Simple and practical algorithms for $\ell_p$-norm low-rank approximation

Anastasios Kyrillidis
IBM T.J. Watson Research Center
Rice University
anastasios@rice.edu

Abstract

We propose practical algorithms for entrywise $\ell_p$-norm low-rank approximation, for $p = 1$ or $p = \infty$. The proposed framework, which is non-convex and gradient-based, is easy to implement and typically attains better approximations, faster, than state of the art.

From a theoretical standpoint, we show that the proposed scheme can attain $(1 + \varepsilon)$-OPT approximations. Our algorithms are not hyperparameter-free: they achieve the desired gaps only assuming algorithm’s hyperparameters are known apriori—or are at least approximate. I.e., our theory indicates what problem quantities need to be known, in order to get a good solution within polynomial time, and does not contradict to recent inapproximability results, as in [45].

1 Introduction

We focus on the following optimization problem:

$$\min_{U \in \mathbb{R}^{m \times r}, V \in \mathbb{R}^{n \times r}} |M - UV^\top|_p, \quad p \in \{1, \infty\}. \quad (1)$$

Here, $M \in \mathbb{R}^{m \times n}$ is a given input matrix of arbitrary rank, $r \leq \{m, n\}$ is the target rank, $(U, V)$ represent the variables such that $\text{rank}(UV^\top) \leq r$, and $|\cdot|_p$ denotes the $p$-th, entrywise, matrix norm. In words, (1) is described as “finding the factors of the best rank-$r$ approximation of $M$, with respect to the $\ell_p$-norm”. We denote such optimal factors $U^*$ and $V^*$, and their product $X^* = U^*V^*$. We focus on $p \in \{1, \infty\}$, since these instances are the most common found in practice, beyond the classic $p = 2$ (Frobenius) norm; we will use the terms “Frobenius” and “$\ell_2$” norm, interchangeably.

There are numerous applications where $\ell_1$- / $\ell_\infty$-norm low rank approximations are useful in practice. First, the $\ell_1$-norm is more robust than the $\ell_2$-norm, and is suited in problem settings where Gaussian assumptions for noise models may not apply. $\ell_1$-norm low rank applications include robust PCA applications [56, 6, 31, 32, 24, 57], computer vision tasks such as background subtraction and motion detection [52, 1, 38], detection of brain activation patterns [44], and detection of anomalous behavior in dynamic networks [44].

For the $\ell_\infty$-norm version of (1), the problem cases are only a few. [43] considers the special case of $m = n$ and $r = \min\{m, n\} - 1$ as the problem of distance to robust non-singularity. [22, 23] use the notion of $\ell_\infty$-norm low rank approximation for the maximal-volume concept in approximation, as well as for the skeleton approximation of a matrix. Finally, [17] identifies that (1) with $p = \infty$ can be used for the recovery of a low-rank matrix from a quantized $M$.

Despite the utility of (1), its solution is not straightforward. While (1) with $\ell_2$-norm has a closed-form solution via the Singular Value Decomposition (SVD), the same does not hold for $p \in \{1, \infty\}$. Additionally, it has been proved that actually finding the exact solution to (1) can be exponentially complex:

1 Closely related to the $\ell_1$-norm low-rank approximation is the problem of $\ell_1$-norm subspace recovery [30]. Briefly, it is well-known that, for $p = 2$ in (1), the SVD solution is also the solution to the dual problem: $U^* = \arg\max_{U \in \mathbb{R}^{m \times r}} |U^\top M|_2$, subject to $U^\top U = I$. $V^*$ is then set as $V^* = U^* \top M$; this can be easily proved due to the orthogonality of $U^*$ [20]. Motivated by this dual formulation, $\ell_1$-norm subspace recovery is defined as

$$U^* = \arg\max_{U \in \mathbb{R}^{m \times r}} |U^\top M|_1, \text{ subject to } U^\top U = I.$$

Algorithmic solutions to this criterion are usually greedy [30], even combinatorial [56, 57]. However, in this case, $U^*$ does not necessarily resemble with that of (1) with $p = 1$ (up to orthogonal rotations).
show that \( \ell_1 \)-norm low rank matrix approximation is NP-hard, even for \( r = 1 \); further, under the exponential time hypothesis for 3SAT problems, \([46]\) provide a \( \left(1 + \frac{1}{\log^{\gamma}(\max\{m,n\})}\right) \)-inapproximability result for some hard instances \( M \), where \( \gamma > 0 \) is an arbitrary small constant. \([17]\) proves the NP-completeness of \((1)\) for \( p = \infty \), using a reduction from not-all-equal-3SAT.

The above restrict research to only approximations of \((1)\). To the best of our knowledge only the works in \([9, 46]\) present polynomial and provably good approximation schemes: \([46]\) focuses mostly on the case of \( \ell_1 \)-norm, and proves the existence of a \( O(\log(\min\{m,n\}) \cdot \text{poly}(r)) \)-approximation scheme with \( O(\text{nnz}(M) + (m + n)\text{poly}(r)) \) computational complexity. \([9]\) extends the ideas in \([46]\) for \( \ell_p \)-norms, where \( p \in [1, \infty] \): there, the authors describe a \( \text{poly}(r) \)-approximation with \( O(\text{poly}(m,n)(r\log\max\{m,n\})^r) \) computational complexity. Both approaches are based on numerical linear algebra and sketching techniques.

Apart from the above provable schemes, there are numerous heuristics proposed for \((1)\), with no rigorous approximation guarantees. Starting with \( \ell_1 \)-norm, \([39]\) propose a coordinate descent algorithm for \((1)\), where a sequence of alternating scalar minimization sub-problems are solved using a (weighted) median filter; see also \([29]\). Previously to that work, \([26, 27]\) follow a similar approach, where each sub-problem is solved using linear or quadratic programming. Inspired by \([55]\), \([13]\) propose a \( \ell_1 \)-norm version of the Wirberg method; the resulting algorithm involves several matrix-matrix multiplications (even of size greater than the input matrix), and the solution of linear programming criteria, per iteration. Cabral et al. use Augmented Lagrange Multipliers (ALM) method and handle the weighted \( \ell_1 \)-norm low rank approximation problem in \([53]\); however, no non-asymptotic convergence guarantees are provided. We note that most of the above heuristics are designed to handle missing data in \( M \) or the case of weighted factorization; we plan to consider such cases for our future research directions. For the \( \ell_\infty \)-norm case, we mention the recent work of Gillis et al. \([17]\) that proposes a block coordinate descent method that operates in an alternating minimization fashion over subsets of variables in \((1)\).

**Our approach and main contributions:** Inspired by the recent advances on smooth non-convex optimization for matrix factorization \([47, 58, 51, 42, 16, 40, 41, 35, 34]\), we study the application of alternating gradient descent in \((1)\). Despite its NP-hardness, this paper follows a more optimistic course and works towards deciphering the components/quantities that, if known a priori, could lead to a \((1 + \epsilon)\)-approximation for \((1)\).

Our approach is based on two techniques from optimization theory: (i) the smoothing technique for non-smooth convex optimization by Nesterov \([39, 12]\) (Section 4), and (ii) the recent theoretical results on finding the global minimum of matrix factorization problems using non-convex smooth methods (Section 5); see also references above. Our theory relies on provably bounding the objective function in \( \ell_1 \) or \( \ell_\infty \)-norm by its smoothing counterpart (Sections 4), using the provable performance of the non-convex algorithm (Section 5), and properly setting up the input parameters (Section 5). Our guarantees assume that we can at least approximate the optimal function value of \((1)\), and that the optimal low-rank solution of the smoothed problem is well-conditioned; the latter assumption is required for a good initialization to be easily found. The above are summarized as:

- Under assumptions, we provide a polynomial approximation algorithm for \( p = \{1, \infty\} \) in \((1)\) that achieves a \((1 + \epsilon)\)-approximation guarantee.
- We experimentally show that our scheme outperforms in practice state-of-the-art approaches.

There are several questions that remain open and need further investigation. In Section 7 we discuss what are the advantages and disadvantages of our approach and point to possible future research directions.

## 2 Notation and assumptions

**Notation.** For matrices \( X, Y \in \mathbb{R}^{m \times n}, (X, Y) = \text{Tr}(X^T Y) \) represents their inner product and \( X \odot Y \) their Hadamard product. We represent matrix norms as follows: \( \|X\|_2 = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |X_{ij}|^2} \) denotes the Frobenius (or \( \ell_2 \)-norm), \( \|X\|_1 = \sum_{i=1}^m \sum_{j=1}^n |X_{ij}| \) denotes the entrywise \( \ell_1 \)-norm, and \( \|X\|_\infty = \max_{i,j} |X_{ij}| \) denotes the entrywise \( \ell_\infty \)-norm. For the spectral norm, we use \( \sigma_1(X) \); this also denotes the largest singular value of \( X \). For vectors, we use \( \|x\|_2 \) to denote its Euclidean \( \ell_2 \)-norm. For a differentiable function \( f(X) \) with \( X = UV^T \), the gradient of \( f \) w.r.t. \( U \) and \( V \) is \( \nabla f(X) V \) and \( \nabla f(X)^T U \), respectively.

**Assumptions.** For our discussion, we will need two well-known notions of convex analysis: (restricted) strong convexity and (restricted) Lipschitz gradient continuity.

**Definition 2.1.** Let \( f : \mathbb{R}^{m \times n} \rightarrow \mathbb{R} \) be a convex differentiable function. Then, \( f \) is (resp. restricted) gradient Lip-
schtiz continuous with parameter $L$ if $\forall X, Y \in \mathbb{R}^{m \times n}$ (resp. $\forall X, Y \in \mathbb{R}^{m \times n}$ that are at most rank-$r$):

$$f(Y) \leq f(X) + \langle \nabla f(X), Y - X \rangle + \frac{L}{2} |Y - X|^2_2.$$  \hspace{1cm} (2)

**Definition 2.2.** Let $f : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ be convex and differentiable. Then, $f$ is (resp. restricted) $\mu$-strongly convex if $\forall X, Y \in \mathbb{R}^{m \times n}$ (resp. $\forall X, Y \in \mathbb{R}^{m \times n}$ that are at most rank-$r$):

$$f(Y) \geq f(X) + \langle \nabla f(X), Y - X \rangle + \frac{\mu}{2} |Y - X|^2_2.$$  \hspace{1cm} (3)

## 3 BFGD for smooth objectives

Let us first succinctly describe the Bi-Factored Gradient Descent (BFGD) algorithm [41], upon which our proposal is based. BFGD is a non-convex gradient descent scheme for smooth problems such as:

$$\min_{U \in \mathbb{R}^{m \times r}, V \in \mathbb{R}^{n \times r}} f(UV^T),$$  \hspace{1cm} (4)

where $f$ is assumed to be convex, differentiable, and at least have Lipschitz continuous gradients. Observe that while $f$ is convex w.r.t. to any input $\in \mathbb{R}^{m \times n}$, motions over $U$ and $V$ jointly lead to non-convex optimization. Such approaches have a long history and different variants have been proposed for [41].

For the rest of this section, we denote $X = UV^T$ as the result of the factorization. Also, let $\tilde{X}^*$ be the optimal point of [41]: if rank$(\tilde{X}^*) = r$, then $\tilde{X}^* = X^*_r$; otherwise, denote its best rank-$r$ approximation (w.r.t. the $\ell_2$-norm) as $\hat{X}^*_r$.

**Algorithm 1 Bi-factored gradient descent (BFGD)**

1: **Input:** $r, T; \gamma (e.g., \frac{1}{2}), C > 0$ (e.g., $C = 1$), $L$.
2: Compute $X_0 := \frac{1}{\sqrt{\gamma} \cdot L} \cdot (-\nabla f(0, \gamma \cdot 0))$.
3: Set $U_0 \in \mathbb{R}^{m \times r}, V_0 \in \mathbb{R}^{n \times r}$ s.t. $X_0 = U_0 V_0^T$, via SVD.
4: for $i = 0$ to $T - 1$ do
5: \hspace{1cm} Set $\eta$ such that: $\eta \leq \frac{C}{\tilde{\Delta} L \|U_i, V_i\|^2_{op} + 4 + \|\nabla f(U_i, V_i)\|_2}$.
6: \hspace{1cm} If $f$ satisfies Definition 2.1
7: \hspace{1.5cm} $[U_{i+1}]_{\cdot i} = \frac{[U_i - \eta \nabla f(U_i, V_i) \cdot V_i]}{[V_i]}$
8: \hspace{1cm} if $f$ satisfies Definition 2.2
9: \hspace{1.5cm} $[U_{i+1}]_{\cdot i} = \frac{[U_i - \eta \nabla f(U_i, V_i) \cdot V_i]}{[V_i]}$
10: for $i = 0$ to $T - 1$ do
11: \hspace{1cm} Output: $\hat{X} = U_T V_T^T$.

The pseudocode for BFGD is provided in Algorithm 1 and obeys the following motions: (i) given a proper ini-

tialization $X_0 = U_0 V_0^T$, and (ii) a proper step size $\eta$. BFGD applies iteratively Rule 1 if $f$ satisfies only Definition 2.1 or Rule 2 if $f$ also satisfies Definition 2.2. The algorithm assumes an approximation of $L$—say $\hat{L}$ and see [4]—and a good initialization point $(U_0, V_0)$. For a more complete discussion of initialization $(U_0, V_0)$, we refer the reader to [41]; we briefly discuss this issue in Section 5.

An important issue in optimizing $f$ over $(U, V)$ is the existence of non-unique possible factorizations for a given $X$. We need a notion of distance to the low-rank solution $X^*_r$ over the factors. Similar to [51][41], we focus on the set of “equally-footed” factorizations:

$$\tilde{X}^*_r = \{ (\tilde{U}^*, \tilde{V}^*) : \tilde{U}^* \in \mathbb{R}^{m \times r}, \tilde{V}^* \in \mathbb{R}^{n \times r}, \tilde{U}^* \tilde{V}^T = \hat{X}^*_r, \sigma_i(\tilde{U}^*) = \sigma_i(\tilde{V}^*) = \sigma_i(\hat{X}^*_r)^{1/2}, \forall i \in [r] \}.$$  \hspace{1cm} (5)

Given a pair $(U, V)$, we define the distance to $\tilde{X}^*_r$ as:

$$\text{DIST} (U, V; \tilde{X}^*_r) = \min_{(\tilde{U}^*, \tilde{V}^*) \in \tilde{X}^*_r} \left| \left| U - \tilde{U}^* \right| \right|_2.$$  \hspace{1cm}

Algorithm 1 has local convergence guarantees, when $f$ is $\mu$-strongly convex and has $L$-Lipschitz continuous gradients, according to the following theorem.

**Theorem 3.1 (Theorem 4.4 in [41]).** Let $\kappa = L/\mu$. If the initial point $X_0 = U_0 V_0^T$, satisfies $\text{DIST}(U_0, V_0; X^*_r) \leq \sqrt{2} \sigma_1(X^*_r)^{1/2}$, then BFGD converges with rate $O(1/T)$:

$$f(U_T V_T^T) - f(\hat{U}^* \hat{V}^*^T) \leq \frac{10 \text{DIST}(U_0, V_0, \hat{X}^*_r)^2}{\eta^2}.$$  \hspace{1cm}

## 4 Charbonnier approximation and the logsumexp function

A key assumption in BFGD is that $f$ is at least once differentiable and has Lipschitz continuous gradients. Therefore, to connect BFGD with our original objective in [1], we will first approximate both the $\ell_1$ and $\ell_\infty$ entrywise matrix norms by smooth functions that have derivatives at least in two degrees. For similar approaches in optimization where non-smooth functions are substituted by smooth ones, we refer to the seminal paper of Nesterov [39] and follow-up works [12][28].

In this work, we do not focus on the most efficient step size selections: e.g., the step size considered in this work varies per iteration, and it is less efficient than a constant step size selection. In this work, we will borrow only the sublinear rate results in [41], since that result alone is sufficient to lead to polynomial algorithms for [1]. Using the linear convergence rate result in [41] is left for the extension of this work.
Function values vs. input variable.

Right: Gradient approximation.

**Approximating the entrywise \( \ell_1 \)-norm.** For the approximation of the \( \ell_1 \)-norm, we will use the Charbonnier loss function \[ h(x, \tau) = \tau \cdot \left( \frac{x}{\tau} \right)^2 + 1 - 1 \right) . \] (6)

To illustrate how a good approximation is \([1] \) to the \( \ell_1 \)-norm, see Figure\([1] \)

We now discuss about the matrix form of (6) and its properties. With a slight overload of notation, we define the matrix version of (6) as follows:

\[
h(X, \tau) = \sum_{i=1}^{m} \sum_{j=1}^{n} h(X_{ij}, \tau) := \tau \cdot \sum_{i=1}^{m} \sum_{j=1}^{n} \left( \frac{X_{ij}}{\tau} \right)^2 + 1 - 1 \right) . \] (7)

The distinction between scalar and matrix \( h \) will be apparent from the text. Gradient and Hessian information of \( h \) satisfy the following lemma; the proof is deferred to the supp. material:

**Lemma 4.1.** For any \( X \in \mathbb{R}^{m \times n} \):

- \( \nabla h(X, \tau) = \frac{1}{\tau} X \odot S \in \mathbb{R}^{m \times n} \), where \( S \in \mathbb{R}^{m \times n} \) and \( S_{ij} := \frac{2}{(X_{ij}/\tau)^2 + 1} \).
- \( \nabla^2 h(X, \tau) = \frac{1}{\tau^2} I \odot Q \in \mathbb{R}^{m \times n} \), where \( Q \in \mathbb{R}^{m \times mn} \) and \( Q_{ij} := \frac{2}{(X_{ij}/\tau)^4 + 1} \).

The above lead to the following lemma; the proof is provided in the supp. material:

**Lemma 4.2.** Function \( h \) is a convex continuously differentiable function and it has Lipschitz continuous gradients with constant \( \frac{2}{\tau} \). Moreover:

\[ |X|_1 - mn\tau \leq h(X, \tau) \leq |X|_1. \]

An alternative to the Charbonnier approximation is the Huber loss function with parameter \( \tau \) [25]:

\[
h(x, \tau) = \begin{cases} 
\frac{x^2}{2\tau}, & \text{if } |x| \leq \tau \\
|x| - \tau/2, & \text{otherwise}.
\end{cases} \] (8)

Huber loss combines a \( \ell_2 \)-norm measure for small values of \( x \) and a \( \ell_1 \)-norm like measure for large \( x \). Observe in (8) that it is only first-order differentiable; thus any computations involving second order derivatives cannot be applied. On the other hand, the Charbonnier loss function, which is also known as the “pseudo-Huber loss function”, is a smooth approximation of the Huber loss that ensures that derivatives are continuous for all degrees. W.l.o.g., we focus on the Charbonnier function.

**Approximating the entrywise \( \ell_{\infty} \)-norm.** Following similar procedure for the entrywise matrix \( \ell_{\infty} \)-norm, we will use the \( \logsumexp \) function, defined as follows:

\[
\sigma(X, \tau) = \tau \cdot \log \left( \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} e^{X_{ij}/\tau} + e^{-X_{ij}/\tau}}{2mn} \right) . \] (9)

Define matrices \( P, N \in \mathbb{R}^{m \times n} \) such that:

- \( P_{ij} = e^{X_{ij}/\tau} + e^{-X_{ij}/\tau} \) and \( N_{ij} = e^{X_{ij}/\tau} - e^{-X_{ij}/\tau} \). Then, the following lemma defines the gradient and Hessian information of the \( \logsumexp \) function; see also the supp. material:

**Lemma 4.3.** For any \( X \in \mathbb{R}^{m \times n} \):

- \( \nabla \sigma(X, \tau) = \frac{1}{\tau \cdot \log(\tau)} \cdot N \in \mathbb{R}^{m \times n} \),
- \( \nabla^2 \sigma(X, \tau) = \left( \text{diag}(\text{vec}(P)) - \frac{\text{vec}(N) \odot \text{vec}(N)^\top}{\tau \cdot \log(\tau)} \right) \in \mathbb{R}^{mn \times mn} \)

where \( \text{diag}(\cdot) : \mathbb{R}^{mn} \to \mathbb{R}^{mn \times mn} \) turns the vector input to a diagonal matrix output, \( \text{vec}(\cdot) : \mathbb{R}^{mn \times n} \to \mathbb{R}^{mn} \) turns a matrix to a vector by “stacking” its columns, and \( 1 \) denotes the all-ones matrix.

Similar to the Charbonnier approximation, we get the following lemma; the proof is in the supp. material:
Lemma 4.4. The \( \logsumexp \) function \( \sigma \) is a convex continuously differentiable function and it has Lipschitz continuous gradients with constant \( \frac{1}{r} \). Moreover:

\[
|X|_\infty - \tau \log(2mn) \leq \sigma(X, \tau) \leq |X|_\infty.
\]

5 An approximate solver for \( \ell_p \)-norm low rank approximation

The proposed schemes are provided in Algorithms 2-3 and are based on Algorithm 1 as a sub-solver. In order to hope for a good initialization, we consider the smooth versions of (1), as described in Section 4, with the added twist that we regularize further the objective and are based on Algorithm 1 as a sub-solver. In order to hope for a good initialization, we consider the smooth versions of (1), as described in Section 4, with the added twist that we regularize further the objective with a strongly convex component. I.e., we approximate (1) for \( p = 1 \) with:

\[
\min_{U \in \mathbb{R}^{m \times r}, V \in \mathbb{R}^{n \times r}} \, h(M - UV^T, \tau) + \frac{\lambda}{2} |UV^T|_2^2, \quad (10)
\]

and the case \( p = \infty \) with

\[
\min_{U \in \mathbb{R}^{m \times r}, V \in \mathbb{R}^{n \times r}} \, \sigma(M - UV^T, \tau) + \frac{\lambda}{2} |UV^T|_2^2. \quad (11)
\]

This modification asserts that both (10), (11) are strongly convex w.r.t. \( X \) with parameter \( \lambda \); see also the proof of Lemma 4.4. Observe that the smaller the \( \lambda \) parameter is, the less the "drift" from the original problem. We remind that the optimal factors of (1) are \( U^* \) and \( V^* \), and their product is denoted as \( X^* = U^*V^*^T \).

Algorithm 2 \( \ell_1 \)-norm low rank approximation solver

1: Parameters: \( r, \text{OPT} \), values of \( |X^*_2|_2 \) and \( \sigma_r(\hat{X}^*_r) \), \( \varepsilon > 0 \).
2: Set \( \tau = \frac{\varepsilon \cdot \text{OPT}}{3mn} \).
3: Set function \( T = O \left( \frac{\sigma_r(\hat{X}^*_r)}{\varepsilon \cdot \text{OPT}} \right) \).
4: Set \( \lambda = \frac{2\varepsilon \cdot \text{OPT}}{3|\hat{X}^*_r|_2^2} \).
5: Compute \( \tilde{L} = (\frac{1}{\tau} + \lambda) \).
6: Set \( f(UV^T) := h(M - UV^T, \tau) + \frac{\lambda}{2} |UV^T|_2^2 \).
7: Run Algorithm 1 \( (U_T, V_T) = \text{BFGD}(r, T, \frac{1}{4}, 1, \tilde{L}) \).

Let us first focus on the case of \( \ell_1 \)-norm and Algorithm 2.

The following theorem states that, under proper configuration of algorithm’s hyperparameters, one can achieve \((1 + \varepsilon)\text{-OPT}\) approximation guarantee.

Theorem 5.1. Let \( \hat{X} = U_TV_T^T \in \mathbb{R}^{m \times n} \) be the solution of Algorithm 2. Let the optimal function value of (1) for \( p = 1 \) be denoted as \( \text{OPT} := \min_{U,V} |M - UV^T|_1 \) and assumed known, or at least approximable. Also, assume we know \( \sigma_r(\hat{X}^*_r) \) and \( |X^*_2|_2^2 \). For user defined parameter \( \varepsilon > 0 \) and setting the Charbonnier parameter \( \tau = \frac{\varepsilon \cdot \text{OPT}}{3mn} \), and the strong convexity parameter as \( \lambda = \frac{2\varepsilon \cdot \text{OPT}}{3|\hat{X}^*_r|_2^2} \), the pair \((U_T, V_T)\) of Algorithm 2 satisfies:

\[
|M - U_TV_T^T|_1 \leq (1 + \varepsilon) \cdot \text{OPT},
\]

after \( T = O \left( \sigma_r(\hat{X}^*_r) \left( \frac{\log(mn)}{(\varepsilon \cdot \text{OPT})^2} + \frac{1}{|\hat{X}^*_r|_2^2} \right) \right) \) iterations.

The proof is provided in the appendix. In the case where OPT is only approximable, straightforward modifications lead to similar performance (where higher number of iterations required).

Analytical complexity: Let us denote the time to compute \( \nabla f(\cdot) \) as \( t_{grad} \). The initialization complexity of Algorithm 1 as well as its per iteration complexity, is \( O(t_{grad} + mnr) \), where the last term is due to either low-rank SVD calculation or matrix-matrix multiplication. Running Algorithm 1 for \( T = O \left( \frac{\sigma_r(\hat{X}^*_r)}{\varepsilon \cdot \text{OPT}} \right) \) iterations leads to an overall \( O \left( \sigma_r(\hat{X}^*_r) \cdot (t_{grad} + mnr) \right) \) time complexity.

Similarly for the case of \( p = \infty \), we use the \( \logsumexp \) function in Algorithm 3 to smooth the objective, and we obtain the following guarantees:

Algorithm 3 \( \ell_\infty \)-norm low rank approximation solver

1: Parameters: \( r, \text{OPT} \), values of \( |X^*_2|_2 \) and \( \sigma_r(\hat{X}^*_r) \), \( \varepsilon > 0 \).
2: Set \( \tau = \frac{\varepsilon \cdot \text{OPT}}{3\log(2mn)} \).
3: Set function \( T = O \left( \frac{\sigma_r(\hat{X}^*_r)}{\varepsilon \cdot \text{OPT}} \right) \).
4: Set \( \lambda = \frac{2\varepsilon \cdot \text{OPT}}{3|\hat{X}^*_r|_2^2} \).
5: Compute \( \tilde{L} = (\frac{1}{\tau} + \lambda) \).
6: Set \( f(UV^T) := h(M - UV^T, \tau) + \frac{\lambda}{2} |UV^T|_2^2 \).
7: Run Algorithm 1 \( (U_T, V_T) = \text{BFGD}(r, T, \frac{1}{4}, 1, \tilde{L}) \).

Corollary 5.2. Let \( \hat{X} = U_TV_T^T \in \mathbb{R}^{m \times n} \) be the solution of Algorithm 2. Let the optimal function value of (1) for \( p = \infty \) be denoted as OPT := \( \min_{U,V} |M - UV^T|_\infty \), and assumed known, or be at least approximable. Also, assume we know \( \sigma_r(\hat{X}^*_r) \) and \( |X^*_2|_2^2 \). For user defined approximation parameter \( \varepsilon > 0 \) and setting the \( \logsumexp \) parameter \( \tau = \frac{\varepsilon \cdot \text{OPT}}{3\log(2mn)} \), and the strong convexity parameter as \( \lambda = \frac{2\varepsilon \cdot \text{OPT}}{3|\hat{X}^*_r|_2^2} \), the pair \((U_T, V_T)\) of Algorithm 2 satisfies:

\[
|M - U_TV_T^T|_\infty \leq (1 + \varepsilon) \cdot \text{OPT},
\]

after \( T = O \left( \sigma_r(\hat{X}^*_r) \left( \frac{\log(mn)}{(\varepsilon \cdot \text{OPT})^2} + \frac{1}{|\hat{X}^*_r|_2^2} \right) \right) \) iterations.

Similar analytical complexity can be derived for Algorithm 3 and is omitted due to lack of space.
Results of similar flavor (and under similar assumptions) can be found in [28] for the problem of maximum flow. There, the authors consider non-Euclidean gradient descent algorithms for the minimization of \( \ell_\infty \)-norm over vectors, where the gradient step takes into consideration the geometry of the non-smooth objective with the use of sharp operators. We applied a similar approach for both \( p \in \{1, \infty\} \) in our setting; however, the empirical performance was prohibitive to consider a similar approach here (despite the fact that one can still achieve \( (1 + \varepsilon) \)-optimal approximation guarantees).

Some remarks regarding the above results.

**Remark 1.** Both algorithms require the knowledge of three quantities: OPT, \(|X^*|_2^2\) and \(\sigma_r(\hat{X}^*)\). While finding these values could be as difficult as the original problem [1], these values do not need to be known exactly: in particular, the algorithms imply that “for sufficiently small \( \tau \) and \( \lambda \) parameters, and for a sufficiently large number of iterations \( T \), we can find a good approximation”.

**Remark 2.** While finding the exact value of OPT is difficult, there are problem cases where this value could be easily upper bounded. E.g., consider the problem of low-rank matrix approximation from quantization, as noted in [17]: there, we know from structure that \(|M - X^*|_\infty = \text{OPT} \leq 0.5.

**Remark 3.** Form a good initialization is a key assumption for Theorem 5.7 and its corollary. Such assumptions are made also in other non-convex matrix factorization results; see [42, 55, 57] [42, 56, 57, 40, 41, 34, 54, 13]. From [41], it is known that we can easily compute such an initialization as the best rank-\( \tau \) approximation of \( M \) w.r.t. the \( \ell_2 \)-norm, via SVD. In particular, such an initialization satisfies \( \text{DIST}(U_0, V_0; \hat{X}^*) \leq \frac{\sqrt{3} \sigma_r(\hat{X}^*)}{10\sqrt{\kappa}} \), as long as \( f \) is strongly convex with condition number \( \kappa \leq 1 + \frac{\sigma_r(\hat{X}^*)^2}{4608|X^*|_2^2} \). While this condition is not easily met in theory (i.e., since \( \kappa = \frac{1 + \lambda}{1 - \lambda} \), this means that \( \tau \) should be large enough compared to \( \lambda \)), our experiments show that such an initialization performs well.

**Remark 4.** As a continuation of the above initialization, the reason we use the regularizer \( \frac{\lambda}{2} |U V^\top|_2^2 \) is to turn the smooth approximations into strongly convex functions (and thus borrow results for initialization). In practice, the proposed schemes work as well without the addition of the regularizer; and thus, knowing a priori the quantity \(|X^*|_2^2\) is not necessary in practice.

**Remark 5.** The approach we follow somewhat resembles with the approach proposed in [27]. There, the authors consider [1] for \( p = 1 \) and propose an alternating minimization scheme. Despite the similarities, there are differences with our approach: among which, we perform a single gradient descent step on \( U \) and \( V \) per iteration, for a smoothed version of [1] instead of minimizing a quadratic programming formulation per each column of \( U \) and \( V \). On the contrary, [27] handles empirically missing values and weighted low-rank matrix factorization cases; we leave this direction for future research.

6 Experiments

Our experiments include synthesized applications, in order to highlight the empirical performance of the proposed framework. We compare the algorithms in Section 5(i) with the algorithms for \( \ell_p \)-low rank approximation in [9], and (ii) with the recent heuristic in [17] for \( \ell_\infty \)-low rank approximation.

Similarly to [9, 17] and in order to guarantee fair comparison, we follow in practice the “folklore” advice for getting an initial estimate for the \( \ell_p \)-norm problem in [1] by beginning with the optimum \( \ell_2 \)-norm solution (i.e., with the low-rank SVD solution).

6.1 \( \ell_1 \)-norm approximation

We perform experiments on both real and synthetic datasets. At first, we generate data according to the recent ICML paper [9]: We use \( 20 \times 30 \) random matrices \( M \), where each entry is a uniformly random value in \([0, 1]\). Such constructions lead to full rank matrices with high-probability. We also construct matrices \( M \) of the same size with \( \{\pm 1\} \) entries, each selected with 0.5 probability. For real datasets, similar to [9], we use the FIDAP dataset\(^3\) and a word frequency dataset from UC Irvine\(^4\). The FIDAP matrix \( M \) is \( 27 \times 27 \) with 279 real asymmetric non-zero entries. The word frequency matrix \( M \) is \( 3430 \times 6906 \) with 353,160 non-zero entries.

For the synthesized datasets, we perform 10 Monte Carlo instantiations and take the median error reported. For all datasets, we are interested in computing the best rank-\( r \) approximation of each \( M \) above, w.r.t. the \( \ell_1 \)-norm and for \( r \in \{1, \ldots, 10\} \). To compare with [9], we use their suggestion and run a simplified version of Algorithm 2 in [9], where we repeatedly sample \( r \) columns, uniformly at random. We then run the \( \ell_p \)-projection (see Lemma 1 in [9]) on each sampled set and finally select the solution with the smallest \( \ell_p \)-error. For a fair contrast between the algorithms, we first run our algorithm and measure the required time; for approximately the same amount of time, we run [9].

To perform the \( \ell_p \)-projection, we use

\[^3\text{http://math.nist.gov/MatrixMarket/}
\[^4\text{data/SPARSKIT/fidap/fidap005.html}
\[^5\text{https://archive.ics.uci.edu/ml/datasets/Bag+of+Words}
In this experiment, we follow the experimental setting in [17]. We generate matrices $M \in \mathbb{R}^{100 \times 75}$ as follows: We generate $\tilde{M} = UV^\top$ where $U \in \mathbb{R}^{100 \times r}$ and $V \in \mathbb{R}^{75 \times r}$. Each $U$ and $V$ is generated i.i.d. from $N(0, 1)$. Given $\tilde{M}$, we compute the rounded version of $\tilde{M}$ such as $M = \text{round}(\tilde{M})$. This procedure guarantees that, given $M$, there is a low-rank matrix $\tilde{M}$ that satisfies $||M - \tilde{M}||_{\infty} \leq 0.5$ (since this is an hard problem, this construction gives an idea how far/close we are to a good solution).

We repeat the above procedure for $r = \{1, \ldots, 10\}$ and for 10 Monte Carlo instances. We report the minimum, mean and median values of the objective function attained and the time required. We compare our algorithms with plain SVD and the heuristics in [17].

The results are reported in Table 1. Our findings show that both our work and the algorithm in [17] perform much better (in terms of quality of solution) than plain SVD (the full set of results can be found in the appendix). Further, the algorithm in [17] has time comparable to the implementation of SVD in Matlab, while our proposed algorithm is much slower; accelerating our proposed algorithm is considered future research direction. However, while our algorithm does not succeed to find solutions with small objective value (see minimum value in table and compare our work with [17]), the median value of objective function values over 10 problem instances is lower than that of [17]. I.e., the “typical” achieved objective value is lower than that of [17].

\[\text{CVX package [14]}\]

In our algorithm, we set $r = \lambda = 10^{-3}$, and the maximum number of iterations as $T = 4 \cdot 10^4$. As mentioned above, we use the SVD initialization, and the step size is set according to Algorithm 1.

The results are provided in Figure 2. Some remarks: (i) for the synthetic cases (two leftmost columns), we observe that our approach attains a better objective function, faster, compared to [9]. Both our work and [9] is much slower than plain SVD; however, the latter gives a worse solution. (ii) for the real case (two rightmost columns), our approach is overall better in terms of objective function values; however, this is not universal; there are cases where [9] (or even SVD) gets to a better result within the same time, especially when $r$ increases. For the large matrix case, [9] with CVX do not scale well; thus omitted.

6.2 $\ell_\infty$-norm approximation

In this experiment, we follow the experimental setting in [17]. We generate matrices $M \in \mathbb{R}^{27 \times 27}$ where each entry is randomly and independently generated. Second column: $M \in \{-1, 1\}^{20 \times 30}$ where each entry is randomly and independently generated. Third column: $M \in \mathbb{R}^{3 \times 3}$ is the FIDAP matrix. Fourth column: $M \in \mathbb{R}^{3430 \times 6906}$ is the word-frequency matrix. In the latter case, the sub-solver for $\ell_p$-projection was not able to complete the task, and thus the algorithm in [9] is omitted.

\[\text{CVX package [14]}\]

\[\text{We are not aware of another standardized package for $\ell_p$-regression. To accelerate the execution of SeDuMi, we use the lowest precision set up in CVX.}\]

\[\text{We ran the algorithm in [17] for more time (repeatedly within allowed time) and picked the best minimum result. However, this did not improve the results of [17].}\]
### Table 1: Attained objective function values and execution time. Table includes minimum, mean and median values for 10 Monte Carlo instances.

| Rank | Time (sec.) | Error |
|------|-------------|-------|
|      | [min, mean, median] |       |
| 1    | [6.81e-02, 2.24e-01, 2.28e-01] | [4.91e-01, 4.93e-01, 4.93e-01] |
| 2    | [1.55e-02, 2.75e-02, 2.31e-02] | [5.33e-01, 6.00e-01, 5.96e-01] |
| 3    | [2.42e-02, 5.89e-02, 4.59e-02] | [5.22e-01, 5.63e-01, 5.44e-01] |
| 4    | [2.69e-02, 4.61e-02, 4.04e-02] | [5.24e-01, 5.66e-01, 5.42e-01] |
| 5    | [4.67e-02, 3.36e-01, 1.48e-01] | [5.04e-01, 5.36e-01, 5.26e-01] |
| 6    | [6.72e-02, 6.24e-01, 1.34e-01] | [4.98e-01, 5.20e-01, 5.22e-01] |
| 7    | [5.46e-02, 8.91e-01, 5.47e-01] | [4.90e-01, 5.14e-01, 5.11e-01] |
| 8    | [1.36e-01, 1.66e+00, 5.39e-01] | [4.81e-01, 5.15e-01, 5.02e-01] |
| 9    | [1.90e-01, 2.91e+00, 5.39e-01] | [4.73e-01, 4.98e-01, 4.89e-01] |
| 10   | [2.30e-01, 9.60e+00, 4.25e+00] | [4.59e-01, 4.97e-01, 4.79e-01] |

7 Conclusion and future work

We consider the problem of low-rank matrix approximation, w.r.t. (entrywise) \( \ell_p \)-norms, and proposed two algorithms that lead to \((1 + \varepsilon)\)-OPT approximations. Our schemes combine ideas from smoothing techniques in convex optimization, as well as recent non-convex gradient descent algorithms. Key assumption is that problem-related quantities are known or at least are approximable. Our experiments show that our scheme performs (at least) competitively with state of the art.

We have provided several possible extensions of this work. A particularly interesting open problem is that of weighted low-rank matrix approximation:

\[
\min_{U \in \mathbb{R}^{m \times r}, V \in \mathbb{R}^{n \times r}} |W \odot (M - UV^T)|_p, \quad p \in \{1, \infty\},
\]

where different assumptions on \( W \) lead to different open research questions.

References

[1] H. Aanas, R. Fisker, K. Astrom, and J. Carstensen. Robust factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1215–1225, 2002.

[2] M. Asteris, A. Kyriillidis, D. Papailiopoulos, and A. Dimakis. Bipartite correlation clustering: Maximizing agreements. In *Artificial Intelligence and Statistics*, pages 121–129, 2016.

[3] J. Barron. A more general robust loss function. *arXiv preprint arXiv:1701.03077*, 2017.

[4] S. Bhojanapalli, A. Kyriillidis, and S. Sanghavi. Dropping convexity for faster semi-definite optimization. In *29th Annual Conference on Learning Theory*, pages 530–582, 2016.

[5] R. Cabral, F. De la Torre, J. Costeira, and A. Bernardino. Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition. In *Proceedings of the IEEE International Conference on Computer Vision*, 2013.

[6] E. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[7] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. *In
state tomography via non-convex methods. arXiv preprint arXiv:1711.02524, 2017.

[34] X. Li, Z. Wang, J. Lu, R. Arora, J. Haupt, H. Liu, and T. Zhao. Symmetry, saddle points, and global geometry of nonconvex matrix factorization. arXiv preprint arXiv:1612.09296, 2016.

[35] Y. Li, Y. Liang, and A. Risteski. Recovery guarantee of non-negative matrix factorization via alternating updates. In Advances in Neural Information Processing Systems, pages 4987–4995, 2016.

[36] P. Markopoulos, G. Karystinos, and D. Pados. Some options for $\ell_1$-subspace signal processing. In Wireless Communication Systems (ISWCS 2013), Proceedings of the Tenth International Symposium on, pages 1–5. VDE, 2013.

[37] P. Markopoulos, G. Karystinos, and D. Pados. Optimal algorithms for $\ell_1$-subspace signal processing. IEEE Transactions on Signal Processing, 62(19):5046–5058, 2014.

[38] D. Meng, Z. Xu, L. Zhang, and J. Zhao. A cyclic weighted median method for $\ell_1$ low-rank matrix factorization with missing entries. In AAAI, volume 4, page 6, 2013.

[39] Y. Nesterov. Smoothing technique and its applications in semidefinite optimization. Mathematical Programming, 110(2):245–259, 2007.

[40] D. Park, A. Kyrillidis, S. Bhousanapalli, C. Caramanis, and S. Sanghavi. Provable Burer-Monteiro factorization for a class of norm-constrained matrix problems. arXiv preprint arXiv:1606.01316, 2016.

[41] D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi. Finding low-rank solutions to matrix problems, efficiently and provably. arXiv preprint arXiv:1606.03168, 2016.

[42] D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi. Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach. arXiv preprint arXiv:1609.03240, 2016.

[43] S. Poljak and J. Rohn. Checking robust nonsingularity is NP-hard. Mathematics of Control, Signals, and Systems (MCSS), 6(1):1–9, 1993.

[44] C. Qiu, N. Vaswani, B. Lois, and L. Hogben. Recursive robust PCA or recursive sparse recovery in large but structured noise. IEEE Transactions on Information Theory, 60(8):5007–5039, 2014.

[45] A. Singh and G. Gordon. A unified view of matrix factorization models. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 358–373. Springer, 2008.

[46] Z. Song, D. Woodruff, and P. Zhong. Low rank approximation with entrywise $\ell_1$-norm error. In Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, pages 688–701. ACM, 2017.

[47] R. Sun and Z.-Q. Luo. Guaranteed matrix completion via nonconvex factorization. In IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, pages 270–289, 2015.

[48] M. Tipping. Probabilistic visualisation of high-dimensional binary data. In Advances in neural information processing systems, pages 592–598, 1999.

[49] M. Tipping and C. Bishop. Probabilistic principal component analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61(3):611–622, 1999.

[50] Q. Tran-Dinh and Z. Zhang. Extended Gauss-Newton and Gauss-Newton-ADMM algorithms for low-rank matrix optimization. arXiv preprint arXiv:1606.03358, 2016.

[51] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via Procrustes flow. arXiv preprint arXiv:1507.03566, 2015.

[52] M. Turk and A. Pentland. Eigenfaces for recognition. Journal of cognitive neuroscience, 3(1):71–86, 1991.

[53] N. Veldt, A. Wirth, and D. Gleich. Correlation clustering with low-rank matrices. In Proceedings of the 26th International Conference on World Wide Web, pages 1025–1034. International World Wide Web Conferences Steering Committee, 2017.

[54] L. Wang, X. Zhang, and Q. Gu. A universal variance reduction-based catalyst for nonconvex low-rank matrix recovery. arXiv preprint arXiv:1701.02301, 2017.

[55] T. Wiberg. Computation of principal components when data are missing. In Proc. of Second Symp. Computational Statistics, pages 229–236, 1976.

[56] H. Xu, C. Caramanis, and S. Sanghavi. Robust PCA via outlier pursuit. In Advances in Neural Information Processing Systems, pages 2496–2504, 2010.

[57] X. Yi, D. Park, Y. Chen, and C. Caramanis. Fast algorithms for robust PCA via gradient descent. In Advances in Neural Information Processing Systems, pages 4152–4160, 2016.

[58] T. Zhao, Z. Wang, and H. Liu. A nonconvex optimization framework for low rank matrix estima-
[59] T. Zhou and D. Tao. GoDec: Randomized low-rank & sparse matrix decomposition in noisy case. In *International conference on machine learning*. Omnipress, 2011.
8 Proofs of lemmata

8.1 Proof of Lemma 4.1

Due to the decomposability of (7), we observe ∀X:

\[
\frac{\partial h(X, \tau)}{\partial X_{ij}} = \frac{2X_{ij}}{\tau} \left( \left( \frac{X_{ij}}{\tau} \right)^2 + 1 \right)^{-1/2} = \frac{X_{ij}}{\tau} \cdot \frac{2}{\sqrt{\left( \frac{X_{ij}}{\tau} \right)^2 + 1}}
\]

Thus, in compact form, \( \nabla h(X, \tau) = \frac{1}{\tau} \cdot X \odot S \), where \( S \) is defined in the lemma.

Regarding the Hessian information, first observe that \( \frac{\partial^2 h(X, \tau)}{\partial X_{ij} \partial X_{kl}} = \frac{\partial \left( \frac{X_{ij}}{\tau} \cdot \frac{2}{\sqrt{\left( \frac{X_{ij}}{\tau} \right)^2 + 1}} \right)}{\partial X_{ij}} = 0 \), for indices \( (i, j) \neq (l, q) \). This means that the off-diagonals of \( \nabla^2 h(X, \tau) \) are zero. For the case where \( (i, j) = (l, q) \), we have:

\[
\frac{\partial^2 h(X, \tau)}{\partial X_{ij}^2} = \frac{2}{\tau} \cdot \frac{\left( X_{ij}/\tau \right)^2 + 1 - (X_{ij}/\tau)^2}{\left( \left( X_{ij}/\tau \right)^2 + 1 \right)^{3/2}} = \frac{2}{\tau} \cdot \frac{1}{\left( \left( X_{ij}/\tau \right)^2 + 1 \right)^{3/2}}
\]

Then, \( \nabla^2 h(X, \tau) = \frac{1}{\tau} \cdot I \odot Q \), where \( Q \) is defined in the lemma.

8.2 Proof of Lemma 4.2

The first part of the lemma is easily deduced from Lemma 4.1. Observe that \( 0 \leq \nabla^2 h(X, \tau) \leq \frac{2}{\tau} I \), ∀X; that is \( h \) function is convex with Lipschitz constant \( \frac{2}{\tau} \). Moreover, by combining \( h \) with any strongly convex function \( \psi(X) \), say \( \psi(X) := \frac{1}{2} |X|^2 \), we easily observe that the composite form \( h(X, \tau) + \psi(X) \) satisfies \( \lambda I \leq \nabla^2 h(X, \tau) + \nabla^2 \psi(X) \leq \left( \frac{2}{\tau} + \lambda \right) I \); i.e., the composite form is also strongly convex.

The last part of the lemma is true because

\[
|X|_1 \geq h(X, \tau) = \sum_{i=1}^{m} \sum_{j=1}^{n} h(X_{ij}, \tau) = \tau \cdot \sum_{i=1}^{m} \sum_{j=1}^{n} \left( \sqrt{\left( \frac{X_{ij}}{\tau} \right)^2 + 1 - 1} \right) = \sum_{i=1}^{m} \sum_{j=1}^{n} \left( \sqrt{X_{ij}^2 + \tau^2 - \tau} \right) \geq \sum_{i=1}^{m} \sum_{j=1}^{n} |X_{ij}| - mn\tau = |X|_1 - mn\tau.
\]

8.3 Proof of Lemma 4.3

The proof is elementary as in Lemma 4.1 and we state it for completeness. First, observe that (9) can be re-written as follows:

\[
\sigma(X, \tau) = \tau \cdot \log \left( \frac{\text{Tr}(\mathbb{I} \cdot P)}{2mn} \right)
\]

Observe that calculating gradients with respect to \( X_{ij} \), the denominator \( 2mn \) plays no role. Following similar motions, we compute partial derivatives as:

\[
\frac{\partial \sigma(X, \tau)}{\partial X_{ij}} = \tau \cdot \frac{1}{\text{Tr}(\mathbb{I} \cdot P)} \cdot \frac{\partial (e^{X_{ij}/\tau} + e^{-X_{ij}/\tau})}{\partial X_{ij}} = \frac{1}{\text{Tr}(\mathbb{I} \cdot P)} \cdot (e^{X_{ij}/\tau} - e^{-X_{ij}/\tau})
\]
Gathering all the partial derivatives in a matrix, we get the reported result.

Computing second-order partial derivatives for \( \sigma(X, \tau) \), we distinct the cases of diagonal and off-diagonal elements. For the former, we have:

\[
\frac{\partial^2 \sigma(X, \tau)}{\partial X_{ij}^2} = \frac{1}{\tau} \frac{\text{Tr}(\mathbb{1} \cdot P) - N_{ij}^2}{\text{Tr}(\mathbb{1} \cdot P)^2}
\]

and for the latter:

\[
\frac{\partial^2 \sigma(X, \tau)}{\partial X_{ij} \partial X_{i'j'}} = -\frac{1}{\tau} \cdot N_{ij} N_{i'j'}
\]

Combining the two, we get the required result.

### 8.4 Proof of Lemma 4.4

Let us first prove convexity. By the definition of the Hessian, we want to prove

\[
\text{Tr}(\mathbb{1} \cdot P) \cdot y^\top \left( \text{diag}(\text{vec}(P)) - \frac{\text{vec}(N) \text{vec}(N)\top}{\text{Tr}(\mathbb{1} \cdot P)} \right) y \geq 0, \quad \forall y \in \mathbb{R}^{mn}.
\]

First, observe that \( \text{Tr}(\mathbb{1} \cdot P) \geq 0 \) since each element of \( P \) is positive by definition. Second, for \( P_{ij} \geq 0 \), \( \forall i, j \), it is obvious that \( \frac{\text{vec}(P) \text{vec}(P)\top}{\text{Tr}(\mathbb{1} \cdot P)} \leq \text{diag}(\text{vec}(P)) \). Thus, what is left is to prove \( y^\top (\text{vec}(N) \text{vec}(N)\top) y \leq y^\top (\text{vec}(P) \text{vec}(P)\top) y \), which is true since:

\[
y^\top (\text{vec}(N) \text{vec}(N)\top) y = \| y^\top \text{vec}(N) \|_2^2 = \sum_{i=1}^{mn} (y_i \cdot \text{vec}(N)_i)^2 \leq \sum_{i=1}^{mn} y_i^2 \cdot \text{vec}(N)_i^2 \leq \sum_{i=1}^{mn} \sum_{j=1}^{mn} y_i y_j \cdot \text{vec}(P)_i \cdot \text{vec}(P)_j \|
\]

since \( P_{ij} \geq N_{ij} \). Upper bounding the Hessian,

\[
y^\top \nabla^2 \sigma(X, \tau) y = y^\top \left( \frac{1}{\tau} \cdot \frac{1}{\text{Tr}(\mathbb{1} \cdot P)} \cdot \left( \text{diag}(\text{vec}(P)) - \frac{\text{vec}(N) \text{vec}(N)\top}{\text{Tr}(\mathbb{1} \cdot P)} \right) \right) y \leq \sum_{i=1}^{mn} \frac{y_i^2 \cdot \text{vec}(P)_i}{\tau \cdot \text{Tr}(\mathbb{1} \cdot P)} \leq \sum_{i=1}^{mn} \frac{|y_i|^2 \cdot (\sum_{i=1}^{mn} \text{vec}(P)_i)}{\tau \cdot \text{Tr}(\mathbb{1} \cdot P)} = \frac{\|y\|_2^2}{\tau}.
\]

This means that \( \sigma \) function is Lipschitz gradient continuous with constant \( \frac{1}{\tau} \). To prove the set of inequalities of the lemma, we observe:

\[
|X|_\infty \geq \sigma(X, \tau) \geq \tau \cdot \log \left( \frac{e^{|X|_\infty / \tau}}{2mn} \right) = |X|_\infty - \tau \cdot \log(2mn).
\]

### 8.5 Proof of Theorem 5.1

Using Lemma 4.2 we bound \( |M - U_T V_T^\top|_1 \) as follows:

\[
|M - U_T V_T^\top|_1 \leq h(M - U_T V_T^\top, \tau) + mn\tau \leq h(M - U_T V_T^\top, \tau) + \frac{\lambda}{2} |U_T V_T^\top|_2^2 + mn\tau
\]
Define \( f : \mathbb{R}^{m \times n} \rightarrow \mathbb{R} \) such as \( f(UV^\top) := h(M - UV^\top, \tau) + \frac{\lambda}{2} |UV^\top|^2 \). Observe that \( f \) is \( \lambda \)-strongly convex with Lipschitz continuous gradients with parameter \( \left( \frac{2}{\tau} + \lambda \right) \). By Theorem 3.1, we know that:

\[
f(U_TV_T^\top) - f(\hat{U}^*\hat{V}^*) \leq \frac{10 \cdot \text{Dist}(U_0, V_0; \hat{X}_r^*)^2}{\eta T},
\]

where \( \text{Dist}(U_0, V_0; \hat{X}_r^*) \leq \frac{\sqrt{2} \cdot \sigma_r(\hat{X}_r^*)^{1/2}}{10 \sqrt{n}} \). Combining this bound with the above, we get:

\[
|M - U_TV_T^\top|_1 \leq h(M - \hat{U}^*\hat{V}^*^\top, \tau) + \frac{\lambda}{2} |\hat{X}_r^*|^2_2 + \frac{10 \cdot \text{Dist}(U_0, V_0; \hat{X}_r^*)^2}{\eta T} + mnT
\]

(12)

We know from Lemma 4.2 that:

\[
h(M - UV^\top, \tau) \leq |M - UV^\top|_1 \quad \implies \quad h(M - UV^\top, \tau) + \frac{\lambda}{2} |UV^\top|^2 \leq |M - UV^\top|_1 + \frac{\lambda}{2} |UV^\top|^2
\]

for every \( U, V \). This further implies that:

\[
\min_{U,V} \left( h(M - UV^\top, \tau) + \frac{\lambda}{2} |UV^\top|^2 \right) \leq \min_{U,V} \left( |M - UV^\top|_1 + \frac{\lambda}{2} |UV^\top|^2 \right)
\]

\[
\quad \implies \quad h(M - \hat{U}^*\hat{V}^*^\top, \tau) + \frac{\lambda}{2} |\hat{U}^*\hat{V}^*|^2 \leq \min_{U,V} \left( |M - UV^\top|_1 + \frac{\lambda}{2} |UV^\top|^2 \right)
\]

\[
\quad \leq \quad \min_{U,V} \left( |M - UV^\top|_1 + \frac{\lambda}{2} |UV^\top|^2 \right)
\]

\[
\quad \min_{U,V} \left( |M - UV^\top|_1 + \frac{\lambda}{2} |UV^\top|^2 \right)
\]

where (i) is due to the optimality of \( \hat{U}^*, \hat{V}^* \) as the minimizer of \( f(UV^\top) := h(M - UV^\top, \tau) + \frac{\lambda}{2} |UV^\top|^2 \). (ii) is due to \( U^*, V^* \) not being necessarily the minimizers of \( \min_{U,V} \left( |M - UV^\top|_1 + \frac{\lambda}{2} |UV^\top|^2 \right) \), and (iii) \( \text{OPT} := \min_{U,V} |M - UV^\top|_1 = |M - U^*V^*^\top|_1 \). Thus, (12) becomes:

\[
|M - U_TV_T^\top|_1 \leq \text{OPT} + \frac{\lambda}{2} |X^*_r|^2_2 + \frac{10 \cdot \text{Dist}(U_0, V_0; \hat{X}_r^*)^2}{\eta T} + mnT
\]

For \( \epsilon > 0 \), setting \( \tau = \frac{\epsilon \cdot \text{OPT}}{3mn} \) we observe that \( mnT = \frac{\epsilon \cdot \text{OPT}}{3} \). Executing Algorithm 1 for \( T \geq \frac{10 \cdot \sigma_r(\hat{X}_r^*)}{3\eta \cdot \text{OPT}} \cdot \frac{3}{\eta \cdot \text{OPT}} \), we can guarantee that:

\[
\frac{10 \cdot \text{Dist}(U_0, V_0; \hat{X}_r^*)^2}{\eta T} \leq \frac{10 \cdot \sigma_r(\hat{X}_r^*)^2}{50 \eta \cdot \text{Opt} \cdot \text{OPT}} = \frac{\epsilon \cdot \text{OPT}}{3}.
\]

Finally, setting \( \lambda = \frac{2 \cdot \text{OPT}}{6 |X^*_r|^2_2} \), we obtain:

\[
|X^*_r|^2_2 = \frac{\epsilon \cdot \text{OPT}}{3}.
\]

Substituting the above in the main recursion, we get:

\[
|M - U_TV_T^\top|_1 \leq \text{OPT} + \frac{\lambda}{2} |X^*_r|^2_2 + \frac{10 \cdot \text{Dist}(U_0, V_0; \hat{X}_r^*)^2}{\eta T} + mnT
\]

\[
\leq \text{OPT} + \frac{\epsilon \cdot \text{OPT}}{3} + \frac{\epsilon \cdot \text{OPT}}{3} + \frac{\epsilon \cdot \text{OPT}}{3}
\]

\[
= (1 + \epsilon) \cdot \text{OPT}.
\]

The number of iterations \( T \) required can be further analyzed to:

\[
T \geq \frac{10 \cdot \sigma_r(\hat{X}_r^*)}{50} \cdot \frac{3}{\eta \cdot \text{OPT}} \quad (i) \quad \frac{10 \cdot \sigma_r(\hat{X}_r^*)}{50} \cdot \frac{3 \cdot O(L)}{\epsilon \cdot \text{OPT}}
\]

\[
= \frac{10 \cdot \sigma_r(\hat{X}_r^*)}{50} \cdot \frac{3 \cdot \sigma_r(\hat{X}_r^*)}{\epsilon \cdot \text{OPT}} \quad (ii) \quad 3 \cdot \sigma_r(\hat{X}_r^*) \cdot \frac{1}{\tau + \lambda}
\]

\[
= \frac{10 \cdot \sigma_r(\hat{X}_r^*)}{50} \cdot \frac{3 \cdot O \left( \frac{3mn}{\epsilon \cdot \text{OPT}} + \frac{2 \cdot \text{OPT}}{6 |X^*_r|^2_2} \right)}{\epsilon \cdot \text{OPT}} \quad (iii) \quad \frac{9mn}{(\epsilon \cdot \text{OPT})^2} + \frac{2}{1 |X^*_r|^2_2}
\]

\[
= O \left( \sigma_r(\hat{X}_r^*) \right) \cdot \left( \frac{mn}{(\epsilon \cdot \text{OPT})^2} + \frac{1}{|X^*_r|^2_2} \right)
\]
where \((i)\) is due to the definition of the step size that \(\eta = O\left(\frac{1}{T}\right)\), \((ii)\) is due to the definition \(L = \frac{1}{T} + \lambda\), \((iii)\) is obtained by substituting \(\lambda\) and \(\tau\).

### 8.6 Proof of Corollary 5.2

The proof is similar to that of Theorem 5.1. Using Lemma 4.4, we bound \(|M - U_TV_T^\top|\_\infty\) as follows:

\[
|M - U_TV_T^\top|\_\infty \leq \sigma(U_TV_T^\top, \tau) + \tau \log(2mn)
\]

Following similar motions with Theorem 5.1 and setting \(\tau = \frac{\varepsilon \cdot \text{OPT}}{3 \log(2mn)}\), and \(T\) and \(\lambda\) similar to the \(p = 1\) case, we get:

\[
|M - U_TV_T^\top|\_\infty \leq (1 + \varepsilon) \cdot \text{OPT}.
\]

The number of iterations \(T\) required follow the same motions as the proof of Theorem 5.1, with a slight difference in the definition of \(\tau\).

### 9 Connections with other related work

[10] considers probabilistic extensions of the PCA problem: starting with various generative probabilistic models, one obtains different matrix factorization objectives. The authors rely on the fundamental work of Csiszar and Tusnady [11], and propose an alternating minimization procedure; see also [49, 48].

[21, 45] show that the differences between many algorithms for matrix factorization can be viewed in terms of a small number of modeling choices. Their view unifies methods for Bregman co-clustering, LSI, non-negative matrix factorization, relational learning, to name a few.

While the bilinear factorization \(UV^\top\) is common across different problems, there are cases where even a trilinear representation is more preferable, from an interpretation perspective. Having constraints over the factors is another differentiation: An illustrative example of this case is that of matrix co-clustering where we are interested in \(M \approx C_1C_2^\top\), with \(C_1\) and \(C_2\) being matrices that denote the participation/indicator matrices. Our work is quite different to this type of factorizations (i.e., with additional constraints on the factors); we defer the reader to [35, 18, 2, 53] for some recent developments on similar subjects.

Finally, there is a recent line of work on robust PCA that further focuses on identifying the (sparse) grossly corrupted elements in \(M\); see [56, 6, 59, 31, 32, 8, 24, 57]. That line of work differs from our problem in that, our approach “models” the corruption through the penalization of the residual \(M - UV^\top\) with an \(\ell_1\)-norm, while in the aforementioned line of works, one optimizes over the residual \(S = M - UV^\top\) in order to minimize the number of “active” corruptions. In that sense our model is “simpler” as we are only interested in identifying the low rank component.

### 10 Supportive experimental results
| Rank | Time (sec.) | Error |
|------|-------------|-------|
|      | [min, mean, median] |        |
| 1    | [2.63e-03, 1.10e-02, 1.08e-02] | [8.36e-01, 9.02e-01, 9.19e-01] |
| 2    | [3.44e-03, 5.58e-03, 4.25e-03] | [7.37e-01, 8.60e-01, 8.74e-01] |
| 3    | [4.08e-03, 8.55e-03, 6.67e-03] | [6.72e-01, 7.51e-01, 7.27e-01] |
| 4    | [2.59e-03, 7.23e-03, 4.47e-03] | [6.60e-01, 7.31e-01, 7.29e-01] |
| 5    | [2.59e-03, 3.69e-03, 3.63e-03] | [6.94e-01, 7.21e-01, 7.21e-01] |
| 6    | [2.52e-03, 3.40e-03, 3.11e-03] | [6.82e-01, 7.22e-01, 7.29e-01] |
| 7    | [2.44e-03, 3.21e-03, 3.29e-03] | [6.87e-01, 7.35e-01, 7.30e-01] |
| 8    | [2.43e-03, 3.58e-03, 3.32e-03] | [6.92e-01, 7.36e-01, 7.32e-01] |
| 9    | [2.56e-03, 3.01e-03, 2.97e-03] | [7.00e-01, 7.27e-01, 7.19e-01] |
| 10   | [1.96e-03, 2.70e-03, 2.84e-03] | [6.97e-01, 7.61e-01, 7.51e-01] |

| Rank | Time (sec.) | Error |
|------|-------------|-------|
|      | [min, mean, median] |        |
| 1    | [6.81e-02, 2.24e-01, 2.28e-01] | [4.91e-01, 4.93e-01, 4.93e-01] |
| 2    | [1.55e-02, 2.75e-02, 2.31e-02] | [5.33e-01, 6.00e-01, 5.96e-01] |
| 3    | [2.42e-02, 5.89e-02, 4.59e-02] | [5.22e-01, 5.63e-01, 5.44e-01] |
| 4    | [2.69e-02, 4.61e-02, 4.04e-02] | [5.24e-01, 5.66e-01, 5.42e-01] |
| 5    | [4.67e-02, 3.36e-01, 1.48e-01] | [5.04e-01, 5.36e-01, 5.26e-01] |
| 6    | [6.72e-02, 6.24e-01, 1.34e-01] | [4.98e-01, 5.20e-01, 5.22e-01] |
| 7    | [5.46e-02, 8.91e-01, 5.47e-01] | [4.90e-01, 5.14e-01, 5.11e-01] |
| 8    | [1.36e-01, 1.66e+00, 5.39e-01] | [4.81e-01, 5.15e-01, 5.02e-01] |
| 9    | [1.90e-01, 2.91e+00, 2.56e+00] | [4.73e-01, 4.98e-01, 4.89e-01] |
| 10   | [2.30e-01, 9.60e+00, 4.25e+00] | [4.59e-01, 4.97e-01, 4.79e-01] |

| Rank | Time (sec.) | Error |
|------|-------------|-------|
|      | [min, mean, median] |        |
| 1    | [2.57e-02, 4.32e+01, 5.44e+01] | [4.99e-01, 5.82e-01, 5.01e-01] |
| 2    | [2.60e-02, 4.95e+01, 5.44e+01] | [5.04e-01, 5.49e-01, 5.07e-01] |
| 3    | [5.20e+01, 5.43e+01, 5.42e+01] | [5.06e-01, 5.10e-01, 5.10e-01] |
| 4    | [1.55e-02, 3.67e+01, 5.15e+01] | [5.05e-01, 5.90e-01, 5.10e-01] |
| 5    | [4.17e-02, 7.92e+01, 8.93e+01] | [5.07e-01, 5.33e-01, 5.13e-01] |
| 6    | [7.27e+01, 8.03e+01, 7.76e+01] | [5.02e-01, 5.08e-01, 5.09e-01] |
| 7    | [1.62e-02, 5.11e+01, 6.52e+01] | [5.08e-01, 5.84e-01, 5.08e-01] |
| 8    | [5.51e+01, 6.55e+01, 6.73e+01] | [4.95e-01, 5.09e-01, 5.02e-01] |
| 9    | [5.36e+01, 5.89e+01, 5.77e+01] | [4.78e-01, 5.06e-01, 5.06e-01] |
| 10   | [1.69e-02, 3.86e+01, 5.23e+01] | [4.69e-01, 5.94e-01, 4.75e-01] |

Table 2: Attained objective function values and execution time. Table includes minimum, mean and median values for 10 Monte Carlo instances.