Journal of Machine Learning Research 25 (2024) 1-44 Submitted 4/23 Revised 8/24; Published 9/24

Efficient Active Manifold Identification via Accelerated Iteratively Reweighted Nuclear Norm Minimization

Hao Wang HAW309@GMAIL.COM
School of Information Science and Technology
ShanghaiTech University
Shanghai, 201210, China

Ye Wang WANGYE_77@OUTLOOK.COM
School of Information Science and Technology
ShanghaiTech University
Shanghai, 201210, China

Xiangyu Yang* YANGXY@HENU.EDU.CN
School of Mathematics and Statistics
Henan University
Kaifeng, 475000, China
Center for Applied Mathematics of Henan Province
Henan University
Zhengzhou, 450046, China

Editor: Lam Nguyen

Abstract

This paper considers the problem of minimizing the sum of a smooth function and the Schatten-p norm of the matrix. Our contribution involves proposing accelerated iteratively reweighted nuclear norm methods designed to solve the nonconvex low-rank minimization problem. Two major novelties characterize our approach. First, the proposed method possesses an active manifold identification property, enabling the provable identification of the correct rank of the stationary point within a finite number of iterations. Second, we introduce an adaptive updating strategy for smoothing parameters. This strategy automatically fixes parameters associated with zero singular values as constants upon detecting the correct rank while quickly driving the remaining parameters to zero. This adaptive behavior transforms the algorithm into one that effectively solves smooth problems after a few iterations, setting our work apart from existing iteratively reweighted methods for low-rank optimization. We prove the global convergence of the proposed algorithm, guaranteeing that every limit point of the iterates is a critical point. Furthermore, a local convergence rate analysis is provided under the Kurdyka-Łojasiewicz property. We conduct numerical experiments using both synthetic and real data to showcase our algorithm's efficiency and superiority over existing methods.

Keywords: Low-rank minimization, nonconvex Schatten-p norm, active manifold identification, extrapolation, Kurdyka-Łojasiewicz property

1. Introduction

We consider the following regularized nonconvex matrix optimization problem

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

where the loss term $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ is twice continuously differentiable and the regularization term $\|\boldsymbol{X}\|_p = \left(\sum_{i=1}^{\min\{m,n\}} \sigma_i(\boldsymbol{X})^p\right)^{1/p}$ is commonly referred to as the nonconvex Schatten-p norm[1] with $p \in (0, 1)$, and $\sigma_i(\boldsymbol{X})$ is the ith element of the singular value vector of $\boldsymbol{X}$. The parameter $\lambda > 0$ is tunable, providing a proper trade-off between the loss and regularization terms. Throughout our discussion, we assume without loss of generality that $m \le n$.

($\mathcal{P}$) is commonly employed to reduce the relaxation gap between the nuclear norm (or the Schatten-1 norm) and the rank function. Specifically, to achieve an optimal low-rank solution, it is reasonable to consider the following rank-regularized formulation

$$\min_{\boldsymbol{X}\in\mathbb{R}^{m\times n}} f(\boldsymbol{X}) + \lambda \cdot \mathrm{Rank}(\boldsymbol{X}), \tag{1.1}$$

where $\mathrm{Rank}(\boldsymbol{X}) = \|\boldsymbol{\sigma}(\boldsymbol{X})\|_0 := \sum_{i=1}^m \mathbb{I}(\sigma_i(\boldsymbol{X}) \neq 0)$ with $\mathbb{I}(\cdot)$ denoting the indicator function. Problem (1.1) models many important problems that emerged in science and engineering fields, including low-rank matrix recovery (Davenport and Romberg, 2016), recommendation systems (Lee et al., 2016), machine learning (Indyk et al., 2019) and image processing (Huang et al., 2014; Zhao et al., 2020). However, such a matrix rank minimization problem is known to be NP-hard due to the combinatorial nature of the rank function (Hu et al., 2021). To mitigate this computational challenge, many studies have proposed relaxing the rank function to its tractable convex counterpart, the nuclear norm (Fazel et al., 2001), resulting in the following convex optimization problem

$$\min_{\boldsymbol{X}\in\mathbb{R}^{m\times n}} f(\boldsymbol{X}) + \lambda\|\boldsymbol{X}\|_*, \tag{1.2}$$

where $\|\boldsymbol{X}\|_* = \|\sigma(\boldsymbol{X})\|_1 := \sum_{i=1}^m \sigma_i(\boldsymbol{X})$. Over the past decade, the nuclear norm has been crucial in algorithm design, theoretical analysis, and practical applications for achieving desired low-rank solutions (Candès and Recht, 2009; Recht et al., 2010). Recall that $\|\boldsymbol{X}\|_* = \|\sigma(\boldsymbol{X})\|_1$, the nuclear norm regularization equally shrinks all singular values (Negahban and Wainwright, 2011), often over-penalizing large singular values and resulting in a solution from a possibly biased solution space that may exclude ground-truth solutions (Zhang, 2010). Given these considerations, problem ($\mathcal{P}$) serves as an efficient alternative to (1.1) for reducing the relaxation gap. The p-th power of Schatten-p norm of $\boldsymbol{X}$ approximates $\mathrm{Rank}(\boldsymbol{X})$ and recovers the nuclear norm in the sense that $\lim_{p\to 0_+} \|\boldsymbol{X}\|_p^p = \mathrm{Rank}(\boldsymbol{X})$ and $\lim_{p\to 1_-} \|\boldsymbol{X}\|_p^p = \|\boldsymbol{X}\|_*$, respectively. Thus, problem ($\mathcal{P}$) incorporates an approximate low-rank assumption for the desired solution (Hu et al., 2021). Empirical evidence demonstrates that the Schatten-p norm outperforms the nuclear norm in terms of the bias-variance trade-off for many problems (Lu et al., 2014; Nie et al., 2012; Marjanovic and Solo, 2012). Moreover, Schatten-p norm minimization needs fewer observations than traditional nuclear norm minimization (Zhang et al., 2013). Under certain restricted isometry property (RIP)

1. It is a matrix quasi-norm when $0 < p < 1$. We call it a norm for convenience.

conditions, it has been shown that the Schatten-p norm minimization over an affine matrix manifold can uniquely recover a low-rank matrix from compressed affine measurements (Yue and So, 2016; Malek-Mohammadi et al., 2015). In this context, problem $(\mathcal{P})$ arises in an incredibly wide range of settings throughout science and applied mathematics (Lee et al., 2016; Chiang et al., 2018; Jun et al., 2019; Tong et al., 2021; Pal and Jain, 2022). In particular, this optimization model $(\mathcal{P})$ is used in many modern machine learning tasks, including low-rank features learning (Wang et al., 2019), multi-view learning (Liu et al., 2015), and transfer learning (Lin et al., 2019), to name just a few.

A commonly used approach to address $(\mathcal{P})$ is the Iteratively Re-Weighted Nuclear (IRWN) norm-type algorithm, which falls under the majorization-minimization algorithmic framework. The nonsmooth and non-Lipschitz properties of the Schatten-p norm typically prompt researchers to initiate their exploration with a smoothed objective function:

$$F_{\boldsymbol{\epsilon}}(\boldsymbol{X}) := f(\boldsymbol{X}) + \lambda \sum_{i=1}^{m} (\sigma_i(\boldsymbol{X}) + \epsilon_i)^p, \tag{1.3}$$

where $\epsilon_i > 0, \forall i \in [m]$ refers to the perturbation parameters. The modified function $F_{\boldsymbol{\epsilon}}(\boldsymbol{X})$ adjusts $F(\boldsymbol{X})$ by introducing a perturbation parameter to each singular value. During the kth iteration with the iterate $\boldsymbol{X}^k$, IRWN effectively generates the new update $\boldsymbol{X}^{k+1}$ by (approximately) solving

$$\boldsymbol{X}^{k+1} \longleftarrow \underset{\boldsymbol{X} \in \mathbb{R}^{m \times n}}{\arg\min}\, F_{\text{surro}}(\boldsymbol{X}; \boldsymbol{X}^k), \tag{1.4}$$

$$F_{\text{surro}}(\boldsymbol{X}; \boldsymbol{X}^k) := \langle \nabla f(\boldsymbol{X}^k), \boldsymbol{X} - \boldsymbol{X}^k \rangle + \frac{\beta}{2} \|\boldsymbol{X} - \boldsymbol{X}^k\|_F^2 + \lambda \|\boldsymbol{X}\|_{*\boldsymbol{w}^k}, \tag{1.5}$$

where the positive number β is generally required to exceed the Lipschitz constant of the smooth loss term f, and $\|\boldsymbol{X}\|_{*\boldsymbol{w}^k} = \sum_{i=1}^m w_i^k \sigma_i(\boldsymbol{X})$ serves as a locally surrogate for $\|\boldsymbol{X}\|_p^p$ at $\boldsymbol{X}^k$. Here, $w_i^k \geq 0, \forall i \in [m]$ represents the weight assigned to $\sigma_i(\boldsymbol{X})$, given by $w_i^k = p(\sigma_i(\boldsymbol{X}^k) + \epsilon_i^k)^{p-1}$.

The major difference between variants of IRWN may be the updating rule for the perturbation $\boldsymbol{\epsilon}^k$, since the values of $\boldsymbol{\epsilon}^k$ are critically linked to the well-posedness and solvability of the subproblem (1.4). As is proved in (Chen et al., 2013, Theorem 2.2), to guarantee $\|\boldsymbol{X}\|_{*\boldsymbol{w}}$ is indeed a convex matrix-norm, the weights should be in descending order. This requirement proves impractical within the context of IRWN, since sufficiently small $\boldsymbol{\epsilon}^k$ leads to weights in ascending order. On the other hand, the subproblem is nonconvex, but boasts an optimal closed-form solution (Lu et al., 2017) when the weights are arranged in ascending order. One simple approach is to maintain $\boldsymbol{\epsilon}$ as sufficiently small positive constants during the iteration (Sun et al., 2017) to maintain the weights in ascending order. As such, it does solve the relaxed problem (1.3) as its objective. It is generally believed that (1.3) approximates the original problem $(\mathcal{P})$ well only for sufficiently small $\boldsymbol{\epsilon}$. Fixing $\boldsymbol{\epsilon}^k$ as sufficiently small values (especially those associated with the zero singular values of the initial points) may cause the algorithm to easily get stuck in unwanted local minimizers near the initial point. A natural idea to remedy this strategy is to use the same perturbation value for each singular value, i.e., $\boldsymbol{\epsilon}^k = \epsilon^k \boldsymbol{e}$, and then decrease $\boldsymbol{\epsilon}^k$ during the iteration. In this way, the performance of the algorithm critically depends on the speed of driving $\boldsymbol{\epsilon}^k$ to zero.

It is conceivable that reducing ϵ_i^k associated with zero singular values of iterates too fast may lead to undesired local minimizers, and reducing ϵ_i^k associated with positive singular values of the iterates too slow may cause the algorithm to be sluggish. An ideal updating strategy should be able to quickly detect those zero singular values in the found optimal solution, and then automatically terminate the decrease for ϵ_i^k assigned to them and at the same time keep driving other ϵ_i^k rapidly to zero. Another benefit of such a strategy is that the ϵ_i^k associated with the zero singular value does not affect the objective value near the optimal solution, and the algorithm's behavior then only depends on the positive singular value and the decreasing speed of the rest ϵ_i^k. However, such an updating strategy may be sophisticated and challenging to design since the user typically lacks prior knowledge of the rank of the final solution until the entire problem is resolved.

In this paper, we propose an Extrapolated Iteratively Reweighted Nuclear norm with Active Manifold Identification (EIRNAMI) to solve ($\mathcal{P}$). We first add perturbation parameters ϵ_i to each singular value of the matrix to smooth the Schatten-p norm. Then we construct the weighted nuclear "norm"[1] subproblem of the approximated function combined with an extrapolation technique. An adaptive updating strategy for $\boldsymbol{\epsilon}$ is also designed, which automatically terminates the update for ϵ_i associated with zero singular values and rapidly reduces those associated with positive singular values to zero. This update strategy keeps the weights in ascending order so that the subproblem is nonconvex but has a closed-form optimal solution (Lu et al., 2017, Theorem 3.1). Our algorithm is designed to automatically identify zero singular values in the optimal solution after a finite number of iterations. This allows the algorithm to effectively transform the problem into one that operates as a smooth optimization on a fixed-rank manifold embedded in the space $\mathbb{R}^{m\times n}$. Based on this, the local convergence rate can be easily derived. It should be mentioned that our work mainly considers applying the proposed algorithm to solve the representative Schatten-p regularized problem ($\mathcal{P}$), however, it is important to note that the proposed algorithm can be extended to other nonconvex regularization functions of the singular values quite straightforwardly, including the Logarithm (Friedman, 2012), exponential-type penalty (Gao et al., 2011), Geman (Geman and Yang, 1995), Laplace (Trzasko and Manduca, 2008), minimax concave penalty (MCP) (Zhang, 2010) and smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), and more.

1.1 Related Work

Over the last decade, significant attention has been directed toward low-rank optimization, resulting in theoretically and practically efficient algorithms applicable to various problems in signal processing and modern machine learning. Within the extensive body of work, we specifically review the most relevant works.

IRWN-type algorithms. The work of (Sun et al., 2017) proposed a proximal iteratively reweighted nuclear norm (PIRNN) algorithm. Their algorithm adds a prescribed positive perturbation parameter ϵ_i to each singular value $\sigma_i(\boldsymbol{X})$ and *fixed* it during the iteration of the algorithm. Therefore, it indeed solves the relaxed problem (1.3) as its goal. In stark contrast, our method is designed for the original problem ($\mathcal{P}$) in the sense that the perturbation parameter is automatically driven to 0 so that the iterations can successfully

1. It is indeed not a norm since it is not convex.

recover the rank of the stationary first-order solution to $(\mathcal{P})$. It should be stressed that our updating strategy is designed such that $\boldsymbol{\epsilon}$ is decreased to zero at an appropriate speed to maintain the well-posedness of the subproblems and the convergence rate of the overall algorithm.

The immediate predecessor of our work, to our knowledge, is the Iteratively Reweighted Nuclear Norm (IRNN) algorithm proposed in (Lu et al., 2014) and its acceleration (AIRNN) introduced in (Phan and Nguyen, 2021; Ge et al., 2022). IRNN considered a general concave singular value function $g(\sigma_i(\boldsymbol{X}))$ as the regularization term. It first calculates the so-called supergradient of Schatten-p norm $w_i^k \in \partial g(\boldsymbol{\sigma}(\boldsymbol{X}^k))$ and uses it as the weight to form the subproblem. In contrast to our method, this method does not involve the perturbation parameter $\boldsymbol{\epsilon}$; therefore, the weight may tend to extreme values as the associated singular value is close to 0. (As for the zero singular value, this method uses an extremely large constant as the weight). We suspect that this might be the reason for the observation that "IRNN may decrease slowly since the upper bound surrogate may be quite loose" reported in (Lu et al., 2015). Then AIRNN used the extrapolation technique and computed the SVD of a smaller matrix at each iteration to accelerate IRNN. The biggest difference between our algorithm with IRNN, AIRNN, and other contemporary reweighted nuclear norm methods is the active manifold identification property possessed by our algorithm, which means that the algorithm can identify the rank of the converged solution after finite iterations. We elaborate on this in the next subsection.

Active manifold identification. The major novelty of our work is the property of identification of the active manifold of the proposed method, which is an extension of the model identification 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 LASSO, a solution x^* typically has 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 of whether an algorithm can identify the low-complexity active manifold in finite iterations. It has become a useful tool in analyzing the behavior of algorithms and has attracted much attention in the past decades in the research of machine learning algorithms in vector optimization. For example, coordinate descent (Massias et al., 2018) for convex sparse regularization problems are proved to have model identification, and the convergence analysis is easily derived under this property. In the last few years, proximal gradient algorithm (Hare, 2011; Liang et al., 2014, 2017) have been shown the model identification for the ℓ_1 regularized problem. Recently, it has also been shown that the iteratively reweighted ℓ_1 minimization for the ℓ_p regularized problem has the model identification property (Wang et al., 2022, 2021a). This property also belongs to the research line on active-manifold identification in nonsmooth optimization (Lewis, 2002; Hare and Lewis, 2007).

Although IRWN-type algorithms have been extensively studied, their capability for active manifold identification has received limited attention. A recent contribution by (pei Lee et al., 2023) explored the use of the proximal gradient method to identify the correct rank for a nuclear norm-regularized problem. In particular, its algorithm can serve as a suitable subproblem solver for our approach. In a related vein, the work of (Zeng, 2023) extended the lower bound theory of nonconvex ℓ_p minimization to Schatten-p norm minimization and

incorporates it as a priori in algorithm design. However, the active manifold identification property of their algorithm remains unverified.

In this paper, we formalize the active manifold identification property as follows.

Definition 1 (Active manifold identification property) *An algorithm is said to possess the active manifold identification property if and only if for a sequence (or at least a subsequence)* $\{\boldsymbol{X}^k\}_{k\in\mathbb{N}_+}$ *generated by the algorithm converges to a solution* $\boldsymbol{X}^*$*, then there exists a finite* $K \in \mathbb{N}_+$ *such that for each* $k \geq K$*,* $\boldsymbol{X}^k \in \mathcal{M}(\boldsymbol{X}^*) := \{\boldsymbol{X} \in \mathbb{R}^{m\times n} \mid \mathrm{Rank}(\boldsymbol{X}) = \mathrm{Rank}(\boldsymbol{X}^*)\}$.

Our algorithm is designed to possess this property, which means that the singular values of the generated iterates satisfy $\sigma_i(\boldsymbol{X}^k) = 0, i \in \mathcal{Z}(\boldsymbol{X}^*)$ and $\sigma_i(\boldsymbol{X}^k) > 0, i \in \mathcal{I}(\boldsymbol{X}^*)$ for all sufficiently large k, where $\mathcal{Z}(\boldsymbol{X}^*)$ is the set of indices that correspond to the zero singular values in the optimal solution and $\mathcal{I}(\boldsymbol{X}^*)$ corresponds to the nonzero singular values. Based on this, an adaptive updating strategy of ϵ can be straightforwardly designed to drive $\epsilon_i, i \in \mathcal{I}(\boldsymbol{X}^*)$ quickly to zero and automatically cease the updating for $\epsilon_i, i \in \mathcal{Z}(\boldsymbol{X}^*)$. In essence, this implies that the algorithm behaves like solving a smooth problem in a low-complexity manifold, facilitating a straightforward derivation of global convergence analysis and application of acceleration techniques. To our knowledge, this idea of designing an algorithm with model/active manifold identification property for the Schatten-p norm is novel in the context of matrix optimization problems.

1.2 Contribution

We summarize our main contributions in the following.

(i) We propose an iteratively reweighted nuclear norm minimization method for the nonconvex regularized problem, and extrapolation techniques are also incorporated into the algorithm to further enhance its performance.

(ii) The key novelty of the proposed method is the adaptively updating strategy for updating the perturbation parameters, bringing two benefits: (i) automatic identification of parameters associated with zero and nonzero singular values, enabling the use of tailored update strategies for each. (ii) consistent maintenance of weights in ascending order, ensuring the explicit computation of a global minimizer for the nonconvex subproblem.

(iii) We show that the algorithm possesses an active manifold identification property, which can successfully identify the rank of the optimal solutions found by the algorithm within finite number of iterations. This property, which is barely studied by the existing related work, signifies a distinct contribution. It implies a transition of the optimization problem to a smoother form in the vicinity of the optimal solution.

(iv) Global convergence and local convergence rate under the Kurdyka-Łojasiewicz (KL) property are derived for the proposed algorithm.

1.3 Notation and Preliminaries

Throughout the paper, we restrict our discussion to the Euclidean space of n-dimensional real vectors, denoted $\mathbb{R}^n$, and the Euclidean space of $m \times n$ real matrices denoted $\mathbb{R}^{m\times n}$, where $m, n \in \mathbb{N}$. $\mathbb{R}^n_+$ represents the nonnegative orthant in $\mathbb{R}^n$ and $\mathbb{R}^n_{++}$ denotes the interior of $\mathbb{R}^n_+$. $\mathbb{R}^n_\uparrow$ and $\mathbb{R}^n_\downarrow$ are used to indicate the set of nondecreasingly ordered vectors and nonincreasingly ordered vectors, respectively. Furthermore, we use the notation $[n] = \{1, 2, \cdots, n\}$ to denote the integer set from 1 to n, for any $n \in \mathbb{N}$. For any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, the element-wise (Hadamard) product between $\boldsymbol{x}$ and $\boldsymbol{y}$ is given by $(\boldsymbol{x} \circ \boldsymbol{y})_i = x_i y_i$ for $i \in [n]$. By abuse of notation, $\circ$ is also used to denote function composition. Define the ℓ_p-(quasi)-norm of $\boldsymbol{x} \in \mathbb{R}^n$ as $\|\boldsymbol{x}\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$.

For any $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{m\times n}$ (assuming $m \le n$ for convenience), the Frobenius norm of $\boldsymbol{X}$ is denoted as $\|\boldsymbol{X}\|_F$, namely $\|\boldsymbol{X}\|_F = \left(\sum_{i=1}^m \sum_{j=1}^n |X_{ij}|^2\right)^{1/2} = \operatorname{tr}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{1/2}$. The Frobenius inner product is $\langle \boldsymbol{X}, \boldsymbol{Y}\rangle = \operatorname{tr}\left(\boldsymbol{X}^\top \boldsymbol{Y}\right)$. Let $\operatorname{diag}(\boldsymbol{x})$ denote the diagonal matrix with vector $\boldsymbol{x}$ on its main diagonal and zeros elsewhere. The full singular value decomposition (SVD) (Van Loan, 1976) of $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ is

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

where $(\boldsymbol{U}, \boldsymbol{V}) \in \overline{\mathcal{M}}(\boldsymbol{X})$ with $\overline{\mathcal{M}}(\boldsymbol{X}) := \{(\boldsymbol{U}, \boldsymbol{V}) \in \mathbb{R}^{m\times m} \times \mathbb{R}^{n\times n} \mid \boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_m, \boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}_n, \boldsymbol{X} = \boldsymbol{U}\operatorname{diag}(\sigma(\boldsymbol{X}))\boldsymbol{V}^\top\}$ and $\boldsymbol{\sigma}(\boldsymbol{X}) \in \mathbb{R}^m_\downarrow \cap \mathbb{R}^m_+$ denotes the singular value vector of $\boldsymbol{X}$. Suppose $\operatorname{Rank}(\boldsymbol{X}) = r \le m$. The associated thin SVD of $\boldsymbol{X}$ is $\boldsymbol{X} = \boldsymbol{U}_r\operatorname{diag}(\boldsymbol{\sigma}_r(\boldsymbol{X}))\boldsymbol{V}_r^\top$, where $\boldsymbol{U}_r$ and $\boldsymbol{V}_r$ are the first r columns of $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively, and $\boldsymbol{\sigma}_r(\boldsymbol{X}) \in \mathbb{R}^r_\downarrow \cap \mathbb{R}^r_+$.

For analysis, we summarize the simultaneous ordered SVD of two matrices introduced in (Lewis and Sendov, 2005).

Definition 2 (Simultaneous ordered SVD) *We say that two real matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ of size $m \times n$ have a simultaneous ordered singular value decomposition if there exist orthogonal matrices $\boldsymbol{U} \in \mathbb{R}^{m\times m}$ and $\boldsymbol{V} \in \mathbb{R}^{n\times n}$ such that*

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

In addition, we define two index sets as follows to track the singular values of the iterates conveniently, which reads

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

For a lower semicontinuous function $J : \mathbb{R}^N \to (-\infty, +\infty]$, its domain is denoted by $\operatorname{dom}(J) := \{x \in \mathbb{R}^N : J(x) < +\infty\}$. We first recall some concepts of subdifferentials that are commonly used in variational analysis and subdifferential calculus, which is a useful tool in developing optimality conditions of the concerned optimization problem in nonsmooth analysis.

Definition 3 (Subdifferentials) *Consider a proper lower semi-continuous function. $\varphi : \mathbb{R}^N \to (-\infty, +\infty]$.*

1. *The Fréchet subdifferential* $\widehat{\partial}\varphi$ *of* φ *at an* $\boldsymbol{x} \in \operatorname{dom} \varphi$ *is analytically defined as*

$$\widehat{\partial}\varphi(\boldsymbol{x}) := \left\{ \boldsymbol{v} \in \mathbb{R}^N \mid \lim_{\boldsymbol{u}\to\boldsymbol{x}} \inf_{\boldsymbol{u}\neq\boldsymbol{x}} \frac{\varphi(\boldsymbol{u}) - \varphi(\boldsymbol{x}) - \langle \boldsymbol{v}, \boldsymbol{u}-\boldsymbol{x}\rangle}{\|\boldsymbol{u}-\boldsymbol{x}\|_2} \geq \boldsymbol{0} \right\}.$$

2. *The (limiting) subdifferential of* $\partial\varphi$ *of* φ *at an* $\boldsymbol{x} \in$ *dom* φ *is defined through the following closure process, which reads,*

$$\partial\varphi(\boldsymbol{x}) := \left\{ \boldsymbol{v} \in \mathbb{R}^N \mid \exists \boldsymbol{v}^k \to \boldsymbol{v}, \boldsymbol{x}^k \overset{\varphi}{\to} \boldsymbol{x} \text{ with } \boldsymbol{v}^k \in \widehat{\partial}\varphi(\boldsymbol{x}^k) \text{ for all } k \right\}.$$

where $\boldsymbol{x}^k \overset{\varphi}{\to} \boldsymbol{x}$ *refers to* φ*-attentive convergence in analysis, meaning* $\boldsymbol{x}^k \to \boldsymbol{x}$ *with* $\varphi(\boldsymbol{x}^k) \to \varphi(\boldsymbol{x})$.

We mention $\widehat{\partial}\varphi(\boldsymbol{x}) = \partial\varphi = \emptyset$ *for* $\boldsymbol{x} \notin \operatorname{dom} \varphi$.

We next collect the result on the limiting subdifferential of the singular value function established in (Lewis and Sendov, 2005). Upon that, we present the limiting subdifferential associated with the nonconvex Schatten-p norm.

Lemma 1 (Limiting subdifferential of singular value function) *Let* $\varphi : \mathbb{R}^m \to \mathbb{R}$ *be an absolutely symmetric function, meaning* $\varphi(x_1, \cdots, x_m) = \varphi(|x_{\pi(1)}|, \cdots, |x_{\pi(m)}|)$ *holds for any permutation* π *of* $[m]$*, and let* $\boldsymbol{\sigma}(\boldsymbol{X})$ *be the singular values of a matrix* $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ *(*$m \leq n$ *is assumed for convenience). Then the limiting subdifferential of singular value function* $\varphi\circ\boldsymbol{\sigma}$ *at a matrix* $\boldsymbol{X}$ *is given by*

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

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

A direct consequence of Lemma 1 is the result of limiting subdifferential of $\|\boldsymbol{X}\|_p^p$.

Proposition 1 *Let* $\operatorname{Rank}(\boldsymbol{X}) = r \leq m \in \mathbb{N}$. *The limiting subdifferential of* $\|\cdot\|_p^p : \mathbb{R}^{m\times n} \to \mathbb{R}$ *at a matrix* $\boldsymbol{X}$ *is given by*

$$\begin{aligned} \partial\|\boldsymbol{X}\|_p^p = \partial\left(\sum_{i=1}^r \sigma_i(\boldsymbol{X})^p\right) &= \partial\left([\|\cdot\|_p^p \circ \boldsymbol{\sigma}](\boldsymbol{X})\right) \\ &= \left\{ \boldsymbol{U}\,diag(\boldsymbol{\Sigma})\boldsymbol{V}^\top \mid \boldsymbol{\Sigma} \in (\partial\|\boldsymbol{\sigma}(\boldsymbol{X})\|_p^p \circ \partial|\sigma_i(\boldsymbol{X})|) \right\}, \end{aligned} \tag{1.6}$$

where $\partial\|\boldsymbol{\sigma}(\boldsymbol{X})\|_p^p = \{\vartheta \in \mathbb{R}^m \mid \vartheta_j = p\sigma_j(\boldsymbol{X})^{p-1}, j \in [r]\}$ *and* $(\boldsymbol{U}, \boldsymbol{V}) \in \overline{\mathcal{M}}(\boldsymbol{X})$.

We use the following stationary principle based on subdifferentials to establish the first-order necessary optimality conditions of $(\mathcal{P})$.

Theorem 2 (Nonsmooth versions of Fermat's rule) *Consider* $(\mathcal{P})$. *If* F *has a local minimum at* $\bar{\boldsymbol{X}}$*, then* $\boldsymbol{0} \in \widehat{\partial}F(\bar{\boldsymbol{X}}) = \partial F(\bar{\boldsymbol{X}})$.

Using Proposition 1 and by Theorem 2, we define the critical point of $(\mathcal{P})$ as follows.

Definition 4 (Critical point) *We say that an $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ is a critical point of F in $(\mathcal{P})$ if it satisfies $\mathbf{0} \in \partial F(\boldsymbol{X})$. Moreover, the set of all critical points is denoted by*

$$\operatorname{crit}(F) := \left\{ \boldsymbol{X} \in \mathbb{R}^{m\times n} \mid \mathbf{0} \in \nabla f(\boldsymbol{X}) + \lambda \boldsymbol{U}\operatorname{diag}(\partial \|\boldsymbol{\sigma}(\boldsymbol{X})\|_p^p)\boldsymbol{V}^\top, (\boldsymbol{U},\boldsymbol{V}) \in \overline{\mathcal{M}}(\boldsymbol{X}) \right\}. \quad (1.7)$$

The Kurdyka-Łojasiewicz (KL) property plays an important role in our convergence analysis. We next recall the essential components as follows. First, let $\Omega \subset \mathbb{R}^{m\times n}$ and $\boldsymbol{X} \in \mathbb{R}^{m\times n}$, the distance from $\boldsymbol{X}$ to Ω is defined by

$$\operatorname{dist}(\boldsymbol{X}, \Omega) := \inf\{\|\boldsymbol{X} - \boldsymbol{Y}\|_F \mid \boldsymbol{Y} \in \Omega\}.$$

In particular, we have $\operatorname{dist}(\boldsymbol{X}, \Omega) = +\infty$ for any $\boldsymbol{X}$ when $\Omega = \emptyset$. Next, we define the desingularizing function.

Definition 5 (Desingularizing function) *(Garrigos, 2015, Section 3.1.2) Let $\eta > 0$. We say that $\Phi : [0, \eta] \to \mathbb{R}_+$ is a desingularizing function if*

(i) $\Phi(0) = 0$;

(ii) Φ *is continuous on* $[0, \eta]$ *and of class* C^1 *on* $(0, \eta)$;

(iii) $\Phi'(s) > 0$ *for all* $s \in (0, \eta)$.

Typical examples of desingularizing functions are the functions of the form $\Phi(t) = cs^{1-\theta}$, for $c > 0$ and KL exponent $\theta \in [0, 1)$.

Now we define the Kurdyka-Łojasiewicz property.

Definition 6 (KL property) *(Bolte et al., 2014, Definition 3) Let $F : \mathbb{R}^{m\times n} \to \mathbb{R}\cup\{+\infty\}$ be proper lower semicontinuous. We say that F satisfies the Kurdyka-Łojasiewicz property at $\bar{\boldsymbol{X}} \in \operatorname{dom}(\partial F) := \{\boldsymbol{X} \in \mathbb{R}^{m\times n} \mid \partial F(\boldsymbol{X}) \neq \emptyset\}$ if there exist $\eta > 0$, a neighborhood $\mathbb{U}(\bar{\boldsymbol{X}}, \rho)$ of $\bar{\boldsymbol{X}}$, and a concave desingularizing function $\Phi : [0, \eta) \to \mathbb{R}_+$, such that the Kurdyka-Łojasiewicz inequality*

$$\Phi'\left(F(\boldsymbol{X}) - F(\bar{\boldsymbol{X}})\right) \operatorname{dist}\left(0, \partial F(\bar{\boldsymbol{X}})\right) \geq 1 \quad (1.8)$$

holds, for all $\boldsymbol{X}$ in the strict local upper level set

$$\operatorname{Lev}_\eta(\bar{\boldsymbol{X}}, \rho) := \{\boldsymbol{X} \in \mathbb{U}(\bar{\boldsymbol{X}}, \rho) \mid F(\bar{\boldsymbol{X}}) < F(\boldsymbol{X}) < F(\bar{\boldsymbol{X}}) + \eta\}.$$

If F satisfies the KL property at any $\boldsymbol{X} \in \operatorname{dom}(\partial F)$, we then call F a KL function.

Moreover, we introduce the more general KL property as follows.

Lemma 3 (Uniform KL property) *(Bolte et al., 2014, Lemma 6) Let Ω be a compact set and let $F : \mathbb{R}^{m\times n} \to \mathbb{R}\cup\{+\infty\}$ be a proper lower semicontinuous function. Assume that F is constant in Ω and satisfies the KL property at each point in Ω. We say that F has uniform KL property on Ω if there exist $\epsilon > 0$, $\eta > 0$ and Φ defined in Definition 6 such that the KL inequality* (1.8) *holds for any $\bar{\boldsymbol{X}} \in \Omega$ and any $\boldsymbol{X} \in \{\boldsymbol{X} \in \mathbb{R}^{m\times n} \mid \operatorname{dist}(\boldsymbol{X}, \Omega) < \epsilon\} \cap \{\boldsymbol{X} \in \mathbb{R}^{m\times n} \mid F(\bar{\boldsymbol{X}}) < F(\boldsymbol{X}) < F(\bar{\boldsymbol{X}}) + \eta\}$.*

The KL property of the objective function is crucial in the convergence analysis of the first-order algorithms for nonconvex and nonsmooth optimization problems, as discussed in (Attouch et al., 2010, 2013). Many functions are known to satisfy the KL property, as introduced in (Bolte et al., 2014, Section 5) and (Garrigos, 2015, Section 3.1). Typical examples include the ℓ_0-(quasi)norm $\|\boldsymbol{x}\|_0$, ℓ_p-(quasi)norm $\|\boldsymbol{x}\|_p$ with $p > 0$ (with some technical conditions required when p is irrational) (Attouch et al., 2010; Bolte et al., 2014), real polynomial functions, indicator functions of polyhedral sets and matrix rank function. For the calculation of the KL exponent, we refer to the recent works by (Li and Pong, 2018; Yu et al., 2022; Ouyang et al., 2024).

2. Proposed Extrapolated Iteratively Reweighed Nuclear Norm Algorithm with Active Manifold Identification

In this section, we provide the details of the proposed EIRNN framework to solve $(\mathcal{P})$ by discussing the solution of the subproblem and a novel update strategy of the perturbation parameter to enable the active manifold identification property.

Before presenting the proposed algorithm, we make the following assumptions on $(\mathcal{P})$ as follows throughout.

Assumption 1 *The function $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ is L_f-smooth, i.e., $\|\nabla f(\boldsymbol{X}) - \nabla f(\boldsymbol{Y})\|_F \leq L_f\|\boldsymbol{X} - \boldsymbol{Y}\|_F$, $\forall \boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{m\times n}$, where the modulus $L_f \geq 0$ refers to the smoothness parameter.*

Assumption 2 *The function F is level-bounded (Rockafellar and Wets, 2009, Definition 1.8) and proper. This assumption about F corresponds to $\lim_{\boldsymbol{X}\in\mathbb{R}^{m\times n}:\|\boldsymbol{X}\|\to\infty} F(\boldsymbol{X}) = +\infty$ and further implies $\min_{\boldsymbol{X}\in\mathbb{R}^{m\times n}} F(\boldsymbol{X}) = \underline{F} > -\infty$ and $\{\boldsymbol{X} \mid \arg\min_{\boldsymbol{X}\in\mathbb{R}^{m\times n}} F(\boldsymbol{X})\} \neq \emptyset$ regarding $(\mathcal{P})$.*

We state the proposed algorithm in Algorithm 1, which consists of solving a sequence of weighted nuclear norm regularized subproblems, an extrapolation technique, and an adaptive perturbation parameter updating strategy.

Algorithm 1 Extrapolated Iteratively Reweighted Nuclear Norm with Active Manifold Identification (EIRNAMI)

Input: $\boldsymbol{X}^0 \in \mathbb{R}^{m\times n}, \boldsymbol{\epsilon}^0 \in \mathbb{R}^m_{++} \cap \mathbb{R}^m_{\downarrow}$, $\mu \in (0, 1)$, and $\alpha_0 \in [0, \bar{\alpha}]$ by (2.3).
Initialize: $k = 0$ and $\boldsymbol{X}^{-1} = \boldsymbol{X}^0$.

1: **repeat**
2: Compute weights $w_i^k = p\left(\sigma_i(\boldsymbol{X}^k) + \boldsymbol{\epsilon}_i^k\right)^{p-1}$, $\forall i \in [m]$.
3: Compute $\boldsymbol{Y}^{k+1}$ according to the extrapolation (2.2).
4: Compute the new iterate as the solution of (2.5).
5: Update $\boldsymbol{\epsilon}^k$ by calling Algorithm 2.
6: Choose $\alpha_k \in [0, \bar{\alpha}]$.
7: Set $k \leftarrow k + 1$.
8: **until** convergence

2.1 A Weighted Nuclear Norm Surrogate with Extrapolation

Our presented approach to solving $(\mathcal{P})$ is primarily motivated by the substantial literature on proximal gradient-type methods employing acceleration techniques (Yu and Pong, 2019) and iteratively reweighted techniques (Wang et al., 2021b). Specifically, we first add perturbation parameters $\boldsymbol{\epsilon} \in \mathbb{R}^n_{++}$ to each singular value of the matrix to smooth the p-th power of the Schatten-p norm,

$$F(\boldsymbol{X};\boldsymbol{\epsilon}) := f(\boldsymbol{X}) + \lambda \sum_{i=1}^{m} (\sigma_i(\boldsymbol{X}) + \epsilon_i)^p. \tag{2.1}$$

Obviously, $F(\boldsymbol{X};\boldsymbol{0}) = F(\boldsymbol{X})$. Drawing upon the Nesterov's acceleration technique (Nesterov, 1983; Phan and Nguyen, 2021), our approach begins by computing an extrapolated $\boldsymbol{Y}^k$, using the current iterate $\boldsymbol{X}^k$ and the previous one $\boldsymbol{X}^{k-1}$, i.e.,

$$\boldsymbol{Y}^k = \boldsymbol{X}^k + \alpha_k(\boldsymbol{X}^k - \boldsymbol{X}^{k-1}), \tag{2.2}$$

where $\alpha_k \in [0, \bar{\alpha})$ refers to the extrapolation parameter and is selected according to the following rule (Wang et al., 2022)

$$\begin{cases} \bar{\alpha} \in (0,1), & \text{if } f(x) \text{ is convex and } L_f\text{-smooth,} \\ \bar{\alpha} \in (0, \sqrt{\frac{\beta}{\beta+3L_f}}), & \text{if } f(x) \text{ is } L_f\text{-smooth,} \end{cases} \tag{2.3}$$

with $\beta \geq L_f$. At $\boldsymbol{Y}^k$, it follows for any feasible $\boldsymbol{X}$ that the perturbed objective $F(\boldsymbol{X};\boldsymbol{\epsilon})$ admits a useful upper bound presented below, i.e.,

$$\begin{aligned} F(\boldsymbol{X},\boldsymbol{\epsilon}) &:= f(\boldsymbol{X}) + \lambda \sum_{i=1}^{m} (\sigma_i(\boldsymbol{X}) + \epsilon_i)^p \\ &\overset{(a)}{\leq} f(\boldsymbol{Y}^k) + \langle \nabla f(\boldsymbol{Y}^k), \boldsymbol{X} - \boldsymbol{Y}^k \rangle + \frac{L_f}{2}\|\boldsymbol{X} - \boldsymbol{Y}^k\|_F^2 \\ &\quad + \lambda \sum_{i=1}^{m} p(\sigma_i(\boldsymbol{X}^k) + \epsilon_i^k)^{p-1}(\sigma_i(\boldsymbol{X}) - \sigma_i(\boldsymbol{X}^k)) \\ &\overset{(b)}{\leq} f(\boldsymbol{Y}^k) + \langle \nabla f(\boldsymbol{Y}^k), \boldsymbol{X} - \boldsymbol{Y}^k \rangle + \frac{L_f}{2}\|\boldsymbol{X} - \boldsymbol{Y}^k\|_F^2 + \frac{L_f}{2}\|\boldsymbol{X} - \boldsymbol{X}^k\|_F^2 \\ &\quad + \lambda \sum_{i=1}^{m} p(\sigma_i(\boldsymbol{X}^k) + \epsilon_i^k)^{p-1}(\sigma_i(\boldsymbol{X}) - \sigma_i(\boldsymbol{X}^k)), \end{aligned} \tag{2.4}$$

where (a) is a direct consequence of the L_f-smoothness of f under Assumption 1 and the concavity of $(\cdot)^p$, and (b) naturally holds due to the nonnegativity of the proximal term $\frac{L_f}{2}\|\boldsymbol{X} - \boldsymbol{X}^k\|_F^2$. Omitting the constants on the right-hand side of (2.4), we obtain the following surrogate function $L(\boldsymbol{X};\boldsymbol{X}^k,\boldsymbol{Y}^k,\boldsymbol{\epsilon}^k)$ to approximate $F(\boldsymbol{X};\boldsymbol{\epsilon})$ at $(\boldsymbol{X}^k,\boldsymbol{\epsilon}^k)$, i.e.,

$$L(\boldsymbol{X};\boldsymbol{X}^k,\boldsymbol{Y}^k,\boldsymbol{\epsilon}^k) := f(\boldsymbol{Y}^k) + \langle \boldsymbol{X}, \nabla f(\boldsymbol{Y}^k)\rangle + \frac{\beta}{2}\|\boldsymbol{X} - \boldsymbol{Y}^k\|_F^2 + \frac{\beta}{2}\|\boldsymbol{X} - \boldsymbol{X}^k\|_F^2 + \lambda \sum_{i=1}^{m} w_i^k \sigma_i(\boldsymbol{X}),$$

where $w_i^k = w(\sigma_i(\boldsymbol{X}^k), \epsilon_i^k) = p\left(\sigma_i(\boldsymbol{X}^k) + \epsilon_i^k\right)^{p-1}, \forall i \in [m]$ and $\beta \geq L_f > 0$. The next iterate $\boldsymbol{X}^{k+1}$ is computed by minimizing $L(\boldsymbol{X}^k; \boldsymbol{X}^k, \boldsymbol{Y}^k, \boldsymbol{\epsilon}^k)$, i.e.,

$$\begin{aligned} \boldsymbol{X}^{k+1} &\in \underset{\boldsymbol{X}\in\mathbb{R}^{m\times n}}{\arg\min}\, L(\boldsymbol{X}; \boldsymbol{X}^k, \boldsymbol{Y}^k, \boldsymbol{\epsilon}^k) \\ &= \underset{\boldsymbol{X}\in\mathbb{R}^{m\times n}}{\arg\min} \left\{ \frac{\beta}{2} \left\| \boldsymbol{X} - \left(\frac{\boldsymbol{X}^k + \boldsymbol{Y}^k}{2} - \frac{\nabla f(\boldsymbol{Y}^k)}{2\beta} \right) \right\|_F^2 + \frac{\lambda}{2} \sum_{i=1}^m w_i^k \sigma_i(\boldsymbol{X}) \right\}. \end{aligned} \tag{2.5}$$

Remark 4 *As demonstrated in (Chen et al., 2013), the subproblem* (2.5) *involving the adaptive nuclear norm is typically nonconvex, given the natural restriction that the weights decrease with the singular values. Despite this nonconvexity, the global optimal solution can be achieved, as shown in our adapted Proposition 2. The authors have further demonstrated the superior statistical properties of the adaptive nuclear norm, which include a continuous solution path, better bias-variance trade-off compared to the nuclear norm, and rank consistency. In addition, they have established prediction and estimation performance bounds for the proposed estimator in the high-dimensional asymptotic regime. In this regard, the adaptive nuclear norm generally enhances the performance of the proposed algorithm. Moreover, for $p \in (0, 1)$, ℓ_p regularization is locally equivalent to a weighted ℓ_1 regularization for vector variables (Wang et al., 2021a, Theorem 9). From this perspective, applying the Schatten-p regularization seeks to determine relatively "optimal" weighting coefficients in terms of the adaptive nuclear norm regularization.*

2.2 Subproblem Solution

It should be mentioned that solving such a weighted nuclear norm minimization problem (2.5) is not a direct extension of solving a weighted ℓ_1 norm minimization counterpart in the vector case. In fact, problem (2.5) is nonconvex and hence generally poses challenges to find the global minimizer. As shown in (Chen et al., 2013, Theorem 2.2), the weighted nuclear norm of $\boldsymbol{X} \in \mathbb{R}^{m\times n}$, $\|\boldsymbol{X}\|_{*\boldsymbol{w}}$, is convex with respect to $\boldsymbol{X}$ if and only if both the singular value $\sigma_i(\boldsymbol{X})$ and its corresponding weights w_i, $\forall i \in [m]$ are in descending order. However, it is unrealistic to design such a strategy, since $w(\sigma_i(\boldsymbol{X}), \epsilon_i) = p(\sigma_i(\boldsymbol{X}) + \epsilon_i)^{p-1} < p(\sigma_j(\boldsymbol{X}) + \epsilon_j)^{p-1} = w(\sigma_j(\boldsymbol{X}), \epsilon_j)$ for $\sigma_i(\boldsymbol{X}) > \sigma_j(\boldsymbol{X})$ as $\epsilon_i \to 0$ and $\epsilon_j \to 0$. However, a closed-form global optimal solution to (2.5) is available by imposing the ascending order on all weights $w_i, \forall i \in [m]$ (Lu et al., 2017; Sun et al., 2017). We restate such a result for (2.5) in the following proposition.

Proposition 2 *Consider* (2.5). *Let* $\boldsymbol{w}^k \in \mathbb{R}^m_{\uparrow} \cap \mathbb{R}^m_{++}$, *that is,*

$$0 < w_1^k \leq w_2^k \leq \ldots \leq w_m^k. \tag{2.6}$$

Then a global optimal solution to (2.5) *reads*

$$\boldsymbol{X}^{k+1} = \boldsymbol{U}^{k+1} diag\left(\left[\boldsymbol{\Sigma}_i^{k+1} - \frac{\lambda w_i^k}{2\beta} \right]_+ \right) \boldsymbol{V}^{k+1^\top} \tag{2.7}$$

with $\boldsymbol{U}^{k+1} diag(\boldsymbol{\Sigma}^{k+1}) \boldsymbol{V}^{k+1^\top}$ *being the SVD of the matrix* $\frac{\boldsymbol{X}^k + \boldsymbol{Y}^k}{2} - \frac{\nabla f(\boldsymbol{Y}^k)}{2\beta}$.

Proof The proof follows the same argument of (Lu et al., 2017, Theorem 3.1). ■

With the help of Proposition 2 and Definition 2 of simultaneous ordered SVD, we establish the following result whose proof follows a similar argument of (Ge et al., 2022, Proposition 2).

Proposition 3 *Consider* (2.5). *Suppose* $\boldsymbol{w}^k \in \mathbb{R}^m_{\uparrow} \cap \mathbb{R}^m_{++}$. *Then there exist* $\boldsymbol{\xi}^{k+1} \in \partial|\boldsymbol{\sigma}(\boldsymbol{X}^{k+1})| \subset \mathbb{R}^m$ *such that*

$$\boldsymbol{U}^{k+1} diag\left(\frac{\lambda}{2\beta}\boldsymbol{w}^k \circ \boldsymbol{\xi}^{k+1}\right) \boldsymbol{V}^{k+1^\top} \in \partial\left\{\sum_{i=1}^m w_i^k \sigma_i(\boldsymbol{X}^{k+1})\right\}, \tag{2.8}$$

where $\boldsymbol{X}^{k+1}$ *and* $\frac{\boldsymbol{X}^k+\boldsymbol{Y}^k}{2} - \frac{\nabla f(\boldsymbol{Y}^k)}{2\beta}$ *have a simultaneous ordered SVD.*

Proof By Theorem 2 and Proposition 2, we have from (2.5) that

$$\boldsymbol{0} \in \beta\left(\boldsymbol{X}^{k+1} - \left(\frac{\boldsymbol{X}^k+\boldsymbol{Y}^k}{2} - \frac{\nabla f(\boldsymbol{Y}^k)}{2\beta}\right)\right) + \frac{\lambda}{2}\partial\left(\sum_{i=1}^m w_i^k \sigma_i(\boldsymbol{X}^{k+1})\right). \tag{2.9}$$

Note also that matrices $\boldsymbol{X}^{k+1}$ and $\frac{\boldsymbol{X}^k+\boldsymbol{Y}^k}{2} - \frac{1}{2\beta}\nabla f(\boldsymbol{Y}^k)$ have the simultaneous ordered SVD. From (2.9), we know there exists $\hat{\boldsymbol{\xi}}^{k+1} \in \partial\left(\sum_{i=1}^m w_i^k \sigma_i(\boldsymbol{X}^{k+1})\right)$ such that

$$\begin{aligned}
\hat{\boldsymbol{\xi}}^{k+1} &= \frac{2\beta}{\lambda}\left(\left(\frac{\boldsymbol{X}^k+\boldsymbol{Y}^k}{2} - \frac{\nabla f(\boldsymbol{Y}^k)}{2\beta}\right) - \boldsymbol{X}^{k+1}\right) \\
&= \frac{2\beta}{\lambda}\left(\boldsymbol{U}^{k+1}\mathrm{diag}(\boldsymbol{\Sigma}^{k+1})\boldsymbol{V}^{k+1^\top} - \boldsymbol{U}^{k+1}\mathrm{diag}\left(\left(\boldsymbol{\Sigma}^{k+1} - \frac{\lambda\boldsymbol{w}^k}{2\beta}\right)_+\right)\boldsymbol{V}^{k+1^\top}\right) \\
&= \frac{2\beta}{\lambda}\left(\boldsymbol{U}^{k+1}\mathrm{diag}\left(\boldsymbol{\Sigma}^{k+1} - \left(\boldsymbol{\Sigma}^{k+1} - \frac{\lambda\boldsymbol{w}^k}{2\beta}\right)_+\right)\boldsymbol{V}^{k+1^\top}\right).
\end{aligned} \tag{2.10}$$

If $\sigma_i(\boldsymbol{X}^{k+1}) = \left(\boldsymbol{\Sigma}_i^{k+1} - \frac{\lambda w_i^k}{2\beta}\right)_+ > 0$ for $i \in [m]$, we know $\partial(\sigma_i(\boldsymbol{X}^{k+1})) = 1$ and $\boldsymbol{\Sigma}^{k+1} - \left(\boldsymbol{\Sigma}^{k+1} - \frac{\lambda\boldsymbol{w}^k}{2\beta}\right)_+ = \frac{\lambda\boldsymbol{w}^k}{2\beta}$. Then, it follows for any $i \in [m]$ such that $\xi_i^{k+1} = 1$ that $\frac{\lambda\boldsymbol{w}^k}{2\beta} \circ \boldsymbol{\xi}^{k+1} = \frac{\lambda\boldsymbol{w}^k}{2\beta} \in \frac{\lambda\boldsymbol{w}^k}{2\beta} \circ \partial(\sigma_i(\boldsymbol{X}^{k+1}))$. If $\sigma_i(\boldsymbol{X}^{k+1}) = \left(\boldsymbol{\Sigma}_i^{k+1} - \frac{\lambda w_i^k}{2\beta}\right)_+ = 0$ for $i \in [m]$, we know $\partial(\sigma_i(\boldsymbol{X}^{k+1})) = [-1, 1]$ and $\boldsymbol{\Sigma}_i^{k+1} - \left(\boldsymbol{\Sigma}_i^{k+1} - \frac{\lambda\boldsymbol{w}_i^k}{2\beta}\right)_+ = \boldsymbol{\Sigma}_i^{k+1} \in [0, \frac{\lambda\boldsymbol{w}_i^k}{2\beta}]$. Then, it follows for any $i \in [m]$ such that $\xi_i^{k+1} = \frac{2\beta\boldsymbol{\Sigma}_i^{k+1}}{\lambda w_i^k} \in [0, 1] \subset [-1, 1]$ that $\frac{\lambda\boldsymbol{w}^k}{2\beta} \circ \boldsymbol{\xi}^{k+1} = \Sigma_i^{k+1} \in \frac{\lambda\boldsymbol{w}^k}{2\beta}\partial(\sigma_i(\boldsymbol{X}^{k+1}))$. Therefore, the proof is completed. ■

Since $\boldsymbol{X}^{k+1}$ is a global minimum for (2.5) by Proposition 2, it follows from Theorem 2 and Proposition 3 that there exists $\boldsymbol{\xi}^{k+1} \in \partial|\boldsymbol{\sigma}(\boldsymbol{X}^{k+1})|$ such that

$$\boldsymbol{0} = \nabla f(\boldsymbol{Y}^k) + \beta(\boldsymbol{X}^{k+1} - \boldsymbol{Y}^k) + \beta(\boldsymbol{X}^{k+1} - \boldsymbol{X}^k) + \lambda\boldsymbol{U}^{k+1}\mathrm{diag}\left(\boldsymbol{w}^k \circ \boldsymbol{\xi}^{k+1}\right)\boldsymbol{V}^{k+1^\top}. \tag{2.11}$$

Algorithm 2 Update perturbation $\boldsymbol{\epsilon}$.

Input: $\mu \in (0,1)$.

1: **if** $\mathcal{I}(\boldsymbol{X}^{k+1}) \subset \mathcal{I}(\boldsymbol{X}^k)$ **then**

2: $\quad \epsilon_i^{k+1} = \mu\epsilon_i^k$, $\forall i \in \mathcal{I}^{k+1}$.

3: $\quad$ Set $\tau_1 = \sigma^{k+1}_{|\mathcal{I}^{k+1}|} + \epsilon^{k+1}_{|\mathcal{I}^{k+1}|}$ and $\tau_2 = \epsilon^k_{|\mathcal{I}^{k+1}|+1}$.

4: $\quad \epsilon_i^{k+1} = \begin{cases} \epsilon_i^k, & \text{if } \tau_1 \geq \tau_2, \\ \min(\epsilon_i^k, \mu\tau_1), & \text{otherwise,} \end{cases} \quad \forall i \in \mathcal{I}(\boldsymbol{X}^k)\backslash\mathcal{I}(\boldsymbol{X}^{k+1})$.

5: $\quad$ Set $\tau_3 = \epsilon^{k+1}_{|\mathcal{I}(\boldsymbol{X}^k)|}$

6: $\quad \epsilon_i^{k+1} = \min\{\epsilon_i^k, \tau_3\}$, $\forall i \in \mathcal{Z}(\boldsymbol{X}^k)$.

7: **end if**

8: **if** $\mathcal{I}(\boldsymbol{X}^k) \subset \mathcal{I}(\boldsymbol{X}^{k+1})$ **then**

9: $\quad \epsilon_i^{k+1} = \mu\epsilon_i^k$, $i \in \mathcal{I}(\boldsymbol{X}^k)$.

10: $\quad$ Set $\tau_3 = \epsilon^k_{|\mathcal{I}(\boldsymbol{X}^k)|}$

11: $\quad \epsilon_i^{k+1} = \mu\min\{\epsilon_i^k, \tau_3\}$, $\forall i \in \mathcal{I}(\boldsymbol{X}^{k+1})\backslash\mathcal{I}(\boldsymbol{X}^k)$.

12: $\quad$ Set $\tau_1 = \sigma^{k+1}_{|\mathcal{I}(\boldsymbol{X}^{k+1})|} + \epsilon^{k+1}_{|\mathcal{I}(\boldsymbol{X}^{k+1})|}$ and $\tau_2 = \epsilon^k_{|\mathcal{I}(\boldsymbol{X}^{k+1})|+1}$.

13: $\quad \epsilon_i^{k+1} = \begin{cases} \epsilon_i^k, & \text{if } \tau_1 \geq \tau_2, \\ \min(\epsilon_i^k, \mu\tau_1), & \text{otherwise,} \end{cases} \quad \forall i \in \mathcal{Z}(\boldsymbol{X}^{k+1})$.

14: **end if**

15: **if** $\mathcal{I}(\boldsymbol{X}^k) = \mathcal{I}(\boldsymbol{X}^{k+1})$ **then**

16: $\quad \epsilon_i^{k+1} = \mu\epsilon_i^k$, $\forall i \in \mathcal{I}(\boldsymbol{X}^{k+1})$.

17: $\quad$ Set $\tau_1 = \sigma^{k+1}_{|\mathcal{I}(\boldsymbol{X}^{k+1})|} + \epsilon^{k+1}_{|\mathcal{I}(\boldsymbol{X}^{k+1})|}$ and $\tau_2 = \epsilon^k_{|\mathcal{I}(\boldsymbol{X}^{k+1})|+1}$.

18: $\quad \epsilon_i^{k+1} = \begin{cases} \epsilon_i^k, & \text{if } \tau_1 \geq \tau_2, \\ \min(\epsilon_i^k, \mu\tau_1), & \text{otherwise,} \end{cases} \quad \forall i \in \mathcal{Z}(\boldsymbol{X}^{k+1})$.

19: **end if**

2.3 An Adaptive Updating Strategy for Perturbation $\boldsymbol{\epsilon}$

A key component of our proposed algorithmic framework is the updating strategy for the perturbation $\boldsymbol{\epsilon}$, which controls how the perturbation evolves during optimization and is critical for analyzing the behavior of our proposed algorithm. The updating strategy should be designed such that we can manipulate the values of $\epsilon_i, \forall i \in [m]$ to maintain the ascending order of $\{w_1, \ldots, w_m\}$ during iteration. This ensures that a global optimal solution to subproblem (2.5) can be obtained according to Proposition 2. Our proposed updating strategy is presented in Algorithm 2.

Assume the initial $\epsilon_i^0, i \in [m]$ are in descending order. Algorithm 2 includes three cases. Our focus is mainly on providing detailed explanations for the first case, as other cases follow similar arguments. **Case 1: $\mathcal{I}(\boldsymbol{X}^{k+1}) \subset \mathcal{I}(\boldsymbol{X}^k)$ holds true in Line 1.** This case corresponds to a situation in which $\boldsymbol{X}^{k+1}$ have more zero singular values than $\boldsymbol{X}^k$, meaning $|\mathcal{I}(\boldsymbol{X}^{k+1})| < |\mathcal{I}(\boldsymbol{X}^k)|$, or, equivalently, $\text{Rank}(\boldsymbol{X}^{k+1}) < \text{Rank}(\boldsymbol{X}^k)$. Notice that all the elements in $\boldsymbol{\sigma}(\boldsymbol{X}^{k+1})$ are naturally organized in descending order. Our goal is to maintain the descending order of $\sigma_i^{k+1} + \epsilon_i^{k+1}, i \in [m]$ (or, equivalently, the ascending order of w_i^{k+1}, $i \in [m]$). To achieve this, we first decrease $\epsilon_i^{k+1}, \forall i \in \mathcal{I}(\boldsymbol{X}^{k+1})$ by a fraction (Line 2), so that

$\sigma_i^{k+1} + \epsilon_i^{k+1}, i \in \mathcal{I}(\boldsymbol{X}^{k+1})$ are descending after the update. Let τ_1 be their smallest value. Lines 3-4 handle the update of $\epsilon_i^{k+1}, \forall i \in \mathcal{I}(\boldsymbol{X}^k)\backslash\mathcal{I}(\boldsymbol{X}^{k+1})$. Let τ_2 be the largest value of $\epsilon_i^k, \forall i \in \mathcal{I}(\boldsymbol{X}^k)\backslash\mathcal{I}(\boldsymbol{X}^{k+1})$ (they are in descending order). If $\tau_1 \geq \tau_2$, there is no need to reduce $\epsilon_i^k, \forall i \in \mathcal{I}(\boldsymbol{X}^k)\backslash\mathcal{I}(\boldsymbol{X}^{k+1})$, since $\sigma_i^{k+1} + \epsilon_i^{k+1}, i \in \mathcal{I}(X^k) = \mathcal{I}(\boldsymbol{X}^{k+1}) \cup (\mathcal{I}(\boldsymbol{X}^k)\backslash\mathcal{I}(\boldsymbol{X}^{k+1}))$ are now descending. Otherwise, $\epsilon_i^k, \forall i \in \mathcal{I}(\boldsymbol{X}^k)\backslash\mathcal{I}(\boldsymbol{X}^{k+1})$ is set to be a fraction of τ_1 to maintain this order. As for $i \in \mathcal{Z}(\boldsymbol{X}^k)$, letting τ_3 be the smallest value of $\sigma_i^{k+1} + \epsilon_i^{k+1}, i \in \mathcal{I}(X^k)$, we then use threshold τ_3 to trim $i \in \mathcal{Z}(\boldsymbol{X}^k)$ (Line 6). After the update, $\sigma_i^{k+1} + \epsilon_i^{k+1}, i \in [m]$ maintain the desired non-decreasing order.

3. Convergence Analysis

We make heavy use of the following auxiliary function in our analysis, which is useful in establishing the convergence properties of the proposed algorithm.

$$H(\boldsymbol{X}, \boldsymbol{Y}, \epsilon) = f(\boldsymbol{X}) + \frac{\beta}{2}\|\boldsymbol{X} - \boldsymbol{Y}\|_F^2 + \lambda \sum_{i=1}^{m} (\sigma_i(\boldsymbol{X}) + \epsilon_i)^p. \tag{3.1}$$

The following lemma indicates that $H(\boldsymbol{X}, \boldsymbol{Y}, \epsilon)$ is monotonically nonincreasing.

Lemma 5 (Sufficient decrease property of EIRNAMI) *Suppose Assumptions 1–2 are satisfied. Let $\{\boldsymbol{X}^k\}$ and $\{\boldsymbol{Y}^k\}$ be the sequences generated by Algorithm 1. Then the following statements hold.*

(i) $\left\{H(\boldsymbol{X}^k, \boldsymbol{X}^{k-1}, \boldsymbol{\epsilon}^k)\right\}$ is monotonically nonincreasing and $\lim\limits_{k\to+\infty} H(\boldsymbol{X}^k, \boldsymbol{X}^{k-1}, \boldsymbol{\epsilon}^k)$ exists. Indeed, we have for each $k \in \mathbb{N}$ that

$$H(\boldsymbol{X}^k, \boldsymbol{X}^{k-1}, \boldsymbol{\epsilon}^k) - H(\boldsymbol{X}^{k+1}, \boldsymbol{X}^k, \boldsymbol{\epsilon}^{k+1}) \geq C\|\boldsymbol{X}^k - \boldsymbol{X}^{k-1}\|_F^2 \tag{3.2}$$

with constant $C = \frac{\beta}{2}(1 - \frac{3L_f+\beta}{\beta}\bar{\alpha}^2) > 0$.

(ii) $\sum_{k=0}^{+\infty} \|\boldsymbol{X}^{k+1} - \boldsymbol{X}^k\|_F^2 < +\infty$, implying $\lim\limits_{k\to+\infty} \|\boldsymbol{X}^{k+1} - \boldsymbol{X}^k\|_F = 0$. In addition, $\lim\limits_{k\to+\infty} \max\{\|\boldsymbol{Y}^k - \boldsymbol{X}^k\|_F, \|\boldsymbol{Y}^k - \boldsymbol{X}^{k+1}\|_F\} = 0$.

(iii) The sequences $\{\boldsymbol{X}^k\}$ and $\{\boldsymbol{Y}^k\}$ are bounded. As a result, there exists constant C_1 such that for any $i \in [m]$

$$\max_i \sigma_i\left(\frac{\boldsymbol{X}^k + \boldsymbol{Y}^k}{2} - \frac{1}{2\beta}\nabla f(\boldsymbol{Y}^k)\right) \leq C_1, \ \forall k \in \mathbb{N}.$$

Proof (i) From (2.5), we know that $\boldsymbol{X}^{k+1} \in \arg\min_{\boldsymbol{X}\in\mathbb{R}^{m\times n}} L(\boldsymbol{X}; \boldsymbol{X}^k, \boldsymbol{Y}^k, \boldsymbol{\epsilon}^k)$, and it follows from $L(\boldsymbol{X}^{k+1}; \boldsymbol{X}^k, \boldsymbol{Y}^k, \boldsymbol{\epsilon}^k) \leq L(\boldsymbol{X}^k; \boldsymbol{X}^k, \boldsymbol{Y}^k, \boldsymbol{\epsilon}^k)$ that

$$\begin{aligned} &\langle \boldsymbol{X}^{k+1}, \nabla f(\boldsymbol{Y}^k)\rangle + \frac{\beta}{2}\|\boldsymbol{X}^{k+1} - \boldsymbol{Y}^k\| + \frac{\beta}{2}\|\boldsymbol{X}^{k+1} - \boldsymbol{X}^k\|_F^2 + \lambda\sum_{i=1}^{m} w_i^k \sigma_i(\boldsymbol{X}^{k+1}) \\ \leq\ &\langle \boldsymbol{X}^k, \nabla f(\boldsymbol{Y}^k)\rangle + \frac{\beta}{2}\|\boldsymbol{X}^k - \boldsymbol{Y}^k\| + \lambda\sum_{i=1}^{m} w_i^k \sigma_i(\boldsymbol{X}^k). \end{aligned} \tag{3.3}$$

By rearranging (3.3) and adding a positive term on both sides, it holds that

$$\begin{aligned}
&\langle \boldsymbol{X}^{k+1}, \nabla f(\boldsymbol{Y}^k)\rangle + \frac{\beta}{2}\|\boldsymbol{X}^{k+1}-\boldsymbol{Y}^k\| + \lambda\sum_{i=1}^m w_i^k(\sigma_i(\boldsymbol{X}^{k+1})-\sigma_i(\boldsymbol{X}^k)) + \lambda\sum_{i=1}^m(\sigma_i(\boldsymbol{X}^k)+\epsilon_i^k)^p \\
\leq\ &\langle \boldsymbol{X}^{k}, \nabla f(\boldsymbol{Y}^k)\rangle + \frac{\beta}{2}\|\boldsymbol{X}^{k}-\boldsymbol{Y}^k\| - \frac{\beta}{2}\|\boldsymbol{X}^{k+1}-\boldsymbol{X}^k\|_F^2 + \lambda\sum_{i=1}^m(\sigma_i(\boldsymbol{X}^k)+\epsilon_i^k)^p.
\end{aligned} \tag{3.4}$$

Denote $\phi(\boldsymbol{X};\boldsymbol{Y}) = f(\boldsymbol{Y}) + \langle \nabla f(\boldsymbol{Y}), \boldsymbol{X}-\boldsymbol{Y}\rangle$ for notional convenience. It leads us to

$$\begin{aligned}
&F(\boldsymbol{X}^{k+1};\boldsymbol{\epsilon}^{k+1}) \\
\overset{(a)}{\leq}\ &\phi(\boldsymbol{X}^{k+1};\boldsymbol{Y}^k) + \frac{L_f}{2}\|\boldsymbol{X}^{k+1}-\boldsymbol{Y}^k\|_F^2 + \lambda\sum_{i=1}^m(\sigma_i(\boldsymbol{X}^{k+1})+\epsilon_i^k)^p \\
\overset{(b)}{\leq}\ &\phi(\boldsymbol{X}^{k+1};\boldsymbol{Y}^k) + \frac{\beta}{2}\|\boldsymbol{X}^{k+1}-\boldsymbol{Y}^k\|_F^2 + \lambda\sum_{i=1}^m(\sigma_i(\boldsymbol{X}^{k})+\epsilon_i^k)^p \\
&+ \lambda\sum_{i=1}^m w_i^k(\sigma_i(\boldsymbol{X}^{k+1})-\sigma_i(\boldsymbol{X}^k)) \\
\overset{(c)}{\leq}\ &\phi(\boldsymbol{X}^{k};\boldsymbol{Y}^k) + \frac{\beta}{2}\|\boldsymbol{X}^{k}-\boldsymbol{Y}^k\|_F^2 - \frac{\beta}{2}\|\boldsymbol{X}^{k+1}-\boldsymbol{X}^k\|_F^2 + \lambda\sum_{i=1}^m(\sigma_i(\boldsymbol{X}^k)+\epsilon_i^k)^p \\
\overset{(d)}{\leq}\ &f(\boldsymbol{X}^k) + \langle \nabla f(\boldsymbol{Y}^k) - \nabla f(\boldsymbol{X}^k), \boldsymbol{X}^k-\boldsymbol{Y}^k\rangle + \frac{L_f+\beta}{2}\|\boldsymbol{X}^k-\boldsymbol{Y}^k\|_F^2 \\
&- \frac{\beta}{2}\|\boldsymbol{X}^{k+1}-\boldsymbol{X}^k\|_F^2 + \lambda\sum_{i=1}^m(\sigma_i(\boldsymbol{X}^k)+\epsilon_i^k)^p \\
\overset{(e)}{\leq}\ &F(\boldsymbol{X}^k,\boldsymbol{\epsilon}^k) + \frac{3L_f+\beta}{2}\|\boldsymbol{X}^k-\boldsymbol{Y}^k\|_F^2 - \frac{\beta}{2}\|\boldsymbol{X}^{k+1}-\boldsymbol{X}^k\|_F^2,
\end{aligned} \tag{3.5}$$

where inequalities (a) follows from the result in (Nesterov, 2003, Lemma 1.2.3) under Assumption 1 and leverages the monotonicity of $(\cdot)^p$ over $\mathbb{R}_+$ resulting from the nonincreasing property of $\sigma_i(\boldsymbol{X}^k)+\epsilon_i^k, \forall i\in[m], k\in\mathbb{N}$ by Algorithm 2, (b) is true because $\beta > L_f$ and the concavity of $(\cdot)^p$ over $\mathbb{R}_+$, (c) makes use of (3.4), (d) again follows from the result in (Nesterov, 2003, Lemma 1.2.3) under Assumption 1, and (e) is by the Cauchy-Schwarz inequality and hence immediately a consequence of Assumption 1.

Rearranging (3.5), together with (2.2), yields

$$F(\boldsymbol{X}^k;\boldsymbol{\epsilon}^k) - F(\boldsymbol{X}^{k+1};\boldsymbol{\epsilon}^{k+1}) \geq \frac{\beta}{2}\|\boldsymbol{X}^k-\boldsymbol{X}^{k+1}\|_F^2 - \frac{3L_f+\beta}{2}\alpha_k^2\|\boldsymbol{X}^k-\boldsymbol{X}^{k-1}\|_F^2. \tag{3.6}$$

This further implies

$$\begin{aligned} & H(\boldsymbol{X}^k, \boldsymbol{X}^{k-1}, \boldsymbol{\epsilon}^k) - H(\boldsymbol{X}^{k+1}, \boldsymbol{X}^k, \boldsymbol{\epsilon}^{k+1}) \\ = & F(\boldsymbol{X}^k; \boldsymbol{\epsilon}^k) + \frac{\beta}{2}\|\boldsymbol{X}^k - \boldsymbol{X}^{k-1}\|_F^2 - \left[F(\boldsymbol{X}^{k+1}; \boldsymbol{\epsilon}^{k+1}) + \frac{\beta}{2}\|\boldsymbol{X}^{k+1} - \boldsymbol{X}^k\|_F^2\right] \\ \overset{(a)}{\geq} & \frac{\beta}{2}\left(1 - \alpha_k^2 \frac{3L_f + \beta}{\beta}\right)\|\boldsymbol{X}^k - \boldsymbol{X}^{k-1}\|_F^2 \\ \overset{(b)}{\geq} & \frac{\beta}{2}\left(1 - \bar{\alpha}^2 \frac{3L_f + \beta}{\beta}\right)\|\boldsymbol{X}^k - \boldsymbol{X}^{k-1}\|_F^2 \geq 0, \end{aligned} \tag{3.7}$$

where inequality (a) holds by (3.6) and (b) is true thanks to $\alpha_k \in [0, \bar{\alpha}], \forall k \in \mathbb{N}$ with $\bar{\alpha} \in (0, \sqrt{\frac{\beta}{3L_f+\beta}})$ imposed in (2.3). Consequently, $\{H(\boldsymbol{X}^k, \boldsymbol{X}^{k-1}, \boldsymbol{\epsilon}^k)\}$ is monotonically decreasing. Moreover, by Assumption 2, we know $\min_{\boldsymbol{X}\in\mathbb{R}^{m\times n}} F(\boldsymbol{X}, \boldsymbol{\epsilon}) > \underline{F} > -\infty$, implying $\{H(\boldsymbol{X}^k, \boldsymbol{X}^{k-1}, \boldsymbol{\epsilon}^k)\}$ is bounded from below, and hence, $\lim_{k\to+\infty} H(\boldsymbol{X}^k, \boldsymbol{X}^{k-1}, \boldsymbol{\epsilon}^k)$ exists. This proves Statement (i).

(ii) Summing both sides of (3.7) over $k = 0, \ldots, t$, we obtain

$$\begin{aligned} \frac{\beta}{2}\left(1 - \bar{\alpha}^2 \frac{3L_f + \beta}{\beta}\right)\sum_{k=0}^{t}\|\boldsymbol{X}^k - \boldsymbol{X}^{k-1}\|_F^2 \leq & \; H(\boldsymbol{X}^0, \boldsymbol{X}^{-1}, \boldsymbol{\epsilon}^0) - H(\boldsymbol{X}^{t+1}, \boldsymbol{X}^t, \boldsymbol{\epsilon}^{t+1}) \\ \leq & \; F(\boldsymbol{X}^0; \boldsymbol{\epsilon}^0) - F(\boldsymbol{X}^{t+1}; \boldsymbol{\epsilon}^{t+1}) < \; +\infty, \end{aligned} \tag{3.8}$$

Let $t \to +\infty$, and it follows from Assumption 2 and $\bar{\alpha} \in (0, \sqrt{\frac{\beta}{3L_f+\beta}})$ that $\lim_{k\to+\infty} \|\boldsymbol{X}^{k+1} - \boldsymbol{X}^k\|_F = 0$. Furthermore,

$$\begin{aligned} \lim_{k\to+\infty} \|\boldsymbol{Y}^k - \boldsymbol{X}^k\|_F = \lim_{k\to+\infty} \alpha_k \|(\boldsymbol{X}^k - \boldsymbol{X}^{k-1})\|_F = 0, \text{ and} \\ \lim_{k\to+\infty} \|\boldsymbol{Y}^k - \boldsymbol{X}^{k+1}\|_F = \lim_{k\to+\infty} \|(\boldsymbol{X}^k - \boldsymbol{X}^{k+1}) + \alpha_k(\boldsymbol{X}^k - \boldsymbol{X}^{k-1})\|_F = 0, \end{aligned} \tag{3.9}$$

as desired. The proof of Statement (ii) is completed.

(iii) For each $k \in \mathbb{N}$, we derive that

$$F(\boldsymbol{X}^k) \leq F(\boldsymbol{X}^k; \boldsymbol{\epsilon}^k) \leq H(\boldsymbol{X}^k, \boldsymbol{X}^{k-1}, \boldsymbol{\epsilon}^k) \overset{(a)}{\leq} H(\boldsymbol{X}^0, \boldsymbol{X}^{-1}, \boldsymbol{\epsilon}^0) = F(\boldsymbol{X}^0; \boldsymbol{\epsilon}^0) \overset{(b)}{<} +\infty, \tag{3.10}$$

where inequality (a) holds by (3.7) and Theorem 7(i), and (b) is true due to Assumption 2. We hence deduce from (3.10) that $\{\boldsymbol{X}^k\}$ is bounded. This, together with the boundedness of $\alpha_k, \forall k \in \mathbb{N}$, shows the boundedness of $\{\boldsymbol{Y}^k\}$. Therefore, it follows from Assumptions 1–2 that there exists $C_1 > 0$ such that for any $i \in [m]$, $\max_i \sigma_i\left(\frac{\boldsymbol{X}^k+\boldsymbol{Y}^k}{2} - \frac{1}{2\beta}\nabla f(\boldsymbol{Y}^k)\right) \leq C_1, \; \forall k \in \mathbb{N}$. This completes the proofs of all statements. ■

3.1 Local Properties: Stable Support and Adaptively Reweighting

The following proposition asserts that the support of the index set of singular vectors remains unchanged after a finite number of iterations.

Proposition 6 *Let Assumptions 1–2 hold. Suppose that the sequence $\{\boldsymbol{X}^k\}$ is generated by Algorithm 1. Then there exists a constant $C_1 > 0$ defined in Lemmas 5 and $\hat{k} \in \mathbb{N}$ such that the following statements hold.*

(i) If $w(\sigma_i(\boldsymbol{X}^{\hat{k}}), \epsilon_i^{\hat{k}}) > \frac{2\beta C_1}{\lambda}$ for $\hat{k} \in \mathbb{N}$, then $\sigma_i(\boldsymbol{X}^k) \equiv 0$ for all $k > \hat{k}$.

(ii) The index sets $\mathcal{I}(\boldsymbol{X}^k)$ and $\mathcal{Z}(\boldsymbol{X}^k)$ remain unchanged for all $k > \hat{k}$. Therefore, there exists index sets $\mathcal{I}^*$ and $\mathcal{Z}^*$ such that $\mathcal{I}(\boldsymbol{X}^k) = \mathcal{I}^*$ and $\mathcal{Z}(\boldsymbol{X}^k) = \mathcal{Z}^*$ for sufficiently large k.

(iii) For any $k > \hat{k}$, σ_i^k is strictly bounded away from 0 for any $i \in \mathcal{I}(\boldsymbol{X}^k)$. Indeed, it holds that $\sigma_i^k > \left(\frac{\lambda p}{2\beta C_1}\right)^{\frac{1}{1-p}} - \epsilon_i^k > 0, \forall i \in \mathcal{I}(\boldsymbol{X}^k)$, implying $\liminf_{k\to+\infty} \sigma_i^k > \left(\frac{\lambda p}{2\beta C_1}\right)^{\frac{1}{1-p}}, \forall i \in \mathcal{I}^*$.

Proof Recall (2.5) and by Proposition 3, we know that matrices $\boldsymbol{X}^{k+1}$ and $\frac{\boldsymbol{X}^k+\boldsymbol{Y}^k}{2} - \frac{1}{2\beta}\nabla f(\boldsymbol{Y}^k)$ have the simultaneous ordered SVD, and

$$\mathbf{0} \in \beta\left(\boldsymbol{X}^{k+1} - \left(\frac{\boldsymbol{X}^k + \boldsymbol{Y}^k}{2} - \frac{\nabla f(\boldsymbol{Y}^k)}{2\beta}\right)\right) + \frac{\lambda}{2}\partial\left(\sum_{i=1}^m w_i^k \sigma_i(\boldsymbol{X}^{k+1})\right),$$

which implies

$$\mathbf{0} = \boldsymbol{\sigma}(\boldsymbol{X}^{k+1}) - \boldsymbol{\sigma}\left(\frac{\boldsymbol{X}^k + \boldsymbol{Y}^k}{2} - \frac{1}{2\beta}\nabla f(\boldsymbol{Y}^k)\right) + \frac{\lambda}{2\beta}\boldsymbol{w}^k \circ \boldsymbol{\xi}^{k+1} \tag{3.11}$$

with $\xi_i^{k+1} \in [0,1]$. Then, we have for each $i \in [m]$ that

$$\sigma_i(\boldsymbol{X}^{k+1}) = \left[\sigma_i\left(\frac{\boldsymbol{X}^k + \boldsymbol{Y}^k}{2} - \frac{1}{2\beta}\nabla f(\boldsymbol{Y}^k)\right) - \frac{\lambda}{2\beta}w_i^k\xi_i^{k+1}\right]_+. \tag{3.12}$$

(i) Suppose that there exists $\hat{k} \in \mathbb{N}$ such that $w_i^{\hat{k}} \geq \frac{2\beta C_1}{\lambda}$ for $i \in [m]$. By Proposition 3 and from (3.12), we know that $\sigma_i(\boldsymbol{X}^{\hat{k}+1}) = 0$. Then, $\sigma_i(\boldsymbol{X}^{\hat{k}+1}) + \epsilon_i^{\hat{k}+1} \leq \sigma_i(\boldsymbol{X}^{\hat{k}}) + \epsilon_i^{\hat{k}}$ by Algorithm 2 and monotonicity of $(\cdot)^{p-1}$ indicate $w_i^{\hat{k}+1} \geq w_i^{\hat{k}} > \frac{2\beta C_1}{\lambda}$. Therefore, we have $\sigma_i^{\hat{k}+2} = 0$. By induction, we know that $\sigma_i^k \equiv 0$ for any $k > \hat{k}$. This completes the proof of statement (i).

(ii) We prove this by contradiction. Suppose this statement is not true. Then there exist $i \in [m]$ and $k \in \mathbb{N}$ such that $\sigma_i(\boldsymbol{X}^k)$ takes a zero and nonzero value both infinitely. We know that 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

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

Hence, there exists subsequence $\mathcal{S}_3 \subset \mathcal{S}_2$ such that $|\mathcal{S}_3| = +\infty$ and $i \in \mathcal{Z}(\boldsymbol{X}^k) \cap \mathcal{I}(\boldsymbol{X}^{k+1})$ for any $k \in \mathcal{S}_3$. In other words, $\sigma_i(\boldsymbol{X}^k) = 0$ and $\sigma_i(\boldsymbol{X}^{k+1}) \neq 0$ for any $k \in \mathcal{S}_3$. Thus, $\mathcal{I}(X^k) \subset \mathcal{I}(X^{k+1})$, $\forall k \in \mathcal{S}_3$. Then it follows from Algorithm 2 that $\lim_{k\to+\infty} \epsilon_i^k = 0$ for $k \in \mathcal{S}_3$ due to $|\mathcal{S}_3| = +\infty$. Hence there exists $\hat{k} \in \mathcal{S}_1$ such that

$$w_i^{\hat{k}} = w(\sigma_i^{\hat{k}}, \epsilon_i^{\hat{k}}) = p\left(\sigma_i(\boldsymbol{X}^{\hat{k}}) + \epsilon_i^{\hat{k}}\right)^{p-1} = p(\epsilon_i^{\hat{k}})^{p-1} \geq \frac{2\beta C_1}{\lambda}.$$

This indicates that $\sigma_i(\boldsymbol{X}^k) = 0$ for any $k > \hat{k}$ by statement (i), implying $\{\hat{k}+1, \hat{k}+2, \hat{k}+3, \cdots\} \subset \mathcal{S}_1$ and $|\mathcal{S}_2|$ is finite. This contradicts $|\mathcal{S}_2| = +\infty$. Consequently, statement (ii) holds.

(iii) We know from statement (i) that if $w_i^k \leq \frac{2\beta C_1}{\lambda}$, $i \in \mathcal{I}(\boldsymbol{X}^k)$ occurs, it then follows that $\sigma_i(\boldsymbol{X}^k) \geq \left(\frac{\lambda p}{2\beta C_1}\right)^{\frac{1}{1-p}} - \epsilon_i^k > 0,\ \forall i \in \mathcal{I}(\boldsymbol{X}^k)$. This completes proofs of all statements. ■

We shall further demonstrate the properties for the update strategy of $\boldsymbol{\epsilon}$. We specifically show that after some k, $\epsilon_i, \forall i \in \mathcal{I}(\boldsymbol{X}^k)$ diminish while $\epsilon_i, \forall i \in \mathcal{Z}(\boldsymbol{X}^k)$ are fixed as constants. Hence, our proposed Algorithm 1 locally behaves as minimizing a smooth problem in a low-dimensional manifold.

Theorem 7 *Suppose Assumption 1-2 hold true. Let $\{\boldsymbol{X}^k\}$ and $\{\boldsymbol{\epsilon}^k\}$ be the sequences generated by Algorithm 1 and $\mathcal{I}^*$ and $\mathcal{Z}^*$ be defined in Proposition 6(ii). Then the following assertions hold.*

(i) The perturbations $\{\epsilon_i^k, \forall i \in \mathcal{I}(\boldsymbol{X}^k)\}$ and perturbed singular values $\{\sigma_i(\boldsymbol{X}^k) + \epsilon_i^k, \forall i \in [m]\}$ are all in strictly descending order for $k \in \mathbb{N}$, while $\{\epsilon_i^k, \forall i \in \mathcal{Z}(\boldsymbol{X}^k)\}$ are in non-increasing order for $k \in \mathbb{N}$. Consequently, for all $k \in \mathbb{N}$, the nonstrictly ascending order constraint on the nonnegative weights in (2.6) can be automatically satisfied.

(ii) There exist $\hat{k} \in \mathbb{N}$ such that for all $k \geq \hat{k}$, the update $\epsilon_i^{k+1} = \mu\epsilon_i^k$ with $\mu \in (0,1)$, $\forall i \in \mathcal{I}^*$ will always be triggered. Consequently, the sequence $\{\epsilon_i^k\}, \forall i \in \mathcal{I}^*$ converges monotonically to 0, i.e., $\epsilon_i^k \to 0$ as $k \to +\infty$ for all $i \in \mathcal{I}^*$.

(iii) There exists $\hat{k} \in \mathbb{N}$ such that for all $k \geq \hat{k}$, the update of $\epsilon_i^k, \forall i \in \mathcal{Z}(\boldsymbol{X}^k)$ will never be triggered. That is, $\epsilon_i^k \equiv \epsilon_i^{\hat{k}}$ after some $\hat{k}$, $\forall i \in \mathcal{Z}^*$. Consequently, the sequence $\{\epsilon_i^k\}$, $\forall i \in \mathcal{Z}^*$ converges to fixed positive constants for all sufficiently large k.

Proof (i) We prove this statement by induction. Indeed, by the setting of $\boldsymbol{\epsilon}^0$ in Algorithm 1 and the property that the singular values are naturally sorted in descending order, the statement is vacuously true at $k = 0$. Suppose now this is also true at the kth iteration. Without loss of generality, we only have to prove the statement (i) in which $\mathcal{I}(\boldsymbol{X}^{k+1}) \subset \mathcal{I}(\boldsymbol{X}^k)$ holds in Algorithm 2, and the proof of other cases follows the similar spirits and arguments.

Consider Algorithm 2. Line 2 and Line 4 indicate that $\{\epsilon_i^{k+1}, \forall i \in \mathcal{I}(\boldsymbol{X}^k)\}$ is in descending order by the descending nature of $\{\epsilon_i^k, \forall i \in \mathcal{I}(\boldsymbol{X}^k)\}$ and $\mu \in (0,1)$. Line 11 guarantees $\{\epsilon_i^{k+1}, i \in \mathcal{Z}(\boldsymbol{X}^k)\}$ is non-increasing, since $\{\epsilon_i^k, \forall i \in \mathcal{Z}(\boldsymbol{X}^k)\}$ are in non-increasing order. This, together with the monotonicity of $(\cdot)^{p-1}$, ensures the satisfaction of (2.6). This finishes the proof of statement (i).

(ii) This statement holds true by Proposition 6(ii) and Line 16 of Algorithm 2.

(iii) In search of a contradiction, suppose that the update of $\epsilon_i^k, \forall i \in \mathcal{Z}(\boldsymbol{X}^k)$ for $k > \hat{k}$ is triggered infinitely many times. By Proposition 6(ii), we note that $\epsilon_i^{k+1} \leq \mu\tau_1, i \in \mathcal{Z}(\boldsymbol{X}^{k+1})$ whenever it is reduced by Line 18 in Algorithm 2. If the update is triggered for infinite times, then $\tau_2 \leq \tau_1$ is always satisfied for any $k > \hat{k}$, which contradicts that τ_1 is strictly bounded below from 0 after some $k > \hat{k}$ by Proposition 6(iii). Therefore, $\epsilon_i^k, i \in \mathcal{Z}(\boldsymbol{X}^k)$ is

never reduced after finite iterations. This contradiction completes the proof. ■

3.2 Active Manifold Identification

The following theorem establishes the active manifold identification property of Algorithm 1, which is a straightforward result from Proposition 6 and Theorem 7. It asserts the rank of the iterates generated by Algorithm 1 will eventually remain fixed, and is equivalent to the rank of the cluster point $\boldsymbol{X}^*$. In addition, all cluster points have the same rank. Consequently, the iterates $\{\boldsymbol{X}^k\}$ will eventually reside in a low-dimensional active manifold $\mathcal{M}(\boldsymbol{X}^*) := \{\boldsymbol{X} \in \mathbb{R}^{m\times n} \mid \text{Rank}(\boldsymbol{X}) = \text{Rank}(\boldsymbol{X}^*)\}$, implying that the original problem eventually reverts to a smooth problem after identifying an active manifold $\mathcal{M}(\boldsymbol{X}^*)$.

Theorem 8 *Suppose Assumptions 1–2 hold. Let $\{\boldsymbol{X}^k\}$ be the sequence generated by Algorithm 1. Then $\text{Rank}(\boldsymbol{X}^k) = r^* := |\mathcal{I}^*|$ for sufficiently large k. Moreover, for any limit point $\boldsymbol{X}^*$ of $\{\boldsymbol{X}^k\}$, $\text{Rank}(\boldsymbol{X}^*) = |\mathcal{I}(\boldsymbol{X}^*)| = r^*$, and $\sigma_i(\boldsymbol{X}^*) \geq \left(\frac{\lambda p}{2\beta C_1}\right)^{\frac{1}{1-p}} > 0$, $i \in \mathcal{I}(\boldsymbol{X}^*)$.*

Remark 9 *Theorem 8 suggests that the proposed Algorithm 1 identifies the rank of the optimal solution after finite iterations. That is, all subsequent iterates $\boldsymbol{X}^k$ satisfy $\boldsymbol{X}^k \in \mathcal{M}(\boldsymbol{X}^*)$. Moreover, all limit points of iterates will be confined to a low-dimensional manifold. This feature of the proposed algorithm represents a stark contrast to the results established in (pei Lee et al., 2023, Theorem 6), where the active manifold identification property merely holds for any convergent subsequence.*

We illustrate the active manifold identification property of Algorithm 1 using a simple example.

Example 1 *Consider*

$$\min_{\boldsymbol{X}\in\mathbb{R}^{15\times 15}} \frac{1}{2}\|\mathcal{P}_\Omega(\boldsymbol{X}) - \mathcal{P}_\Omega(\widehat{\boldsymbol{X}})\| + \lambda\|\boldsymbol{X}\|_p^p, \tag{3.13}$$

where $\widehat{\boldsymbol{X}} \in \mathbb{R}^{15\times 15}$ with $\text{Rank}(\widehat{\boldsymbol{X}}) = 3$ is the ground-truth matrix to be found, Ω denotes the random sample set with sampling ratio (SR) 0.5 and $|\Omega|$ satisfies $|\Omega| = \lceil 15^2 * SR \rceil$, $\mathcal{P}_\Omega$ is the projection onto the subspace of sparse matrices with nonzeros restricted to the index set Ω, $\lambda = 0.1$, and $p = 0.5$.

The y-axis on the left represents the rank and the one on the right represents the relative residual. We use solid lines and dashed lines to show the rank of the iterates and the relative residual $\|\boldsymbol{X}^k - \widehat{\boldsymbol{X}}\|_F^2/\|\boldsymbol{X}^0 - \widehat{\boldsymbol{X}}\|_F^2$ of our method, respectively. The gray line represents the rank at the optimum $\widehat{\boldsymbol{X}}$. As observed in Figure 1, when the rank of the iterate reaches $\text{Rank}(\widehat{\boldsymbol{X}})$, the relative distance also decreases significantly. This indicates the algorithm finds a high-precision solution on the smooth fixed-rank manifold.

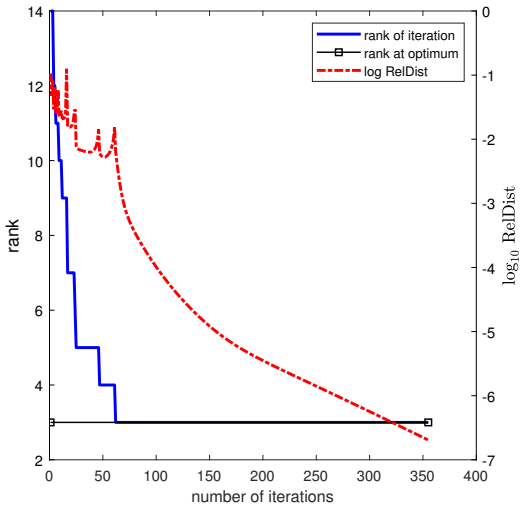


Figure 1: A sample example to show the active manifold identification property.

3.3 Global Convergence

By Proposition 1, the necessary optimality condition of $(\mathcal{P})$ is given by:

$$\nabla f(\boldsymbol{X}^*) + \lambda \boldsymbol{U}^* \mathrm{diag}(\boldsymbol{w}^* \circ \boldsymbol{\xi}^*) {\boldsymbol{V}^*}^\top = \boldsymbol{0}, \tag{3.14}$$

where $\bar{\boldsymbol{w}}^* \in \partial \|\boldsymbol{\sigma}(\boldsymbol{X}^*)\|_p^p$, $\boldsymbol{\xi}^* \in \partial |\boldsymbol{\sigma}(\boldsymbol{X}^*)|$, and $(\boldsymbol{U}^*, \boldsymbol{V}^*) \in \overline{\mathcal{M}}(\boldsymbol{X}^*)$

To show the global convergence properties of Algorithm 1, we investigate the optimality error at $\boldsymbol{X}^{k+1}$.

Theorem 10 (Bounded subgradients) *Suppose Assumptions 1–2 hold. Let $\{\boldsymbol{X}^k\}$ be the sequence generated by Algorithm 1. For each $k \in \mathbb{N}$, define the optimality error associated with $(\mathcal{P})$ as*

$$E^{k+1} = \nabla f(\boldsymbol{X}^{k+1}) + \lambda \boldsymbol{U}^{k+1} \mathrm{diag}(\bar{\boldsymbol{w}}^{k+1} \circ \boldsymbol{\xi}^{k+1}) {\boldsymbol{V}^{k+1}}^\top, \tag{3.15}$$

where $\bar{w}_i^{k+1} = p(\sigma_i(\boldsymbol{X}^{k+1}))^{p-1}, i \in \mathcal{I}(\boldsymbol{X}^{k+1})$, $\bar{w}_i^{k+1} = w_i^k, i \in \mathcal{Z}(\boldsymbol{X}^{k+1})$ and $\boldsymbol{\xi}^{k+1} \in \partial |\sigma(\boldsymbol{X}^{k+1})|$. Then it holds that $E^{k+1} \in \partial F(\boldsymbol{X}^{k+1})$ and there exists $C_2 > 0$ such that

$$\|E^{k+1}\|_F \leq (L_f + 2\beta + C_2)\|\boldsymbol{X}^{k+1} - \boldsymbol{X}^k\|_F + (L_f + \beta)\bar{\alpha}\|\boldsymbol{X}^k - \boldsymbol{X}^{k-1}\|_F + C_2 \epsilon_1^0 \tag{3.16}$$

with $C_2 = mp(1-p)\left(\frac{2\beta C_1}{\lambda p}\right)^{\frac{2-p}{1-p}}$.

Proof *By Proposition 1, we know that $\bar{\boldsymbol{w}}^{k+1} \in \partial \|\boldsymbol{\sigma}(\boldsymbol{X}^{k+1})\|_p^p$. This, together with differentiability of f (Rockafellar and Wets, 2009)[10.10 Exercise], leads to the desired result $E^{k+1} \in \partial F(\boldsymbol{X}^{k+1})$. On the other hand, by subtracting (2.11) from both sides of (3.15), we have*

$$\begin{aligned} E^{k+1} = {} & [\nabla f(\boldsymbol{X}^{k+1}) - \nabla f(\boldsymbol{Y}^k)] - \beta[(\boldsymbol{X}^{k+1} - \boldsymbol{Y}^k) + (\boldsymbol{X}^{k+1} - \boldsymbol{X}^k)] \\ & + \lambda \boldsymbol{U}^{k+1} \mathrm{diag}((\bar{\boldsymbol{w}}^{k+1} - \boldsymbol{w}^k) \circ \boldsymbol{\xi}^{k+1}) {\boldsymbol{V}^{k+1}}^\top. \end{aligned} \tag{3.17}$$

It follows from Assumption 1 and (2.2) *that the first term in* (3.17)

$$\begin{aligned}\|\nabla f(\boldsymbol{X}^{k+1}) - \nabla f(\boldsymbol{Y}^k)\|_F &\leq L_f\|\boldsymbol{X}^{k+1} - \boldsymbol{Y}^k\|_F \\ &\overset{(a)}{\leq} L_f\|\boldsymbol{X}^{k+1} - \boldsymbol{X}^k\|_F + L_f\bar{\alpha}\|\boldsymbol{X}^k - \boldsymbol{X}^{k-1}\|_F,\end{aligned} \tag{3.18}$$

where (a) *holds due to the triangle inequality and* (2.3)*. Similarly, we have from the second term in* (3.17) *that*

$$\|\beta[(\boldsymbol{X}^{k+1} - \boldsymbol{Y}^k) + (\boldsymbol{X}^{k+1} - \boldsymbol{X}^k)]\|_F \leq 2\beta\|\boldsymbol{X}^{k+1} - \boldsymbol{X}^k\|_F + \beta\bar{\alpha}\|\boldsymbol{X}^k - \boldsymbol{X}^{k-1}\|_F. \tag{3.19}$$

As for the third term in (3.17)*, we have*

$$\begin{aligned}&\|\boldsymbol{U}^{k+1}\mathrm{diag}((\bar{\boldsymbol{w}}^{k+1} - \boldsymbol{w}^k) \circ \boldsymbol{\xi}^{k+1}){\boldsymbol{V}^{k+1}}^\top\|_F = \|\mathrm{diag}((\bar{\boldsymbol{w}}^{k+1} - \boldsymbol{w}^k) \circ \boldsymbol{\xi}^{k+1})\|_F \\ &\overset{(a)}{\leq} \|\bar{\boldsymbol{w}}^{k+1} - \boldsymbol{w}^k\|_2 \leq \|\bar{\boldsymbol{w}}^{k+1} - \boldsymbol{w}^k\|_1 \\ &\overset{(b)}{=} \sum_{i=1}^{|\mathcal{I}(\boldsymbol{X}^{k+1})|} p(1-p)\left(\hat{\sigma}_i(\boldsymbol{X}^k)\right)^{p-2}\left(\sigma_i(\boldsymbol{X}^{k+1}) - (\sigma_i(\boldsymbol{X}^k) + \epsilon_i^k)\right) \\ &\overset{(c)}{\leq} \sum_{i=1}^{|\mathcal{I}(\boldsymbol{X}^{k+1})|} p(1-p)\left(\frac{2\beta C_1}{\lambda p}\right)^{\frac{2-p}{1-p}}\left(\left|\sigma_i(\boldsymbol{X}^k) - \sigma_i(\boldsymbol{X}^{k+1})\right| + |\epsilon_i^k|\right) \\ &\overset{(d)}{\leq} \sum_{i=1}^{|\mathcal{I}(\boldsymbol{X}^{k+1})|} p(1-p)\left(\frac{2\beta C_1}{\lambda p}\right)^{\frac{2-p}{1-p}}\left(\|\boldsymbol{X}^k - \boldsymbol{X}^{k+1}\|_F + \epsilon_i^k\right),\end{aligned} \tag{3.20}$$

where (a) *holds due to* $\xi_i^{k+1} \in [-1, 1], \forall i \in [m]$*, equality* (b) *makes use of the mean value theorem with* $\hat{\sigma}_i(\boldsymbol{X}^k)$ *lying between* $\sigma_i(\boldsymbol{X}^k) + \epsilon_i^k$ *and* $\sigma_i(\boldsymbol{X}^{k+1})$ *for each* $i \in [m]$*, inequality* (c) *makes the uses of Proposition 6(iii), the monotonicity of* $(\cdot)^{p-2}$ *over* $\mathbb{R}_{++}$ *and triangle inequality, and inequality* (d) *is true because of the conclusion drawn by (Horn and Johnson, 2012, Corollary 7.3.5).*

Therefore, combining (3.18), (3.19) *with* (3.20) *leads to*

$$\begin{aligned}\|E^{k+1}\|_F &\leq \|\nabla f(\boldsymbol{X}^{k+1}) - \nabla f(\boldsymbol{Y}^k)\|_F + \|\beta[(\boldsymbol{X}^{k+1} - \boldsymbol{Y}^k) + (\boldsymbol{X}^{k+1} - \boldsymbol{X}^k)]\|_F \\ &\quad + \|\boldsymbol{U}^{k+1}\mathrm{diag}((\bar{\boldsymbol{w}}^{k+1} - \boldsymbol{w}^k) \circ \boldsymbol{\xi}^{k+1}){\boldsymbol{V}^{k+1}}^\top\|_F \\ &\leq (L_f + 2\beta)\|\boldsymbol{X}^{k+1} - \boldsymbol{X}^k\|_F + (L_f + \beta)\bar{\alpha}\|\boldsymbol{X}^k - \boldsymbol{X}^{k-1}\|_F \\ &\quad + \sum_{i=1}^{|\mathcal{I}(\boldsymbol{X}^{k+1})|} p(1-p)\left(\frac{2\beta C_1}{\lambda p}\right)^{\frac{2-p}{1-p}}\left(\|\boldsymbol{X}^k - \boldsymbol{X}^{k+1}\|_F + \epsilon_i^k\right) \\ &\overset{(a)}{\leq} \left(L_f + 2\beta + mp(1-p)\left(\frac{2\beta C_1}{\lambda p}\right)^{\frac{2-p}{1-p}}\right)\|\boldsymbol{X}^{k+1} - \boldsymbol{X}^k\|_F \\ &\quad + (L_f + \beta)\bar{\alpha}\|\boldsymbol{X}^k - \boldsymbol{X}^{k-1}\|_F + mp(1-p)\left(\frac{2\beta C_1}{\lambda p}\right)^{\frac{2-p}{1-p}}\epsilon_1^0,\end{aligned} \tag{3.21}$$

where the inequality (a) *is true because* $|\mathcal{I}(\boldsymbol{X}^{k+1})| \leq m, \forall k \in \mathbb{N}$ *and* $\boldsymbol{\epsilon}^0 \in \mathbb{R}^m_\downarrow \cap \mathbb{R}^m_{++}$*. This completes the proof.* ■

Based on the previous analysis, we now demonstrate the global convergence properties of the EIRNAMI algorithm. Before proceeding, we use χ^∞ to denote the cluster point of $\{\boldsymbol{X}^k\}$ generated by Algorithm 1.

Theorem 11 *Suppose Assumptions 1–2 hold. Let $\{\boldsymbol{X}^k\}$ be the sequence generated by Algorithm 1. Then the following assertions hold.*

(i) The set of cluster points χ^∞ is nonempty, compact, and connected.

(ii) $\{F(\boldsymbol{X}^k)\}$ is convergent. Moreover, the objective function F is constant on χ^∞.

(iii) $\lim_{k\to+\infty} \|E^{k+1}\|_F = 0$. Therefore, any cluster point $\boldsymbol{X}^* \in \chi^\infty$ is a critical point of F, meaning $\chi^\infty \subset crit(F)$.

Proof (i) Lemma 5(iii) implies χ^∞ is nonempty. On the other hand, Lemma 5(ii)-(iii), combined with the classical Ostrowski result (Ostrowski, 1973, Theorem 26.1), leads to the desired results. This proves Statement (i).

(ii) The convergence of $\{F(\boldsymbol{X}^k)\}$ is a direct consequence of (3.1) by invoking Theorem 7(ii) and Lemma 5(ii). On the other hand, it holds from Theorem 8 and Lemma 5(i)-(iii) that for each $\boldsymbol{X}^* \in \chi^\infty$ with $\mathrm{Rank}(\boldsymbol{X}^*) = r^*$,

$$\begin{aligned} F(\boldsymbol{X}^*) &= f(\boldsymbol{X}^*) + \lambda \sum_{i=1}^{r^*} (\sigma_i(\boldsymbol{X}^*))^p \\ &\overset{(a)}{=} \lim_{k\to+\infty} \left[f(\boldsymbol{X}^{k+1}) + \lambda \sum_{i=1}^{m} (\sigma_i(\boldsymbol{X}^{k+1}) + \epsilon_i^{k+1})^p + \beta \|\boldsymbol{X}^{k+1} - \boldsymbol{X}^k\|_F^2 \right] - \lambda \sum_{i=r^*+1}^{m} (\epsilon_i^{\hat{k}})^p \\ &= \lim_{k\to+\infty} H(\boldsymbol{X}^{k+1}, \boldsymbol{X}^k, \epsilon^{k+1}) - \lambda \sum_{i=r^*+1}^{m} (\epsilon_i^{\hat{k}})^p \\ &\overset{(b)}{=} H^* - \lambda \sum_{i=r^*+1}^{m} (\epsilon_i^{\hat{k}})^p, \end{aligned}$$

where (a) is true because Theorem 7 guarantees that there exists $\hat{k} \in \mathbb{N}$ such that $\boldsymbol{\epsilon}^k \to \boldsymbol{\epsilon}^* = [0, \cdots, 0, \epsilon_{r^*+1}^{\hat{k}}, \cdots, \epsilon_m^{\hat{k}}]^\top$ as $k \to +\infty$, and also by Lemma 5(ii), and (c) holds simply by Lemma 5(i).

(iii) Recall (3.18) and (3.19), we can deduce from Lemma 5(iii) that

$$\lim_{k\to+\infty} \|\nabla f(\boldsymbol{X}^{k+1}) - \nabla f(\boldsymbol{Y}^k)\|_F = 0 \tag{3.22}$$

and

$$\lim_{k\to+\infty} \|\beta[(\boldsymbol{X}^{k+1} - \boldsymbol{Y}^k) + (\boldsymbol{X}^{k+1} - \boldsymbol{X}^k)]\|_F = 0. \tag{3.23}$$

On the other hand, by Proposition 6(iii) and monotonicity of $(\cdot)^{p-2}$ over $\mathbb{R}_{++}$, we have from (3.20) that for sufficiently large $k \in \mathbb{N}$

$$\begin{aligned} & \|\boldsymbol{U}^{k+1}\mathrm{diag}((\bar{\boldsymbol{w}}^{k+1} - \boldsymbol{w}^k) \circ \boldsymbol{\xi}^{k+1})\boldsymbol{V}^{k+1^\top}\|_F^2 \\ \leq & \sum_{i=1}^{|\mathcal{I}(\boldsymbol{X}^{k+1})|} p(1-p)\left(\hat{\sigma}_{|\mathcal{I}(\boldsymbol{X}^{k+1})|}(\boldsymbol{X}^k)\right)^{p-2}\left(\|\boldsymbol{X}^k - \boldsymbol{X}^{k+1}\|_F + \epsilon_i^k\right) \\ < & \sum_{i=1}^{|\mathcal{I}(\boldsymbol{X}^{k+1})|} p(1-p)\left(\left(\frac{\lambda p}{2\beta C_1}\right)^{1/(1-p)} - \epsilon_i^k\right)^{p-2}\left(\|\boldsymbol{X}^k - \boldsymbol{X}^{k+1}\|_F + \epsilon_i^k\right). \end{aligned} \tag{3.24}$$

By Theorem 7(ii) and Lemma 5(iii), we know from (3.24) that

$$\lim_{k\to+\infty} \|\boldsymbol{U}^{k+1}\mathrm{diag}((\bar{\boldsymbol{w}}^{k+1} - \boldsymbol{w}^k) \circ \boldsymbol{\xi}^{k+1})\boldsymbol{V}^{k+1^\top}\|_F^2 = 0. \tag{3.25}$$

Combining (3.22), (3.23) and (3.25) yields $\lim_{k\to+\infty} \|E^{k+1}\|_F = 0$, meaning $\mathbf{0} \in \partial F(\boldsymbol{X}^*)$ for any $\boldsymbol{X}^* \in \chi^\infty$, as desired. This completes the proof. ■

4. Convergence Analysis Under KL Property

In this section, we analyze the convergence properties of the sequence $\{(\boldsymbol{X}^k, \boldsymbol{Y}^k, \boldsymbol{\epsilon}^k)\}$ generated by Algorithm 1 under the KL property of F for sufficiently large k. We first define a reduced form of (3.1) as

$$\hat{H}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\delta}) = f(\boldsymbol{X}) + \sum_{i=1}^{r^*} \left(\sigma_i(\boldsymbol{X}) + (\delta_i)^2\right)^p + \frac{\beta}{2}\|\boldsymbol{X} - \boldsymbol{Y}\|_F^2, \tag{4.1}$$

where $\epsilon_i = \delta_i^2$ with $\delta_i \geq 0$ since ϵ_i is restricted to be non-negative. Notice that by Theorem 7, $\delta_i^k \equiv \delta_i^{\hat{k}}, i \in \mathcal{Z}^*$ and $\delta_i^{k+1} = \sqrt{\mu}\delta_i^k, i \in \mathcal{I}^*$ for all sufficiently large k. We consider the Cartesian product of triplets $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{z}) \in \mathbb{R}^{m\times n} \times \mathbb{R}^{m\times n} \times \mathbb{R}^{r^*}$.

Definition 7 *Consider a set* $\mathbb{S} = \{(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{z}) \mid (\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{z}) \in \mathbb{R}^{m\times n} \times \mathbb{R}^{m\times n} \times \mathbb{R}^{r^*}\}$. *Define the Cartesian product of any* $\boldsymbol{X} = (X_1, X_2, \boldsymbol{x_3}) \in \mathbb{S}$ *and* $\boldsymbol{Y} = (Y_1, Y_2, \boldsymbol{y_3}) \in \mathbb{S}$ *as*

$$\boldsymbol{X} \times \boldsymbol{Y} = \langle X_1, Y_1\rangle + \langle X_2, Y_2\rangle + \langle \boldsymbol{x_3}, \boldsymbol{y_3}\rangle.$$

The norm of any $\boldsymbol{X} \in \mathbb{S}$ *is defined as*

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

and thus the distance between $\boldsymbol{X} \in \mathbb{S}$ *and* $\boldsymbol{Y} \in \mathbb{S}$ *is*

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

We assume that the uniform KL property holds for $\hat{H}$ in this Cartesian product space.

Assumption 3 *Suppose $\hat{H} : \mathbb{S} \to \mathbb{R}$ satisfies the uniform KL property on $\Omega := \{(\boldsymbol{X}^*, \boldsymbol{X}^*, \boldsymbol{0}_{r^*}) \mid \boldsymbol{X}^* \in \mathrm{crit}(F)\}$.*

Remark 12 *The assumption that $\hat{H}$ satisfies the uniform KL property at any point of Ω is generally not stronger than the assumption that it satisfies the KL property. The difference between the uniform KL and the original KL property is whether the cluster points share the same parameters c and θ in the desingularizing function $\Phi = cs^{1-\theta}$ with $c > 0$ and $\theta \in [0, 1)$. It should be noted that Theorem 11(i) shows the set of cluster points χ^∞ is nonempty, compact, and connected. If there are different c and θ for each $\boldsymbol{X}^* \in \chi^\infty$, we can choose the smallest c and the largest θ among them. This ensures that the KL inequality holds uniformly $dist(\boldsymbol{0}, \partial\hat{H}(\boldsymbol{X})) \geq c(\hat{H}(\boldsymbol{X}) - \hat{H}(\boldsymbol{X}^*))^\theta$ for any $\boldsymbol{X}^* \in \chi^\infty$ whenever $\boldsymbol{X}$ is sufficiently close to $\boldsymbol{X}^*$. Additionally, the equivalence of the KL exponents of the potential function $\hat{H}$ and the objective function F is studied in (Li and Pong, 2018, Theorem 3.6). The equivalence of the KL exponents of F on the entire space and F on an active manifold is studied in (Li and Pong, 2018, Theorem 3.7). Therefore, Assumption 3 can be considered reasonable and mild.*

Now we prove the convergence properties of $\{(\boldsymbol{X}^k, \boldsymbol{Y}^k, \boldsymbol{\delta}^k)\}$ using KL property.

Lemma 13 (Uniqueness of convergence properties under KL condition) *Suppose Assumptions 1-3 hold. Let $\{\boldsymbol{X}^k\}$ be the sequence generated by Algorithm 1. Then there exists $\hat{k} \in \mathbb{N}$ such that the following statements hold.*

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

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

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

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

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

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

(iv) For any $t > 0$, $T^t := \sum_{k=t}^{+\infty} \|\boldsymbol{X}^{k-1} - \boldsymbol{X}^k\|_F < +\infty$. Therefore, $\lim_{k\to+\infty} \boldsymbol{X}^k = \boldsymbol{X}^*$.

Proof Since $\hat{H}$ is differentiable with respect to its all input variables $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\delta})$ separately, we have

$$\begin{aligned}
\nabla_{\boldsymbol{X}}\hat{H}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\delta}) &= \nabla f(\boldsymbol{X}) + \beta(\boldsymbol{X} - \boldsymbol{Y}) + \lambda\boldsymbol{U}\mathrm{diag}(\hat{\boldsymbol{w}})\boldsymbol{V}^\top, \\
\nabla_{\boldsymbol{Y}}\hat{H}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\delta}) &= \beta(\boldsymbol{X} - \boldsymbol{Y}), \\
\nabla_{\delta_i}\hat{H}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\delta}) &= 2\lambda p\delta_i(\sigma_i(\boldsymbol{X}) + \delta_i^2)^{p-1}, \ \forall i \in [r^*],
\end{aligned} \tag{4.2}$$

where $\hat{\boldsymbol{w}} = (p(\sigma_1(\boldsymbol{X}) + (\delta_1)^2)^{p-1}, \cdots, p(\sigma_{r^*}(\boldsymbol{X}) + (\delta_{r^*})^2)^{p-1}, 0, \cdots, 0)^\top \in \mathbb{R}^m$ by Proposition 1.

For each $k > \hat{k}$, we have from (2.11) and Proposition 6(ii)-(iii) that

$$\mathbf{0} = \nabla f(\boldsymbol{Y}^k) + \beta(\boldsymbol{X}^{k+1} - \boldsymbol{Y}^k) + \beta(\boldsymbol{X}^{k+1} - \boldsymbol{X}^k) + \lambda \boldsymbol{U}^{k+1} \mathrm{diag}\left(\hat{\boldsymbol{w}}^k\right) \boldsymbol{V}^{k+1^\top}, \tag{4.3}$$

where $\hat{\boldsymbol{w}}^k = (p(\sigma_1(\boldsymbol{X}^k) + (\delta_1^k)^2)^{p-1}, \cdots, p(\sigma_{r^*}(\boldsymbol{X}^k) + (\delta_{r^*}^k)^2)^{p-1}, 0, \cdots, 0)^\top$.

(i) Combining the first expression in (4.2) and (4.3), we have

$$\begin{aligned} &\nabla_{\boldsymbol{X}} \hat{H}(\boldsymbol{X}^{k+1}, \boldsymbol{X}^k, \boldsymbol{\delta}^{k+1}) \\ =&\nabla f(\boldsymbol{X}^{k+1}) - \nabla f(\boldsymbol{Y}^k) - \beta(\boldsymbol{X}^{k+1} - \boldsymbol{Y}^k) + \lambda \boldsymbol{U}^{k+1} \mathrm{diag}\left(\hat{\boldsymbol{w}}^{k+1} - \hat{\boldsymbol{w}}^k\right) \boldsymbol{V}^{k+1^\top}. \end{aligned} \tag{4.4}$$

It follows from (2.2) that

$$\|\beta(\boldsymbol{X}^k - \boldsymbol{Y}^k)\|_F \le \beta\bar{\alpha}\|\boldsymbol{X}^k - \boldsymbol{X}^{k-1}\|_F. \tag{4.5}$$

For any $k \ge \hat{k}$, we then have

$$\begin{aligned} &\|\boldsymbol{U}^{k+1} \mathrm{diag}(\hat{\boldsymbol{w}}^{k+1} - \hat{\boldsymbol{w}}^k) \boldsymbol{V}^{k+1^\top}\|_F \\ \le& \|\mathrm{diag}(\hat{\boldsymbol{w}}^{k+1} - \hat{\boldsymbol{w}}^k)\|_F = \|\hat{\boldsymbol{w}}^{k+1} - \hat{\boldsymbol{w}}^k\|_2 \le \|\hat{\boldsymbol{w}}^{k+1} - \hat{\boldsymbol{w}}^k\|_1 \\ \le& \sum_{i=1}^{r^*} p(1-p)(\sigma_i(\boldsymbol{X}^k))^{p-2}(|\sigma_i(\boldsymbol{X}^k) - \sigma_i(\boldsymbol{X}^{k+1})| + |(\delta_i^k)^2 - (\delta_i^{k+1})^2|) \\ \overset{(a)}{\le}& \sum_{i=1}^{r^*} p(1-p)\left(\frac{2\beta C_1}{\lambda p}\right)^{\frac{2-p}{1-p}} \left(\|\boldsymbol{X}^k - \boldsymbol{X}^{k+1}\|_F + |(\delta_i^k)^2 - (\delta_i^{k+1})^2|\right), \\ \le& D_p \left(\|\boldsymbol{X}^k - \boldsymbol{X}^{k+1}\|_F + \max_i(\delta_i^k + \delta_i^{k+1})\|\boldsymbol{\delta}^k - \boldsymbol{\delta}^{k+1}\|_1\right) \\ \le& D_p \left[\|\boldsymbol{X}^k - \boldsymbol{X}^{k+1}\|_F + 2\|\boldsymbol{\delta}^0\|_\infty(\|\boldsymbol{\delta}^k\|_1 - \|\boldsymbol{\delta}^{k+1}\|_1)\right], \end{aligned} \tag{4.6}$$

where $D_p = p(1-p)\left(\frac{2\beta C_1}{\lambda p}\right)^{\frac{2-p}{1-p}}$, and inequality (a) holds by Proposition 6(iii) and conclusions drawn by (Horn and Johnson, 2012, Corollary 7.3.5). This, together with (4.4), (3.18), (4.5) and (4.6), yields

$$\begin{aligned} &\| \nabla_{\boldsymbol{X}} \hat{H}(\boldsymbol{X}^{k+1}, \boldsymbol{X}^k, \boldsymbol{\delta}^{k+1})\|_F \\ \le& \|\nabla f(\boldsymbol{X}^{k+1}) - f(\boldsymbol{Y}^k)\|_F + \beta\|\boldsymbol{X}^k - \boldsymbol{Y}^k\|_F + \lambda\|\boldsymbol{U}^{k+1}\mathrm{diag}(\hat{\boldsymbol{w}}^{k+1} - \hat{\boldsymbol{w}}^k)\boldsymbol{V}^{k+1^\top}\|_F \\ \le& (\lambda D_p + L_f)\|\boldsymbol{X}^{k+1} - \boldsymbol{X}^k\|_F + (L_f + \beta)\bar{\alpha}\|\boldsymbol{X}^k - \boldsymbol{X}^{k-1}\|_F \\ &+ 2\lambda D_p\|\boldsymbol{\delta}^0\|_\infty(\|\boldsymbol{\delta}^k\|_1 - \|\boldsymbol{\delta}^{k+1}\|_1). \end{aligned} \tag{4.7}$$

Similarly, we have

$$\|\nabla_{\boldsymbol{Y}} \hat{H}(\boldsymbol{X}^{k+1}, \boldsymbol{X}^k, \boldsymbol{\delta}^{k+1})\|_F = \|\beta(\boldsymbol{X}^{k+1} - \boldsymbol{X}^k)\|_F \le \beta\|\boldsymbol{X}^{k+1} - \boldsymbol{X}^k\|_F, \tag{4.8}$$

and note that

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

It then follows that

$$\begin{aligned}\|\nabla_{\boldsymbol{\delta}}\hat{H}(\boldsymbol{X}^{k+1},\boldsymbol{X}^k,\boldsymbol{\delta}^{k+1})\|_2 \le \|\nabla_{\boldsymbol{\delta}}\hat{H}(\boldsymbol{X}^{k+1},\boldsymbol{X}^k,\boldsymbol{\delta}^{k+1})\|_1 &= \sum_{i=1}^{r^*} 2\lambda\hat{w}_i^{k+1}\delta_i^{k+1} \\ &\overset{(a)}{\le} \sum_{i=1}^{r^*} 2\lambda(\tfrac{2\beta C_1}{\lambda})\tfrac{\sqrt{\mu}}{1-\sqrt{\mu}}\left(\delta_i^k-\delta_i^{k+1}\right) \\ &\le \tfrac{4\beta C_1\sqrt{\mu}}{1-\sqrt{\mu}}\left(\|\boldsymbol{\delta}^k\|_1-\|\boldsymbol{\delta}^{k+1}\|_1\right),\end{aligned} \tag{4.9}$$

where inequality (a) is due to Proposition 6 (i) and $\delta_i^{k+1}\le\sqrt{\mu}\delta_i^k, i\in[r^*]$.

Therefore, combining (4.7), (4.8) and (4.9) gives

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

with $D=\max\left(\lambda D_p+L_f+\beta,(L_f+\beta)\bar{\alpha},2D_p\lambda\|\boldsymbol{\delta}^0\|_\infty+\frac{4\beta C_1\sqrt{\mu}}{1-\sqrt{\mu}}\right)<+\infty$. Furthermore, by Lemma 5 and Theorem 7, we know that $\lim_{k\to+\infty}\|\nabla\hat{H}(\boldsymbol{X}^{k+1},\boldsymbol{X}^k,\boldsymbol{\delta}^{k+1})\|_F=0$. This completes the proof of statement (i).

(ii) It is straightforward from $C=\frac{\beta}{2}\left(1-\frac{3L_f+\beta}{\beta}\bar{\alpha}^2\right)>0$ by Lemma 5(i).

(iii) Theorem 11(ii) shows $H(\boldsymbol{X}^{k+1},\boldsymbol{X}^k,\boldsymbol{\delta}^{k+1})\to\zeta$ as $k\to+\infty$ and it follows from Lemma 5(iii) and Theorem 11(i)-(ii) that $\{\hat{H}(\boldsymbol{X}^{k+1},\boldsymbol{Y}^k,\boldsymbol{\delta}^{k+1})\}$ uniquely converges.

(iv) By statement (iii), we know that

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

If $\hat{H}(\boldsymbol{X}^{k+1},\boldsymbol{X}^k,\boldsymbol{\delta}^{k+1})=\zeta$ for each $k>\hat{k}$, then we know $\boldsymbol{X}^{k+1}=\boldsymbol{X}^k$ by statement (ii) after $\hat{k}$, indicating $\boldsymbol{X}^k=\boldsymbol{X}^{\hat{k}}\in\chi^\infty$. The proof is then complete. We otherwise have to consider the case in which $\hat{H}(\boldsymbol{X}^{k+1},\boldsymbol{X}^k,\boldsymbol{\delta}^{k+1})>\zeta$ after $\hat{k}$.

By Theorem 3 and Assumption 3, we know that there exists a desingularizing function Φ, $\eta>0$ and $\rho>0$ such that

$$\Phi'\left(\hat{H}(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\delta})-\zeta\right)\|\nabla\hat{H}(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\delta})\|\ge 1, \tag{4.10}$$

for all $(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\delta})\in\mathbb{U}((\boldsymbol{X}^*,\boldsymbol{X}^*,\boldsymbol{0});\rho)\cap\left\{(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\delta})\in\mathbb{S}:\zeta<\hat{H}(\boldsymbol{X}^{k+1},\boldsymbol{X}^k,\boldsymbol{\delta}^{k+1})<\zeta+\eta\right\}$. Note that $\boldsymbol{X}^*\in\chi^\infty$, we have $\lim_{k\to+\infty}\mathrm{dist}((\boldsymbol{X}^k,\boldsymbol{X}^{k-1},\boldsymbol{\delta}^k),\Gamma)=0$, meaning $\forall\rho>0$, there exists $k_1\in\mathbb{N}$ such that $\mathrm{dist}((\boldsymbol{X}^k,\boldsymbol{X}^{k-1},\boldsymbol{\delta}^k),\Gamma)<\rho$ for all $k>k_1$. Since $\{\hat{H}(\boldsymbol{X}^{k+1},\boldsymbol{X}^k,\boldsymbol{\delta}^{k+1})\}\to\zeta$ as $k\to+\infty$, we know that there exists $k_2\in\mathbb{N}$ such that $\zeta<\hat{H}(\boldsymbol{X}^{k+1},\boldsymbol{X}^k,\boldsymbol{\delta}^{k+1})<\zeta+\eta$ for all $k>k_2$. Thus, we see from the smoothness of $\hat{H}$ that for any $k>\max(k_1,k_2)$,

$$\Phi'\left(\hat{H}(\boldsymbol{X}^{k+1},\boldsymbol{X}^k,\boldsymbol{\delta}^{k+1})-\zeta\right)\|\nabla\hat{H}(\boldsymbol{X}^{k+1},\boldsymbol{X}^k,\boldsymbol{\delta}^{k+1})\|\ge 1. \tag{4.11}$$

Then, we have for any $k > \hat{k}$ that

$$\begin{aligned}
&\left[\Phi\left(\hat{H}(\boldsymbol{X}^k, \boldsymbol{X}^{k-1}, \boldsymbol{\delta}^k) - \zeta\right) - \Phi\left(\hat{H}(\boldsymbol{X}^{k+1}, \boldsymbol{X}^k, \boldsymbol{\delta}^{k+1}) - \zeta\right)\right] \\
&\times D\left(\|\boldsymbol{X}^{k-2} - \boldsymbol{X}^{k-1}\|_F + \|\boldsymbol{X}^{k-1} - \boldsymbol{X}^k\|_F + \|\boldsymbol{\delta}^{k-1}\|_1 - \|\boldsymbol{\delta}^k\|_1\right) \\
\overset{(a)}{\geq} &\left[\Phi\left(\hat{H}(\boldsymbol{X}^k, \boldsymbol{X}^{k-1}, \boldsymbol{\delta}^k) - \zeta\right) - \Phi\left(\hat{H}(\boldsymbol{X}^{k+1}, \boldsymbol{X}^k, \boldsymbol{\delta}^{k+1}) - \zeta\right)\right] \\
&\times \|\nabla \hat{H}(\boldsymbol{X}^k, \boldsymbol{X}^{k-1}, \boldsymbol{\delta}^k)\| \\
\overset{(b)}{\geq} &\left[\hat{H}(\boldsymbol{X}^k, \boldsymbol{X}^{k-1}, \boldsymbol{\delta}^k) - \hat{H}(\boldsymbol{X}^{k+1}, \boldsymbol{X}^k, \boldsymbol{\delta}^{k+1})\right] \Phi'\left(\hat{H}(\boldsymbol{X}^k, \boldsymbol{X}^{k-1}, \boldsymbol{\delta}^k) - \zeta\right) \\
&\times \|\nabla \hat{H}(\boldsymbol{X}^k, \boldsymbol{X}^{k-1}, \boldsymbol{\delta}^k)\| \overset{(c)}{\geq} C\|\boldsymbol{X}^k - \boldsymbol{X}^{k-1}\|_F^2,
\end{aligned} \tag{4.12}$$

where the inequality (a) holds by Lemma 13(i), the inequality (b) makes use of the concavity of Φ and inequality (c) follows from (4.10) and Lemma 13(ii). Therefore,

$$\begin{aligned}
&\|\boldsymbol{X}^k - \boldsymbol{X}^{k-1}\|_F \\
\overset{(a)}{\leq} &\sqrt{\frac{2D}{C}\left[\Phi\left(\hat{H}(\boldsymbol{X}^k, \boldsymbol{X}^{k-1}, \boldsymbol{\delta}^k) - \zeta\right) - \Phi\left(\hat{H}(\boldsymbol{X}^{k+1}, \boldsymbol{X}^k, \boldsymbol{\delta}^{k+1}) - \zeta\right)\right]} \\
&\times \sqrt{\frac{1}{2}\left(\|\boldsymbol{X}^{k-2} - \boldsymbol{X}^{k-1}\|_F + \|\boldsymbol{X}^{k-1} - \boldsymbol{X}^k\|_F + \|\boldsymbol{\delta}^{k-1}\|_1 - \|\boldsymbol{\delta}^k\|_1\right)} \\
\overset{(b)}{\leq} &\frac{D}{C}\left[\Phi\left(\hat{H}(\boldsymbol{X}^k, \boldsymbol{X}^{k-1}, \boldsymbol{\delta}^k) - \zeta\right) - \Phi\left(\hat{H}(\boldsymbol{X}^{k+1}, \boldsymbol{X}^k, \boldsymbol{\delta}^{k+1}) - \zeta\right)\right] \\
&+ \frac{1}{4}\left(\left\|\boldsymbol{X}^{k-2} - \boldsymbol{X}^{k-1}\right\|_F + \left\|\boldsymbol{X}^{k-1} - \boldsymbol{X}^k\right\|_F + \|\boldsymbol{\delta}^{k-1}\|_1 - \|\boldsymbol{\delta}^k\|_1\right),
\end{aligned} \tag{4.13}$$

where the inequality (a) holds by (4.12) and the inequality (b) is true because of the AM-GM inequality. Then, subtracting $\frac{1}{2}\|\boldsymbol{X}^k - \boldsymbol{X}^{k-1}\|_F$ from both sides of (4.13), we obtain

$$\begin{aligned}
\frac{1}{2}\|\boldsymbol{X}^k - \boldsymbol{X}^{k-1}\|_F \leq &\frac{D}{C}\left[\Phi\left(\hat{H}(\boldsymbol{X}^k, \boldsymbol{X}^{k-1}, \boldsymbol{\delta}^k) - \zeta\right) - \Phi\left(\hat{H}(\boldsymbol{X}^{k+1}, \boldsymbol{X}^k, \boldsymbol{\delta}^{k+1}) - \zeta\right)\right] \\
&+ \frac{1}{4}\left(\|\boldsymbol{X}^{k-2} - \boldsymbol{X}^{k-1}\|_F - \|\boldsymbol{X}^{k-1} - \boldsymbol{X}^k\|_F + \|\boldsymbol{\delta}^{k-1}\|_1 - \|\boldsymbol{\delta}^k\|_1\right).
\end{aligned} \tag{4.14}$$

Summing up both sides of (4.14) from $l = t$ to k, we have

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

We then have $\boldsymbol{\delta}^k \to 0$ and $\|\boldsymbol{X}^k - \boldsymbol{X}^{k-1}\|_F \to 0$ by taking $k \to +\infty$ according to Lemma 5 and Theorem 7(ii), respectively. Furthermore, we find, by taking $k \to +\infty$ and the continuity of Φ that $\Phi\left(\hat{H}(\boldsymbol{X}^{k+1}, \boldsymbol{X}^k, \boldsymbol{\delta}^{k+1}) - \zeta\right) \to \Phi(\zeta - \zeta) = \Phi(0) = 0$ by Definition 5(i). Therefore,

for any $t > 0$, it holds that

$$\begin{aligned} T^t &= \sum_{l=t}^{+\infty} \|\boldsymbol{X}^l - \boldsymbol{X}^{l-1}\|_F \\ &\leq \frac{2D}{C}\Phi\left(\hat{H}(\boldsymbol{X}^t, \boldsymbol{X}^{t-1}, \boldsymbol{\delta}^t) - \zeta\right) + \frac{1}{2}\left(\|\boldsymbol{X}^{t-2} - \boldsymbol{X}^{t-1}\|_F + \|\boldsymbol{\delta}^{t-1}\|_1\right) < +\infty. \end{aligned} \tag{4.15}$$

On the other hand, $\forall i > j > l$, we know that $\boldsymbol{X}^i - \boldsymbol{X}^j = \sum_{k=j}^{i-1}(\boldsymbol{X}^{k+1} - \boldsymbol{X}^k)$. Then

$$\|\boldsymbol{X}^i - \boldsymbol{X}^j\|_F = \left\|\sum_{k=j}^{i-1}(\boldsymbol{X}^{k+1} - \boldsymbol{X}^k)\right\|_F \overset{(a)}{\leq} \sum_{k=j}^{i-1}\|\boldsymbol{X}^{k+1} - \boldsymbol{X}^k\|_F, \tag{4.16}$$

where the inequality (a) holds by the Triangle inequality. This, together with (4.15), implies that $\{\boldsymbol{X}^k\}$ is a Cauchy sequence and hence converges uniquely. The proof is complete. ■

Now we are ready to prove the convergence rate under the KL property. We mention that the ideas for the proof much follow those presented in (Attouch and Bolte, 2009; Wen et al., 2018).

Theorem 14 (Local convergence rate) *Let Assumptions 1-3 hold. Let $\{\boldsymbol{X}^k\}$ be generated by Algorithm 1 and converge to a critical point $\boldsymbol{X}^* \in \mathrm{crit}(F)$. Consider a desingularizing function of the form $\Phi(s) = cs^{1-\theta}$ where $c > 0$ and Łojasiewicz exponent $\theta \in [0, 1)$. Then the following statements hold.*

(i) If $\theta = 0$, then there exists $\hat{k} \in \mathbb{N}$ such that $\boldsymbol{X}^k \equiv \boldsymbol{X}^*$ for all $k > \hat{k}$.

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

$$\|\boldsymbol{X}^k - \boldsymbol{X}^*\|_F \leq c_0\gamma^k - c_1\|\boldsymbol{\delta}^k_{\mathcal{I}^*}\|_1 \tag{4.17}$$

for all sufficiently large k, where $c_0 = \frac{T^{\hat{k}} + \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^{\hat{k}}\|_1}{\gamma^{\hat{k}+1}}$ and $c_1 = \frac{\sqrt{\mu}}{1-\mu}$ with

$$\gamma = \sqrt{\frac{\nu_1 + \nu_2}{\nu_1 + \nu_2 + 1}},\ \nu_1 = \frac{2cD}{C}[cD(1-\theta)]^{\frac{1-\theta}{\theta}} > 0,\ \nu_2 = \frac{1}{2} + \frac{\mu}{1-\mu} > 0,$$

$D = \max\left(\lambda D_p + L_f + \beta, (L_f + \beta)\bar{\alpha}, 2D_p\lambda\|\boldsymbol{\delta}^0\|_\infty + \frac{4\beta C_1\sqrt{\mu}}{1-\sqrt{\mu}}\right) < +\infty$ and $D_p = p(1-p)\left(\frac{2\beta C_1}{\lambda p}\right)^{\frac{2-p}{1-p}}$.

(iii) If $\theta \in (\frac{1}{2}, 1)$, there exist $d_0, c_1 > 0$ such that

$$\|\boldsymbol{X}^k - \boldsymbol{X}^*\|_F \leq d_0 k^{-\frac{1-\theta}{2\theta-1}} - c_1\|\boldsymbol{\delta}^k_{\mathcal{I}^*}\|_1 \tag{4.18}$$

for all sufficiently large k, where $d_0 = 2^{\frac{1-\theta}{2\theta-1}}d_2\max\left(1, 2^{\frac{1-\theta}{2\theta-1}}\right)$ for some finite positive scalar d_2.

Proof Since $\boldsymbol{X}^k \to \boldsymbol{X}^*$ as $k \to +\infty$, we have from Lemma 13(iv) and (4.15) that

$$\|\boldsymbol{X}^k - \boldsymbol{X}^*\|_F = \|\boldsymbol{X}^k - \lim_{t\to+\infty} \boldsymbol{X}^t\|_F = \|\lim_{t\to+\infty} \sum_{l=k}^{t} (\boldsymbol{X}^l - \boldsymbol{X}^{l+1})\|_F \le T^k, \tag{4.19}$$

and

$$T^k \le \frac{2D}{C}\Phi\left(\hat{H}(\boldsymbol{X}^k, \boldsymbol{X}^{k-1}, \boldsymbol{\delta}^k) - \zeta\right) + \frac{1}{2}(T^{k-2} - T^{k-1}) + \frac{1}{2}\|\boldsymbol{\delta}^{k-1}\|_1. \tag{4.20}$$

(i) If $\theta = 0$, then $\Phi(s) = cs$ and $\Phi'(s) = c$. we claim that there exists $\hat{k} \in \mathbb{N}$ such that $\hat{H}(\boldsymbol{X}^{\hat{k}}, \boldsymbol{X}^{\hat{k}-1}, \boldsymbol{\delta}^{\hat{k}}) = \zeta$. Seeking a contradiction, suppose this is not true, so $\hat{H}(\boldsymbol{X}^k, \boldsymbol{X}^{k-1}, \boldsymbol{\delta}^k) > \zeta$ for all $k \in \mathbb{N}$. Since $\lim_{k\to+\infty} \boldsymbol{X}^k = \boldsymbol{X}^*$ and $\{\hat{H}(\boldsymbol{X}^k, \boldsymbol{X}^{k-1}, \boldsymbol{\delta}^k)\}$ monotonically decrease to ζ by Lemma 13(ii). We have from the KL inequality that for all sufficiently large k

$$\|\nabla \hat{H}(\boldsymbol{X}^k, \boldsymbol{X}^{k-1}, \boldsymbol{\delta}^k)\|_F \ge \frac{1}{c} > 0 \tag{4.21}$$

with $\Phi'(s) = c$. This is a contradiction with $\|\nabla \hat{H}(\boldsymbol{X}^k, \boldsymbol{X}^{k-1}, \boldsymbol{\delta}^k)\| \to 0$ by Lemma 13(i). Thus, there exists $\hat{k} \in \mathbb{N}$ such that $\hat{H}(\boldsymbol{X}^k, \boldsymbol{X}^{k-1}, \boldsymbol{\delta}^k) = \hat{H}(\boldsymbol{X}^{\hat{k}}, \boldsymbol{X}^{\hat{k}-1}, \boldsymbol{\delta}^{\hat{k}}) \equiv \zeta$ for all $k \ge \hat{k}$. Hence, we conclude from Lemma 13(ii) that $\boldsymbol{X}^k = \boldsymbol{X}^* = \boldsymbol{X}^{\hat{k}}$ for all $k > \hat{k}$, i.e., the sequence converges in a finite number of iterations. This proves statement (i).

(ii)-(iii) If $\theta \in (0, 1)$, then $\Phi'(s) = c(1-\theta)s^{-\theta}$. If there exists $\hat{k} \in \mathbb{N}$ such that $\hat{H}(\boldsymbol{X}^{\hat{k}}, \boldsymbol{X}^{\hat{k}-1}, \boldsymbol{\delta}^{\hat{k}}) = \zeta$, then we know from Lemma 13(ii) that $\boldsymbol{X}^{k+1} = \boldsymbol{X}^k$ for all $k > \hat{k}$, indicating $\boldsymbol{X}^k \equiv \boldsymbol{X}^{\hat{k}} \in \chi^\infty$. Therefore, we need only to consider the case in which $\hat{H}(\boldsymbol{X}^k, \boldsymbol{X}^{k-1}, \boldsymbol{\delta}^k) > \zeta$ for all $k \in \mathbb{N}$.

Note that Assumption 3 implies that

$$c(1-\theta)(\hat{H}(\boldsymbol{X}^k, \boldsymbol{X}^{k-1}, \boldsymbol{\delta}^k) - \zeta)^{-\theta}\|\nabla \hat{H}(\boldsymbol{X}^k, \boldsymbol{X}^{k-1}, \boldsymbol{\delta}^k)\|_F \ge 1, \tag{4.22}$$

for all $k > \hat{k}$ from Lemma 13(iv). On the other hand, we obtain from Lemma 13(i) that

$$\|\nabla \hat{H}(\boldsymbol{X}^k, \boldsymbol{X}^{k-1}, \boldsymbol{\delta}^k)\|_F \le D\left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1 - \|\boldsymbol{\delta}^k\|_1\right). \tag{4.23}$$

This, together with (4.22), yields

$$(\hat{H}(\boldsymbol{X}^k, \boldsymbol{X}^{k-1}, \boldsymbol{\delta}^k) - \zeta)^\theta \le cD(1-\theta)\left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1 - \|\boldsymbol{\delta}^k\|_1\right). \tag{4.24}$$

Taking a power of $\frac{1-\theta}{\theta}$ to both sides of (4.24), we have for all $k \ge \hat{k}$ that

$$\begin{aligned}\Phi(\hat{H}(\boldsymbol{X}^k, \boldsymbol{X}^{k-1}, \boldsymbol{\delta}^k) - \zeta) &= c(\hat{H}(\boldsymbol{X}^k, \boldsymbol{X}^{k-1}, \boldsymbol{\delta}^k) - \zeta)^{1-\theta} \\ &\le c\left[cD(1-\theta)\left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1 - \|\boldsymbol{\delta}^k\|_1\right)\right]^{\frac{1-\theta}{\theta}} \\ &\le c\left[cD(1-\theta)\left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1\right)\right]^{\frac{1-\theta}{\theta}}.\end{aligned} \tag{4.25}$$

This, together with (4.20), yields

$$T^k \le \nu_1\left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1\right)^{\frac{1-\theta}{\theta}} + \frac{1}{2}\left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1\right) \tag{4.26}$$

with $\nu_1 = \frac{2cD}{C}[cD(1-\theta)]^{\frac{1-\theta}{\theta}} > 0$.

It then follows from (4.26) that

$$\begin{aligned}
&T^k + \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^k\|_1 \\
&\leq \nu_1 \left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1\right)^{\frac{1-\theta}{\theta}} + \frac{1}{2}\left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1\right) + \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^k\|_1 \\
&\overset{(a)}{\leq} \nu_1 \left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1\right)^{\frac{1-\theta}{\theta}} + \frac{1}{2}\left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1\right) + \frac{\mu}{1-\mu}\|\boldsymbol{\delta}^{k-1}\|_1 \\
&\overset{(b)}{\leq} \nu_1 \left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1\right)^{\frac{1-\theta}{\theta}} + \nu_2 \left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1\right),
\end{aligned} \tag{4.27}$$

where inequality (a) holds simply due to $\delta_i^k \leq \sqrt{\mu}\delta_i^{k-1}, i \in [r^*]$ and inequality (b) with $\nu_2 = \frac{1}{2} + \frac{\mu}{1-\mu} > 0$ is true because $T^{k-2} - T^k \geq 0$ and $\mu \in (0,1)$.

Now consider $\theta \in (0, \frac{1}{2}]$. It follows from Lemma 5(ii) and Theorem 7(ii) that $\lim_{k\to+\infty} T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1 = 0$. Since $\frac{1-\theta}{\theta} \geq 1$, we thus know for sufficiently large k that

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

This, together with (4.27), yields

$$T^k + \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^k\|_1 \leq (\nu_1 + \nu_2)\left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1\right) \tag{4.28}$$

for sufficiently large k.

On the other hand, by Theorem 7(ii), we have $\delta_i^k \leq \sqrt{\mu}\delta_i^{k-1}$ and $\delta_i^{k-1} \leq \sqrt{\mu}\delta_i^{k-2}$, $i \in [r^*]$, which implies that

$$\delta_i^{k-1} \leq \frac{\sqrt{\mu}}{1-\mu}(\delta_i^{k-2} - \delta_i^k),\ i \in [r^*], \tag{4.29}$$

leading to

$$\|\boldsymbol{\delta}^{k-1}\|_1 \leq \frac{\sqrt{\mu}}{1-\mu}(\|\boldsymbol{\delta}^{k-2}\|_1 - \|\boldsymbol{\delta}^k\|_1). \tag{4.30}$$

Therefore, we have from (4.28) that for any $k \geq \hat{k}$

$$T^k + \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^k\|_1 \overset{(a)}{\leq} (\nu_1 + \nu_2)\left(T^{k-2} + \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^{k-2}\|_1 - \left(T^k + \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^k\|_1\right)\right) \tag{4.31}$$

where inequality (a) holds by (4.30). This implies

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

We hence have for any $k \geq \hat{k}$ that

$$\|\boldsymbol{X}^k - \boldsymbol{X}^*\|_F \leq T^k \leq c_0\gamma^k - c_1\|\boldsymbol{\delta}^k\|_1 \tag{4.32}$$

with

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

This completes the proof of statement (ii).

Consider now $\theta \in (\frac{1}{2}, 1)$. We have from $\frac{1-\theta}{\theta} < 1$ and $\lim_{k\to+\infty} T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1 = 0$ for sufficiently large k that

$$\left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1\right) \leq \left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1\right)^{\frac{1-\theta}{\theta}}. \tag{4.33}$$

This, together with (4.27), gives

$$T^k + \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^k\|_1 \leq (\nu_1+\nu_2)\left(T^{k-2} - T^k + \|\boldsymbol{\delta}^{k-1}\|_1\right)^{\frac{1-\theta}{\theta}}. \tag{4.34}$$

Raising a power of $\frac{\theta}{1-\theta}$ to the both sides of (4.34) and considering (4.30) gives

$$\left(T^k + \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^k\|_1\right)^{\frac{\theta}{1-\theta}} \leq \nu_3\left[(T^{k-2} + \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^{k-2}\|_1) - (T^k + \|\boldsymbol{\delta}^k\|_1)\right], \tag{4.35}$$

where $\nu_3 = (\nu_1 + v_2)^{\frac{\theta}{1-\theta}}$.

Split the sequence $\{k_2, k_2+1, \cdots\}$ into even and odd subsequences. For the even subsequnce, define $\Delta_t := T^{2t} + \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^{2t}\|_1$ for $t \geq \lceil \frac{\hat{k}}{2} \rceil := N_1$. Following from the techniques presented in the proofs of (Wang et al., 2022, Theorem 4) and (Attouch and Bolte, 2009, Theorem 2), we have

$$\Delta_k \leq \left(\Delta_{N_1-1}^{\frac{1-2\theta}{1-\theta}} + \nu(k-N_1)\right)^{-\frac{1-\theta}{2\theta-1}} \leq d_2 k^{-\frac{1-\theta}{2\theta-1}}, \tag{4.36}$$

for some $d_2 > 0$. As for the odd subsequence of $\{k_2, k_2+1, \cdots\}$, define $\Delta_t := T^{2t+1} + \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^{2t+1}\|_1$. We know that (4.36) still holds. Therefore, for all sufficiently large and even number k, it holds that

$$\|\boldsymbol{X}^k - \boldsymbol{X}^*\|_F \leq T^k = \Delta_{\frac{k}{2}} - \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^k\|_1 \leq 2^{\frac{1-\theta}{2\theta-1}} d_2 k^{-\frac{1-\theta}{2\theta-1}} - \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^k\|_1. \tag{4.37}$$

For all sufficiently large and odd numbers k,

$$\|\boldsymbol{X}^k - \boldsymbol{X}^*\|_F \leq T^k = \Delta_{\frac{k-1}{2}} - \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^k\|_1 \leq 2^{\frac{1-\theta}{2\theta-1}} d_2 (k-1)^{-\frac{1-\theta}{2\theta-1}} - \frac{\sqrt{\mu}}{1-\mu}\|\boldsymbol{\delta}^k\|_1. \tag{4.38}$$

Overall, we thus have for any sufficiently large k that

$$\|\boldsymbol{X}^k - \boldsymbol{X}^*\|_F \leq d_0 k^{-\frac{1-\theta}{2\theta-1}} - c_1\|\boldsymbol{\delta}^k\|_1, \tag{4.39}$$

where $d_0 = 2^{\frac{1-\theta}{2\theta-1}} d_2 \max\left(1, 2^{\frac{1-\theta}{2\theta-1}}\right)$ and $c_1 = \frac{\sqrt{\mu}}{1-\mu}$. The proof is complete. ■

5. Numerical Experiments

In this section, we conduct low-rank matrix completion tasks using synthetic data and natural color images to demonstrate the effectiveness and efficiency of the proposed EIRNAMI algorithm. Specifically, numerical experiments address the low-rank matrix completion problem, applied with $f(\boldsymbol{X}) = \frac{1}{2}\|\boldsymbol{M} - \mathcal{P}_\Omega(\boldsymbol{X})\|_F^2$, defined as

$$\min_{\boldsymbol{X}\in\mathbb{R}^{m\times n}} \quad F(\boldsymbol{X}) = \frac{1}{2}\|\boldsymbol{M} - \mathcal{P}_\Omega(\boldsymbol{X})\|_F^2 + \lambda\|\boldsymbol{X}\|_p^p, \tag{5.1}$$

where $\boldsymbol{M} = \mathcal{P}_\Omega(\widehat{\boldsymbol{X}})$ is the given incomplete observation and $\widehat{\boldsymbol{X}} \in \mathbb{R}^{m\times n}$ is the original matrix, The set $\Omega \subseteq [m]\times[n]$ is an index set of observed entries, and $\mathcal{P}_\Omega : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ is a linear operator that retains the entries of $\widehat{X}$ in Ω unchanged and sets the entries outside Ω to zero. Here, $\lambda > 0$. The goal of problem (5.1) is to reconstruct the missing entries of a partially observed low-rank matrix $\boldsymbol{M} \in \mathbb{R}^{m\times n}$ from the known entries $\{M_{ij} \mid (i,j) \in \Omega\}$. The Lipschitz constant of f is $L_f = 1$. All methods tested in this section were implemented in MATLAB on a desktop equipped with an Intel(R) Xeon(R) CPU E5-2620 v2 (2.10 GHz) and 64GB RAM, running 64-bit Windows 10 Enterprise.[1]

To evaluate the numerical performance of the algorithms concerned, we follow (Sun et al., 2017) to define the relative error of the original matrix $\widehat{\boldsymbol{X}} \in \mathbb{R}^{m\times n}$ and the recovered matrix $\boldsymbol{X}^* \in \mathbb{R}^{m\times n}$ as $\mathrm{RelErr}(\boldsymbol{X}^*) = \frac{\|\boldsymbol{X}^* - \widehat{\boldsymbol{X}}\|_F}{\|\widehat{\boldsymbol{X}}\|_F}$. Moreover, to verify whether a given point $\boldsymbol{X} \in \mathrm{crit}(F)$ regarding problem (5.1), we adopt the following relative distance error:

$$\mathrm{RelDist}(\boldsymbol{X}) = \frac{\mathrm{dist}(\boldsymbol{0}, \partial F(\boldsymbol{X}))}{\|\boldsymbol{X}\|_F},$$

where $\mathrm{dist}(\boldsymbol{0}, \partial F(\boldsymbol{X})) = \|\mathcal{P}_\Omega(\boldsymbol{M} - \boldsymbol{X}) + \lambda \boldsymbol{U}\mathrm{diag}(\partial\|\boldsymbol{\sigma}(\boldsymbol{X})\|_p^p)\boldsymbol{V}^\top\|_F$ with $(\boldsymbol{U}, \boldsymbol{V}) \in \overline{\mathcal{M}}(\boldsymbol{X})$ according to Definition 4 and Proposition 1. We consider a matrix $\widehat{\boldsymbol{X}}$ is successfully recovered by $\boldsymbol{X}^*$ if the corresponding relative error $\mathrm{RelErr}(\boldsymbol{X}^*)$ or the relative distance $\mathrm{RelDist}(\boldsymbol{X}^*)$ is less than the prespecified tolerance 10^{-5}, similar to the criterion used in (Sun et al., 2017). We define the sampling ratio so that $|\Omega| = \lceil \mathrm{SR} \times (mn) \rceil$, where $\lceil\cdot\rceil$ denotes the ceiling function.

For synthetic data, we adopt $\lambda = 10^{-1}\|\boldsymbol{M}\|_\infty$. For natural color images, we begin by testing the regularization parameter λ with an initial large value $\lambda_0 = 2^8$. Specifically, we determine a candidate by $\lambda = (2^{-4})^k\lambda_0$, where k ranges from 0 to 6. After identifying the candidate λ, we further refine the selection by exploring a finer range of λ values around the chosen λ.

In all experiments, we terminate the proposed algorithm if $\mathrm{RelErr} \le tol_1$ or $\mathrm{RelDist} \le tol_1$ or the number of iterations exceeds the prespecified maximum number of iterations $IterMax = 3\times10^3$. In addition, we also use another termination condition $\|\boldsymbol{X}^{k+1} - \boldsymbol{X}^k\|_\infty \le tol_2$ according to the criterion (3.16).

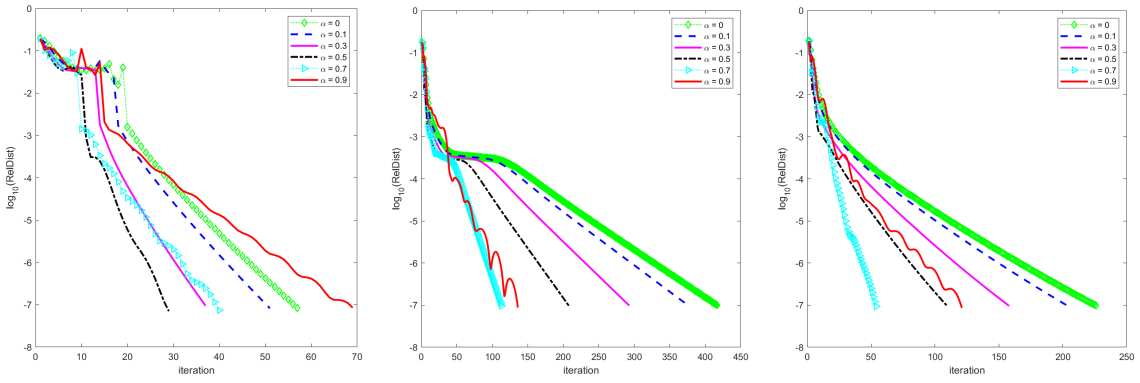
1. We have made the source code publicly available at the GitHub repository `https://github.com/Optimizater/Low-rank-optimization-with-active-manifold-identification`.

5.1 Synthetic Data

In this experiment, we compare the proposed EIRNAMI with the state-of-the-art method PIRNN (Sun et al., 2017) using synthetic data. Furthermore, we include the IRNAMI algorithm that does not use the extrapolation technique presented in (2.2). To empirically evaluate the ability of EIRNAMI to recover the correct rank of the solution, we call a correct low-rank detection (CLD) holds for an algorithm when the rank of the recovered solution matches the rank of the original matrix $\widehat{\boldsymbol{X}}$.

We now specify the experimental setup for data generation. We generate a low-rank matrix $\widehat{\boldsymbol{X}}$ with $\text{Rank}(\widehat{\boldsymbol{X}}) = r^*$, where $\widehat{\boldsymbol{X}} = \boldsymbol{BC}$. Here, $\boldsymbol{B} \in \mathbb{R}^{m\times r^*}$ and $\boldsymbol{C} \in \mathbb{R}^{r^*\times n}$ are generated randomly with i.i.d. standard Gaussian entries. We consider $r^* \in \{5, 10, 15\}$ for the original matrix $\widehat{\boldsymbol{X}}$ in our tests. We then uniformly sample a subset Ω with $\text{SR} = 0.5$, and then form the observed matrix $\boldsymbol{M} = \mathcal{P}_\Omega(\widehat{\boldsymbol{X}})$. In this test, we set $m = n = 150$.

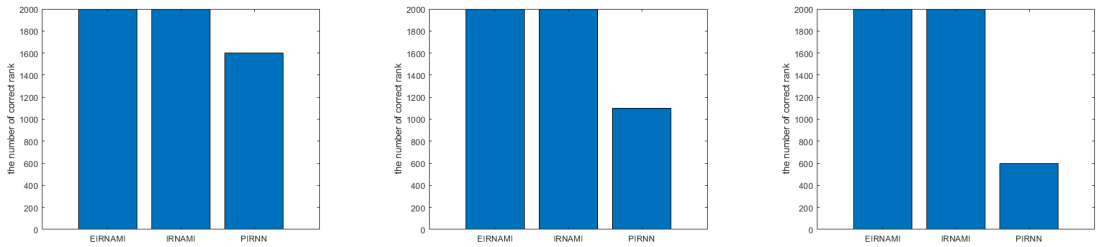
All algorithms compared in this section are initialized with a random Gaussian matrix $\boldsymbol{X}^0$. For IRNAMI and EIRNAMI, we set the parameters as follows: $p = 0.5$, $\beta = 1.1 > L_f$, $\mu = 0.1$, $tol_1 = 10^{-5}$ and $tol_2 = 10^{-7}$. In addition, we initialize $\boldsymbol{\epsilon}^0 = 10^{-3}\boldsymbol{e}$. For the extrapolation parameter α, we consider values in the range $\alpha \in \{0, 0.1, 0.3, 0.5, 0.7, 0.9\}$. We then select the values that yield the best performance in most cases. Our experimental results shown in Figure 2 indicate that $\alpha = 0.7$ is a reasonable choice.



(a) $r^* = 5$ (b) $r^* = 10$ (c) $r^* = 15$

Figure 2: The performance of EIRNAMI with different values of extrapolation parameter α for various ranks r^*.

Next, we initialize $\boldsymbol{X}^0$ with $\text{Rank}(\boldsymbol{X}^0) = r \in \{1, 2, 3, \ldots, 20\}$. For each r, we solve 100 independent realizations with the initialized data. For PIRNN, the perturbation parameter is fixed as $\epsilon_i = 10^{-3}, i \in [m]$, while other parameters follow the default settings suggested in their paper. From a total of 2,000 problems, Figure 3 shows the number of problems converging to a solution $\boldsymbol{X}^*$ with the correct rank, $r^* = \text{Rank}(\widehat{\boldsymbol{X}}) = \text{Rank}(\boldsymbol{X}^*)$, for three considered algorithms. We can observe from Figure 3 that PIRNN does not find the correct rank in some cases, mainly due to the fixed perturbation strategy, as discussed in §1. In addition, we present the average relative errors of these two algorithms for different values of r^* in Table 1. The average is calculated as the arithmetic mean of the total relative errors across 2000 problems for each r.



(a) The number of CLD with $r^* = 5$.

(b) The number of CLD with $r^* = 10$.

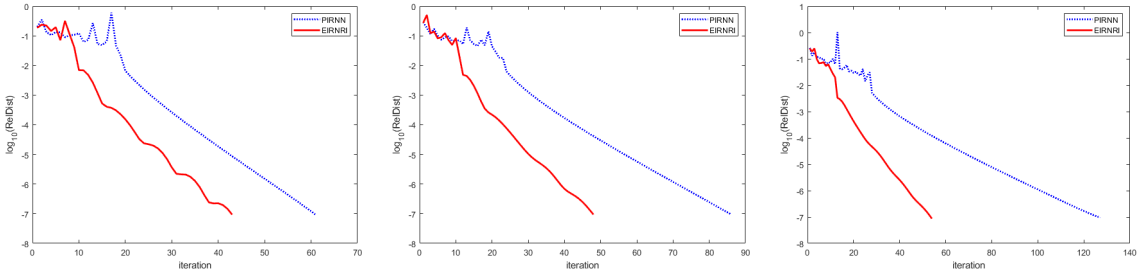
(c) The number of CLD with $r^* = 15$.

Figure 3: The number of problems that achieve r^*.

	$r^* = 5$	$r^* = 10$	$r^* = 15$
PIRNN	5.13×10^{-6}	3.00×10^{-4}	1.10×10^{-2}
EIRNAMI	$6.42 \times 10{-}7$	2.67×10^{-5}	9.15×10^{-4}

Table 1: The average of the relative errors.

We also plot the evolution of the relative distance for different values of r^* in Figure 4. We can see that the proposed EIRNAMI significantly outperforms PIRNN in terms of the convergence speed.



(a) $r^* = 5$

(b) $r^* = 10$

(c) $r^* = 15$

Figure 4: The performance comparison between PIRNN and EIRNAMI with initialization $\boldsymbol{X}^0$ satisfying $\text{Rank}(\boldsymbol{X}^0) = r^*$.

5.2 Natural Color Images

The second experiment is conducted using natural color images. In this section, we compare the proposed method with some state-of-the-art methods, including PIRNN (Sun et al., 2017), AIRNN (Phan and Nguyen, 2021)[1], SCp (Li et al., 2020)[2], and FGSRp (Fan et al.,

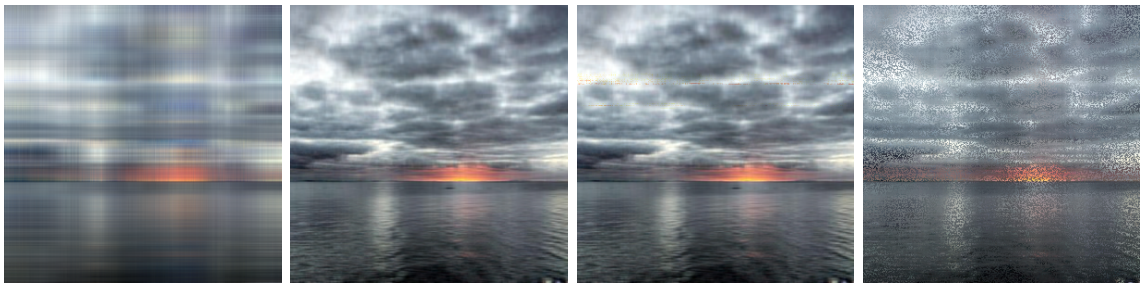
1. The code is available at https://github.com/ngocntkt/AIRNN
2. The code is available at https://github.com/liguorui77/scpnorm

2019)[1]. While not all natural color images are inherently low-rank, the primary information is captured by the top singular values (Sun et al., 2017). Hence, a natural color image can be well recovered by the low-rank approximation. The considered natural images are sized $300 \times 300 \times 3$. We apply matrix recovery for each channel independently.

We first conduct a set of random mask experiments similar to those described in (Li et al., 2020). In these experiments, the row and column indices of the missing entries in each image channel are randomly selected and the corresponding pixel values are set to zero, resulting in a missing rate of 50%. Additionally, we perform block mask experiments using a block column mask, where four increasingly sized blocks are arranged diagonally across the image. In this test, we set $p = 0.5$.

The performance of all algorithms evaluated is evaluated using two metrics: (i) the difference in rank between the recovered $\boldsymbol{X}^*$ and the low-rank ground truth $\widehat{\boldsymbol{X}}$, and (ii) the peak signal-to-noise ratio (PSNR), defined as $\text{PSNR}(\widehat{\boldsymbol{X}}, \boldsymbol{X}^*) = 10\log_{10}\left(\frac{mn\ 255^2}{\|\widehat{\boldsymbol{X}}-\boldsymbol{X}^*\|_F^2}\right)$. We mention that higher PSNR values indicate greater accuracy of the recovered result.

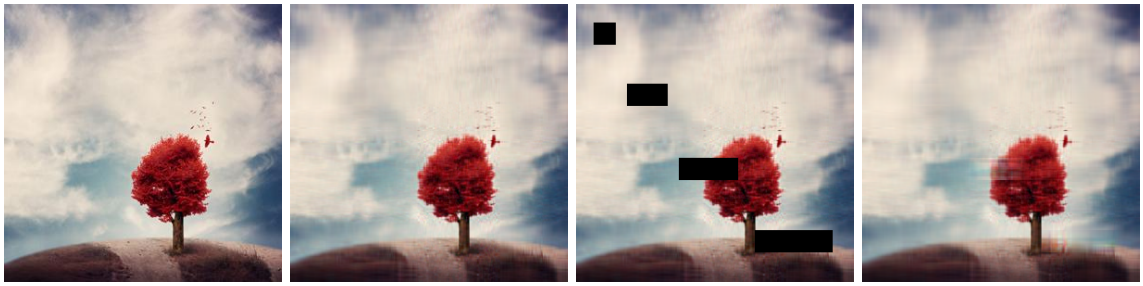
(a) Original image (b) Low-rank image (c) Random mask (d) PIRNN



(e) AIRNN (f) EIRNAMI (g) SCp (h) FGSRp

Figure 5: The performance of considered algorithms with a random mask and SR= 0.5. (a) Original image: $\text{Rank}(\boldsymbol{X}) = 300$; (b) Low-rank image: $\text{Rank}(\widehat{\boldsymbol{X}}) = 30$; (c) Noisy image; (d) PIRNN: $\text{Rank}(\boldsymbol{X}^*) = 30$, PSNR=34.47; (e) AIRNN: $\text{Rank}(\boldsymbol{X}^*) = 3$, PSNR=22.68; (f) EIRNAMI: $\text{Rank}(\boldsymbol{X}^*) = 30$, PSNR=34.47; (g) SC$p$: $\text{Rank}(\boldsymbol{X}^*) = 300$, PSNR=33.20; (h) FGSR$p$: $\text{Rank}(\boldsymbol{X}^*) = 300$, PSNR=20.68.

1. The code is available at https://github.com/udellgroup/Codes-of-FGSR-for-effecient-low-rank-matrix-recovery



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

(e) AIRNN (f) EIRNAMI (g) SCp (h) FGSRp

Figure 6: The performance of considered algorithms with the block mask. (a) Original image: Rank($\boldsymbol{X}$) = 300; (b) Low-rank image: Rank($\boldsymbol{X}^*$) = 30; (c) Noised picture; (d) PIRNN: Rank($\boldsymbol{X}$) = 25, PSNR=31.00; (e) AIRNN: Rank($\boldsymbol{X}$) = 4, PSNR=24.24; (f) EIRNAMI: Rank($\boldsymbol{X}$) = 30, PSNR=34.74; (g) SCp: Rank($\boldsymbol{X}$) = 126, PSNR=31.76; (h) FGSRp: Rank($\boldsymbol{X}$) = 126, PSNR=26.41.

	PIRNN		AIRNN		EIRNAMI		SCp		FGSRp	
	PSNR	Rank	PSNR	Rank	PSNR	Rank	PSNR	Rank	PSNR	Rank
$r^* = 15$	**30.67**	**15**	22.69	3	**30.67**	**15**	30.39	300	20.53	300
$r^* = 20$	**32.19**	**20**	22.69	3	**32.19**	**20**	31.64	300	20.60	300
$r^* = 25$	**33.39**	**25**	22.69	3	**33.39**	**25**	32.54	300	20.63	300
$r^* = 30$	**34.47**	**30**	22.68	3	**34.47**	**30**	33.20	300	20.66	300
$r^* = 35$	**35.46**	**35**	22.68	3	**35.46**	**35**	33.64	300	20.68	300
$r^* = 40$	**36.37**	**40**	22.68	3	**36.37**	**40**	33.87	300	20.69	300

Table 2: The performance of the considered algorithms with the random mask is evaluated. Bold values indicate the best results for each task.

	PIRNN		AIRNN		EIRNAMI		SCp		FGSRp	
	PSNR	Rank	PSNR	Rank	PSNR	Rank	PSNR	Rank	PSNR	Rank
$r^* = 15$	30.46	**15**	24.24	4	**31.22**	**15**	30.75	75	25.70	75
$r^* = 20$	31.24	**20**	24.24	4	**32.66**	**20**	31.73	100	26.03	100
$r^* = 25$	31.34	24	24.24	4	**33.83**	**25**	31.93	121	26.25	121
$r^* = 30$	31.00	25	24.24	4	**34.74**	**30**	31.76	126	26.41	126
$r^* = 35$	30.97	26	24.24	4	**35.45**	**35**	31.65	131	26.53	131
$r^* = 40$	30.93	27	24.24	4	**35.89**	**40**	31.81	136	26.62	136

Table 3: The performance of the considered algorithms with the block mask is evaluated. Bold values indicate the best results for each task.

For the presented results, we use the default rank function in MATLAB without specifying any tolerance to calculate the rank of the recovered results. Figure 5 displays the recovered images with the random mask. Most of the considered algorithms achieve relatively high-quality visual recovery. Table 2 records the results for different r^* across considered algorithms. Notably, the recovered results by EIRNAMI and PIRNN match the rank of the ground-truth solution. However, this is not the case for PIRNN with the block mask, as illustrated in Figure 6 and Table 3.

6. Conclusion and Discussion

In this paper, we have proposed, analyzed, and implemented iteratively reweighted Nuclear norm methods for solving the Schatten-p norm regularized low-rank optimization. Our work features two main novelties. The first is the exhibition of an active manifold identification property, enabling the algorithm to identify the rank of stationary points of the concerned problems in finite iterations. Leveraging this property, we have designed a novel updating strategy for ϵ_i, so that ϵ_i associated with the positive singular values can be driven to zero rapidly and those associated with zero singular values can be automatically fixed as constants after a finite number of 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 be directly applied including acceleration techniques and convergence analysis.

The convergence properties established for our algorithm are illustrated empirically on test sets comprising both synthetic and real data sets. We remark, however, that several practical considerations remain to be addressed for enhancing the performance of the proposed method. One potential avenue for improvement involves integrating the active manifold identification property into the implementation. Once the correct rank has been identified within a finite number of iterations, the algorithm can be terminated and subsequently switched to a traditional Frobenius recovery with a fixed rank to further enhance the quality of the recovered solution. We defer the exploration of this aspect to future research efforts. On the other hand, as suggested by a referee and motivated by (Yukawa and Amari, 2015), we believe that it is worthwhile to discuss the selection of the regularization param-

eter λ. Inspired by (Yukawa and Amari, 2015), we can consider a critical path, where a curve consists of critical points of the problem for different values of λ. This is realistic since the (local) optimal solution changes continuously to form a path of the critical points as λ varies continuously. A path-following algorithm could be developed by iteratively solving the regularization problem for a series of increasing values of $\lambda_1 < \cdots < \lambda_N$. The optimal solution $\boldsymbol{X}^*(\lambda_j)$ for λ_j would be used to warm start the solution $\boldsymbol{X}^*(\lambda_{j+1})$ for λ_{j+1}, ensuring that $\boldsymbol{X}^*(\lambda_{j+1})$ lies in a neighborhood of $\boldsymbol{X}^*(\lambda_j)$. We are well aware that there may be multiple paths of (local) solutions as λ varies. Generally, it is not an easy task to obtain the global solution path due to the highly nonconvex nature of both loss term $f(\boldsymbol{X})$ and the regularizer $\|\boldsymbol{X}\|_p^p$. Moreover, the global solution path may be discontinuous. This phenomenon is studied in (Yukawa and Amari, 2015) for the ℓ_p-regularized least squares problem for vector variables.

Acknowledgments

Xiangyu Yang is the corresponding author. We would like to acknowledge the support in part for this paper from the Natural Science Foundation of Shanghai under Grant 21ZR1442800 and the Young Scientists Fund of the National Natural Science Foundation of China No. 12301398. Xiangyu Yang also acknowledges the financial support from the China Scholarship Council under Grant No. 202308410343.

References

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

Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward—backward splitting, and regularized Gauss—Seidel methods. *Mathematical Programming*, 137(1):91–129, 2013.

Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2): 459–494, 2014.

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

Kun Chen, Hongbo Dong, and Kung-Sik Chan. Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, 100(4):901–920, 2013.

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

Mark A Davenport and Justin Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4): 608–622, 2016.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

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, Haitham Hindi, and Stephen P Boyd. A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the 2001 American Control Conference.*, volume 6, pages 4734–4739. IEEE, 2001.

Jerome H Friedman. Fast sparse regression and classification. *International Journal of Forecasting*, 28(3):722–738, 2012.

Cuixia Gao, Naiyan Wang, Qi Yu, and Zhihua Zhang. A feasible nonconvex relaxation approach to feature selection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 25(1):356–361, 2011.

Guillaume Garrigos. *Descent dynamical systems and algorithms for tame optimization and multi-objective problems.* PhD thesis, Université de Montpellier; Universidad Tecnica Federico Santa Maria, 2015.

Zhili Ge, Xin Zhang, and Zhongming Wu. A fast proximal iteratively reweighted nuclear norm algorithm for nonconvex low-rank matrix minimization problems. *Applied Numerical Mathematics*, 179:66–86, 2022.

Donald Geman and Chengda Yang. Nonlinear image recovery with half-quadratic regularization. *IEEE Transactions on Image Processing*, 4(7):932–946, 1995.

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.

Warren L Hare and Adrian S Lewis. Identifying active manifolds. *Algorithmic Operations Research*, 2(2):75–82, 2007.

Roger A Horn and Charles R Johnson. *Matrix analysis.* Cambridge University Press, 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.

Joonseok Lee, Seungyeon Kim, Guy Lebanon, Yoram Singer, and Samy Bengio. LLORMA: Local low-rank matrix approximation. *Journal of Machine Learning Research*, 17(15): 1–24, 2016.

Adrian S Lewis. Active sets, nonsmoothness, and sensitivity. *SIAM Journal on Optimization*, 13(3):702–725, 2002.

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

Guorui Li, Guang Guo, Sancheng Peng, Cong Wang, Shui Yu, Jianwei Niu, and Jianli Mo. Matrix completion via Schatten capped p norm. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

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.

Zhipeng Lin, Zhenyu Zhao, Tingjin Luo, Wenjing Yang, Yongjun Zhang, and Yuhua Tang. Non-convex transfer subspace learning for unsupervised domain adaptation. *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1468–1473, 2019.

Meng Liu, Yong Luo, Dacheng Tao, Chao Xu, and Yonggang Wen. Low-rank multi-view learning in matrix completion for multi-label image classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), 2015.

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.

Canyi Lu, Changbo Zhu, Chunyan Xu, Shuicheng Yan, and Zhouchen Lin. Generalized singular value thresholding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Zhaosong Lu, Yong Zhang, and Jian Lu. ℓ_p regularized low-rank approximation via iterative reweighted singular value minimization. *Computational Optimization and Applications*, 68(3):619–642, 2017.

Mohammadreza Malek-Mohammadi, Massoud Babaie-Zadeh, and Mikael Skoglund. Performance guarantees for Schatten-p quasi-norm minimization in recovery of low-rank matrices. *Signal Processing*, 114:225–230, 2015.

Goran Marjanovic and Victor Solo. On ℓ_q optimization and matrix completion. *IEEE Transactions on Signal Processing*, 60(11):5714–5724, 2012.

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.

Sahand Negahban and Martin J Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097, 2011.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

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, Hua Wang, Xiao Cai, Heng Huang, and Chris Ding. Robust matrix completion via joint Schatten p-norm and ℓ_p-norm minimization. In *2012 IEEE 12th International Conference on Data Mining*, pages 566–574. IEEE, 2012.

Alexander M Ostrowski. *Solution of equations in Euclidean and Banach spaces*. Academic Press, 1973.

Wenqing Ouyang, Yuncheng Liu, Ting Kei Pong, and Hao Wang. Kurdyka–Łojasiewicz exponent via Hadamard parametrization. *arXiv preprint arXiv:2402.00377*, 2024.

Soumyabrata Pal and Prateek Jain. Online low rank matrix completion. In *The Eleventh International Conference on Learning Representations*, 2022.

Ching pei Lee, Ling Liang, Tianyun Tang, and Kim-Chuan Toh. Accelerating nuclear-norm regularized low-rank matrix optimization through Burer-Monteiro decomposition. *arXiv preprint arXiv:2204.14067*, 2023.

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.

R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

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.

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

Joshua Trzasko and Armando Manduca. Highly undersampled magnetic resonance image reconstruction via homotopic ℓ_0-minimization. *IEEE Transactions on Medical Imaging*, 28(1):106–121, 2008.

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.

Lei Wang, Bangjun Wang, Zhao Zhang, Qiaolin Ye, Liyong Fu, Guangcan Liu, and Meng Wang. Robust auto-weighted projective low-rank and sparse recovery for visual representation. *Neural Networks*, 117:201–215, 2019.

Bo Wen, Xiaojun Chen, and Ting Kei Pong. A proximal difference-of-convex algorithm with extrapolation. *Computational Optimization and Applications*, 69:297–324, 2018.

Peiran Yu and Ting Kei Pong. Iteratively reweighted ℓ_1 algorithms with extrapolation. *Computational Optimization and Applications*, 73(2):353–386, 2019.

Peiran Yu, Guoyin Li, and Ting Kei Pong. Kurdyka–Łojasiewicz exponent via inf-projection. *Foundations of Computational Mathematics*, 22(4):1171–1217, 2022.

Man-Chung Yue and Anthony Man-Cho So. A perturbation inequality for concave functions of singular values and its applications in low-rank matrix recovery. *Applied and Computational Harmonic Analysis*, 40(2):396–416, 2016.

Masahiro Yukawa and Shun-ichi Amari. ℓ_p-regularized least squares $(0 < p < 1)$ and critical path. *IEEE Transactions on Information Theory*, 62(1):488–502, 2015.

Chao Zeng. Proximal linearization methods for Schatten p-quasi-norm minimization. *Numerische Mathematik*, 153(1):213–248, 2023.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38(2):894–942, 2010.

Min Zhang, Zheng-Hai Huang, and Ying Zhang. Restricted p-isometry properties of non-convex matrix recovery. *IEEE Transactions on Information Theory*, 59(7):4316–4323, 2013.

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