Journal of Machine Learning Research 23 (2022) 1-34 Submitted 1/21; Revised 5/22; Published 9/22

Iteratively Reweighted Nuclear Norm Methods for Low-Rank Minimization with Rank Identification

Hao Wang haw309@gmail.com
Department of Statistics
University of Washington
Seattle, WA 98195-4322, USA

Ye Wang* two@cs.berkeley.edu
Division of Computer Science
University of California
Berkeley, CA 94720-1776, USA

Editor: My editor

Abstract

The nonsmooth and nonconvex regularized low-rank matrix minimization problems arise in numerous important applications in imaging science and machine learning research due to their excellent recovery performance. A popular form is the Schatten-p norm regularized problem, which typically consist of minimizing a sum of a smooth function and the Schatten-p norm of the matrix. In this paper, we propose iteratively reweighted nuclear norm methods for solving Schatten-p regularization problems, which have two major novelties. The first novelty of this work is the rank identification property possessed by the proposed methods, meaning the correct rank can be detected in finite iterations. The second one is the adaptively updating strategy for smoothing parameters to automatically fix those parameters associated with the 0 singular values as constants after detecting the correct rank, and drive the rest parameters quickly to 0. In this way, the algorithm behaves like solving smooth problems. This distinguishes our work from all other existing iteratively reweighted methods for low-rank optimization. Based on this, an extrapolated accelerated version of our method is proposed. Global convergence ensures that every limit point of the iterates is a critical point, and local convergence rate analysis is also derived under Kurdyka-Łojasiewicz property. Numerical experiments on synthetic and real data are performed to demonstrate the efficiency of our algorithm and its superiority over contemporary methods.

Keywords: low-rank minimization, weighted Nuclear norm, Schatten-p norm, Kurdyka-Łojasiewicz property, extrapolation acceleration, rank identification

1 Introduction

In this paper, we mainly consider the Schatten-p norm minimization problem

$$\min_{X\in\mathbb{R}^{m\times n}} F(X) := f(X) + \lambda\|X\|_p^p, \tag{P}$$

where $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ is continuously differentiable and $p \in (0, 1)$, $\lambda > 0$ is a regularization parameter. Notice that the nonconvex Schatten-p norm $\|X\|_p$ with $p \in (0, 1)$ is not a real matrix norm. This problem arises in an incredibly wide range of settings throughout science and applied mathematics Chiang et al. (2018); Jun et al. (2019); Lee and Kim (2016); Pal and Jain (2023); Tong et al. (2021), and is a general approximation to the well-known low-rank matrix minimization problem

$$\min_{X \in \mathbb{R}^{m \times n}} \operatorname{rank}(X) \quad \text{s.t.}\, \mathcal{A}(X) = b, \tag{1.1}$$

where X is the considered low-rank matrix in $\mathbb{R}^{m \times n}$, $\mathcal{A}(\cdot)$ is a linear operator and b is a given measurement in $\mathbb{R}^N$ or $\mathbb{R}^{m \times n}$. Problem (1.1) has appeared in the literature of diverse fields including quality-of-service (QoS) prediction Luo et al. (2017), recommender systems Lee and Kim (2016) machine learning Amit et al. (2007); Indyk et al. (2019) and image processing Huang et al. (2014); Zhao et al. (2020). Despite of various applications, (1.1) is a NP-hard problem Hu et al. (2021)

Traditionally, many researchers suggested to use the nuclear norm to the rank of X, leading to the convex relaxation introduced in Recht et al. (2010) which minimizes the nuclear norm over the same constraints:

$$\min_{X \in \mathbb{R}^{m \times n}} \|X\|_* \quad \text{s.t.}\, \mathcal{A}(X) = b. \tag{1.2}$$

Theoretical analysis Fazel (2002) shows that the low-rank matrix can be exactly recovered under mild conditions with high probability by solving (1.2). Thus, nuclear norm based model has recently attracted significant attention Wang et al. (2021c) and numerous numerical methods have been proposed for (1.2), such as singular value thresholding (SVT) Cai et al. (2010), accelerated proximal gradient algorithms (APG) Toh and Yun (2010) and accelerated inexact soft-impute (AIS-Impute) Wang et al. (2019).

While nuclear norm has achieved success in several practical applications, it suffers a well-documented shortcoming that all singular values are simultaneously minimized. While in real data, larger singular values generally quantify the main information users want to preserve Hu et al. (2021). In addition, nuclear norm could result in significantly biased estimates that cannot achieve reliable recovery with the least observations Wen et al. (2018). To further improve the generalization of the prediction performance, and/or enhance the robustness of the solution the sparsity or low-rankness of the solution, nonconvex sparsity-inducing techniques have been employed in the past decades Wang et al. (2021b). The advantages of non-convex relaxations over nuclear norm are first shown in Mohan and Fazel (2012); Nie et al. (2012) for dealing with the matrix completion problems, which generalize the nuclear norm minimization to Schatten-p norm[1] minimization Hu et al. (2021). Moreover, many other nonconvex relaxation functions of the similar property with Schatten-p norm are also considered in Lu et al. (2014); Sun et al. (2017). Another nonconvex relaxation technique over nuclear norm only penalizes the larger singular values or a part of the chosen singular values, such as the Capped Nuclear Norm (CNN) Sun et al. (2013) the Truncated Nuclear Norm (TNN) Hu et al. (2012); Zhang et al. (2012), and the Partial Sum Nuclear Norm (PSNN) developed in Oh et al. (2013). Gu et al. Gu et al. (2014) propose a Weighted Nuclear Norm (WNN):

$$\|X\|_{\mathbf{w}*} = \sum_{i=1}^{m} w_i \sigma_i(X), \tag{1.3}$$

1. Schatten-p norm is a quasi-norm if $0 < p < 1$.

where $\mathbf{w} = [w_1, \cdots, w_m]^\top \in \mathbb{R}^m_{++}$ is a vector of weights value assigned to $\boldsymbol{\sigma}(X)$ and $\boldsymbol{\sigma}(X)$ is the vector of eigenvalues of X. In this case, TNN, PSNN and CNN can be considered as the special cases of WNN Hu et al. (2021). However, $\|X\|_{\mathbf{w}*}$ is convex and forms a matrix norm if and only if $w_1, \cdots, w_m$ are arranged in descending order Sun et al. (2017).

In this paper, we propose an Iteratively Reweighted Nuclear norm algorithm with Rank Identification (IRNRI) to solve (P). We first add perturbation parameters ϵ_i to each singular value of the matrix to smooth the Schatten-p norm,

$$F(X;\epsilon) := f(X) + \lambda \sum_{i=1}^{m} (\sigma_i(X) + \epsilon_i)^p. \tag{1.4}$$

Obviously, $F(X;0) = F(X)$. At X^k, we select ϵ^k and solve the following subproblem

$$\min_{X \in \mathbb{R}^{m\times n}} L(X; X^k, \epsilon^k) := \frac{\beta}{2}\|X - (X^k - \frac{1}{\beta}\nabla f(X^k))\|_F^2 + \lambda \sum_{i=1}^{m} w_i^k \sigma_i(X), \tag{1.5}$$

where $\beta > 0$ and $w_i^k = w(\sigma_i(X^k), \epsilon_i^k) = p(\sigma_i(X^k) + \epsilon_i^k)^{p-1}$. The perturbation vector ϵ is then driven to 0 as the algorithm proceeds. An adaptively updating strategy for ϵ is also designed, such that it can automatically terminate the update for ϵ_i associated with the zero singular value and drive those associated with the positive eigenvalues quickly to zero. Most importantly, this updating strategy keep the weights in a ascending order. In this case, the subproblem is nonconvex, but it has a closed-form optimal solution Sun et al. (2017). Our algorithm is designed in a way such that after finite iterations, the algorithm can automatically detect those zero eigenvalues in the optimal solution, meaning the algorithm eventually behaves like solving a smooth problem in the manifold.. Based on this, the local convergence rate can be easily derived and an accelerated version using extrapolation technique is also proposed to further improve the local convergence rate.

Let us review the related literatures to our work. Several iteratively reweighted nuclear norm algorithms have appeared in the past decade for solving low-rank minimization problems. The first related work Lai et al. (2013) is the iterative reweighted method proposed by Yin et al for solving unconstrained Schatten-p regularized least squares minimization with prescribed rank K, i.e., $f(X) = \frac{1}{2}\|\mathcal{A}(X) - b\|_2^2$, where $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^{mn}$ is a linear operator. At the kth iteration, it relaxes Schatten-p norm $\|X\|_p^p = \operatorname{tr}(X^\top X)^{p/2}$ to $\operatorname{tr}(X^\top X + \epsilon_k^2 I)^{p/2}$ by adding the perturbation parameter $\epsilon_k > 0$ and then at each iteration approximates $\operatorname{tr}\left((X^\top X + \epsilon_k^2 I)^{p/2}\right) \approx \operatorname{tr}(W^k X^\top X)$ with $W^k = ((X^k)^\top X^k + \epsilon_k^2 I)^{p/2-1}$. The perturbation parameter is updated by the $(K+1)$th singular value $\epsilon_{k+1} = \min\{\epsilon_k, \alpha\sigma_{K+1}(X^{k+1})\}$. A upper bound between the limit points and the K-rank optimal solution is then derived under RIP condition. Wu's et al. Wu et al. (2018) also use the same reweighting technique. Different from these methods, our work uses $\|X\|_{\mathbf{w}*}$ instead of $\operatorname{tr}(X^\top X + \epsilon_k^2 I)^{p/2}$ to approximate $\|X\|_p^p$, and we propose a novel adaptively updating strategy for ϵ.

The most related work is the proximal iteratively reweighted nuclear norm algorithm which proposed by Sun et al. in Sun et al. (2017). PIRNN adds a prescribed perturbation parameter $\epsilon > 0$ to each singular value $\sigma_i(X)$ and *fixed* it during the iteration of the algorithm. Therefore, it indeed solves the relaxed problem (1.4) as its goal, and large values of ϵ can smooth out many local minimizers Wang et al. (2022) and may not approximate the original problem (P) well. It is believed that (1.4) may approximate the original problem (P) well for sufficiently small ϵ; however, in this case, the subproblems may suffer from ill-condition for too small ϵ, causing the algorithm easily trapped into bad local minimizers. As a stark contrast, our method is designed for the original problem (P) in the sense that the perturbation parameter is automatically driven to 0 so that the iterates can successfully recover the optimal solution of (P). It should be stressed that our updating strategy is designed such that ϵ is decreased to zero in an appropriate speed to maintain the convergence rate of the overall algorithm and the well-posedness of the subproblems.

The immediate predecessor of our work, to the best of our knowledge, is the Iteratively Reweighted Nuclear Norm (IRNN) Lu et al. (2014) algorithm proposed by Lu et al. and its acceleration (AIRNN) Phan and Nguyen (2021) introduced by Phan et al., whose core strategy is the extrapolation technique and the computation the SVD of a smaller matrix at each iteration. IRNN considers a general concave singular value function $g(\sigma_i(X))$ for the regularization term. It first calculates the supergradient of Schatten-p norm $w_i^k \in \partial g(\boldsymbol{\sigma}(X^k))$ and uses it as the weight to form the subproblem (1.5). In contrast to our method, this method does not involve the perturbation parameter ϵ; therefore, the weight may tend to extreme values as the associated singular value is close to zero. (As for zero singular value, this method uses an extreme large constant as the weight). We suspect this might be the reason for the observation "IRNN may decrease slowly since the upper bound surrogate may be quite loose" reported in Lu et al. (2014). The biggest difference of our algorithm to IRNN, AIRNN and other contemporary reweighted nuclear norm methods is the model identification property possessed by our algorithm, meaning the algorithm can identify the optimal rank after finite iterations. We elaborate this in the next subsection.

1.1 Rank identification

The major novelty of our work is the rank identification property of the proposed method, which is an extention for vector optimization. In sparse optimization such as the Lasso or the support-vector machine, problems generally generate solutions onto a low-complexity model such as solutions of the same supports. For example of LASSO, a solution x^* has typically only a few nonzeros coefficients: it lies on the reduced space composed of the nonzeros components (the support) of x^*. model identification relates to answering the question that whether an algorithm can identify the low-complexity model in finite iterations. model identification is a useful tool in analyzing the behavior of algorithms and has attracted many attentions in the past decades in the research of machine learning algorithms in vector optimization. For example, coordinate descent for convex sparse regularization problems Klopfenstein et al. (2020); Massias et al. (2018) are proved to have model identification and the convergence analysis is easily derived under this property. In the last few years, proximal gradient algorithm have been shown Hare (2011); Liang et al. (2014, 2017) the model identification for the ℓ_1 regularized problem. Recently, the iteratively reweighted ℓ_1 minimization for the ℓ_p regularized problem are also shown Wang et al. (2022, 2021a) to have model identification property.

As for matrix optimization, however, model/rank identification property has not appeared as a major tool for designing algorithms and analyzing the properties—the major motivation of our work. Our algorithm is designed to possess this property, meaning the singular values of the generated iterates satisfy $\sigma_i(X^k) = 0, i \in \mathcal{Z}^*$ and $\sigma_i(X^k) > 0, i \in \mathcal{I}^*$ for all sufficiently large k, where $\mathcal{Z}^*$ is the set of indices corresponding the the zero eigenvalues in the optimal solution and $\mathcal{I}^*$ corresponds the nonzero singular values. Based on this, the adaptively updating strategy of ϵ can be easily designed to drive $\epsilon_i, i \in \mathcal{I}^*$ quickly to zero and automatically cease the updating for $\epsilon_i, i \in \mathcal{Z}^*$. Intuitively, this means the algorithm behaves like solving a smooth problem in the low-complexity model, making the convergence analysis easily derived and the acceleration technique easily applied. To the best of our knowledge, this idea of designing an algorithm with model/rank identification property is the first for matrix optimization problems.

1.2 Contribution

We summarize our main contributions in the following.

- We design algorithms of model identification property for solving low-rank matrix optimization, which can successfully identify the rank of optimal solutions within finite iterations.
- We design an adaptively updating strategy for driving the perturbation parameters ϵ_i^k to zero, which can automatically identify the parameters associated with the zero and nonzero singular values and use different update strategies accordingly.
- We incorporate the extrapolation techniques into our algorithm to further improve the performance of it.
- Global convergence and local convergence rate under the Kurdyka-Łojasiewicz (KL) property are derived for proposed algorithms.

1.3 Notation

Before introducing the proposed algorithm, let us first recall some basic notations that will be used in the sequel.

For any $N \in \mathbb{N}$ we use $[N]$ to represent the set of integers $\{1, 2, \cdots, N\}$. The set $\mathbb{R}^N$ is the real N-dimensional Educlidean space with $\mathbb{R}^N_+$ being the positive orthant in $\mathbb{R}^N$ and $\mathbb{R}^N_{++}$ the interior of $\mathbb{R}_+$. The ℓ_p-norm of a vector is $\|x\|_p = \left(\sum_{i=1}^N |x_i|^p\right)^{1/p}$. The Hadamard product is $(x \circ y)_i = x_i y_i, x, y \in \mathbb{R}^N$. For $X \in \mathbb{R}^{m\times n}$ (assuming $m \le n$ for convenience), the Frobenius norm of X is denoted by $\|X\|_F$, namely, $\|X\|_F = \sqrt{\operatorname{tr}(X^\top X)}$, where $\operatorname{tr}(\cdot)$ denotes the trace of a matrix. The Frobenius inner product is $\langle X, Y\rangle = \operatorname{tr}\left(X^\top Y\right)$. For $x \in \mathbb{R}^N$, let $\operatorname{diag}(x)$ denote the diagonal matrix with entries $\operatorname{diag}(x)_{i,i} = x_i$ for $i = 1, \cdots, N$. The full singular value decomposition (SVD) Van Loan (1976) of X is

$$X = U \operatorname{diag}(\boldsymbol{\sigma}(X))\, V^\top, \tag{1.6}$$

where $U \in \mathbb{R}^{m\times m}, V \in \mathbb{R}^{n\times n}$ are unitary matrices, $\boldsymbol{\sigma}(X)$ is the singular value vector of X, and the singular value vector satisfies $\sigma_1(X) \ge \cdots \ge \sigma_r(X) > 0$, $\sigma_{r+1}(X) = \cdots = \sigma_m(X) = 0$ with $\operatorname{rank}(X) = r$. The thin SVD of X is

$$X = U_r \operatorname{diag}(\boldsymbol{\sigma}_r(X)) V_r^\top, \tag{1.7}$$

where U_r and V_r are the first r columns of U and V, respectively, and $\boldsymbol{\sigma}_r(X) = \mathrm{diag}([\sigma_1(X), \cdots, \sigma_r(X)])$.

For a lower semi-continuous function $J : \mathbb{R}^N \to (-\infty, +\infty]$, its domain denoted by $\mathrm{dom}(J) := \{x \in \mathbb{R}^N : J(x) \leq +\infty\}$. We introduce following definitions as well as some useful properties in variational analysis.

Definition 1 (subdifferentials Lewis and Sendov (2005)) *Let $J : \mathbb{R}^N \to (-\infty, +\infty]$ be a proper lower semi-continuous function,*

1. *For a given $x \in \mathrm{dom}(J)$, the Fréchet subdifferential of J at x, written as $\hat{\partial}J(x)$, is the set of all vectors $z \in \mathbb{R}^N$ which satisfy*

$$\lim_{y \neq x} \inf_{y \to x} \frac{J(y) - J(x) - \langle z, y - x\rangle}{\|y - x\|_2} \geq 0.$$

When $x \notin \mathrm{dom}(J)$, we set $\hat{\partial}J(x) = \emptyset$.

2. *The (limiting) subdifferential (or simply the subdifferential) of J at $x \in \mathbb{R}^N$, is defined through the following closure process*

$$\partial J(x) := \left\{ z \in \mathbb{R}^N : \exists x^k \to x, J(x^k) \to J(x) \text{ and } z^k \in \hat{\partial} J(x^k) \to z \text{ as } k \to \infty \right\}.$$

The necessary condition Sun et al. (2017) for $x \in \mathbb{R}^N$ to be a minimizer of $J(x)$ is

$$0 \in \partial J(x). \tag{1.8}$$

Lemma 1 (Limiting subdifferential of singular value function Lewis and Sendov (2005)) *Let $f : \mathbb{R}^N \to \mathbb{R}$ be an absolutely symmetric function i.e., $f(x_1; \cdots; x_N) = f(|x_{\pi(1)}|; \cdots; |x_{\pi(N)}|)$ holds for any permutation π for $[N]$. Then the limiting subdifferential of a singular value function $f \circ \sigma$ at a matrix $X \in \mathbb{R}^{m \times n}$ is given by the formula*

$$\partial[f \circ \sigma](X) = U \mathrm{diag}\left(\partial f[\boldsymbol{\sigma}(X)]\right) V^\top,$$

with $X = U \mathrm{diag}(\boldsymbol{\sigma}(X)) V^\top$ being the SVD of X.

proposition 1 *Sun et al. (2017) Let $g : \mathbb{R}_+ \to \mathbb{R}$ be an absolutely symmetric function, the subdifferential of $\sum_i g[\sigma_i(X)]$ at a matrix $X \in \mathbb{R}^{m \times n}$ is given by*

$$\partial \left\{ \sum_i g[\sigma_i(X)] \right\} = U \mathrm{diag}\left(\xi_1 g'[\sigma_1(X)], \cdots, \xi_m g'[\sigma_m(X)]\right) V^\top,$$

with $X = U \mathrm{diag}(\boldsymbol{\sigma}(X)) V^\top$ being the SVD of X, and $\xi_i \in \partial|\sigma_i(X)|$.

Since the $|x|^p$ is absolutely symmetric, the limiting subdifferential of $\|X\|_p^p$ is given by

$$\partial \|X\|_p^p = \left\{ U \Sigma V^\top : \Sigma = \mathrm{diag}(\partial |\boldsymbol{\sigma}(X)|_p^p) \right\},$$

where $\partial|\boldsymbol{\sigma}(X)|_p^p = (\sigma_1(X)^{p-1}, \cdots, \sigma_r(X)^{p-1}, \xi_1, \cdots, \xi_{m-r}), \xi_j \in \mathbb{R}, j = 1 \cdots, m-r$, and $r = \mathrm{rank}(X)$.

Definition 2 (Critical point Sun et al. (2017)) *We call X as a critical point of $F(\cdot)$ if it satisfies $\mathbf{0}_{m\times n} \in \partial F(X)$. The set of critical points of $F(X)$ at X, denoted by* $\text{crit}(F)$, *is given by*

$$\text{crit}(F) := \left\{ X : \mathbf{0}^{m\times n} \in \nabla f(X) + \lambda p U \text{diag}(\partial |\boldsymbol{\sigma}(X)|_p^p) V^\top \right\}. \tag{1.9}$$

The Kurdyka-Łojasiewicz (KL) property is a useful tool for analyzing the behavior of non-convex optimization algorithms Li and Pong (2018) , as it provides a quantitative estimate of the rate at which a real-valued function decreases toward its critical points. Many algorithms analyze the convergence based on the KL property such as proximal gradient descent Nikolova and Tan (2018), iteratively reweighted ℓ_1 Wang et al. (2021a). The definition of Kurdyka-ojasiewicz property is given below.

Definition 3 (Kurdyka-Łojasiewicz inequality Garrigos (2015)) *A locally Lipschitz and proper lower semi-continuous function over a Hilbert space $F : \mathcal{H} \to (-\infty, \infty)$ is said to have the Kurdyka-Łojasiewicz property at $X \in \text{dom}(\partial F)$ if and only if there exist $\eta \in (0, +\infty]$, a neighborhood $\mathbb{U}(X;\rho)$ of X and a continuous function $\Phi(s) = cs^{1-\theta}$ for some $c > 0$ and $\theta \in [0, 1)$ such that for all $\bar{X} \in \mathbb{U}(X;\rho) \cap \left\{F(X) < F(\bar{X}) < F(X) + \eta\right\}$, the Kurdyka-Łojasiewicz inequality holds*

$$\Psi' \left(F(X) - F(\hat{X}) \right) \text{dist}\left(0, \partial F(X)\right) \geq 1. \tag{1.10}$$

If $F(\cdot)$ satisfies the KL inequality at every point $X \in \mathcal{H}$ we call $F(\cdot)$ a KL function. ξ

2 Algorithm description

In this section, we introduce the framework of Iteratively Reweighted Nuclear norm with Rank Identification (IRNRI) algorithm for solving (P). We make the following assumption about $f(\cdot)$ at first.

Assumption 1 *$f(\cdot) : \mathbb{R}^{m\times n} \to \mathbb{R}$ is L-Lipschitz differentiable with*

$$\|\nabla f(X) - \nabla f(Y)\|_F \leq L_f \|X - Y\|_F,$$

for any $X, Y \in \mathbb{R}^{m\times n}$.

Assumption 2 *The initial point (X^0, ϵ^0) and β are chosen such that the level set $\mathcal{L}(F^0) := \{X \,|F(X) \leq F^0\}$ is bounded where $F^0 = F(X^0;\epsilon^0)$ and $\beta \geq L_f$, and the initial perturbation vector ϵ^0 satisfies $\epsilon_1^0 \geq \cdots \geq \epsilon_N^0 > 0$.*

At the kth iteration, we denote $\boldsymbol{\sigma}^k = \boldsymbol{\sigma}(X^k)$ and $\sigma_i^k = \sigma_i(X^k)$ for convenience. In order to track the singular values of the iterates, we define two index sets:

$$\mathcal{I}^k = \mathcal{I}(X^k) = \left\{ i : \sigma_i^k > 0 \right\} \quad \text{and} \quad \mathcal{Z}^k = \mathcal{Z}(X^k) = \left\{ i : \sigma_i^k = 0 \right\}.$$

The next iterate X^{k+1} is then obtained by solving the subproblem

$$\min_{X\in\mathbb{R}^{m\times n}} L(X; X^k, \epsilon^k) := G_k(X) + \lambda \sum_{i=1}^{m} w_i^k \sigma_i(X), \tag{2.1}$$

where $G_k(X) := \nabla f(X^k)^\top (X - X^k) + \frac{\beta}{2}\|X - X^k\|_F^2$ with $\beta > 0$. The weights are defined by $\boldsymbol{w}^k = [w_1^1, \ldots, w_m^k]^T$ with $w_i^k = w\left(\sigma_i^k, \epsilon_i^k\right) := p\left(\sigma_i^k + \epsilon_i^k\right)^{p-1}$.

We state the framework of IRNRI in Algorithm 1.

Algorithm 1 Iteratively Reweighted Nuclear norm with Rank Identification (IRNRI)

Input: Input $X^0 \in \mathbb{R}^{m\times m}, \epsilon^0 \in \mathbb{R}^N_{++}$ and $\mu \in (0,1)$.
Compute the SVD of X^0.
Initialize: set $k = 0$.
repeat
Reweight: $w_i^k = p\left(\sigma_i^k + \epsilon_i^k\right)^{p-1}$.
Solve the subproblem to obtain the new iterate:

$$X^{k+1} = \arg\min_{X\in\mathbb{R}^{m\times n}} L(X; X^k, \epsilon^k) \tag{2.2}$$

Adaptively update the parameter according to Subroutine 3.
Set $k \leftarrow k+1$.
until convergence

With the SVD of $X^k - \frac{\nabla f(X^k)}{\beta} = U^{k+1}S^{k+1}{V^{k+1}}^\top$, it is shown Lu et al. (2014); Sun et al. (2017) that (2.1) has a global optimal solution as

$$X^{k+1} = U^{k+1}\mathrm{diag}(\boldsymbol{\sigma}^k){V^{k+1}}^\top, \tag{2.3}$$

where $\sigma_i^{k+1} = \max\left(S_{ii}^{k+1} - \frac{\lambda w_i^k}{\beta}, 0\right)$.

By taking inspiration from the Nesterov's acceleration techniques Nesterov (1983); Phan and Nguyen (2021), we use the extrapolation techniques to accelerate IRNRI, which is stated in Algorithm 2 and is named Extrapolated Iteratively Reweighted Nuclear norm with Rank Identification (EIRNRI).

We select $\bar{\alpha}$ according to the following

$$\begin{cases} \bar{\alpha} \in (0,1), & \text{if } f(x) \text{ is convex and Lipschitz differentiable;} \\ \bar{\alpha} \in (0, \sqrt{\frac{\beta}{\beta+3L_f}}), & \text{if } f(x) \text{ is Lipschitz differentiable.} \end{cases} \tag{2.7}$$

Similar to the solution of subproblem (2.2), the optimal solution for (2.6) can be derived by (2.3)

$$X^{k+1} = U^{k+1}\mathrm{diag}(\boldsymbol{\sigma}^{k+1})V^{k+1}, \tag{2.8}$$

where $\mathrm{diag}(\boldsymbol{\sigma}^{k+1})$ is a diagonal matrix $\sigma_i^{k+1} = \max\left(S_{ii}^{k+1} - \frac{\lambda w_i^k}{2\beta}, 0\right)$, and S^{k+1} comes from the SVD of $\frac{Y^k + X^k}{2} - \frac{\nabla f(Y^k)}{2\beta} = U^{k+1}S^{k+1}{V^{k+1}}^\top$. Since X^{k+1} is optimal to (2.6), there exists $\boldsymbol{\xi}^{k+1} \in \partial|\boldsymbol{\sigma}^{k+1}|$ such that

$$0 = \nabla f(Y^k) + \beta(X^{k+1} - Y^k) + \beta(X^{k+1} - X^k) + \lambda U^{k+1}\mathrm{diag}\left(\boldsymbol{w}^k \circ \boldsymbol{\xi}^{k+1}\right){V^{k+1}}^\top. \tag{2.9}$$

Algorithm 2 Extrapolated Iteratively Reweighted Nuclear norm with Rank Identification (EIRNRI)

Input: Input point $X^0 \in \mathbb{R}^{m\times n}, \epsilon^0 \in \mathbb{R}^N_{++}$ and $\mu \in (0,1)$, $0 \le \alpha_0 \le \bar{\alpha} < 1$, , where $\bar{\alpha}$ is selected in (2.7).

Initialize: set $k = 0, X^{-1} = X^0$.

Compute the SVD of X^0.

repeat

Compute new iterate :

$$w_i^k = p\left(\sigma_i^k + \epsilon_i^k\right)^{p-1}, \tag{2.4}$$

$$Y^k = X^k + \alpha_k(X^k - X^{k-1}). \tag{2.5}$$

Updating X^{k+1} by solving:

$$\arg\min_X \langle X, \nabla f(Y^k)\rangle + \frac{\beta}{2}\|X - Y^k\|_F^2 + \frac{\beta}{2}\|X - X^k\|_F^2 + \lambda \sum_{i=1}^m w_i^k \sigma_i(X), \tag{2.6}$$

Choose $0 \le \alpha_k \le \bar{\alpha}$ and update the parameter according to Subroutine 3.

Set $k \leftarrow k+1$.

until convergence

The updating strategy of ϵ is given in the following subroutine, which plays a crucial role in analyzing the behavior of the proposed algorithms.

Algorithm 3 Reweighting subroutine.

if $\mathcal{I}^k \supset \mathcal{I}^{k+1}$ **then**
  $\epsilon_i^{k+1} = \mu\epsilon_i^k,\ i \in \mathcal{I}^{k+1}$.
  Set $\tau_\epsilon = \epsilon^k_{|\mathcal{I}^k|}$ if $\mathcal{I}^k \ne \emptyset$; otherwise $\tau_\epsilon = \infty$.
  $\epsilon_i^{k+1} = \min\{\epsilon_i^k, \tau_\epsilon\},\ i \in \mathcal{Z}^k$.
end if
if $\mathcal{I}^k \subset \mathcal{I}^{k+1}$ **then**
  $\epsilon_i^{k+1} = \mu\epsilon_i^k,\ i \in \mathcal{I}^k$.
  Set $\tau_\epsilon = \epsilon^k_{|\mathcal{I}^k|}$ if $\mathcal{I}^k \ne \emptyset$; otherwise $\tau_\epsilon = \infty$.
  $\epsilon_i^{k+1} = \mu\min\{\epsilon_i^k, \tau_\epsilon\}, i \in \mathcal{Z}^k \cap \mathcal{I}^{k+1}$.
end if
Set $\tau_\sigma = \sigma^{k+1}_{|\mathcal{I}^{k+1}|} + \epsilon^{k+1}_{|\mathcal{I}^{k+1}|}$ if $\mathcal{I}^{k+1} \ne \emptyset$; otherwise $\tau_\sigma = \infty$.
Set $\tau_z = \epsilon^{k+1}_{|\mathcal{I}^{k+1}|+1}$ if $\mathcal{Z}^{k+1} \ne \emptyset$.
For any $i \in \mathcal{Z}^{k+1}$, $\epsilon_i^{k+1} = \begin{cases} \epsilon_i^k & \text{if } \tau_z \le \tau_\sigma \\ \min(\mu\epsilon_i^k, \mu\tau_\sigma) & \text{if } \tau_z > \tau_\sigma. \end{cases}$

Obviously, if $\mathcal{I}^{k+1}=\mathcal{I}^k$ and $\mathcal{Z}^{k+1}=\mathcal{Z}^k$, Subroutine (3) then reverts to

$$\epsilon_i^{k+1}=\mu\epsilon_i^k,\quad i\in\mathcal{I}^k, \tag{2.10}$$

and for $i\in\mathcal{Z}^{k+1}$,

$$\epsilon_i^{k+1}=\begin{cases}\epsilon_i^k & \text{if } \epsilon_i^k\le\tau_\sigma\\ \min(\mu\epsilon_i^k,\mu\tau_\sigma) & \text{if } \epsilon_i^k>\tau_\sigma.\end{cases} \tag{2.11}$$

2.1 Convergence analysis

IRNRI is a special case of EIRNRI if we select $\alpha_k\equiv 0$ for all k. Thus we study the convergence properties for EIRNRI. The following auxiliary function is used in our analysis.

$$H(X,Y,\epsilon)=f(X)+\frac{\beta}{2}\|X-Y\|_F^2+\lambda\sum_{i=1}^m(\sigma_i(X)+\epsilon_i)^p. \tag{2.12}$$

We show $H(X,Y,\epsilon)$ is monotonically decreasing in the following lemma.

Lemma 2 *Suppose Assumptions 1 and 2 hold. Let $\{X^k\}$ be the sequence generated by Algorithm 2 with $\beta>L_f$, then the following statements hold.*

(i). $\{H(X^k,X^{k-1},\epsilon^k)\}$ is monotonically decreasing; furthermore

$$H(X^k,X^{k-1},\epsilon^k)-H(X^{k+1},X^k,\epsilon^{k+1})\ge\bar{\beta}\|X^k-X^{k-1}\|_F^2,$$

for $k>0$, where

$$\bar{\beta}=\begin{cases}\frac{\beta}{2}(1-\bar{\alpha}^2), & \textit{if } f(x) \textit{ is convex and Lipschitz differentiable on } \mathcal{L}(F^0);\\ \frac{\beta}{2}(1-\frac{\beta+3L_f}{\beta}\bar{\alpha}^2), & \textit{if } f(x) \textit{is Lipschitz differentiable on } \mathcal{L}(F^0).\end{cases}$$

(ii). The sequence $\{X^k\}\subset\mathcal{L}(F^0)$ is bounded.

(iii). $\lim_{k\to\infty}\|X^{k+1}-X^k\|_F=0$, $\lim_{k\to\infty}\|Y^k-X^k\|_F=0$ and $\lim_{k\to\infty}\|Y^k-X^{k+1}\|_F=0$.

Proof

(i) Since X^{k+1} is the optimal solution of (2.6), yields

$$\begin{aligned}&\langle X^{k+1},\nabla f(Y^k)\rangle+\frac{\beta}{2}\|X^{k+1}-Y^k\|+\lambda\sum_{i=1}^m w_i^k\sigma_i^{k+1}+\frac{\beta}{2}\|X^{k+1}-X^k\|_F^2\\ &\le\langle X^k,\nabla f(Y^k)\rangle+\frac{\beta}{2}\|X^k-Y^k\|+\lambda\sum_{i=1}^m w_i^k\sigma_i^k.\end{aligned} \tag{2.13}$$

The concavity of p-norm over $\mathbb{R}_+: x^p\le x_0^p+px_0^{p-1}(x-x_0)$, i.e., $x^p-x_0^p\le px_0^{p-1}(x-x_0)$ indicates that

$$\left(\sigma_i^{k+1}+\epsilon_i^k\right)^p-\left(\sigma_i^k+\epsilon_i^k\right)^p\le w_i^k\left(\sigma_i^{k+1}-\sigma_i^k\right),$$

holds for each singular value of X^{k+1}. Rearranging, we have

$$-w_i^k \sigma_i^{k+1} + \left(\sigma_i^{k+1} + \epsilon_i^k\right)^p \le -w_i^k \sigma_i^k + \left(\sigma_i^k + \epsilon_i^k\right)^p,$$

Summing up both sides from 1 to m

$$-\sum_{i=1}^m w_i^k \sigma_i^{k+1} + \sum_{i=1}^m \left(\sigma_i^{k+1} + \epsilon_i^k\right)^p \le -\sum_{i=1}^m w_i^k \sigma_i^k + \sum_{i=1}^m \left(\sigma_i^k + \epsilon_i^k\right)^p. \tag{2.14}$$

Combining (2.13) and (2.14) yields

$$\begin{aligned} &\langle X^{k+1}, \nabla f(Y^k)\rangle + \frac{\beta}{2}\|X^{k+1} - Y^k\| + \lambda \sum_{i=1}^m (\sigma_i^{k+1} + \epsilon_i^k)^p \\ \le\ &\langle X^k, \nabla f(Y^k)\rangle + \frac{\beta}{2}\|X^k - Y^k\| + \lambda \sum_{i=1}^m (\sigma_i^k + \epsilon_i^k)^p - \frac{\beta}{2}\|X^{k+1} - X^k\|_F^2. \end{aligned} \tag{2.15}$$

It follows that

$$\begin{aligned} &F(X^{k+1};\epsilon^{k+1}) \\ \le\ & f(Y^k) + \langle \nabla f(Y^k), X^{k+1} - Y^k\rangle + \frac{L_f}{2}\|X^{k+1} - Y^k\|_F^2 + \lambda \sum_{i=1}^m (\sigma_i^{k+1} + \epsilon_i^{k+1})^p \\ \le\ & f(Y^k) + \langle \nabla f(Y^k), X^{k+1} - Y^k\rangle + \frac{\beta}{2}\|X^{k+1} - Y^k\|_F^2 + \lambda \sum_{i=1}^m (\sigma_i^{k+1} + \epsilon_i^{k})^p \\ \le\ & f(Y^k) + \langle \nabla f(Y^k), X^{k} - Y^k\rangle + \frac{\beta}{2}\|X^{k} - Y^k\|_F^2 \\ &+ \lambda \sum_{i=1}^m (\sigma_i^k + \epsilon_i^k)^p - \frac{\beta}{2}\|X^{k+1} - X^k\|_F^2 \\ \le\ & f(X^k) + \langle \nabla f(Y^k) - \nabla f(X^k), X^{k} - Y^k\rangle + \frac{L_f + \beta}{2}\|X^{k} - Y^k\|_F^2 \\ &+ \lambda \sum_{i=1}^m (\sigma_i^k + \epsilon_i^k)^p - \frac{\beta}{2}\|X^{k+1} - X^k\|_F^2 \\ \le\ & F(X^k;\epsilon^k) + \frac{3L_f + \beta}{2}\|X^k - Y^k\|_F^2 - \frac{\beta}{2}\|X^{k+1} - X^k\|_F^2, \end{aligned} \tag{2.16}$$

where the first, fourth and last inequalities hold by Assumption 1, the second inequality is by $\beta \ge L_f$ and the monotonicity of p-norm, the third inequality follows (2.15). It then follows from (2.5) that

$$F(X^{k+1};\epsilon^{k+1}) - F(X^k;\epsilon^k) \le \frac{3L_f + \beta}{2}\alpha_k^2\|X^k - X^{k-1}\|_F^2 - \frac{\beta}{2}\|X^k - X^{k+1}\|_F^2, \tag{2.17}$$

This implies that

$$\begin{aligned}
&H(X^k, X^{k-1}, \epsilon^k) - H(X^{k+1}, X^k, \epsilon^{k+1}) \\
= \ &F(X^k; \epsilon^k) + \frac{\beta}{2}\|X^k - X^{k-1}\|_F^2 - \left[F(X^{k+1}; \epsilon^{k+1}) + \frac{\beta}{2}\|X^{k+1} - X^k\|_F^2\right] \\
\geq \ &\frac{\beta}{2}\left(1 - \alpha_k^2 \frac{\beta + 3L_f}{\beta}\right)\|X^k - X^{k-1}\|_F^2 \\
\geq \ &\frac{\beta}{2}\left(1 - \bar{\alpha}^2 \frac{\beta + 3L_f}{\beta}\right)\|X^k - X^{k-1}\|_F^2,
\end{aligned} \tag{2.18}$$

where the last inequality holds with $\alpha_k \in [0, \bar{\alpha}]$ and $0 < \bar{\alpha} < \sqrt{\frac{\beta}{\beta + 3L_f}}$. Consequently, $\{H(X^k, X^{k-1}, \epsilon^k)\}$ is monotonically decreasing. If f is convex, the fourth inequality in (2.16) becomes

$$\begin{aligned}
&F(X^{k+1}; \epsilon^{k+1}) \\
\leq \ &f(Y^k) + \langle \nabla f(Y^k), X^k - Y^k \rangle + \frac{\beta}{2}\|X^k - Y^k\|_F^2 \\
&+ \lambda \sum_{i=1}^{m} (\sigma_i^k + \epsilon_i^k)^p - \frac{\beta}{2}\|X^{k+1} - X^k\|_F^2 \\
\leq \ &F(X^k; \epsilon^k) + \frac{\beta}{2}\|X^k - Y^k\|_F^2 - \frac{\beta}{2}\|X^{k+1} - X^k\|_F^2.
\end{aligned} \tag{2.19}$$

$$H(X^k, X^{k-1}, \epsilon^k) - H(X^{k+1}, X^k, \epsilon^{k+1}) \geq \frac{\beta}{2}(1 - \bar{\alpha}^2)\|X^k - X^{k-1}\|_F^2. \tag{2.20}$$

Thus the sequence $\{H(X^k, X^{k-1}, \epsilon^k)\}$ is also monotonically decreasing with $\bar{\alpha} \in (0, 1)$. This proves part (i).

(ii) For all $k > 0$, we have

$$F(X^k) \leq F(X^k, \epsilon^k) \leq H(X^k, X^{k-1}, \epsilon^k) \leq H(X^0, X^{-1}, \epsilon^0) = F(X^0; \epsilon^0). \tag{2.21}$$

From Assumption 2 we know that $\{X^k\} \subset \mathcal{L}(F^0)$ is bounded. This completes the proof of part (ii).

(iii) Summing both side of (2.18) from 0 to t, we obtain

$$\begin{aligned}
&\frac{\beta}{2}\left(1 - \bar{\alpha}^2 \frac{\beta + 3L_f}{\beta}\right)\sum_{i=0}^{t} \|X^k - X^{k-1}\|_F^2 \\
\leq \ &H(X^0, X^{-1}, \epsilon^0) - H(X^{t+1}, X^k, \epsilon^{t+1}) \\
\leq \ &F(X^0; \epsilon^0) - F(X^{t+1}; \epsilon^{t+1}) \\
< \ &+\infty,
\end{aligned} \tag{2.22}$$

where the second inequality comes from (2.21). Taking $t \to \infty$, it follows from Assumption 2 that $\lim_{k \to \infty} \|X^{k+1} - X^k\|_F = 0$. Furthermore,

$$\begin{aligned}
Y^k - X^k &= \alpha_k (X^k - X^{k-1}) \to 0, \\
Y^k - X^{k+1} &= (X^k - X^{k+1}) + \alpha_k (X^k - X^{k-1}) \to 0.
\end{aligned} \tag{2.23}$$

Therefore, part (iii) is true. ■

The boundedness of $\{X^k\}$ and $\{Y^k\}$ by Lemma 2(ii)-(iii) implies that there exists a constant $C > 0$ such that

$$\max_i \sigma_i \left(\frac{X^k + Y^k}{2} - \frac{1}{2\beta} \nabla f(X^k) \right) \leq C, \forall k \geq 0.$$

2.2 Rank identification

Notice that the SVD sorts the singular value from largest to smallest value, as shown in (1.6).

Definition 4 (Simultaneous ordered SVD Lewis and Sendov (2005)) *We say that two matrices X and Y in $\mathbb{R}^{m\times n}$ have a simultaneous ordered singular value decomposition if there are unitary matrices $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ such that*

$$X = U\mathrm{diag}(\boldsymbol{\sigma}(X))V^\top \quad \textit{and} \quad Y = U\mathrm{diag}(\boldsymbol{\sigma}(Y))V^\top.$$

Now we summarize the property of each iteration after a large number of iteration and show the model identification as follows.

Theorem 3 *Suppose Assumptions 1 and 2 hold. Let $\{(X^k; \epsilon^k)\}$ be the sequence generated by Algorithm 2. Then the following statements hold true*

i). If $w(\sigma_i(X^{\hat{k}}), \epsilon_i^{\hat{k}}) > \frac{2\beta C}{\lambda}$, for some $\hat{k} \in \mathbb{N}$, then $\sigma_i^k = 0$ for all $k > \hat{k}$.

ii). The rank of iteration points are stable, i.e., $r \in \mathbb{N}$ make $\mathrm{rank}(X^k) = r$ for all sufficiently large k. Thus the index set $\mathcal{I}(X^k)$ and $\mathcal{Z}(X^k)$ are fixed for all $k > \hat{k}$. We denote $\mathcal{I}^* = \mathcal{I}(X^k)$ and $\mathcal{Z}^* = \mathcal{Z}(X^k)$ for any $k > \hat{k}$.

iii). $\sigma_i^k > \left(\frac{\lambda p}{\beta C}\right)^{\frac{1}{1-p}} - \epsilon_i^k > 0, i \in \mathcal{I}(X^k)$ for sufficient large k. Furthermore, for every limit point X^* of $\{X^k\}$, $\mathcal{I}(X^*) = \mathcal{I}^*$ and $\sigma_i^k \geq \left(\frac{\lambda p}{\beta C}\right)^{\frac{1}{1-p}}$, $i \in \mathcal{I}(X^*)$, meaning rank$(X^*) = r$.

Proof It is shown in Sun et al. (2017) that in the optimality condition (2.9), both X^{k+1} and $\frac{X^k+Y^k}{2} - \frac{1}{2\beta}\nabla f(X^k)$ have the simultaneous ordered SVD, thus

$$0 = \boldsymbol{\sigma}^{k+1} - \boldsymbol{\sigma} \left(\frac{X^k + Y^k}{2} - \frac{1}{2\beta} \nabla f(X^k) \right) + \frac{\lambda}{2\beta} \boldsymbol{w}^k \circ \boldsymbol{\xi}^{k+1}, \tag{2.24}$$

with $\xi_i^{k+1} \in [-1, 1]$. Hence

$$\sigma_i^{k+1} = \max \left\{ \sigma_i \left(\frac{X^k + Y^k}{2} - \frac{1}{2\beta} \nabla f(X^k)) \right) - \frac{\lambda}{2\beta} w_i^k \xi_i^{k+1}, 0 \right\}. \tag{2.25}$$

(i) Suppose there exists $\hat{k}$ such that $w_i^{\hat{k}} \geq \frac{2\beta C}{\lambda}$ for some $i \in [m]$. To guarantee

$$\sigma_i^{\hat{k}+1} = \sigma_i \left(\frac{X^{\hat{k}} + Y^{\hat{k}}}{2} - \frac{1}{2\beta} \nabla f(X^{\hat{k}}) \right) - \frac{\lambda}{2\beta} w_i^{\hat{k}} \xi_i^{\hat{k}+1} \geq 0, \tag{2.26}$$

it holds that $\xi_i^{\hat{k}+1} < 1$, implying $\sigma_i^{\hat{k}+1} = 0$. Monotonicity of $(\cdot)^{p-1}$ and $\epsilon_i^{\hat{k}+1} \leq \epsilon_i^{\hat{k}} \leq \sigma_i(X^{\hat{k}}) + \epsilon_i^{\hat{k}}$ yield $w(\sigma_i^{\hat{k}+1}, \epsilon^{\hat{k}+1}) > \frac{2\beta C}{\lambda}$ is also true. Therefore $\sigma_i^{\hat{k}+2} = 0$. By induction we know that $\sigma_i^k \equiv 0$ for any $k > \hat{k}$.

(ii) Suppose by contradiction this statement is not true. Then for some $i \in [m]$ such that σ_i^k takes zero and nonzero value both for infinite times. Hence, there are two subsequences $\mathcal{S}_1 \cup \mathcal{S}_2 = \mathbb{N}$ such that $|\mathcal{S}_1| = \infty$, $|\mathcal{S}_2| = \infty$ and that

$$\sigma_i^k = 0, \ \forall k \in \mathcal{S}_1 \text{ and } \sigma_i^k > 0, \ \forall k \in \mathcal{S}_2.$$

Therefore, there exists subsequence $\mathcal{S}_3 \subset \mathcal{S}_2$ such that $|\mathcal{S}_3| = \infty$ and $i \in \mathcal{Z}^k \cap \mathcal{I}^{k+1}$ for any $k \in \mathcal{S}_3$. In other words, $\sigma_i^k = 0$ and $\sigma_i^{k+1} \neq 0$ for any $k \in \mathcal{S}_3$. Hence, $\mathcal{I}^k \subset \mathcal{I}^{k+1}$ for any $k \in \mathcal{S}_3$ and the third line of this case in the update strategy 3 implies that $\lim\limits_{k\to\infty} \epsilon_i^k = 0$ due to $|\mathcal{S}_3| = \infty$. Hence there exists $\hat{k} \in \mathcal{S}_1$ such that

$$w(\sigma_i^{\hat{k}}, \epsilon_i^{\hat{k}}) = p\left(\sigma_i(X^{\hat{k}}) + \epsilon_i^{\hat{k}}\right)^{p-1} = p(\epsilon_i^{\hat{k}})^{p-1} \geq \frac{2\beta C}{\lambda}.$$

It follows that $\sigma_i(X^k) = 0$ for any $k > \hat{k}$ by (i), which implies $\{\hat{k}+1, \hat{k}+2, \hat{k}+3, \cdots\} \subset \mathcal{S}_1$ and $|\mathcal{S}_2| < \infty$. This contradicts $|\mathcal{S}_2| = \infty$. Hence (ii) is true.

(iii) Part (ii) implies that $\mathcal{I}^k \equiv \mathcal{I}^*$, indicating

$$\epsilon_i^k \to 0, \ \text{for any } i \in \mathcal{I}^k \tag{2.27}$$

by (2.10). From (i) we know that if $w_i^k < \frac{2\beta C}{\lambda}$, $i \in \mathcal{I}^k$ for sufficiently large k. It follows that

$$\sigma_i^k > \left(\frac{\lambda p}{2\beta C}\right)^{\frac{1}{1-p}} - \epsilon_i^k > 0, \ i \in \mathcal{I}^k.$$

Letting $\epsilon_i^k \to 0$, we have

$$\lim_{\substack{k\to\infty \\ k\in\mathcal{S}_1}} \sigma_i^k \geq \left(\frac{\lambda p}{2\beta C}\right)^{\frac{1}{1-p}} > 0, i \in \mathcal{I}^*.$$

■

2.3 Adaptively reweighting

Besides the rank identification property proved in the previous section, we further design our algorithm such that the $\epsilon_i, i \in \mathcal{Z}^k = \mathcal{Z}^*$ does not affect the perturbed objective value $F(X, \epsilon)$ for sufficiently large iterations. If this happens, the algorithm behaves like minimizing $F(X, \epsilon)$ by eliminating the $(\sigma_i(X) + \epsilon_i)^p, i \in \mathcal{Z}^k$ since it is fixed as a constant in $F(X, \epsilon)$. The crucial point is to manipulate the values of ϵ_i to maintain $\{w_1^k, \ldots, w_n^k\}$ is arranged in ascending order for each $k \in \mathbb{N}$, so that each subproblem solution can be easily obtained Lu et al. (2014). Though putting $\{w_1^k, \ldots, w_n^k\}$ in descending order would make the subproblem convex and easy to handle, it is unrealistic to design such a strategy to keep $w(\sigma_i(X), \epsilon_i) = p(\sigma_i(X), \epsilon_i)^{p-1} > p(\sigma_j(X), \epsilon_j)^{p-1} = w(\sigma_j(X), \epsilon_j)$ for $\sigma_i(X) > \sigma_j(X)$ as $\epsilon_i \to 0$ and $\epsilon_j \to 0$.

Theorem 4 *Suppose Assumptions 1 and 2 hold. Let $\{(X^k;\epsilon^k)\}$ be the sequence generated by Algorithm 1. Then the following statements hold*

(i) $\{\epsilon_i^k, i\in\mathcal{I}^k\}$, $\{\epsilon_i^k, i\in\mathcal{Z}^k\}$, and $\{\sigma_i^k+\epsilon_i^k, i\in[m]\}$ are all in descending order for every $k\in\mathbb{N}$. Therefore $w_1^k,\dots,w_m^k$ is in ascending order for $k\in\mathbb{N}$.

(ii) For sufficiently large k, the update of $\epsilon_i^k, i\in\mathcal{I}^*$ is always triggered, i.e., $\epsilon_i^k=\mu\epsilon_i^{k+1}, i\in\mathcal{I}^*$. Therefore, $\{\epsilon_i^k\}\searrow 0, i\in\mathcal{I}^*$.

(iii) For sufficiently large k, the update of $\epsilon_i^k, i\in\mathcal{Z}^*$ is never triggered. In other words, there exists $\hat{k}$ such that $\epsilon_i^k=\epsilon_i^{\hat{k}}, i\in\mathcal{Z}^*$ for all $k>\hat{k}$.

Proof

(i) We prove this by induction. First of all, the statement is obviously true at $k=0$ since $\epsilon_1^0,\dots,\epsilon_m^0$ is descending. Suppose this is also true at the kth iteration. Now we delve into the case at $(k+1)$th iteration.

Case (a): Line 2 of Subroutine (3) implies that $\{\epsilon_i^{k+1}, i\in\mathcal{I}^{k+1}\}$ is descending since $\{\epsilon_i^k, i\in\mathcal{I}^k\cap\mathcal{I}^{k+1}\}$ is descending. Line 4 guarantees $\{\epsilon_i^{k+1}, i\in\mathcal{Z}^{k+1}\}$ is descending, since $\mathcal{Z}^{k+1}=\mathcal{Z}^k\cup(\mathcal{I}^k\setminus\mathcal{I}^{k+1})$.

Case (b): Line 7-9 guarantees that $\{\epsilon_i^{k+1}, i\in\mathcal{I}^k\cap\mathcal{I}^{k+1}\}$ is descending since $\mathcal{I}^{k+1}=\mathcal{I}^k\cup(\mathcal{Z}^k\cap\mathcal{I}^{k+1})$.

Finally, Line 13 again ensures that $\{\epsilon_i^{k+1}, i\in\mathcal{Z}^{k+1}\}$ and $\{\sigma_i^k+\epsilon_i^k, i\in[m]\}$ is descending.

(ii) This statement is obviously true by (2.10).

(iii) Notice that $\epsilon_i^{k+1}\le\mu\epsilon_i^k, i\in\mathcal{Z}^*$ whenever it is reduced by Line 13 in Subroutine 3. On the other hand, τ_z is the largest number in $\{\epsilon_i^{k+1}, i\in\mathcal{Z}^*\}$ and τ_σ is bounded below from 0 for sufficiently large k by Theorem 3(iii). Therefore, if the update is triggered for infinite many times, then eventually $\tau_z\le\tau_\sigma$ is always satisfied for sufficiently large k—a contradiction. Therefore, $\epsilon_i^{k+1}\le\mu\epsilon_i^k, i\in\mathcal{Z}^*$ is never reduced after finite iterations. ■

2.4 Global convergence

Theorem 3 shows that X^k remain in manifold $\mathbb{R}_r^{m\times n}:=\{X:\mathrm{rank}(X)=r, X\in\mathbb{R}^{m\times n}\}$ for sufficiently large k. Denote χ as the cluster point of $\{X^k\}$; then χ is nonempty by Lemma 2(ii). The necessary optimality condition of (P) (Sun et al., 2017, Proposition 1) is given by:

$$\nabla f(X^*)+\lambda U^*\mathrm{diag}(\boldsymbol{w}^*\circ\boldsymbol{\xi}^*){V^*}^\top=0, \tag{2.28}$$

where $\boldsymbol{\xi}^*\in\partial|\boldsymbol{\sigma}(X^*)|$, $\boldsymbol{w}^*\in\partial|\boldsymbol{\sigma}(X^*)|_p^p$ and U^*,V^* being the SVD unitary matrices of X^*. To show the global convergence of Algorithm 1 and 2, we investigate the optimality error at X^k

$$\mathcal{E}^{k+1}=\nabla f(X^{k+1})+\lambda U^{k+1}\mathrm{diag}(\bar{\boldsymbol{w}}^{k+1}\circ\boldsymbol{\xi}^{k+1}){V^{k+1}}^\top, \tag{2.29}$$

where $\bar{w}_i^{k+1}=p(\sigma_i^{k+1})^{p-1}, i\in\mathcal{I}^{k+1}$ and $\bar{w}_i^{k+1}=w_i^k$, $i\in\mathcal{Z}^{k+1}$.

Theorem 5 (Global convergence) *Suppose Assumptions 1 and 2 are satisfied. Let $\{X^k\}$ be a sequence generated by Algorithm 2, the following statements hold true* *(i). F attains the same value at every cluster point of $\{X^k\}$, i.e., there exists $\zeta \in \mathbb{R}$ such that $F(X^*) = \zeta$ for any $X^* \in \chi$.*

(ii). $\lim_{k\to\infty} \|\mathcal{E}^{k+1}\|_F \to 0$. Therefore, each point $X^* \in \chi$ is a critical point of $F(X)$.

Proof (i) Theorem 4 implies that $\boldsymbol{\epsilon}^k \to \boldsymbol{\epsilon}^* = [0, 0, \cdots, 0, \epsilon_{r+1}^{\hat{k}}, \cdots, \epsilon_n^{\hat{k}}]$. Lemma 2(i) implies that $H^* := \lim_{k\to\infty} H(X^{k+1}, X^k, \epsilon^{k+1})$. Let $X^* \in \chi$. It holds from Lemma 2(i)(iii) that

$$
\begin{aligned}
F(X^*) &= f(X^*) + \lambda \sum_{i=1}^{r} (\sigma_i^*)^p \\
&= \lim_{k\to\infty} f(X^{k+1}) + \lambda \sum_{i=1}^{m} (\sigma_i^{k+1} + \epsilon_i^{k+1})^p + \beta \|X^{k+1} - X^k\|_F^2 - \lambda \sum_{i=r+1}^{n} (\epsilon_i^{\hat{k}})^p \\
&= \lim_{k\to\infty} H(X^{k+1}, X^k, \epsilon^{k+1}) - \lambda \sum_{i=r+1}^{n} (\epsilon_i^{\hat{k}})^p \\
&= H^* - \lambda \sum_{i=r+1}^{n} (\epsilon_i^{\hat{k}})^p.
\end{aligned}
$$

Therefore, (i) is proved with $\zeta := H^* - \lambda \sum_{i=r+1}^{n} (\epsilon_i^{\hat{k}})^p$.

(ii) Subtracting both side of (2.29) by (2.9), we obtain

$$
\begin{aligned}
\mathcal{E}^{k+1} = &[\nabla f(X^{k+1}) - \nabla f(Y^k)] - \beta[(X^{k+1} - Y^k) + (X^{k+1} - X^k)] \\
&+ \lambda U^{k+1} \mathrm{diag}((\bar{w}^{k+1} - w^k) \circ \boldsymbol{\xi}^{k+1}) {V^{k+1}}^\top.
\end{aligned} \tag{2.30}
$$

The first term in (2.30) vanishes as $k \to \infty$ since

$$
\begin{aligned}
\|\nabla f(X^{k+1}) - \nabla f(Y^k)\|_F &\le L_f \|X^{k+1} - Y^k\|_F \\
&= L_f \|X^{k+1} - X^k\|_F + L_f \bar{\alpha} \|X^k - X^{k-1}\|_F,
\end{aligned} \tag{2.31}
$$

by Assumption 1 and Lemma 2(iii). The second term in (2.30) also vanishes by Lemma 2(iii). As for the third term,

$$\begin{aligned}
&\|U^{k+1}\mathrm{diag}((\bar{\boldsymbol{w}}^{k+1}-\boldsymbol{w}^k)\circ\boldsymbol{\xi}^{k+1}){V^{k+1}}^\top\|_F\\
=&\|\mathrm{diag}((\bar{\boldsymbol{w}}^{k+1}-\boldsymbol{w}^k)\circ\boldsymbol{\xi}^{k+1})\|_F\\
\le&\|\bar{\boldsymbol{w}}^{k+1}-\boldsymbol{w}^k\|_2\\
=&\sum_{i=1}^r|p(\sigma_i^{k+1})^{p-1}-p(\sigma_i^k+\epsilon_i^k)^{p-1}|\\
=&\sum_{i=1}^r p(1-p)\left(\hat{\sigma}_i^k\right)^{\frac{2-p}{1-p}}\left(\left|\sigma_i^k-\sigma_i^{k+1}\right|+\epsilon_i^k\right)\\
\le&\sum_{i=1}^r p(1-p)\left(\frac{2\beta C}{\lambda p}\right)^{\frac{2-p}{1-p}}\left(\left|\sigma_i^k-\sigma_i^{k+1}\right|+\epsilon_i^k\right)\\
\le&\sum_{i=1}^r p(1-p)\left(\frac{2\beta C}{\lambda p}\right)^{\frac{2-p}{1-p}}\left(\|X^k-X^{k+1}\|_F+\epsilon_i^k\right),
\end{aligned}\tag{2.32}$$

where the first equality is by the unitarity of U^{k+1} and V^{k+1}, the third equality is by the Mean Value Theorem with $\hat{\sigma}_i^k$ between σ_i^k and σ_i^{k+1}, the second inequality is by Theorem 3(iii), and the last inequality is by (Horn and Johnson, 2012, Problem 7.3.P16). It then follows Lemma 2(iii) and Theorem 4(ii) that $\|U^{k+1}\mathrm{diag}((\bar{\boldsymbol{w}}^{k+1}-\boldsymbol{w}^k)\circ\boldsymbol{\xi}^{k+1}){V^{k+1}}^\top\|_F\to 0$. This completes the proof of (ii). ■

3 Convergence ananlysis under KL property

To further analyze the property of $\{(X^k,\epsilon^k)\}$ for sufficiently large k, we denote $\epsilon_i=\delta_i^2$ with $\delta_i\ge 0$ since ϵ_i is restricted to be non-negative. Theorem 4 guarantees that $\delta_i^k, i\in\mathcal{Z}^k$ remain constant and $\delta_i^{k+1}=\sqrt{\mu}\delta_i^k, i\in\mathcal{I}^k$ for sufficiently large k. Consider the reduced form of (2.12)

$$\hat{H}(X,Y,\delta)=f(X)+\frac{\beta}{2}\|X-Y\|_F^2+\sum_{i=1}^r\left(\sigma_i(X)+(\delta_i)^2\right)^p,\tag{3.1}$$

where r is the optimal rank, or equivalently, $r=|\mathcal{I}^*|$. Notice that $\hat{H}$ is differentiable for any $X\in\mathbb{R}_{sr}$. The partial derivative of (3.1) with respect to X is

$$\nabla_X\hat{H}(X,Y,\delta)=\nabla f(X)+\beta(X-Y)+\lambda USV^\top,\tag{3.2}$$

with $S=\mathrm{diag}([p(\sigma_1(X)+\delta_1^2)^{p-1},\cdots,p(\sigma_r(X)+\delta_r^2)^{p-1},0,\cdots,0])$ and $X=USV^\top$. The partial derivative of (3.1) with respect to Y is

$$\nabla_Y\hat{H}(X,Y,\delta)=\beta(Y-X),\tag{3.3}$$

and the partial derivative of (3.1) with respect to δ is

$$\nabla_{\delta_i}\hat{H}(X,Y,\delta)=2\lambda p\delta_i(\sigma_i(X)+\delta_i^2)^{p-1}.\tag{3.4}$$

Definition 5 (matrix measures on Cartesian product) *For the Cartesian product space* $\mathbb{S} := \mathbb{R}^{m\times n} \times \mathbb{R}^{m\times n} \times \mathbb{R}^r$, *the product over* $\mathbb{S}$ *is defined as*

$$\langle X, Y\rangle = \langle X_1, Y_1\rangle + \langle X_2, Y_2\rangle + \langle x_3, y_3\rangle,$$

for any $X = (X_1, X_2, x_3), Y = (Y_1, Y_2, y_3) \in \mathbb{S}$, *then norm of* $X \in \mathbb{S}$ *is defined as*

$$\|X\| = (\|X_1\|_F^2 + \|X_2\|_F^2 + \|x_3\|_2^2)^{1/2},$$

and distance of two elements on the product metric is defined as

$$\operatorname{dist}(X, Y) = \sqrt{\|X_1 - Y_1\|_F^2 + \|X_2 - Y_2\|_F^2 + \|x_3 - y_3\|_2^2}.$$

Since lower semi-continuous function $f(\cdot)$ over the smooth manifold $\mathbb{R}_r^{m\times n}$ is KL function Xu and Yin (2013), and the Schatten-p norm is also KL function Zhang et al. (2019), thus we make the following assumption.

Assumption 3 *Suppose there exists* c, θ *such that the function* $\hat{H} : \mathbb{S} \to (-\infty, \infty)$ *satisfies the KL property in Definition 3 at each point*

$$(X^*, X^*, 0_r), \quad \forall X^* \in \operatorname{crit}(F), \tag{3.5}$$

with the neighborhood

$$\mathbb{U}((X^*, X^*, 0); \rho) := \{(X, X, \boldsymbol{\delta}) \in \mathbb{S}_{sr} : \operatorname{dist}((X, X, \boldsymbol{\delta}), (X^*, X^*, 0)) < \rho\}$$

and the continuous function Φ, *where* $\mathbb{S}_{sr}$ *denotes the Cartesian product set* $\mathbb{R}_r^{m\times n} \times \mathbb{R}_r^{m\times n} \times \mathbb{R}^r$ *and* $r = \operatorname{rank}(X^*)$.

Now we consider the convergence of under KL property.

Lemma 6 (uniqueness of convergence) *Let* $\{X^k\}$ *be a sequence generated by Algorithm 2, Assumptions 1-3 are satisfied. There exists* $\hat{k}$, *such that the following statements hold*

(i). There exists $D_1 > 0$ *such that for all* $k \geq \hat{k}$

$$\left\|\nabla \hat{H}(X^{k+1}, X^k, \boldsymbol{\delta}^{k+1})\right\| \leq D_1 \left(\left\|X^k - X^{k+1}\right\|_F + \left\|X^{k-1} - X^k\right\|_F + \|\boldsymbol{\delta}^k\|_1 - \|\boldsymbol{\delta}^{k+1}\|_1\right).$$

Moreover, $\lim_{k\to\infty} \nabla \hat{H}(X^{k+1}, X^k, \boldsymbol{\delta}^{k+1}) = 0$.

(ii). $\{\hat{H}(X^k, Y^{k-1}, \boldsymbol{\delta}^k)\}$ *is monotonically decreasing and there exists* D_2 *such that*

$$\hat{H}(X^k, Y^{k-1}, \boldsymbol{\delta}^k) - \hat{H}(X^{k+1}, Y^k, \boldsymbol{\delta}^{k+1}) \geq D_2\|X^k - X^{k-1}\|_F^2.$$

(iii). $\hat{H}(X^*, X^*, 0) = \zeta := \lim_{k\to\infty} \hat{H}(X^k, Y^{k-1}, \boldsymbol{\delta}^k)$, *where* $(X^*, X^*, 0) \in \Gamma$ *and* Γ *is the set of cluster points of* $\{(X^k, Y^{k-1}, \boldsymbol{\delta}^k)\}$, *i.e.,* $\Gamma := \{(X^*, X^*, 0) : X^* \in \chi\}$.

(iv). For any t, $T^t := \sum_{k=t}^{\infty} \|X^{k-1} - X^k\|_F < +\infty$. Therefore, $\lim_{k\to\infty} X^k = X^*$.

Proof For the sufficiently large k, (2.9) has reformulation as

$$0 = \nabla f(Y^k) + \beta(X^{k+1} - Y^k) + \lambda U^{k+1}\mathrm{diag}\left(\hat{w}^k\right) {V^{k+1}}^\top + \beta(X^{k+1} - X^k), \tag{3.6}$$

where $\hat{w}^k = (w_1^k, \cdots, w_r^k, 0, \cdots, 0)$.

(i) Combining (3.2) and (3.6), one has

$$\begin{aligned} &\nabla_X \hat{H}(X^{k+1}, X^k, \boldsymbol{\delta}^{k+1}) \\ =& \nabla f(X^{k+1}) - f(Y^k) - \beta(X^k - Y^k) + \lambda U_r^{k+1}\left(\hat{w}^{k+1} - \hat{w}^k\right) {V_r^{k+1}}^\top \end{aligned} \tag{3.7}$$

with $\hat{w}^k = \left(w_1^k, \cdots, w_r^k, 0, \cdots, 0\right)^\top$.

The second term in (3.7) obeys

$$\|\beta(X^k - Y^k)\|_F \le \beta\bar{\alpha}\|X^k - X^{k-1}\|_F, \tag{3.8}$$

from (2.5).

Similarly to (2.31), we have

$$\begin{aligned} &\|U^{k+1}\mathrm{diag}((\hat{\boldsymbol{w}}^{k+1} - \hat{\boldsymbol{w}}^k) \circ \boldsymbol{\xi}^{k+1}) {V^{k+1}}^\top\|_F \\ \le& \sum_{i=1}^r p(1-p)\left(\frac{2\beta C}{\lambda p}\right)^{\frac{2-p}{1-p}} \left(\|X^k - X^{k+1}\|_F + (\delta_i^k)^2 - (\delta_i^{k+1})^2\right), \\ \le& C_L\left(\|X^k - X^{k+1}\|_F + \max_i(\delta_i^k + \delta_i^{k+1})\|\boldsymbol{\delta}^k - \boldsymbol{\delta}^{k+1}\|_1\right) \\ \le& C_L\left[\|X^k - X^{k+1}\|_F + 2\|\boldsymbol{\delta}^0\|_\infty(\|\boldsymbol{\delta}^k\|_1 - \|\boldsymbol{\delta}^{k+1}\|_1)\right], \end{aligned} \tag{3.9}$$

with $C_L := p(1-p)\left(\frac{2\beta C}{\lambda p}\right)^{\frac{2-p}{1-p}}$.

Combining (2.31) (3.8) and (3.9), we obtain

$$\begin{aligned} &\| \nabla_X \hat{H}(X^{k+1}, X^k, \boldsymbol{\delta}^{k+1})\|_F \\ \le& \|\nabla f(X^{k+1}) - f(Y^k)\|_F + \beta\|X^k - Y^k\|_F + \|\lambda U_r^{k+1}\left(\hat{\boldsymbol{w}}^{k+1} - \hat{\boldsymbol{w}}^k\right) {V_r^{k+1}}^\top\|_F \\ \le& (\lambda C_L + L_f)\|X^{k+1} - X^k\|_F + (L_f + \beta)\bar{\alpha}\|X^k - X^{k-1}\|_F \\ &+ 2\lambda C_L\|\boldsymbol{\delta}^0\|_\infty(\|\boldsymbol{\delta}^k\|_1 - \|\boldsymbol{\delta}^{k+1}\|_1). \end{aligned} \tag{3.10}$$

Similarly, we have $\|\nabla_Y \hat{H}(X^{k+1}, X^k, \boldsymbol{\delta}^{k+1})\|_F = \|\beta(X^{k+1} - X^k)\|_F \le \beta\|X^{k+1} - X^k\|_F$. As for $\boldsymbol{\delta}$,

$$\nabla_{\boldsymbol{\delta}} \hat{H}(X^{k+1}, X^k, \boldsymbol{\delta}^{k+1}) = 2\lambda\hat{\boldsymbol{w}}^{k+1} \circ \boldsymbol{\delta}^{k+1}. \tag{3.11}$$

It follows that

$$\begin{aligned}\|\nabla_{\boldsymbol{\delta}} \hat{H}(X^{k+1}, X^k, \boldsymbol{\delta}^{k+1})\|_2 &\leq \|\nabla_{\boldsymbol{\delta}} \hat{H}(X^{k+1}, X^k, \boldsymbol{\delta}^{k+1})\|_1 \\ &= \sum_{i=1}^{r} 2\lambda \hat{w}_i^{k+1} \delta_i^{k+1} \\ &\leq \sum_{i=1}^{r} 2\lambda \tfrac{2\beta C}{\lambda} \tfrac{\sqrt{\mu}}{1-\sqrt{\mu}} \left(\delta_i^k - \delta_i^{k+1}\right) \\ &\leq \tfrac{4\beta C\sqrt{\mu}}{1-\sqrt{\mu}} \left(\|\boldsymbol{\delta}^k\|_1 - \|\boldsymbol{\delta}^{k+1}\|_1\right),\end{aligned} \tag{3.12}$$

where the second inequality holds by Theorem 3 and $\delta_i^{k+1} \leq \sqrt{\mu}\delta_i^k, i \in [r]$.

Overall, it holds that

$$\|\nabla \hat{H}(X^{k+1}, X^k, \boldsymbol{\delta}^{k+1})\|_F \leq D_1 \left(\left\|X^k - X^{k+1}\right\|_F + \left\|X^{k-1} - X^k\right\|_F + \|\boldsymbol{\delta}^k\|_1 - \|\boldsymbol{\delta}^{k+1}\|_1\right),$$

with $D_1 = \max\left(\lambda C_L + L_f + \beta, (L_f + \beta)\bar{\alpha}, 2C_L\lambda\|\boldsymbol{\delta}^0\|_\infty + \frac{4\beta C\sqrt{\mu}}{1-\sqrt{\mu}}\right)$.

(ii) It holds with $D_2 = \frac{\beta}{2}\left(1 - \bar{\alpha}^2\frac{\beta+3L_f}{\beta}\right)$ from Lemma 2 (i).

(iii) It is obviously that $\{\hat{H}(X^{k+1}, Y^k, \boldsymbol{\delta}^{k+1})\}$ has unique limit point from Lemma 2 (ii), since Theorem 5 shows $H(X^{k+1}, X^k, \boldsymbol{\delta}^{k+1}) \to \zeta$ as $k \to \infty$.

(iv) Lemma 2 (i) and Theorem 5 (ii) show $\{\hat{H}(X^{k+1}, X^k, \boldsymbol{\delta}^{k+1})\} \downarrow \zeta$. If $\hat{H}(X^{\hat{k}+1}, X^{\hat{k}}, \boldsymbol{\delta}^{\hat{k}+1}) = \zeta$, then we have $X^{k+1} = X^k$ for all $k > \hat{k}$ from (ii). We are done. We only have to consider the case $\hat{H}(X^{k+1}, X^k, \boldsymbol{\delta}^{k+1}) > \zeta$ for all sufficiently large k. Since Assumption 3 holds, there exists Φ, η, ρ satisfy

$$\Phi'\left(\hat{H}(X, Y, \boldsymbol{\delta}) - \zeta\right) \operatorname{dist}\left(0, \partial\hat{H}(X, Y, \boldsymbol{\delta})\right) \geq 1, \tag{3.13}$$

for all $(X, Y, \boldsymbol{\delta}) \in \mathbb{U}((X^*, X^*, 0); \rho) \cap \left\{(X, Y, \boldsymbol{\delta}) \in \mathbb{S}_{sr} : \zeta < \hat{H}(X^{k+1}, X^k, \boldsymbol{\delta}^{k+1}) < \zeta + \eta\right\}$. Notice that χ is the cluster point of $\{X^k\}$, thus

$$\lim_{k\to\infty} \operatorname{dist}((X^k, X^{k-1}, \boldsymbol{\delta}^k), \Gamma) = 0.$$

That is to say there exists $k_1 \in \mathbb{N}$ such that $\operatorname{dist}((X^k, X^{k-1}, \boldsymbol{\delta}^k), \Gamma) < \rho$ for all $k > k_1$. Since $\{\hat{H}(X^{k+1}, X^k, \boldsymbol{\delta}^{k+1})\} \downarrow \zeta$, we know there exists $k_2 \in \mathbb{N}$ such that $\zeta < \hat{H}(X^{k+1}, X^k, \boldsymbol{\delta}^{k+1}) < \zeta + \eta$ for all $k > k_2$. From the smoothness of Φ, we have for any $k > \max(k_1, k_2)$

$$\Phi'\left(\hat{H}(X^{k+1}, X^k, \boldsymbol{\delta}^{k+1}) - \zeta\right) \operatorname{dist}\left(0, \partial\hat{H}(X^{k+1}, X^k, \boldsymbol{\delta}^{k+1})\right) \geq 1, \tag{3.14}$$

from (3.13). It follows the KL property that

$$\begin{aligned}
&\left[\Phi\left(\hat{H}(X^k, X^{k-1}, \boldsymbol{\delta}^k) - \zeta\right) - \Phi\left(\hat{H}(X^{k+1}, X^k, \boldsymbol{\delta}^{k+1}) - \zeta\right)\right] \\
&\times D_1\left(\|X^{k-2} - X^{k-1}\|_F + \|X^{k-1} - X^k\|_F + \|\boldsymbol{\delta}^{k-1}\|_1 - \|\boldsymbol{\delta}^k\|_1\right) \\
\geq&\left[\Phi\left(\hat{H}(X^k, X^{k-1}, \boldsymbol{\delta}^k) - \zeta\right) - \Phi\left(\hat{H}(X^{k+1}, X^k, \boldsymbol{\delta}^{k+1}) - \zeta\right)\right] \\
&\times \|\hat{H}(X^k, X^{k-1}, \boldsymbol{\delta}^k)\| \\
\geq&\left[\hat{H}(X^k, X^{k-1}, \boldsymbol{\delta}^k) - \hat{H}(X^{k+1}, X^k, \boldsymbol{\delta}^{k+1})\right]\Phi'\left(\hat{H}(X^k, X^{k-1}, \boldsymbol{\delta}^k) - \zeta\right) \\
&\times \|\hat{H}(X^k, X^{k-1}, \boldsymbol{\delta}^k)\| \\
\geq& D_2\|X^k - X^{k-1}\|_F^2,
\end{aligned} \tag{3.15}$$

where the first inequality holds by Lemma 6(i), the second inequality comes from the concavity and the last inequality follows from (3.13) and Lemma 6(ii) directly. Then it follows that

$$\begin{aligned}
&\|X^k - X^{k-1}\|_F \\
\leq&\sqrt{\frac{2D_1}{D_2}\left[\Phi\left(\hat{H}(X^k, X^{k-1}, \boldsymbol{\delta}^k) - \zeta\right) - \Phi\left(\hat{H}(X^{k+1}, X^k, \boldsymbol{\delta}^{k+1}) - \zeta\right)\right]} \\
&\times\sqrt{\frac{1}{2}\left(\|X^{k-2} - X^{k-1}\|_F + \|X^{k-1} - X^k\|_F + \|\boldsymbol{\delta}^{k-1}\|_1 - \|\boldsymbol{\delta}^k\|_1\right)} \\
\leq&\frac{D_1}{D_2}\left[\Phi\left(\hat{H}(X^k, X^{k-1}, \boldsymbol{\delta}^k) - \zeta\right) - \Phi\left(\hat{H}(X^{k+1}, X^k, \boldsymbol{\delta}^{k+1}) - \zeta\right)\right] \\
&+\frac{1}{4}\left(\left\|X^{k-2} - X^{k-1}\right\|_F + \left\|X^{k-1} - X^k\right\|_F + \|\boldsymbol{\delta}^{k-1}\|_1 - \|\boldsymbol{\delta}^k\|_1\right).
\end{aligned} \tag{3.16}$$

By subtracting $\frac{1}{2}\|X^k - X^{k-1}\|_F$ from both sides, we obtain

$$\begin{aligned}
\frac{1}{2}\|X^k - X^{k-1}\|_F \leq&\frac{D_1}{D_2}\left[\Phi\left(\hat{H}(X^k, X^{k-1}, \boldsymbol{\delta}^k) - \zeta\right) - \Phi\left(\hat{H}(X^{k+1}, X^k, \boldsymbol{\delta}^{k+1}) - \zeta\right)\right] \\
&+\frac{1}{4}\left(\|X^{k-2} - X^{k-1}\|_F - \|X^{k-1} - X^k\|_F + \|\boldsymbol{\delta}^{k-1}\|_1 - \|\boldsymbol{\delta}^k\|_1\right).
\end{aligned} \tag{3.17}$$

Summing up both sides from t to k, we have

$$\begin{aligned}
&\frac{1}{2}\sum_{l=t}^{k}\|X^l - X^{l-1}\|_F \\
\leq&\frac{D_1}{D_2}\left[\Phi\left(\hat{H}(X^t, X^{t-1}, \boldsymbol{\delta}^t) - \zeta\right) - \Phi\left(\hat{H}(X^{k+1}, X^k; \boldsymbol{\delta}^{k+1}) - \zeta\right)\right] \\
&+\frac{1}{4}\left(\|X^{t-2} - X^{t-1}\|_F - \|X^{k-1} - X^k\|_F + \|\boldsymbol{\delta}^{t-1}\|_1 - \|\boldsymbol{\delta}^k\|_1\right).
\end{aligned} \tag{3.18}$$

Taking $k \to \infty$, we know $\boldsymbol{\delta}^k \to 0$ and $\|X^k - X^{k-1}\|_F \to 0$, and that $\Phi\left(\hat{H}(X^{k+1}, X^k; \boldsymbol{\delta}^k) - \zeta\right) \to \Phi(\zeta - \zeta) = 0$. Therefore,

$$\begin{aligned} T^t &= \sum_{l=t}^{\infty} \|X^l - X^{l-1}\|_F \\ &\leq \frac{2D_1}{D_2}\Phi\left(\hat{H}(X^t, X^{t-1}, \boldsymbol{\delta}^t) - \zeta\right) + \frac{1}{2}\left(\|X^{t-2} - X^{t-1}\|_F + \|\boldsymbol{\delta}^{t-1}\|_1\right), \\ &< +\infty. \end{aligned} \tag{3.19}$$

■

Now we are ready to prove the convergence rate under KL property.

Theorem 7 (Local convergence rate) *Suppose Assumptions 1-3 are satisfied. Let $\{X^k\}$ be generated by Algorithm 2 and converges to a critical point X^* of $F(X)$. Then the following statements hold.*

(i). If $\theta = 0$, then $X^k \equiv X^*$ for sufficient large k,

(ii). If $\theta \in (0, \frac{1}{2}]$, then there exists $\gamma \in (0,1)$ and $c_0, c_1 > 0$ such that

$$\|X^k - X^*\|_F \leq c_0\gamma^k - c_1\|\boldsymbol{\delta}^k\|_1 \tag{3.20}$$

for sufficient large k,

(iii). If $\theta \in (\frac{1}{2}, 1)$, then there exists $d_0, d_1 > 0$ such that

$$\|X^k - X^*\|_F \leq d_0 k^{-\frac{1-\theta}{2\theta-1}} - d_1\|\boldsymbol{\delta}^k\|_1 \tag{3.21}$$

for sufficient large k.

Proof Since $X^k \to X^*$, we have

$$\|X^k - X^*\|_F = \|X^k - \lim_{t\to\infty} X^t\|_F = \|\sum_{l=k}^{\infty}(X^l - X^{l+1})\|_F \leq T^k, \tag{3.22}$$

and

$$T^k \leq \frac{2D_1}{D_2}\Phi\left(\hat{H}(X^k, X^{k-1}, \boldsymbol{\delta}^k) - \zeta\right) + \frac{1}{2}(T^{k-2} - T^{k-1}) + \frac{1}{2}\|\boldsymbol{\delta}^{k-1}\|_1, \tag{3.23}$$

by (3.19). Thus we only need to prove T^k has the same upper bound for case (ii) and (iii).

(i) If $\theta = 0$: then $\Phi'(s) = cs$ and $\Phi'(s) = c$, we claim that there exists $\hat{k}$ such that $\hat{H}(X^{\hat{k}}, X^{\hat{k}-1}, \boldsymbol{\delta}^{\hat{k}},, X^{\hat{k}-1}) = \zeta$. Suppose by contradiction this is not true so that $\hat{H}(X^k, X^{k-1}, \boldsymbol{\delta}^k) > \zeta$ for all k. Since X^k converges to X^* and $\{\hat{H}(X^k, X^{k-1}, \boldsymbol{\delta}^k)\}$ is monotonically decreasing to ζ by Lemma 6(ii). We have the KL inequality

$$\|\nabla\hat{H}(X^k, X^{k-1}, \boldsymbol{\delta}^k)\| \geq \frac{1}{c}, \tag{3.24}$$

with $\Phi'(s) = c$. This is a contradiction with $\|\hat{H}(X^k, X^{k-1}, \boldsymbol{\delta}^k)\|_2 \to 0$ by Lemma 6(i). Thus, there exists $\hat{k} \in \mathbb{N}$ such that $\hat{H}(X^k, X^{k-1}, \boldsymbol{\delta}^k) = \hat{H}(X^{\hat{k}}, X^{\hat{k}-1}, \boldsymbol{\delta}^{\hat{k}}) = \zeta$ for all $k \geq \hat{k}$. Hence, we conclude from Lemma 6(ii) that $X^{k+1} = X^k$ for all $k > \hat{k}$, i.e., the sequence converges in a finite numbers of steps. This proves (i).

(ii)-(iii) For $\theta \in (0,1)$: if there exists $\hat{k} \in \mathbb{N}$ such that $\hat{H}(X^{\hat{k}}, X^{\hat{k}-1}, \boldsymbol{\delta}^{\hat{k}}, X^{\hat{k}-1}) = \zeta$, then this reverts to (i). Therefore, we only need to consider the case that $\hat{H}(X^k, X^{k-1}, \boldsymbol{\delta}^k) > \zeta$ for all k.

Assumption 3 implies

$$c(1-\theta)(\hat{H}(X^k, X^{k-1}, \boldsymbol{\delta}^k) - \zeta)^{-\theta} \mathrm{dist}\left(0, \partial \hat{H}(X^k, X^{k-1}, \boldsymbol{\delta}^k)\right) \geq 1, \tag{3.25}$$

for all $k > \bar{k}$ from Lemma 6(iv) with $\Phi'(s) = c(1-s)^{-\theta}$. On the other hand, from Lemma 6(i) we obtain

$$\mathrm{dist}\left(0, \partial \hat{H}(X^k, X^{k-1}, \boldsymbol{\delta}^k)\right) \leq d_1 \left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1 - \|\boldsymbol{\delta}^k\|_1\right). \tag{3.26}$$

This, combined with (3.25), yields

$$(\hat{H}(X^k, X^{k-1}, \boldsymbol{\delta}^k) - \zeta)^{\theta} \leq c(1-\theta) d_1 \left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1 - \|\boldsymbol{\delta}^k\|_1\right). \tag{3.27}$$

Taking a power of $\frac{1-\theta}{\theta}$ to both sides of the above inequality and scaling both sides by c, we have

$$\begin{aligned} \Phi(\hat{H}(X^k, X^{k-1}, \boldsymbol{\delta}^k) - \zeta) =& c(\hat{H}(X^k, X^{k-1}, \boldsymbol{\delta}^k) - \zeta)^{1-\theta} \\ &\leq c\left[c(1-\theta) d_1 \left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1 - \|\boldsymbol{\delta}^k\|_1\right)\right]^{\frac{1-\theta}{\theta}} \\ &\leq c\left[c(1-\theta) d_1 \left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1\right)\right]^{\frac{1-\theta}{\theta}}. \end{aligned} \tag{3.28}$$

This, together with (3.23), yields

$$T^k \leq \nu_1 \left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1\right)^{\frac{1-\theta}{\theta}} + \frac{1}{2}\left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1\right) \tag{3.29}$$

with $\nu_1 = \frac{2cd_1}{d_2}[c(1-\theta)d_1]^{\frac{1-\theta}{\theta}}$. From Theorem 4(ii), we have that $\delta_i^k \leq \sqrt{\mu}\delta_i^{k-1}$, which indicates

$$\delta_i^{k-1} \leq \frac{\sqrt{\mu}}{1-\mu}(\delta_i^{k-2} - \delta_i^k), \tag{3.30}$$

resulting in

$$\|\boldsymbol{\delta}^{k-1}\|_1 \leq \frac{\sqrt{\mu}}{1-\mu}(\|\boldsymbol{\delta}^{k-2}\|_1 - \|\boldsymbol{\delta}^k\|_1). \tag{3.31}$$

It follows from (3.29) that

$$
\begin{aligned}
&T^k + \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^k\|_1 \\
&\leq \nu_1 \left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1\right)^{\frac{1-\theta}{\theta}} + \frac{1}{2}\left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1\right) + \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^k\|_1 \\
&\leq \nu_1 \left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1\right)^{\frac{1-\theta}{\theta}} + \frac{1}{2}\left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1\right) + \frac{\mu}{1-\mu}\|\boldsymbol{\delta}^{k-1}\|_1 \\
&\leq \nu_1 \left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1\right)^{\frac{1-\theta}{\theta}} + \nu_2\left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1\right),
\end{aligned}
\tag{3.32}
$$

with $\nu_2 = \frac{1}{2} + \frac{\mu}{1-\mu}$, where the second inequality holds by the norm inequality, the last inequality is by $T^{k-2} - T^k \geq 0$.

For part (ii), $\theta \in (0, \frac{1}{2}]$. Noticed that

$$
\frac{1-\theta}{\theta} \geq 1 \text{ and } \lim_{k\to\infty} T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1 = 0.
$$

It holds from Lemma 6(iii) that for sufficiently large k,

$$
\left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1\right)^{\frac{1-\theta}{\theta}} \leq \left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1\right), \tag{3.33}
$$

This, combined with (3.32), yields

$$
T^k + \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^k\|_1 \leq (\nu_1 + \nu_2)\left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1\right). \tag{3.34}
$$

Therefore, we have

$$
\begin{aligned}
&T^k + \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^k\|_1 \\
&\leq \frac{\nu_1+\nu_2}{\nu_1+\nu_2+1}(T^{k-2} + \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^{k-2}\|_1) \\
&\leq \left(\frac{\nu_1+\nu_2}{\nu_1+\nu_2+1}\right)^{\lfloor\frac{k-k_1}{2}\rfloor}\left(T^{[(k-k_1)\bmod 2]+k_1} + \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^{[(k-k_1)\bmod 2]+k_1}\|_1\right) \\
&\leq \left(\frac{\nu_1+\nu_2}{\nu_1+\nu_2+1}\right)^{\frac{k-k_1}{2}}\left(T^{k_1} + \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^{k_1}\|_1\right),
\end{aligned}
\tag{3.35}
$$

where the first inequality holds by (3.30). Hence,

$$
\|X^k - X^*\|_F \leq T^k \leq \nu_1\gamma^k - \nu_2\|\boldsymbol{\delta}^k\|_1, \tag{3.36}
$$

holds for all $k \geq k_1$, with

$$
\gamma = \sqrt{\frac{\nu_1+\nu_2}{\nu_1+\nu_2+1}}, c_0 = \frac{T^{k_1} + \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^{k_1}\|_1}{\gamma^{k_1}}, c_1 = \frac{\sqrt{\mu}}{1-\mu}.
$$

This completes the proof of (ii). For part (iii), $\theta \in (\frac{1}{2}, 1)$. Notice that

$$\frac{1-\theta}{\theta} < 1 \text{ and } \lim_{k\to\infty} T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1 = 0.$$

Hence, for sufficiently large k

$$\left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1\right) \le \left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1\right)^{\frac{1-\theta}{\theta}}. \tag{3.37}$$

This, combined with (3.32), results in

$$T^k + \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^k\|_1 \le (\nu_1 + \nu_2)\left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1\right)^{\frac{1-\theta}{\theta}}. \tag{3.38}$$

Raising a power of $\frac{1-\theta}{\theta}$ to the both sides of (3.38),

$$\left(T^k + \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^k\|_1\right)^{\frac{\theta}{1-\theta}} \le \nu_3\left[(T^{k-2} + \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^{k-2}\|_1) - (T^k + \|\boldsymbol{\delta}^k\|_1)\right], \tag{3.39}$$

by (3.31), where $\nu_3 = (\nu_1 + v_2)^{\frac{\theta}{1-\theta}}$. For the "even" subsequence of $\{k_2, k_2+1, \cdots\}$, define $\delta_t := T^{2t} + \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^{2t}\|_1$ for $t \ge \lceil \frac{k_2}{2} \rceil := N_1$. Following from the proof for (Wang et al., 2022; Attouch and Bolte, 2009, Theorem 4,Theorem 2), we have

$$\delta_k \le \left(\delta_{N_1-1}^{\frac{1-2\theta}{1-\theta}} + \nu(k - N_1)\right)^{-\frac{1-\theta}{2\theta-1}} \le d_2 k^{-\frac{1-\theta}{2\theta-1}}, \tag{3.40}$$

for some $d_2 > 0$. As for the "odd" subsequence of $\{k_2, k_2+1, \cdots\}$, define $\delta_t := T^{2t+1} + \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^{2t+1}\|_1$. Then (3.40) still holds true. Therefore, for all sufficiently large and even number k it holds that

$$\|X^k - X^*\|_F \le T^k = \delta_{\frac{k}{2}} - \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^k\|_1 \le 2^{\frac{1-\theta}{2\theta-1}} d_2 k^{-\frac{1-\theta}{2\theta-1}} - \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^k\|_1. \tag{3.41}$$

For all sufficiently large and odd number k,

$$\|X^k - X^*\|_F \le T^k = \delta_{\frac{k-1}{2}} - \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^k\|_1 \le 2^{\frac{1-\theta}{2\theta-1}} d_2 (k-1)^{-\frac{1-\theta}{2\theta-1}} - \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^k\|_1. \tag{3.42}$$

Overall, we have

$$\|X^k - X^*\|_F \le d_0 k^{-\frac{1-\theta}{2\theta-1}} - d_1\|\boldsymbol{\delta}^k\|_1, \tag{3.43}$$

for sufficiently large k, where $d_0 = 2^{\frac{1-\theta}{2\theta-1}} d_2 \max\left(1, 2^{\frac{1-\theta}{2\theta-1}}\right)$ and $d_1 = \frac{\sqrt{\mu}}{1-\mu}$. This completes the proof of (iii). ■

4 Numerical results

In this section, we perform low rank matrix completion experiments on both synthetic data and nature images to demonstrate the performance of the IRNRI and the EIRNRI algorithm. The examples are all concerned with solving the matrix completion problem, i.e.,

$$\min_{X} \frac{1}{2}\|M - \mathbf{P}_\Omega(X)\|_F^2 + \lambda\|X\|_p^p, \tag{4.1}$$

where $M \in \mathbb{R}^{m\times n}$, Ω is the set of indices of samples, and $\mathbf{P}_\Omega$ is the projection onto the subspace of sparse matrices with nonzeros restricted to the index set Ω. In this case, the Lipschitz constant $L_f = 1$.

The codes of all methods tested in this section are implemented in Matlab on a desktop with an Intel Core i5-8500 CPU (3.00 GHz) and 24GB RAM running 64-bit Windows 10 Enterprise. In the experiments, the relative error is recorded as

$$\text{RelErr} = \frac{\|X^k - X^*\|_F}{\|X^*\|_F}, \tag{4.2}$$

For a given point X, the SVD of $X = U_r\Sigma V_r$. To verify whether X is a critical point of problem (4.1), i.e., $\text{dist}(0, \partial F(X)) = 0$, we calculate Sun et al. (2017) the distance by

$$\text{dist}(0, \partial F(X)) = \|U_r^\top \nabla f(X) V_r + \lambda p \Sigma^{p-1}\|_F. \tag{4.3}$$

and the relative distance error as

$$\text{Reldist} \le \frac{\text{dist}(0, \partial F(X))}{\|M\|_F}. \tag{4.4}$$

We adopt the same criterion as used in Recht et al. (2010); Sun et al. (2017), and say a matrix X^* is successfully recovered by X^k if the corresponding relative error or the relative distance is less than 10^{-5}.

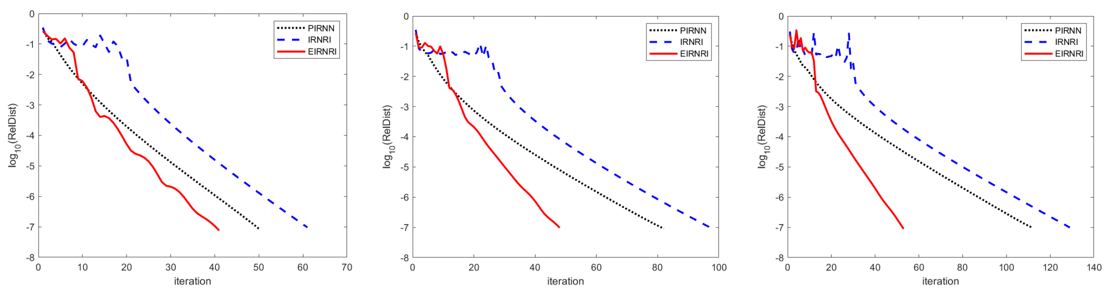
4.1 Synthetic data

For synthetic data, our target is to recover the generated matrix X^* from M. Thus, we call a correct low-rank detecting (CLD) is that the rank of termination point is the same with rank of X^*. In the experiments, we compare the performance of IRNRI and EIRNRI with PIRNN Sun et al. (2017) for solving (4.1) on random data. We set $m = n = 150, r = 5, 10, 15$ for the matrix X^*, select $p = 0.5$ for Schatten-p norm. The rank r matrix X^* is generated by $X^* = BC$, where $B \in \mathbb{R}^{m\times r}$ and $C \in \mathbb{R}^{r\times n}$ are generated randomly with i.i.d. standard Gaussian entries. We randomly sample a subset Ω with sampling ratio $\text{SR} = |\Omega|/(mn)$, and select three values $0.2, 0.5$ and 0.8, and then the observed matrix is $M = \mathbf{P}_\Omega(X^*)$.

All algorithms start from a random Gaussian matrix X^0, and terminate when $\text{RelErr} \le$ *opttol* or $\text{Reldist} \le$ *opttol* or the number of iteration exceeds the preset limit *itmax*. Furthermore, we add another termination condition $\|X^{k+1} - X^k\|_\infty \le KLopt$ by Lemma 6. Unless otherwise mentioned, we use the following parameters to run the experiments: $\beta = 1.1 > L_f$, $\epsilon_i^0 = 1, i \in [m]$ for IRNRI and EIRNRI, $opttol = 10^{-5}$, $itmax = 1\times 10^3$ and $KLopt = 10^{-7}$. For EIRNRI, $\alpha = 0.7$ is roughly tuned for the best performance. For PIRNN, $\epsilon_i, i \in [m]$ are fixed as sufficiently small values: $\epsilon_i = 10^{-3}$.

The number of problems (out of 2000 problems in total) converging to a solution with the correct rank satisfying $\mathrm{rank}(X_{true}) = \mathrm{rank}(X^*)$ are depicted in Figure ??. The average of the final relative errors of each algorithm for each case are shown in Table ??. We can observe from these results that our proposed algorithms can always return a solution with low relative error and the correct rank.

To compare the efficiency of the algorithms, we also plot the evolution of the relative error for each case in Figure ??. We can see that the proposed IRNRI achieves roughly the same speed as PIRNN, while the proposed EIRNRI significantly outperforms PIRNN.

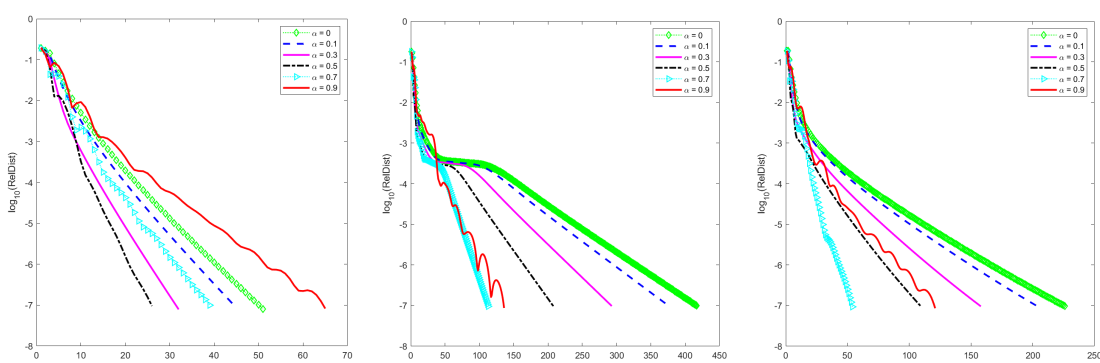


(a) $\mathrm{rank}(X*) = 5$ (b) $\mathrm{rank}(X*) = 10$ (c) $\mathrm{rank}(X*) = 15$

Figure 1: Comparison of matrix recovery on synthetic data initial point with $\mathrm{rank}(X^0) = \mathrm{rank}(X^*)$. We set the perturbation value to be the same for PIRNN, IRNRI, and EIRNRI at $\epsilon = 10^{-16}$, while keeping the other parameters at their default values.

Figure 1 shows IRNRI has the similar performance with PIRNN, [?] EIRNRI has better performance than the both if the perturbation is the same.

Since Algorithm 2 introduced a new parameter α to accelerate the convergence, so we set the different values of $\alpha = \{0, 0.1, 0.3, 0.5, 0.7, 0.9\}$ to test the sensitivity to α_k for the EIRNRI algorithm. The experiments shows the constant momentum coefficient accelerate the convergence of the IRNRI, and when $\alpha \approx 0.7$ EIRNRI has the best performance.



(a) the relative error (b) the relative distance (c) the value of primal problem

Figure 2: The performance for different α with $\mathrm{rank}(X^0) = 15$, SR=0.5, $\mu = 0.8$. It presents that α around 0.7 has better performance.

4.2 Application to Image Recovery

In this section, we compare our method with three solvers, PIRNN Sun et al. (2017) and Scp Sun et al. (2013)[2], FGSRp Fan et al. (2019)[3]. We consider the real images though they are usually are not low rank, but the top singular values of the real images dominate the main information. The nature images are of scale $300 \times 300 \times 3$, we observe the top singular value firstly, and sample the 80% of the elements uniformly for the random mask. The other mask ????. We set $\lambda = 1$ as the regularization parameter. The iteration is terminated when RelDist $\leq 10^{-5}$ or KLopt $\leq 10^{-5}$ or the number of iteration exceeds *itmax*$= 10^3$. We set $\beta = 1.1$ for the three algorithms. In order to make the PIRNN has the accurate solution we set $\epsilon = 0.0001$ for PIRNN, then set $\epsilon^0 = 1, \mu = 0.1$ for IRNRI and EIRNRI, set $\alpha = 0.8$. The performance of all algorithms are compared on two measurements (1) the difference of the rank of the final iterate and X^*, (2) peak signal-to-noise ratio (PSNR)

Figure **??** depicts ....

$$\mathrm{PSNR}(M, X^*) = 10 \times \log_{10} \frac{255^2}{\frac{1}{3mn} \sum_{i=1}^{3} \|X_i^* - M_i\|_F^2}, \tag{4.5}$$

where M is the original image and X^* is the restored image,

Although Figure 3 and Figure 4 demonstrate that all of these methods can recover the image, the iterative reweighted method appears to achieve a low-rank recovery. Scp makes the biggest PSNR in both task, but it may have a small flaw in Figure 4. Since FGSRp is equivalent the Schatten-p norm, the parameters λ may not appropriate in this task. To provide more information, Table 1 includes additional low-rank tasks.

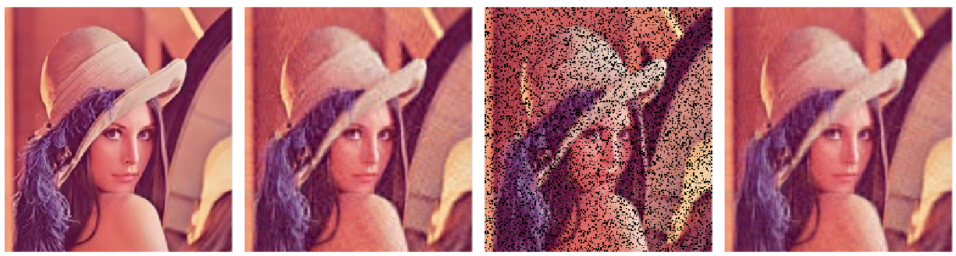
	PIRNN		IRNRI		EIRNRI		Scp		FGSR	
	psnr	rank	psnr	rank	psnr	rank	psnr	rank	psnr	rank
$\mathrm{rank}(X^*) = 15$	24.932	15	24.932	15	24.932	15	24.973	187	18.703	187
$\mathrm{rank}(X^*) = 20$	26.444	20	26.444	20	26.444	20	26.552	187	18.949	187
$\mathrm{rank}(X^*) = 25$	27.57	25	27.57	25	27.57	25	27.812	187	19.081	187
$\mathrm{rank}(X^*) = 30$	28.359	30	28.359	30	28.359	30	28.818	187	19.313	187
$\mathrm{rank}(X^*) = 35$	28.546	32	28.546	32	28.546	32	29.432	187	19.268	187
$\mathrm{rank}(X^*) = 40$	28.599	32	28.599	32	28.599	32	29.848	187	19.416	187

Table 1: The performance for different mmethod with different low-rank target, bold values correspond to the best results for each algorithm. It presents that the proper perturbation affecting the algorithm less, A bigger ϵ makes PIRNN converge slower, a too small ϵ may make PIRNN get a terrible solution. The experiment shows IRNRI and EIRNRI are more robust than PIRNN when the perturbation bigger than 10^{-4}.

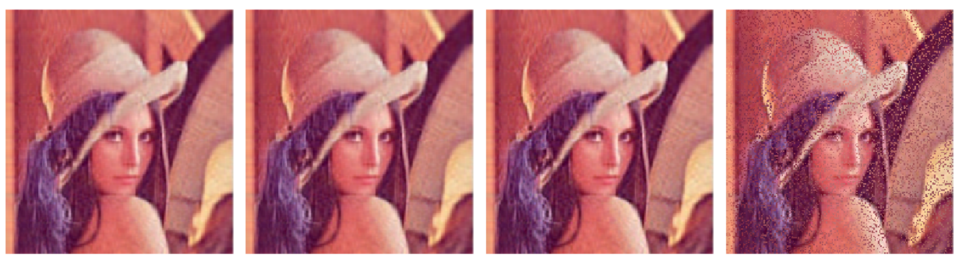
5 Conclusion and Future Work

In this paper, we proposed, analyzed and implemented iteratively reweighted Nuclear norm methods for solving the Schatten-p norm regularized low-rank optimization. Therefore two

2. Available: https://github.com/liguorui77/scpnorm
3. Available: https://github.com/udellgroup/Codes-of-FGSR-for-effecient-low-rank-matrix-recovery

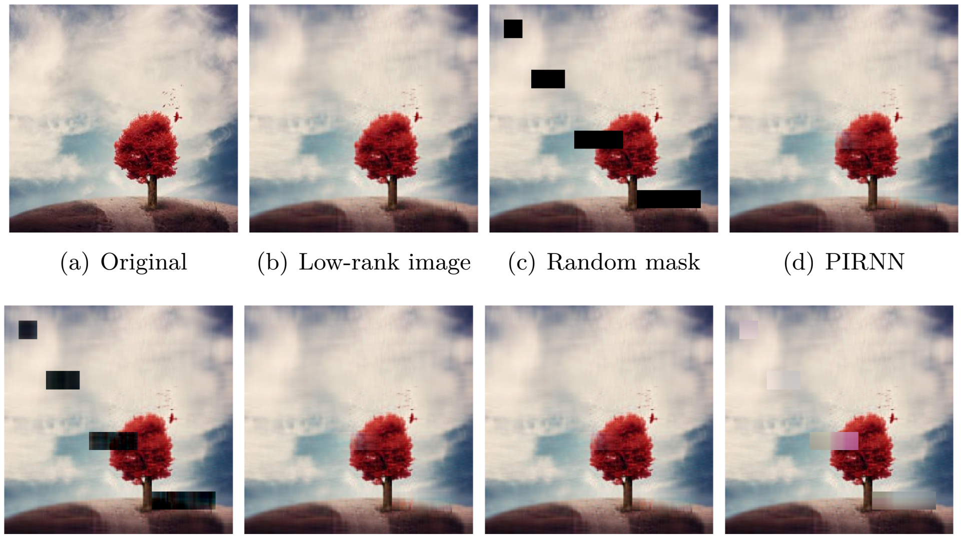


(a) Original (b) Low-rank image (c) Random mask (d) PIRNN



(e) AIRNN (f) EIRNRI (g) Scp (h) FGSRp

Figure 3: The performance of different methods with a random mask in image recovery with SR= 0.8. The performance for different methods with random mask in image recovery, the (a) Original image, $\text{rank}(X) = 300$; (b) Low-rank image, $\text{rank}(X^*) = 30$ (c) Noised picture; (d) PIRNN: $\text{rank}(X) = 30$, PSNR=28.855; (e) IRNRI: $\text{rank}(X) = 30$, PSNR=28.854; (f) EIRNRI: $\text{rank}(X) = 30$, PSNR=28.855; (g) Scp: $\text{rank}(X) = 30$, PSNR=28.971; (h) FGSR: $\text{rank}(X) = 30$, PSNR=21.806;



(a) Original (b) Low-rank image (c) Random mask (d) PIRNN

(e) AIRNN (f) EIRNRI (g) Scp (h) FGSRp

Figure 4: The performance for different methods with block mask, the (a) Original image, $\text{rank}(X) = 300$; (b) Low-rank image, $\text{rank}(X^*) = 30$ (c) Noised picture; (d) PIRNN: $\text{rank}(X) = 30$, PSNR=28.855; (e) IRNRI: $\text{rank}(X) = 30$, PSNR=28.854; (f) EIRNRI: $\text{rank}(X) = 30$, PSNR=28.855; (g) Scp: $\text{rank}(X) = 30$, PSNR=28.971; (h) FGSR: $\text{rank}(X) = 30$, PSNR=21.806;

main novel features of our work. The first is the rank identification property possessed by the proposed methods, meaning the algorithm can find the correct rank in finite iterations. We believe this is the first work of pointing out the model identification property (a famous property in vector optimization) in matrix optimization. Based on this property, we also designed a novel updating strategy for ϵ_i to smooth the Schatten-p norm, so that ϵ_i associated with the positive singular values can be driven to 0 rapidly and those associated with the 0 singular values can be automatically fixed as constants after finite iterations. The crucial role of this strategy is that the algorithm eventually behaves like a truncated weighted Nuclear norm method, so that the techniques for smooth algorithms can directly applied including acceleration techniques and convergence analysis.

The convergence properties that we have proved for our algorithm were illustrated empirically on test sets of synthetic and real data sets. We remark, however, that many remaining practical issues can be resolved to further improve the performance of the proposed method. For example, one can incorporate the rank identification property into the implementation. Since the correct rank has been detected after finite iterations, the algorithm can be terminated and switch to a traditional Frobenius recovery with fixed rank to further improve the quality of the recovered solution. We would like leave it to the further work.

Acknowledgments and Disclosure of Funding

We would like to acknowledge the support for this paper from the and the Natural Science Foundation of Shanghai under Grant 21ZR1442800 and the Young Scientists Fund of the National Natural Science Foundation of China No. 12001367.

References

Yonatan Amit, Michael Fink, Nathan Srebro, and Shimon Ullman. Uncovering shared structures in multiclass classification. In *Proceedings of the 24th international conference on Machine learning*, pages 17–24, 2007.

Hedy Attouch and Jérôme Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116:5–16, 2009.

Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, 20(4):1956–1982, 2010.

Kai-Yang Chiang, Inderjit S Dhillon, and Cho-Jui Hsieh. Using side information to reliably learn low-rank matrices from missing and corrupted observations. *The Journal of Machine Learning Research*, 19(1):3005–3039, 2018.

Jicong Fan, Lijun Ding, Yudong Chen, and Madeleine Udell. Factor group-sparse regularization for efficient low-rank matrix recovery. *Advances in Neural Information Processing Systems*, 32:5104–5114, 2019.

Maryam Fazel. *Matrix rank minimization with applications*. PhD thesis, PhD thesis, Stanford University, 2002.

Guillaume Garrigos. *Descent dynamical systems and algorithms for tame optimization, and multi-objective problems*. PhD thesis, 2015.

Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. Weighted nuclear norm minimization with application to image denoising. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2862–2869, 2014.

Warren L Hare. Identifying active manifolds in regularization problems. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 261–271. Springer, 2011.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Yao Hu, Debing Zhang, Jieping Ye, Xuelong Li, and Xiaofei He. Fast and accurate matrix completion via truncated nuclear norm regularization. *IEEE transactions on pattern analysis and machine intelligence*, 35(9):2117–2130, 2012.

Zhanxuan Hu, Feiping Nie, Rong Wang, and Xuelong Li. Low rank regularization: A review. *Neural Networks*, 136:218–232, 2021.

Chen Huang, Xiaoqing Ding, Chi Fang, and Di Wen. Robust image restoration via adaptive low-rank approximation and joint kernel regression. *IEEE Transactions on Image Processing*, 23(12):5284–5297, 2014.

Piotr Indyk, Ali Vakilian, and Yang Yuan. Learning-based low-rank approximations. *Advances in Neural Information Processing Systems*, 32, 2019.

KS Jun, R Willett, R Nowak, and S Wright. Bilinear bandits with low-rank structure. *Proceedings of Machine Learning Research*, 97, 2019.

Quentin Klopfenstein, Quentin Bertrand, Alexandre Gramfort, Joseph Salmon, and Samuel Vaiter. Model identification and local linear convergence of coordinate descent. *arXiv preprint arXiv:2010.11825*, 2020.

Ming-Jun Lai, Yangyang Xu, and Wotao Yin. Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization. *SIAM Journal on Numerical Analysis*, 51 (2):927–957, 2013.

Joonseok Lee and Seungyeon Kim. Llorma: Local low-rank matrix approximation. *Journal of Machine Learning Research*, 17:1–24, 2016.

Adrian S Lewis and Hristo S Sendov. Nonsmooth analysis of singular values. part i: Theory. *Set-Valued Analysis*, 13(3):213–241, 2005.

Guoyin Li and Ting Kei Pong. Calculus of the exponent of kurdyka–łojasiewicz inequality and its applications to linear convergence of first-order methods. *Foundations of computational mathematics*, 18(5):1199–1232, 2018.

Jingwei Liang, Jalal Fadili, and Gabriel Peyré. Local linear convergence of forward–backward under partial smoothness. *Advances in neural information processing systems*, 27, 2014.

Jingwei Liang, Jalal Fadili, and Gabriel Peyré. Activity identification and local linear convergence of forward–backward-type methods. *SIAM Journal on Optimization*, 27(1): 408–437, 2017.

Canyi Lu, Jinhui Tang, Shuicheng Yan, and Zhouchen Lin. Generalized nonconvex nonsmooth low-rank minimization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4130–4137, 2014.

Xin Luo, MengChu Zhou, Shuai Li, Yunni Xia, Zhu-Hong You, QingSheng Zhu, and Hareton Leung. Incorporation of efficient second-order solvers into latent factor models for accurate prediction of missing qos data. *IEEE transactions on cybernetics*, 48(4):1216–1228, 2017.

Mathurin Massias, Alexandre Gramfort, and Joseph Salmon. Celer: a fast solver for the lasso with dual extrapolation. In *International Conference on Machine Learning*, pages 3315–3324. PMLR, 2018.

Karthik Mohan and Maryam Fazel. Iterative reweighted algorithms for matrix rank minimization. *The Journal of Machine Learning Research*, 13(1):3441–3473, 2012.

Yurii E Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547, 1983.

Feiping Nie, Heng Huang, and Chris Ding. Low-rank matrix recovery via efficient schatten p-norm minimization. In *Twenty-sixth AAAI conference on artificial intelligence*, 2012.

Mila Nikolova and Pauline Tan. Alternating structure-adapted proximal gradient descent for nonconvex block-regularised problems. 2018.

Tae-Hyun Oh, Hyeongwoo Kim, Yu-Wing Tai, Jean-Charles Bazin, and In So Kweon. Partial sum minimization of singular values in rpca for low-level vision. *methods*, 19:24, 2013.

Soumyabrata Pal and Prateek Jain. Online low rank matrix completion. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=47KG_AvNqeZ.

Duy Nhat Phan and Thuy Ngoc Nguyen. An accelerated irnn-iteratively reweighted nuclear norm algorithm for nonconvex nonsmooth low-rank minimization problems. *Journal of Computational and Applied Mathematics*, 396:113602, 2021.

Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Qian Sun, Shuo Xiang, and Jieping Ye. Robust principal component analysis via capped norms. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 311–319, 2013.

Tao Sun, Hao Jiang, and Lizhi Cheng. Convergence of proximal iteratively reweighted nuclear norm algorithm for image processing. *IEEE Transactions on Image Processing*, 26(12):5632–5644, 2017.

Kim-Chuan Toh and Sangwoon Yun. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of optimization*, 6(615-640):15, 2010.

Tian Tong, Cong Ma, and Yuejie Chi. Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *The Journal of Machine Learning Research*, 22(1): 6639–6701, 2021.

Charles F Van Loan. Generalizing the singular value decomposition. *SIAM Journal on numerical Analysis*, 13(1):76–83, 1976.

Hao Wang, Hao Zeng, Jiashan Wang, and Qiong Wu. Relating ℓ_p regularization and reweighted ℓ_1 regularization. *Optimization Letters*, 15(8):2639–2660, 2021a.

Hao Wang, Fan Zhang, Yuanming Shi, and Yaohua Hu. Nonconvex and nonsmooth sparse optimization via adaptively iterative reweighted methods. *Journal of Global Optimization*, 81:717–748, 2021b.

Hao Wang, Hao Zeng, and Jiashan Wang. An extrapolated iteratively reweighted ℓ_1 method with complexity analysis. *Computational Optimization and Applications*, 83(3):967–997, 2022.

Zhi Wang, Wendong Wang, Jianjun Wang, and Siqi Chen. Fast and efficient algorithm for matrix completion via closed-form 2/3-thresholding operator. *Neurocomputing*, 330: 212–222, 2019.

Zhi Wang, Dong Hu, Xiaohu Luo, Wendong Wang, Jianjun Wang, and Wu Chen. Performance guarantees of transformed schatten-1 regularization for exact low-rank matrix recovery. *International Journal of Machine Learning and Cybernetics*, 12(12):3379–3395, 2021c.

Fei Wen, Lei Chu, Peilin Liu, and Robert C Qiu. A survey on nonconvex regularization-based sparse and low-rank recovery in signal processing, statistics, and machine learning. *IEEE Access*, 6:69883–69906, 2018.

Qiong Wu, Fan Zhang, Hao Wang, and Yuanming Shi. Generalized low-rank matrix completion via nonconvex schatten p-norm minimization. In *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*, pages 1–5. IEEE, 2018.

Yangyang Xu and Wotao Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on imaging sciences*, 6(3):1758–1789, 2013.

Debing Zhang, Yao Hu, Jieping Ye, Xuelong Li, and Xiaofei He. Matrix completion by truncated nuclear norm regularization. In *2012 IEEE Conference on computer vision and pattern recognition*, pages 2192–2199. IEEE, 2012.

Hengmin Zhang, Jianjun Qian, Bob Zhang, Jian Yang, Chen Gong, and Yang Wei. Low-rank matrix recovery via modified schatten-p norm minimization with convergence guarantees. *IEEE Transactions on Image Processing*, 29:3132–3142, 2019.

Fujun Zhao, Jigen Peng, and Angang Cui. Design strategy of thresholding operator for low-rank matrix recovery problem. *Signal Processing*, 171:107510, 2020.