants.

- When $1 \leq j_1 < j_2 < \cdots < j_k \leq m$, i.e. $J \in \binom{[m]}{k}$, $b_J$ is already defined.
- When there exists $s \neq t, j_s = j_t$, define $b_J = 0$.
- Otherwise, there exists $1 \leq j'_1 < j'_2 < \cdots < j'_k \leq m$ and a permutation $\pi$, such that

$$\pi(j_1, \ldots, j_k) = (j'_1, j'_2, \ldots, j'_k).$$

Let $J' = (j'_1, j'_2, \ldots, j'_k)$. In such case, we define

$$b_J = \text{sgn}(\pi) b_{J'},$$

where $\text{sgn}(\pi)$ is the parity of $\pi$, i.e. $\text{sgn}(\pi) = 1$ if $\pi$ is an even permutation, and $\text{sgn}(\pi) = -1$ otherwise.

Note that if $J$ is a transposition (2-element exchanges) of $\tilde{J}$, then $b_J = -b_{\tilde{J}}$.

We can also define $[\Lambda(a, b)]_I$ for all $I \in [m]^{k+1}$, by

$$[\Lambda(a, b)]_I = \sum_{t=1}^{k+1} (-1)^{t+1} a_t b_{I-t}.$$ 

Here, $I-t = (i_1, \ldots, i_{t-1}, i_{t+1}, \ldots, i_{k+1}) \in [m]^{k}$. Similarly, if $I$ is a transposition (2-element exchanges) of $\tilde{I}$, then $[\Lambda(a, b)]_I = -[\Lambda(a, b)]_{\tilde{I}}$.

As mentioned before, we only need to verify (4) for the special cases $p = 1, 2, \infty$.

**Case 1:** $p = 1$. The inequality is equivalent to

$$\|\Lambda(a, b)\|_1 \leq \|a\|_1 \|b\|_1.$$ 

In fact, by the definition, we always have

$$\|\Lambda(a, b)\|_1 = \sum_{I \in \binom{[m]}{k+1}} \|[\Lambda(a, b)]_I\| = \frac{1}{(k+1)!} \sum_{I \in [m]^{k+1}} \|[\Lambda(a, b)]_I\|. $$
Therefore,

\[
\|\Lambda(a, b)\|_1 = \frac{1}{(k+1)!} \sum_{I \in [m]^{k+1}} \left| \sum_{t=1}^{k+1} (-1)^{t+1} a_i b_{I-t} \right|
\]

\[
\leq \frac{1}{(k+1)!} \sum_{I \in [m]^{k+1}} \sum_{t=1}^{k+1} |a_i| |b_{I-t}|
\]

\[
= \frac{1}{(k+1)!} (k+1) \sum_{I \in [m]^{k+1}} |a_i| |b_{I-1}|
\]

\[
= \frac{1}{k!} \sum_{i \in [m]} |a_i| \sum_{J \in \binom{m}{k}} |b_J|
\]

\[
= \sum_{i \in [m]} |a_i| \sum_{J \in \binom{m}{k}} |b_J|
\]

\[
= \|a\|_1 \|b\|_1.
\]

**Case 2:** \( p = \infty \). The inequality is equivalent to

\[
\|\Lambda(a, b)\|_\infty \leq (k+1) \|a\|_\infty \|b\|_\infty
\]

\[
\|\Lambda(a, b)\|_\infty = \max_{I \in \binom{m}{k+1}} \|\Lambda(a, b)\|_I
\]

\[
= \max_{I \in \binom{m}{k+1}} \left| \sum_{t=1}^{k+1} (-1)^{t+1} a_i b_{I-t} \right|
\]

\[
\leq \max_{I \in \binom{m}{k+1}} \sum_{t=1}^{k+1} |a_i| |b_{I-t}|
\]

\[
\leq \sum_{t=1}^{k+1} \max_{i \in [m]} |a_i| \max_{J \in \binom{m}{k}} |b_J|
\]

\[
= (k+1) \max_{i \in [m]} |a_i| \max_{J \in \binom{m}{k}} |b_J|
\]

\[
= (k+1) \|a\|_\infty \|b\|_\infty.
\]

**Case 3:** \( p = 2 \). The inequality is equivalent to

\[
\|\Lambda(a, b)\|_2 \leq \|a\|_2 \|b\|_2.
\]

\[
\|\Lambda(a, b)\|_2^2 = \sum_{I \in \binom{m}{k+1}} \|\Lambda(a, b)\|_I^2
\]

\[
= \frac{1}{(k+1)!} \sum_{I \in [m]^{k+1}} \|\Lambda(a, b)\|_I^2
\]

\[
= \frac{1}{(k+1)!} \sum_{I \in [m]^{k+1}} \left| \sum_{t=1}^{k+1} (-1)^{t+1} a_i b_{I-t} \right|^2.
\]
Note that
\[
\left| \sum_{t=1}^{k+1} (-1)^{t+1} \alpha_t b_{l,t} \right|^2 = \left( \sum_{t=1}^{k+1} (-1)^{t+1} \alpha_t b_{l,t} \right) \left( \sum_{s=1}^{k+1} (-1)^{s+1} \bar{\alpha}_s \bar{b}_{l,s} \right)
\]
\[
= \sum_{t=1}^{k+1} |\alpha_t|^2 |b_{l,t}|^2 + 2 \sum_{1 \leq t \neq s \leq k+1} (-1)^{t+s} \alpha_t b_{l,t} \bar{\alpha}_s \bar{b}_{l,s}.
\]

Therefore,
\[
(k+1)! \| \Lambda(a, b) \|_2^2 = \sum_{l \in [m]^{k+1}} \sum_{t=1}^{k+1} |\alpha_t|^2 |b_{l,t}|^2 + 2 \sum_{l \in [m]^{k+1}} \sum_{1 \leq t \neq s \leq k+1} (-1)^{t+s} \alpha_t b_{l,t} \bar{\alpha}_s \bar{b}_{l,s}.
\]

The first term can be simplified as
\[
\sum_{l \in [m]^{k+1}} \sum_{t=1}^{k+1} |\alpha_t|^2 |b_{l,t}|^2 = (k+1) \sum_{t_1 \in [m]} |\alpha_{t_1}|^2 \sum_{J \in [m]^k} |b_J|^2 = (k+1)! \|a\|_2^2 \|b\|_2^2.
\]

Therefore, we only need to prove that the second term is non-positive.

When \( t < s \),
\[
b_{l,s} = b_{(i_1, \ldots, i_{s-1}, \overline{i}_{s+1}, \ldots, i_{k+1})}
= (-1)^{t+1} b_{(i_1, \ldots, i_{s-1}, \overline{i}_{s+1}, \ldots, i_{k+1})}
= (-1)^{t+1} b_{(i_1, \overline{L}_{(t,s)})},
\]
and
\[
b_{l,t} = b_{(i_1, \ldots, i_{t-1}, \overline{i}_{t+1}, \ldots, i_{k+1})}
= (-1)^{s-2} b_{(i_1, \ldots, i_{t-1}, \overline{i}_{t+1}, \ldots, i_{k+1})}
= (-1)^{s-2} b_{(i_1, \overline{L}_{(t,s)})}.
\]

Therefore,
\[
(-1)^{t+s} b_{l,t} \bar{b}_{l,s} = -b_{(i_1, \overline{L}_{(t,s)})} \bar{b}_{(i_1, \overline{L}_{(t,s)})}.
\]

The same argument holds for the case \( t > s \). Thus, for each pair of \((t,s)\), we have
\[
\sum_{l \in [m]^{k+1}} (-1)^{t+s} \alpha_t b_{l,t} \bar{\alpha}_s \bar{b}_{l,s}
= - \sum_{l \in [m]^{k+1}} a_t \bar{\alpha}_s \bar{b}_{(i_1, \overline{L}_{(t,s)})} \bar{b}_{(i_1, \overline{L}_{(t,s)})}
= - \sum_{J \in [m]^{k+1}} \sum_{i_1=1}^{m} a_{i_1} \bar{\alpha}_s \bar{b}_{(i_1, J)} \bar{b}_{(i_1, J)}
= - \sum_{J \in [m]^{k+1}} \left( \sum_{i_1=1}^{m} a_{i_1} \bar{b}_{(i_1, J)} \right) \left( \sum_{i_2=1}^{m} \bar{a}_{i_2} \bar{b}_{(i_2, J)} \right)
= - \sum_{J \in [m]^{k+1}} \left| \sum_{i_1=1}^{m} a_{i_1} \bar{b}_{(i_1, J)} \right|^2.
\]

Thus, the second term can be simplified as
We are now ready to introduce Algorithm 2. As mentioned in [1], we can without loss of generality
We first show that if we select a set 
We can prove that Algorithm 2 from [1] runs in time 
Algorithm 2

\[
\sum_{I \in [m]^{k+1}} \sum_{1 \leq \ell \neq s \leq k+1} (-1)^{\ell+s} 2a_i b_{I-\ell} \bar{a}_s b_{I-s}
\]
\[
= 2 \sum_{1 \leq \ell \neq s \leq k+1} \sum_{I \in [m]^{k+1}} (-1)^{\ell+s} a_i b_{I-\ell} \bar{a}_s b_{I-s}
\]
\[
= -4k(k+1) \sum_{J \in [m]^{k-1}} \left| \sum_{i=1}^m a_i \bar{b}_{(i,i,J)} \right|^2 \leq 0.
\]

\[\square\]

D Analysis for A poly\((nm)\)-Time Bi-Criteria Algorithm

We can prove that Algorithm 2 from [1] runs in time poly\((nm)\) but returns O\((k \log m)\) columns of A
that can be used in place of \(U\), with an error \(O(c_{p,k})\) times the error of the best \(k\)-factorization. In
other words, it obtains more than \(k\) columns but achieves a polynomial running time. The analysis
can be derived by slightly modifying the definition and proof in [1].

Definition 1 (Approximate coverage). Let \(S\) be a subset of \(k\) column indices. We say that column
\(A_i\) is \(\lambda_{p}\)-approximately covered by \(S\) if for \(p \in [1, \infty)\) we have \(\min_{x \in \mathbb{R}^{k+1}} \|A_{x} - A_{i}\|_{p} \leq \lambda_{p}\|\Delta\|_{p}\), and for \(p = \infty, \min_{x \in \mathbb{R}^{k+1}} \|A_{x} - A_{i}\|_{\infty} \leq \lambda(k+1)\|\Delta\|_{\infty}\). If \(\lambda = 1\), we say \(A_i\) is
covered by \(S\).

We first show that if we select a set \(R\) columns of size \(2k\) uniformly at random in \(\{1, \ldots, n\} \setminus 2k\), with constant
probability we cover a constant fraction of columns of \(A\).

Lemma D.1. Suppose \(R\) is a set of \(2k\) uniformly random chosen columns of \(A\). With probability at
least \(2/9\), \(R\) covers at least a \(1/10\) fraction of columns of \(A\).

Proof. Same as the proof of Lemma 6 in [1] except that we use \(c_{p,k}\) instead of \((k+1)\) in the
approximation bounds.

\[\square\]

We are now ready to introduce Algorithm 2. As mentioned in [1], we can without loss of generality
assume that the algorithm knows a number \(N\) for which \(\|\Delta\|_{p} \leq N \leq 2\|\Delta\|_{p}\). Indeed, such a value
can be obtained by first computing \(\|\Delta\|_{2}\) using the SVD. Note that although one does not know \(\Delta\),
one does know \(\|\Delta\|_{2}\) since this is the Euclidean norm of all but the top \(k\) singular values of \(A\), which
one can compute from the SVD of \(A\). Then, note that for \(p < 2, \|\Delta\|_{2} \leq \|\Delta\|_{p} \leq n^{2-p}\|\Delta\|_{2}\), while for
\(p \geq 2, \|\Delta\|_{p} \leq \|\Delta\|_{2} \leq n^{1-2/p}\|\Delta\|_{p}\). Hence, there are only \(O(\log n)\) values of \(N\) to try, given \(\|\Delta\|_{2}\),
one of which will satisfy \(\|\Delta\|_{p} \leq N \leq 2\|\Delta\|_{p}\). One can take the best solution found by Algorithm 2
for each of the \(O(\log n)\) guesses to \(N\).

Theorem D.2. With probability at least \(9/10\), Algorithm 2 runs in time \(\text{poly}(nm)\) and returns
\(O(k \log m)\) columns that can be used as a factor of the whole matrix inducing \(\ell_{p}\) error \(O(c_{p,k}\|\Delta\|_{p}).\)
Algorithm 3: An algorithm that transforms an $O(k \log m)$-rank matrix factorization into a $k$-rank matrix factorization without inflating the error too much.

1: Input: $U \in \mathbb{R}^{n \times O(k \log m)}$, $V \in \mathbb{R}^{O(k \log m) \times m}$
2: Output: $W \in \mathbb{R}^{n \times k}$, $Z \in \mathbb{R}^{k \times m}$
3: Apply Lemma E.1 to $U$ to obtain matrix $W^0$
4: Run $\ell_p$ linear regression over $Z^0$, s.t. $\|W^0 Z^0 - UV\|_p$ is minimized
5: Apply Algorithm 1 with input $(Z^0)^T \in \mathbb{R}^{n \times O(k \log m)}$ and $k$ to obtain $X$ and $Y$
6: Set $Z \leftarrow X^T$
7: Set $W \leftarrow W^0 Y^T$
8: Output $W$ and $Z$

**Proof.** Same as the proof of Theorem 7 in [1] except that we use $e_{p,k}^m$ instead of $(k + 1)$ in the approximation bounds.

### E Analysis for $A ((k \log n)^k \text{poly}(mn))$-Time Algorithm

In this section we show how to get a rank-$k$, $O(c_{p,k} k \log m)$-approximation efficiently starting from a rank-$O(k \log m)$ approximation. This algorithm runs in polynomial time as long as $k = O \left( \frac{\log n}{\log \log n} \right)$.

Let $U$ be the columns of $A$ selected by Algorithm 2.

#### E.1 An Isoperimetric Transformation

The first step of the proof is to show that we can modify the selected columns of $A$ to span the same space but to have small distortion. For this, we need the following notion of isoperimetry.

**Definition 2** (Almost isoperimetry). A matrix $B \in \mathbb{R}^{n \times m}$ is almost-$\ell_p$-isoperimetric if for all $x$, we have

$$\frac{\|x\|_p}{2m} \leq \|Bx\|_p \leq \|x\|_p.$$  

The following lemma from [1] show that given a full rank $A \in \mathbb{R}^{n \times m}$, it is possible to construct in polynomial time a matrix $B \in \mathbb{R}^{n \times m}$ such that $A$ and $B$ span the same space and $B$ is almost-$\ell_p$-isoperimetric.

**Lemma E.1** (Lemma 10 in [1]). Given a full (column) rank $A \in \mathbb{R}^{n \times m}$, there is an algorithm that transforms $A$ into a matrix $B$ such that $\text{span} A = \text{span} B$ and $B$ is almost-$\ell_p$-isoperimetric. Furthermore the running time of the algorithm is $\text{poly}(nm)$.

#### E.2 Reducing the Rank to $k$

Here we give an analysis of Algorithm 3 from [1]. It reduces the rank of our low-rank approximation from $O(k \log m)$ to $k$. Let $\delta = \|\Delta\|_p = \text{OPT}$.

**Theorem E.2.** Let $A \in \mathbb{R}^{n \times m}$, $U \in \mathbb{R}^{n \times O(k \log m)}$, $V \in \mathbb{R}^{O(k \log m) \times m}$ be such that $\|A - UV\|_p = O(\delta)$. Then, Algorithm 3 runs in time $O((k \log m)^k (mn)^{O(1)})$ and outputs $W \in \mathbb{R}^{n \times k}$, $Z \in \mathbb{R}^{k \times m}$ such that $\|A - WZ\|_p = O((c_{p,k} k \log m) \delta)$.

**Proof.** We start by bounding the running time. Step 3 is computationally the most expensive since it requires to execute a brute-force search on the $O(k \log m)$ columns of $(Z^0)^T$. So the running time is $O((k \log m)^k (mn)^{O(1)})$.

Now we have to show that the algorithm returns a good approximation. The main idea behind the proof is that $UV$ is a low-rank approximable matrix. So after applying Lemma E.1 to $U$ to obtain a low-rank approximation for $UV$ we can simply focus on $Z^0 \in \mathbb{R}^{O(k \log m) \times n}$. Next, by applying Algorithm 1 to $Z^0$, we obtain a low-rank approximation in time $O(k \log m)^k (mn)^{O(1)}$. Finally we can use this solution to construct the solution to our initial problem.
We know by assumption that \( \|A - UV\|_p = O(c_{p,k} \delta) \). Therefore, it suffices by the triangle inequality to show \( \|UV - WZ\|_p = O(c_{p,k}^3 k \log m \delta) \). First note that \( UV = W^0 Z^0 \) since Lemma \[1\] guarantees that \( \text{span}U = \text{span}W^0 \). Hence we can focus on proving \( \|W^0 Z^0 - WZ\|_p \leq O((c_{p,k}^3 k \log m) \delta) \).

We first prove two useful intermediate steps.

**Lemma E.3.** There exist matrices \( U^* \in \mathbb{R}^{n \times k} \), \( V^* \in \mathbb{R}^{k \times m} \) such that \( \|W^0 Z^0 - U^* V^*\|_p = O(c_{p,k} \delta) \).

**Proof.** Same as the proof of Lemma 12 in \[1\] except that we use \( O(c_{p,k} \delta) \) instead of \( O(k \delta) \).

**Lemma E.4.** There exist matrices \( F \in \mathbb{R}^{O(k \log m) \times k} \), \( D \in \mathbb{R}^{k \times n} \) such that \( \|W^0 ((Z^0 - FD))\|_p = O(c_{p,k}^2 \delta) \).

**Proof.** Same as the proof of Lemma 13 in \[1\] except that we use \( O(c_{p,k} \delta) \) and \( O(c_{p,k}^2 \delta) \) instead of \( O(k \delta) \) and \( O(k^2 \delta) \).

Now from the guarantees of Lemma \[E.1\] we know that for any vector \( y \), \( \|W^0 y\|_p \leq \|y\|_p / (k \log m) \). So we have \( \|Z^0 - FD\|_p \leq O((c_{p,k}^2 k \log m) \delta) \). Thus \( \| (Z^0)^T - DT^T F^T \|_p \leq O((c_{p,k}^2 k \log m) \delta) \), so \( (Z^0)^T \) has a low-rank approximation with error at most \( O((c_{p,k}^2 k \log m) \delta) \). So we can apply Theorem 1.1 again and we know that there are \( k \) columns of \( (Z^0)^T \) such that the low-rank approximation obtained starting from those columns has error at most \( O((c_{p,k}^2 k \log m) \delta) \). We obtain such a low-rank approximation from Algorithm \[1\] with input \( (Z^0)^T \in \mathbb{R}^{n \times O(k \log m)} \) and \( k \). More precisely, we obtain an \( X \in \mathbb{R}^{n \times k} \) and \( Y \in \mathbb{R}^{k \times O(k \log m)} \) such that \( \| (Z^0)^T - XY \|_p \leq O((c_{p,k}^3 k \log m) \delta) \). Thus \( \|Z^0 - Y^T X^T\|_p \leq O((c_{p,k}^3 k \log m) \delta) \).

Now using again the guarantees of Lemma \[E.1\] for \( W^0 \), we get \( \|W^0 ((Z^0 - Y^T X^T))\|_p \leq O((c_{p,k}^3 k \log m) \delta) \). So \( \|W^0 (Z^0 - Y^T X^T)\|_p = \|W^0 Z^0 - WZ\|_p = \|UV - WZ\|_p \leq O((c_{p,k}^3 k \log m) \delta) \). By combining it with \( \|A - UV\|_p = O(c_{p,k} \delta) \) and using the Minkowski inequality, the proof is complete.