Achieving Oracle Statistical Property for Low-Rank Matrix Recovery with Application to Image Denoising

Anonymous CVPR submission

Paper ID 10048

Abstract

This paper studies the classical low-rank matrix recovery problem, a topic of significant interest in computer vision. The nuclear norm minimization (NNM) method is a commonly used convex approach to addressing this problem. However, equal regularization of singular values in NNM restricts its effectiveness in practical applications, such as image restoration, where singular values carry distinct physical meanings and should be treated differently. This limitation led to the development of the weighted nuclear norm minimization (WNNM) approach, which offers enhanced recovery capabilities. Despite the advantages of NNM and WNNM, their reliance on ℓ_1 homotopy penalties can introduce a non-negligible recovery bias. In response to this challenge, we introduce non-convex penalties for low-rank matrix recovery in our work. We propose an algorithm within the majorization-minimization framework, leading to the adaptive weighted nuclear norm minimization (AWNNM) method. Despite the non-convex nature of the problem, we rigorously analyze the statistical properties of the iterations produced by AWNNM and demonstrate its ability to achieve the optimal statistical rate, known as the oracle rate. Our theoretical results are validated through numerical experiments on synthetic data. Extensive experiments on real-world datasets for image denoising demonstrate the efficacy of AWNNM.

1. Introduction

The problem of low-rank matrix recovery is focused on estimating a provided matrix with a low-rank alternative. This process aids in simplifying a complex dataset by uncovering inherent patterns, thus maintaining crucial details [19, 29, 30, 32, 38]. By enhancing data analysis efficiency and accuracy, low-rank matrix recovery reduces computational complexity and minimizes storage requirements, making it particularly beneficial for handling large-scale datasets.

Low-rank matrix recovery is a fundamental concept with broad applications in various fields such as computer vision [7, 27, 30, 47, 50], natural language processing [21], pattern recognition [44], and collaborative filtering [35]. For example, low-rank matrix recovery is employed in image matting to separate video sequences into the background and move foreground components [26]. Similarly, in image compression, low-rank matrix recovery reduces data storage requirements without compromising image quality [45, 49]. Furthermore, research indicates that matrices generated by non-local similar blocks in natural images demonstrate low-rank structures, which are advantageous for enhancing image restoration performance [39].

Over the past few decades, various techniques for low-rank matrix recovery have been developed, typically falling into two primary categories: matrix factorization methods [1, 10, 18, 20, 32, 36, 38] and rank penalty methods [2, 4, 9, 12, 22–25, 34, 42]. Matrix factorization methods involve breaking down a matrix into the product of two lower-dimensional matrices. However, these methods face challenges due to their non-convex nature, leading to complex optimization landscapes with various local minima and saddle points, making it difficult to achieve global optimality. Furthermore, the non-uniqueness of the decomposition and the requirement to specify the matrix rank in advance create challenges, mainly when the actual rank is unknown or inaccurately estimated. On the other hand, rank penalty methods like nuclear norm minimization (NNM) provide a convex relaxation of the rank minimization problem, making them more appealing in theoretical contexts because of their advantageous optimization properties.

Among the various methods based on rank penalties, NNM is arguably the most widely used for low-rank matrix recovery, as it provides the tightest convex relaxation of the matrix rank [31]. While convex formulations guarantee global optimality, they treat all singular values uniformly, which may limit their effectiveness in specific applications [13, 14, 28]. This uniform treatment overlooks prior knowledge that larger singular values can be more critical in representing the latent data. To address this limita-

tion, [13] introduced the weighted nuclear norm minimization (WNNM) method to incorporate varying importance across singular values. However, both WNNM and NNM are based on ℓ_1 homotopy palatines, which introduces non-negligible estimation bias that compromises the estimation accuracy. In contrast, non-convex penalties, such as the smoothly clipped absolute deviation (SACD) penalty [11] and minimax concave penalty (MCP) [48], are preferred for their superior estimation accuracy and variable selection consistency [40]. Despite encouraging empirical findings in various studies [13–15, 33], the theoretical underpinnings of non-convex penalties for low-rank matrix recovery remain largely unexplored. The quest for a theoretical rationale supporting non-convex substitutes for matrix rank continues to be an unresolved issue.

In this paper, we provide a comprehensive investigation of the non-convex low-rank matrix recovery problem, aiming to bridge the gap between practical applications and theoretical foundations. We introduce a novel low-rank matrix recovery that incorporates non-convex penalties. We devise an algorithm based on the majorization-minimization (MM) framework to address the proposed problem, leading to the adaptive weighted nuclear norm minimization (AWNNM) method. Our investigation illustrates that our estimator effectively utilizes singular values of significant magnitudes, leading to enhanced statistical convergence rates. Furthermore, under a mild assumption on the magnitude of the singular values, we establish that the proposed estimator possesses the oracle property. This property enables accurate recovery of the actual rank of the underlying matrix and attaining improved convergence rates. The primary contributions of this paper are threefold:

- We introduce a novel approach for low-rank matrix recovery based on non-convex penalties called AWNNM. We develop an MM-based algorithm that iteratively solves a sequence of subproblems, each admitting a closed-form solution.
- We theoretically analyze the statistical properties of all iterates generated by the proposed MM-based algorithm, proving that our method achieves the oracle statistical rate in the Frobenius norm under weak assumptions. These advancements offer both computational efficiency and strong theoretical guarantees, enhancing the robustness and applicability of matrix estimation techniques.
- We validate our theoretical results through comprehensive numerical experiments on synthetic and real-world datasets, further demonstrating the effectiveness of our method in image restoration tasks.

2. Low-Rank Matrix Recovery

In this section, we first revisit the classical NNM and WNNM methods. We then introduce our proposed method, AWNNM, which is based on non-convex penalties. Following this, we provide an overview of the MM algorithm framework and present our MM-based multistage relaxation algorithm. We also outline the necessary technical assumptions for the theoretical analysis before establishing the statistical convergence rate of the proposed method.

2.1. Existing Methods

The rank minimization problem is typically given by

$$\min_{\mathbf{X}} \frac{1}{2}\|\mathbf{Y}-\mathbf{X}\|_{\mathrm{F}}^2 + \lambda \operatorname{rank}(\mathbf{X}), \tag{1}$$

where $\mathbf{Y} \in \mathbb{R}^{m\times n}$ is the observed noisy matrix, $\mathbf{X}$ is the matrix to be recovered, $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm, and λ is a regularization parameter that balances the data fidelity and rank minimization. However, solving (1) is NP-hard. To enhance tractability, the NNM method utilizes the nuclear norm, defined as the sum of the singular values of the matrix $\mathbf{X}$, to approximate the rank $\operatorname{rank}(\mathbf{X})$. Under certain incoherence assumptions, NNM has been shown to recover a near-optimal low-rank solution [3]. The NNM formulation is:

$$\min_{\mathbf{X}} \frac{1}{2}\|\mathbf{Y}-\mathbf{X}\|_{\mathrm{F}}^2 + \lambda \sum_{i=1}^{d} \sigma_i(\mathbf{X}),$$

where $\sigma_i(\mathbf{X})$ representing the i-th singular values of $\mathbf{X}$ and $d = \min\{m, n\}$.

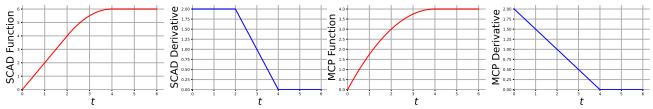
In many cases, the rows (or columns) of the matrix lie within a low-dimensional subspace, where the primary singular values capture the dominant projection directions. These prominent singular values are crucial for preserving the data structure and should, thus, be assigned smaller penalty coefficients. To address this, WNNM [13] introduces a weighting scheme for the singular values. This modification provides greater flexibility and can significantly improve matrix recovery, particularly in scenarios with structured noise or matrices with varying significance. The optimization problem of WNNM is defined as:

$$\min_{\mathbf{X}} \frac{1}{2}\|\mathbf{Y}-\mathbf{X}\|_{\mathrm{F}}^2 + \lambda \sum_{i=1}^{d} w_i \sigma_i(\mathbf{X}),$$

where w_i denotes the non-negative weight for $\sigma_i(\mathbf{X})$. This weighted approach enables more precise rank recovery, enabling a more accurate recovery of the actual low-rank structure compared to the standard nuclear norm minimization method [17].

2.2. Proposed Method via Non-Convex Penalties

Both the NNM and WNNM methods are ℓ_1-based method, which introduces non-negligible estimation bias, thus compromising solution accuracy [28]. In contrast, non-convex penalties are preferred due to their superior estimation accuracy and variable selection consistency, as they adaptively



(a) SCAD penalty (b) MCP penalty

Figure 1. Illustration of the sparsity-inducing functions p_λ with their derivatives($\lambda = 2$ and $a = 2$ for both SCAD and MCP)

adjust weights during the optimization process [40]. To mitigate the estimation bias inherent in ℓ_1 homotopy penalties, we propose a low-rank minimization approach using the non-convex penalty, also known as the adaptive weighted nuclear norm penalty. Specifically, we aim to solve the following non-convex optimization problem:

$$\min_{\mathbf{X}} \frac{1}{2} \|\mathbf{X} - \mathbf{Y}\|_{\mathrm{F}}^2 + \sum_{i=1}^{d} p_\lambda \left(\sigma_i \left(\mathbf{X}\right)\right), \tag{2}$$

where p_λ is a univariate non-convex function with parameter λ. In this paper, we consider a class of functions p_λ satisfying the following assumption.

Assumption 1. *The penalty function p_λ: $\mathbb{R}^+ \to \mathbb{R}^+$ satisfies the following conditions:*

- *$p_\lambda(t)$ is symmetric and non-decreasing on $[0, +\infty)$, with $p_\lambda(0) = 0$, and is differentiable almost everywhere on $(0, +\infty)$.*
- *For all $t_1 \geq t_2 \geq 0$, it holds that $0 \leq p'_\lambda(t_1) \leq p'_\lambda(t_2) \leq \lambda$, and $\lim_{t \to 0} p'_\lambda(t) = \lambda$;*
- *There exists an $\alpha > 0$ such that $p'_\lambda(t) = 0$ for $t \geq \alpha\lambda$.*
- *There exists some $c \in (0, \alpha)$ such that $p'_\lambda(c\lambda) \geq \frac{\lambda}{2}$.*

In Assumption 1, the first three conditions ensure sparsity and unbiasedness, while the last condition is introduced for the following theoretical analysis, which can always hold due to $p'_\lambda(0) = \lambda$ and $p'_\lambda(\alpha\lambda) = 0$. Typical examples of the non-convex penalty function p_λ in Assumption 1 include:

- Smoothly clipped absolute deviation (SCAD) penalty: SCAD function is given by

$$p_\lambda(t) = \begin{cases} \lambda t, & 0 \leq t \leq \lambda \\ \frac{-t^2 + 2\lambda a t - \lambda^2}{2(a-1)}, & \lambda < t \leq a\lambda \\ \frac{\lambda^2(a+1)}{2}, & t > a\lambda \end{cases}$$

  for some $a > 2$. The choice $a = 3.7$ is suggested in [11] based on a Bayesian argument.
- Minimax concave penalty (MCP): MCP function $p_\lambda(\cdot)$ is defined as follows:

$$p_\lambda(t) = \begin{cases} \lambda t - \frac{t^2}{2a}, & 0 \leq t \leq a\lambda \\ \frac{1}{2}at^2, & t > a\lambda \end{cases}$$

  for some $a > 1$.

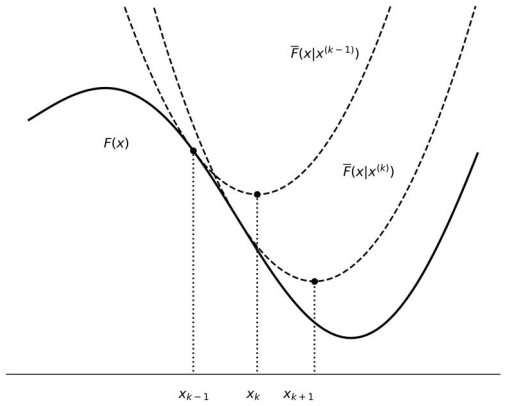


Figure 2. The MM procedure.

We give a pictorial illustration of these two functions together with their derivatives in Figure 1.

3. Optimization Algorithm

3.1. Brief Review of the MM Algorithm Framework

The MM algorithm [16, 37] is an iterative optimization procedure involving two primary steps: majorization and minimization. Given a loss function $F(\mathbf{x})$ to minimize, initialized at $\mathbf{x}^{(0)}$, the MM algorithm generates a sequence of feasible points $\{\mathbf{x}^{(k)}\}_{k \geq 1}$. At each iteration k, starting from $\mathbf{x}^{(k-1)}$, the majorization step involves constructing a surrogate function $\overline{F}\left(\mathbf{x} \mid \mathbf{x}^{(k-1)}\right)$ that locally approximates the objective function $F(\mathbf{x})$, satisfying the following conditions:

$$\overline{F}\left(\mathbf{x} \mid \mathbf{x}^{(k-1)}\right) \geq F(\mathbf{x}),$$
$$\overline{F}\left(\mathbf{x}^{(k-1)} \mid \mathbf{x}^{(k-1)}\right) = F\left(\mathbf{x}^{(k-1)}\right).$$

Next, in the minimization step, the algorithm update $\mathbf{x}^{(k)}$ by minimizing the surrogate function:

$$\mathbf{x}^{(k)} \in \arg\min_{\mathbf{x}} \overline{F}\left(\mathbf{x} \mid \mathbf{x}^{(k-1)}\right).$$

Starting from an initial point $\mathbf{x}^{(0)}$, the alternating process of MM generates a sequence of iterates $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots$ with non-increasing objective function values. The algorithm can be executed until a convergence criterion is met. See Figure 2 for an illustration.

3.2. An MM-based Algorithm for Problem (2)

In this section, we derive an MM algorithm for problem (2). In the majorization step, we construct a surrogate function, which locally linearizes the non-convex penalty function p_λ. Specifically, we find an upper bound for $p_\lambda(t)$ using its first-order Taylor expansion around the iterate x_0:

$$p_\lambda(t) \le p_\lambda(t_0) + p'_\lambda(t_0)(t - t_0).$$

Based on the above result, we have the following inequality at iterate $\mathbf{X}^{(k-1)}$:

$$\sum_{i=1}^{d} p_\lambda(\sigma_i(\mathbf{X})) \le \sum_{i=1}^{d} p_\lambda\left(\sigma_i\left(\mathbf{X}^{(k-1)}\right)\right) + \sum_{i=1}^{d} w_i^{(k-1)}\left(\sigma_i(\mathbf{X}) - \sigma_i\left(\widehat{\mathbf{X}}^{(k-1)}\right)\right),$$

where $w_i^{(k-1)} = p'_\lambda\left(\sigma_i\left(\widehat{\mathbf{X}}^{(k-1)}\right)\right)$. This leads to the following surrogate function:

$$\overline{F}\left(\mathbf{X} \mid \widehat{\mathbf{X}}^{(k-1)}\right) = \frac{1}{2}\|\mathbf{X} - \mathbf{Y}\|_{\mathrm{F}}^2 + \sum_{i=1}^{d} w_i^{(k-1)}\sigma_i(\mathbf{X}).$$

In the minimization step, we compute the iterate that minimizes the surrogate function derived in the majorization step. This is achieved by solving the following optimization problem:

$$\widehat{\mathbf{X}}^{(k)} \in \arg\min_{\mathbf{X}} \left\{ \frac{1}{2}\|\mathbf{X} - \mathbf{Y}\|_{\mathrm{F}}^2 + \sum_{i=1}^{d} w_i^{(k-1)}\sigma_i(\mathbf{X}) \right\}. \tag{3}$$

The problem (3) is inherently non-convex. However, a closed-form solution can be obtained under Assumption 1, for the reason it states that $p'(t)$ is non-increasing for $t \ge 0$.

Let $\mathbf{Y} \in \mathbb{R}^{m\times n}$ be a matrix with singular value decomposition (SVD) given by $\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{m\times d}$ contains the left singular vectors, $\mathbf{V} \in \mathbb{R}^{n\times d}$ contains the right singular vectors, and $\mathbf{\Sigma} = \mathrm{diag}(\sigma_i(\mathbf{Y}), i = 1, \dots d) \in \mathbb{R}^{d\times d}$ is a diagonal matrix of singular values. The singular values are ordered such that $\sigma_1(\mathbf{Y}) \ge \sigma_2(\mathbf{Y}) \ge \cdots \ge \sigma_d(\mathbf{Y}) \ge 0$. Due to this ordering, the weight sequence $\{w_1, w_2, \cdots, w_d\}$ satisfies $0 \le w_1 \le w_2 \le \cdots \le w_d$, following the anti-monotonic property. This ordering enables a closed-form solution using a weighted singular value thresholding approach in the following lemma.

Lemma 2. *[[6], Theorem 2.3] For $\mathbf{X} \in \mathbb{R}^{m\times n}$ and $0 \le w_1 \le w_2 \le \cdots \le w_d$, a global solution to the optimization problem in (3) is given by the weighted singular value thresholding*

$$\widehat{\mathbf{X}} \coloneqq \mathcal{T}(\mathbf{Y}), \tag{4}$$

where $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ is the SVD of $\mathbf{Y}$, $\mathcal{T}(\mathbf{Y}) = \mathbf{U}\mathcal{T}(\mathbf{\Sigma})\mathbf{V}^\top$, $\mathcal{T}(\mathbf{\Sigma}) = \mathrm{Diag}(\max(0, \sigma_i(\mathbf{Y}) - w_i), i = 1, \dots, d)$.

The MM-based multistage convex relaxation algorithm is summarized in Algorithm 1. For simplicity, we start with a trivial initial value $\widehat{\mathbf{X}}^{(0)} = \mathbf{0}$.

Algorithm 1: Multistage Relaxation Algorithm for Solving (2).

Input: Data matrix $\mathbf{Y}$, tuning parameter λ;
Initialize $\widehat{\mathbf{X}}^{(0)} = \mathbf{0}$, $k = 1$;
for $k = 1, 2, \dots, K$ **do**
  update $w_i^{(k-1)} = p'_\lambda\left(\sigma_i\left(\widehat{\mathbf{X}}^{(k-1)}\right)\right)$
  update $\widehat{\mathbf{X}}^{(k)}$ by solving Problem (3) via Lemma 2;
  $k = k + 1$;
end for
Output: $\widehat{\mathbf{X}}^{(K)}$.

4. Theoretical Analysis

We define the true low-rank matrix as $\mathbf{X}^*$. Let $\mathcal{F}$ and its orthogonal complement $\mathcal{F}^\perp$ be the subspaces corresponding to the SVD of $\mathbf{X}^*$, defined as follows:

$$\mathcal{F}(\mathbf{X}^*) = \{\mathbf{W} \mid \mathcal{R}(\mathbf{W}) \subseteq \mathrm{sp}(\mathbf{V}^*), \mathcal{C}(\mathbf{W}) \subseteq \mathrm{sp}(\mathbf{U}^*)\},$$
$$\mathcal{F}^\perp(\mathbf{X}^*) = \{\mathbf{W} \mid \mathcal{R}(\mathbf{W}) \perp \mathrm{sp}(\mathbf{V}^*), \mathcal{C}(\mathbf{W}) \perp \mathrm{sp}(\mathbf{U}^*)\},$$

where $\mathcal{R}(\mathbf{W})$ and $\mathcal{C}(\mathbf{W})$ represent the row and column space of $\mathbf{W}$, respectively, and $\mathrm{sp}(\mathbf{V}^*), \mathrm{sp}(\mathbf{U}^*)$ denote the subspace spanned by $\mathbf{V}^*$ and $\mathbf{U}^*$, respectively. For brevity, we will use the shorthand notations of $\mathcal{F}$ and $\mathcal{F}^\perp$ when the dependence on $\mathbf{X}^*$ is clear from the context. The projection operators onto the subspace $\mathcal{F}$ and $\mathcal{F}^\perp$ are denoted by $\Pi_{\mathcal{F}}(\cdot)$ and $\Pi_{\mathcal{F}^\perp}(\cdot)$, respectively. Given matrix $\mathbf{A}$, the projections are defined as:

$$\Pi_{\mathcal{F}}(\mathbf{A}) = \mathbf{U}^*\mathbf{U}^{*\top}\mathbf{A}\mathbf{V}^*\mathbf{V}^{*\top},$$
$$\Pi_{\mathcal{F}^\perp}(\mathbf{A}) = \left(\mathbf{I} - \mathbf{U}^*\mathbf{U}^{*\top}\right)\mathbf{A}\left(\mathbf{I} - \mathbf{V}^*\mathbf{V}^{*\top}\right),$$

where $\mathbf{I}$ is an identity matrix of the appropriate dimension.

Next, we introduce a useful assumption.

Assumption 3 (Minimal Signal Strength)**.** *The true matrix $\mathbf{X}^*$ satisfies*

$$\min\{\sigma_i(\mathbf{X}^*) \mid i \in \mathcal{S}^*\} \ge (\alpha + c)\lambda \gtrsim \lambda,$$

where $\mathcal{S}^* = \{i \mid \sigma_i(\Pi_{\mathcal{F}}(\mathbf{X}^*)) \ne 0\}$ is the index set associated with the true rank subspace and α and c are constants defined in Assumption 1.

Assumption 3, referred to as the minimum signal strength condition, is mild and commonly employed [14]. We now present the main theorem, which establishes the contraction property of the solution path $\{\widehat{\mathbf{X}}^{(k)}\}_{k\ge1}$. Define $f(\mathbf{X}) = \frac{1}{2}\|\mathbf{X} - \mathbf{Y}\|_{\mathrm{F}}^2$.

Theorem 4. *Suppose that Assumptions 1 and 3 hold. If $\lambda \ge 4\|\nabla f(\mathbf{X}^*)\|_2$, then the optimal solution $\widehat{\mathbf{X}}^{(k)}$ satisfies the following δ-contraction property:*

$$\|\widehat{\mathbf{X}}^{(k)} - \mathbf{X}^*\|_{\mathrm{F}} \le \underbrace{\|\Pi_{\mathcal{F}_{\mathcal{S}^*}}(\nabla f(\mathbf{X}^*))\|_{\mathrm{F}}}_{\textit{oracle rate}} + \underbrace{\delta\left\|\widehat{\mathbf{X}}^{(k-1)} - \mathbf{X}^*\right\|_{\mathrm{F}}}_{\textit{contraction}},$$

for $1 \leq k < K$, *where* $\mathcal{F}_{\mathcal{S}^*}$ *is a subspace of* $\mathcal{F}$ *associated with* $\mathcal{S}^*$ *and* $\delta \in (0,1)$ *is the contraction parameter.*

Remark 5. The oracle rate refers to the statistical convergence rate of the oracle estimator, which has prior knowledge of the true rank subspace $\mathcal{F}(\mathbf{X}^*)$. The oracle estimator $\widehat{\mathbf{X}}^O$ is defined as

$$\widehat{\mathbf{X}}^O = \arg \min_{\mathbf{X} \in \mathcal{F}(\mathbf{X}^*)} f(\mathbf{X}).$$

According to the definition, it is easy to obtain that $\widehat{\mathbf{X}}^O$ satisfies $\left\|\widehat{\mathbf{X}}^O - \mathbf{X}^*\right\|_{\mathrm{F}} \lesssim \|\Pi_{\mathcal{F}_{\mathcal{S}^*}}(\nabla f(\mathbf{X}^*))\|_{\mathrm{F}}$.

Theorem 4 demonstrates that the estimation error between the optimal solution $\widehat{\mathbf{X}}^{(k)}$ and the true value $\mathbf{X}^*$ is bounded by two terms, namely, the oracle rate and a contraction term. We now provide the explicit statistical rate of convergence under the sub-Gaussian design.

Corollary 6. *Let* $\mathbf{Y} = \mathbf{X}^* + \mathbf{E}$ *and* $\mathbf{E}$ *be a sub-Gaussian random matrix with entries* E_{ij} *that are sub-Gaussian random variables with zero mean and covariance* ε. *Suppose that Assumptions 1 and 3 hold. If* $\lambda \asymp \sqrt{\log mn}$, *then the optimal solution* $\widehat{\mathbf{X}}^{(1)}$ *satisfies*

$$\left\|\widehat{\mathbf{X}}^{(1)} - \mathbf{X}^*\right\|_{\mathrm{F}} \lesssim \sqrt{r^* \log mn},$$

with high probability.

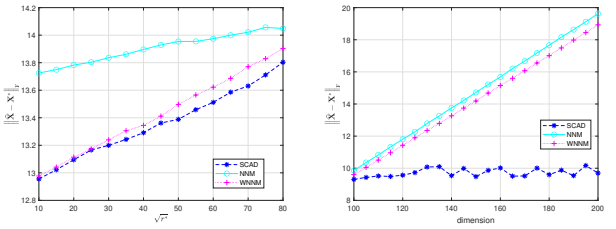
Corollary 6 follows directly from Theorem 4 for $k = 1$. Notably, the first subproblem (i.e., $k = 1$) corresponds to the nuclear norm penalized matrix recovery problem, which results in a convergence rate of $\sqrt{r^* \log mn}$. In Theorem 4, the contraction property is induced by the MM-based multistage convex relaxation algorithm. To achieve the oracle rate, it is crucial to appropriately select a sufficiently large value for K. The following result extends this analysis.

Corollary 7. *Let* $\mathbf{Y} = \mathbf{X}^* + \mathbf{E}$ *and* $\mathbf{E}$ *be a sub-Gaussian random matrix with entries* E_{ij} *are sub-Gaussian random variables with zero mean and covariance* ε. *Suppose that Assumptions 1 and 3 hold. If* $\lambda \asymp \sqrt{\log mn}$, *and* $K \gtrsim \log \lambda \gtrsim \log \log mn$, *then the optimal solution* $\widehat{\mathbf{X}}^{(K)}$ *satisfies*

$$\left\|\widehat{\mathbf{X}}^{(K)} - \mathbf{X}^*\right\|_{\mathrm{F}} \lesssim \sqrt{r^*},$$

with high probability.

Corollary 7 follows directly from Theorem 4, which implies that under weak assumptions, at most approximately $\log \log mn$ convex problems need to be solved to achieve the oracle rate $\sqrt{r^*}$. This result highlights that the proposed method attains a faster statistical rate of convergence, matching the oracle rate, compared to existing methods that employ nuclear norm penalties in the Frobenius norm.



(a) Estimation error via varying r^* (b) Estimation error via varying d

Figure 3. The oracle rate for low-rank matrix recovery

5. Numerical Simulations

5.1. Simulations on Theoretical Properties

In this section, we evaluate the practical recovery performance of the proposed method and present numerical results for the low-rank matrix recovery problem. The oracle rate derived for the proposed method is provided in 3. Figure 3a illustrates the recovery performance for a matrix with dimensions 100×200, where each entry is sampled from a normal distribution. A scaled noise with a parameter of 0.1 is added to the matrix. We examine the effect of varying rank, and the results demonstrate that the proposed non-convex penalty outperforms both the weighted nuclear norm and the nuclear norm. This observation is consistent with our theoretical findings. Figure 3b presents results for a square matrix of different dimensions, with each entry generated from a normal distribution. A scaled noise with a parameter of 0.1 is also added, and the matrix rank is fixed at 50. The results confirm that the proposed non-convex penalty maintains superior performance compared to the weighted nuclear norm and the nuclear norm, highlighting that the recovery error remains stable as the matrix dimension increases. This contrasts with the weighted nuclear norm and nuclear norm, where the error increases with higher matrix dimensions.

5.2. Image Denoising Applications

To validate the proposed estimator, we apply it to a classical low-level vision task: image denoising. Image denoising is a critical pre-processing step in a variety of vision applications and serves as an effective framework for evaluating statistical image modeling techniques. This task aims to reconstruct the original image denoted as $\mathbf{x}$, from its noisy observation $\mathbf{y}$, where the noise component $\mathbf{e}$ is assumed to be additive white Gaussian noise, such that $\mathbf{y} = \mathbf{x} + \mathbf{e}$.

Natural images often display significant redundancy, with recurring patterns and textures across different regions. This prior fact motivates the use of nonlocal self-similarity (NSS) techniques in image denoising. NSS relies on the observation that many patches within a natural image share similar structures, allowing denoising algorithms [ref] to re-

Algorithm 2: Image Denoising by AWNNM.

Input: Noisy image $\mathbf{y}$, tuning parameter λ;
Initialize $\hat{\mathbf{x}}^{(0)} = \mathbf{y}$, $\mathbf{y}^{(0)} = \mathbf{y}$;
for $k = 1, 2, \ldots, K$ **do**
  Residual correction $\mathbf{y}^{(k)} = \hat{\mathbf{x}}^{(k-1)} + \delta\left(\mathbf{y} - \hat{\mathbf{y}}^{(k-1)}\right)$;
  for each patch $\mathbf{y}_i$ in $\mathbf{y}^{(k)}$ **do**
    Search similar patches and group them into $\mathbf{Y}_i$;
    Apply Algorithm 1 to $\mathbf{Y}_i$ to estimate $\mathbf{X}_i$;
  end for
  Elementwisely average the estimates $\mathbf{X}_i$ for each pixel to form the denoised image $\hat{\mathbf{x}}^{(k)}$;
end for
Output: $\hat{\mathbf{x}}^{(K)}$

construct each patch effectively by referencing information from similar patches. By aggregating these reconstructed patches, NSS techniques improve the overall quality of denoising.

The structural redundancy among similar patches introduces a low-rank tendency within each group of patches, allowing specific patch matrices to be represented with fewer components, retaining essential details while reducing noise. Specifically, for a given local patch $\mathbf{y}_i$ in the image, similar patches from other regions can be grouped into a matrix $\mathbf{Y}_i$. Each group of similar patches, denoted as $\mathbf{Y}_i$, can be approximated by a low-rank component $\mathbf{X}_i$ addition to Gaussian noise $\mathbf{E}_i$. We apply the proposed estimator to $\mathbf{Y}_i$ to estimate $\mathbf{X}_i$, thereby performing image denoising. This process is formulated as follows:

$$\widehat{\mathbf{X}}_i \in \arg\min_{\mathbf{X}_i} \frac{1}{2} \|\mathbf{X}_i - \mathbf{Y}_i\|_{\mathrm{F}}^2 + \sum_j p_{\lambda_i}\left(\sigma_j\left(\mathbf{X}_i\right)\right). \quad (5)$$

By applying this denoising procedure to each patch and subsequently aggregating the denoised patches, we reconstruct the image $\mathbf{x}$. In practice, multiple iterations of this reconstruction process across all image patches can further enhance the denoising results. The whole denoising algorithm is summarized in Algorithm 2.

We compare the proposed AWNNM based image restoration algorithm with several classical restoration methods, including NNM, WNNM, BM3D [8], EPLL [51], PCLR [5], PGPD [43], STROLLR [41], and SNSS [46].

- BM3D: Groups similar patches into 3D blocks for collaborative filtering, effectively preserving details, but may struggle with complex noise patterns.
- NNM minimizes the nuclear norm to promote low-rank structures. It works well in low-rank areas but introduces bias in high-rank regions.
- EPLL maximizes the expected log-likelihood of patches under a probabilistic prior, yielding good results but can be computationally intensive.
- WNNM extends NNM by weighting singular values, providing flexibility for different noise levels but still limited by non-uniform noise.
- PCLR clusters patches for localized low-rank regularization, effective but sensitive to cluster accuracy in complex regions.
- PGPD uses patch group priors to reduce noise, preserving structure but reliant on effective priors.
- STROLLR maintains non-local similarities for structural fidelity yet may struggle with unique features.
- SNSS preserves self-similarity, ideal for textures but less effective in isolated features.
- AWNNM, the proposed algorithm, extends WNNM with adaptive weights via non-convex penalties, which enhances flexibility and noise handling across various structures.

5.2.1. Parameter Setting

The proposed AWNNM algorithm involves several vital parameters, namely the regularization parameter δ, the number of iterations K, and patch size, each tuned based on the noise level. The iterative regularization parameter δ is fixed to 0.1 for all noise levels. Both K and patch size vary with noise level: larger patches and a higher number of iterations are required for higher noise levels to achieve optimal results. In practice, we set the patch size and iteration number as follows: for $\sigma_n \leq 20$, $20 < \sigma_n \leq 40$, $40 < \sigma_n \leq 60$ and $60 < \sigma_n$, the patch sizes are 6×6, 7×7, 8×8 and 9×9, respectively. The number of iterations K is set to 8, 10,12, and 12, according to the noise levels, on the noise levels. For competing methods, we obtained source codes from the original authors and used default parameter settings to ensure a fair comparison across all methods. Our non-convex approach supports various penalty functions. In this experiment, we use the SCAD function with $a = 3.7$ to generate adaptive weights. The tuning parameter λ is chosen by five-fold cross-validation.

5.2.2. Results on Synthetic Images

We benchmark against several classical denoising methods on 20 commonly used test images with synthetic noises. Additive White Gaussian Noise (AWGN) with zero mean and variance σ_n^2 is applied to these test images to simulate various noise conditions. We test four different noise levels: low noise ($\sigma_n = 10$), medium noise ($\sigma_n = 30$ and $\sigma_n = 50$), and high noise ($\sigma_n = 100$).

The Peak Signal-to-Noise Ratio (PSNR) is the primary metric to assess the quality of the denoised images quantitatively. PSNR is defined as follows:

$$\mathrm{PSNR} = 10 \cdot \log\left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}}\right),$$

where MAX represents the maximum possible pixel value of the image, and MSE (Mean Squared Error) is the average

Table 1. Denoising results (PSNR) by different methods.

	$\sigma_n = 10$										$\sigma_n = 30$									
	BM3D	NNM	EPLL	WNNM	PCLR	PGPD	STROLLR	TWSC	SNSS	AWNNM	BM3D	NNM	EPLL	WNNM	PCLR	PGPD	STROLLR	TWSC	SNSS	AWNNM
Alley	32.29	26.47	32.45	32.76	32.67	32.45	32.22	32.78	32.41	**32.79**	26.47	24.70	27.11	27.46	27.36	27.19	26.88	27.29	27.27	**27.47**
Baboon	30.43	23.99	30.32	30.50	30.47	30.23	30.29	30.42	30.07	**30.52**	24.56	22.61	24.04	24.64	24.56	24.23	24.56	24.42	24.30	**24.66**
Barbara	30.92	28.34	32.63	34.54	34.23	33.60	33.89	33.79	34.49	**34.57**	25.87	25.69	26.91	28.60	28.09	27.79	28.37	27.94	28.54	**28.63**
Boat	33.49	27.27	33.18	33.46	33.47	33.12	33.00	33.45	33.17	**33.58**	27.05	25.06	27.60	27.70	27.77	27.48	27.32	26.95	27.49	**27.80**
Book	33.69	30.48	36.28	36.78	36.71	36.22	35.97	36.39	36.23	**36.79**	28.25	28.19	31.31	31.58	31.83	31.27	30.98	31.35	31.70	**31.86**
Cameraman	33.08	26.53	34.40	34.46	34.46	34.14	33.79	34.04	34.30	**34.50**	27.54	24.60	28.39	28.79	28.83	28.53	28.16	27.39	28.57	**28.81**
Couple	34.14	27.12	33.11	33.44	33.44	33.24	32.75	32.62	33.21	**33.45**	28.53	24.81	27.37	27.48	27.48	27.32	27.17	26.87	27.28	**27.49**
Dice	36.06	35.77	40.54	41.08	41.20	40.04	40.34	40.74	41.02	**41.24**	31.17	32.04	34.58	34.92	35.38	35.12	34.50	28.68	35.28	**35.39**
F16	34.04	28.06	34.32	34.59	34.63	34.40	33.96	33.94	34.42	**34.64**	28.31	25.73	28.65	28.74	28.79	28.62	28.27	27.12	28.62	**28.83**
Fingerprint	30.17	27.27	30.08	30.66	30.45	30.45	30.36	29.91	30.42	**30.69**	24.16	23.34	24.02	24.67	24.54	24.57	24.69	24.49	24.57	**24.80**
Girl	33.08	36.18	39.35	39.59	39.66	38.75	39.44	39.44	**39.73**	39.67	27.27	32.06	34.12	34.03	34.53	34.52	34.51	31.80	34.34	**34.56**
Hallway	39.05	34.21	38.85	39.58	39.45	38.89	38.97	39.79	39.43	**39.81**	34.49	30.22	33.23	33.62	33.83	33.59	33.08	32.07	**34.02**	33.96
House	32.82	32.36	35.76	36.92	36.83	36.56	36.67	36.35	36.89	**36.94**	26.94	29.29	31.41	32.30	32.17	32.24	31.67	31.31	**32.68**	32.61
Pentagon	33.35	26.68	30.77	31.69	31.40	30.85	30.85	31.18	31.19	**33.70**	26.38	24.40	25.50	26.29	25.84	25.63	25.69	24.67	25.94	**26.41**
Peppers	33.45	28.16	34.58	34.98	34.96	34.65	34.32	34.94	34.86	**35.01**	28.35	25.81	29.27	29.43	29.56	26.80	28.90	28.10	29.39	**29.58**
Plaza	34.19	30.62	34.52	34.68	34.64	34.27	34.20	34.37	34.33	**34.71**	28.55	28.23	29.95	29.98	30.03	29.87	29.52	28.03	29.78	**30.10**
Statue	30.32	28.65	35.48	35.68	35.74	35.30	35.02	35.34	35.45	**35.69**	24.77	26.58	29.59	29.57	29.68	29.55	29.11	26.44	29.45	**29.72**
Traffic	32.73	26.59	33.01	33.32	33.30	32.88	32.52	32.79	32.91	**33.33**	25.86	24.52	27.38	27.55	27.56	25.30	26.87	25.21	27.29	**27.57**
Validemossa	32.79	26.05	33.49	33.79	33.82	33.49	33.15	33.81	33.54	**33.85**	27.84	24.18	27.51	27.69	27.72	27.50	27.15	27.13	27.50	**27.85**
Yard	34.36	28.90	34.16	34.77	34.53	34.36	34.26	34.19	34.53	**34.79**	29.97	26.53	29.07	29.21	29.22	29.09	29.02	28.86	29.00	**30.01**
AVE.	33.22	28.98	34.36	34.86	34.80	34.39	34.31	34.51	34.63	**35.01**	27.62	26.43	28,87	29.22	29.24	28.81	28.82	27.81	29.15	**29.40**
	$\sigma_n = 50$										$\sigma_n = 100$									
	BM3D	NNM	EPLL	WNNM	PCLR	PGPD	STROLLR	TWSC	SNSS	AWNNM	BM3D	NNM	EPLL	WNNM	PCLR	PGPD	STROLLR	TWSC	SNSS	AWNNM
Alley	23.84	23.25	25.05	25.36	25.27	25.17	24.91	25.35	25.23	**25.42**	20.67	21.10	22.63	22.94	22.84	22.83	22.49	22.36	23.00	**23.01**
Baboon	22.81	21.60	22.47	22.70	22.68	22.51	22.64	22.60	22.51	**22.72**	20.56	19.99	20.85	20.92	20.87	20.97	20.94	20.86	20.85	**20.98**
Barbara	24.02	23.59	25.11	26.13	25.61	22.72	25.85	25.97	26.07	**26.19**	21.94	20.54	22.52	22.82	22.78	22.72	22.57	22.07	22.85	**22.91**
Boat	25.95	23.16	25.24	25.38	25.36	25.17	25.06	24.94	25.29	**25.96**	21.54	20.92	22.48	22.61	22.63	22.62	22.37	21.97	**22.68**	22.67
Book	25.68	26.45	28.99	29.39	29.48	29.05	28.56	29.24	29.35	**29.51**	21.39	23.54	25.83	26.59	26.34	26.09	25.53	26.30	**26.75**	26.69
Cameraman	25.91	23.12	26.09	26.45	26.56	26.46	25.83	26.33	26.38	**26.67**	21.71	21.05	22.94	23.37	23.48	23.23	22.71	23.24	23.39	**23.51**
Couple	25.39	23.08	24.97	25.10	25.04	25.01	24.86	24.86	25.00	**25.66**	20.21	20.78	22.35	22.52	22.47	22.48	22.13	21.91	22.42	**22.56**
Dice	31.69	29.49	31.65	32.76	32.52	32.52	31.44	32.64	33.04	**33.30**	23.52	26.06	27.99	29.17	28.54	28.73	27.53	28.80	**29.67**	29.61
F16	25.04	23.96	26.10	26.34	26.21	26.12	25.71	26.28	26.24	**26.37**	20.46	21.52	23.06	23.39	23.24	23.30	23.09	22.51	23.27	**23.47**
Fingerprint	26.03	20.62	21.52	22.54	22.33	22.30	22.51	22.58	22.38	**22.61**	19.79	18.11	17.59	20.23	19.53	19.57	20.25	19.86	19.88	**20.41**
Girl	24.68	29.96	31.66	32.05	32.30	32.39	31.79	30.57	32.16	**32.32**	21.51	27.84	28.41	29.13	29.03	29.11	28.22	28.32	**29.54**	29.48
Hallway	30.24	28.00	30.60	31.35	31.09	31.06	30.12	28.18	31.48	**31.38**	22.35	25.04	27.15	27.92	27.46	27.64	26.54	27.14	**28.31**	**28.31**
House	24.26	26.59	29.02	30.30	29.78	29.93	29.06	30.14	**30.52**	30.43	24.09	23.34	25.50	26.71	25.96	26.16	25.35	26.10	**27.09**	26.98
Pentagon	24.59	22.53	23.71	24.35	23.96	24.02	23.96	23.69	24.26	**24.37**	20.82	21.17	21.83	22.10	21.95	22.13	21.85	21.77	22.15	**22.16**
Peppers	26.97	23.60	26.76	26.93	27.03	26.80	26.51	27.04	26.94	**27.11**	22.23	20.50	23.20	23.50	23.69	23.44	22.79	22.10	23.71	**23.72**
Plaza	25.65	26.68	28.18	28.22	28.30	28.22	27.78	28.01	28.24	**28.33**	20.67	24.98	25.85	25.83	25.83	26.03	25.50	25.81	25.83	**25.86**
Statue	26.57	25.02	27.25	27.41	27.35	27.31	27.12	27.30	27.31	**27.42**	19.26	22.72	24.28	24.66	24.43	24.58	24.09	24.13	24.64	**24.69**
Traffic	28.95	22.84	25.25	25.45	25.42	25.30	24.84	25.27	25.30	**25.49**	22.96	20.51	22.47	22.70	22.68	22.54	21.98	22.27	22.63	**22.72**
Validemossa	27.28	22.57	25.06	25.28	25.29	25.15	24.88	25.27	25.19	**25.35**	20.32	20.32	21.89	22.30	22.26	22.15	21.94	22.26	22.32	**22.36**
Yard	29.51	24.96	27.02	27.13	27.11	27.08	26.80	27.09	26.97	**27.17**	24.02	22.66	24.12	24.43	24.41	24.40	23.85	23.06	24.38	**24.49**
AVE.	26.25	24.55	26.59	27.04	26.93	26.71	26.51	26.67	26.99	**27.18**	21.60	22.13	23.64	24.19	24.02	24.04	23.59	23.64	24.25	**24.33**

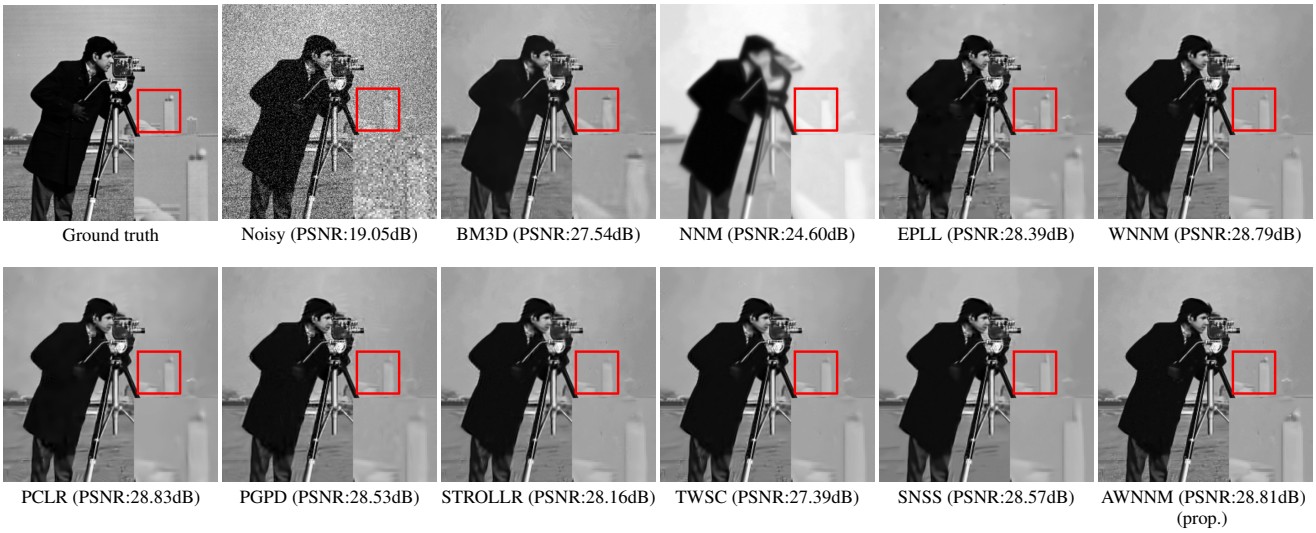


Figure 4. Denoising results on image *Cameraman* by different methods (noise level $\sigma_n = 30$).

squared difference between the pixels of the original and denoised images.

Table 1 shows that the proposed AWNNM method almost achieves the highest PSNR, demonstrating a significant average improvement over the other methods. Fig. 4 and Fig. 5 illustrate a visual comparison, where AWNNM effectively recovers subtle features, such as the crack on the dice, which are missed by the other methods.

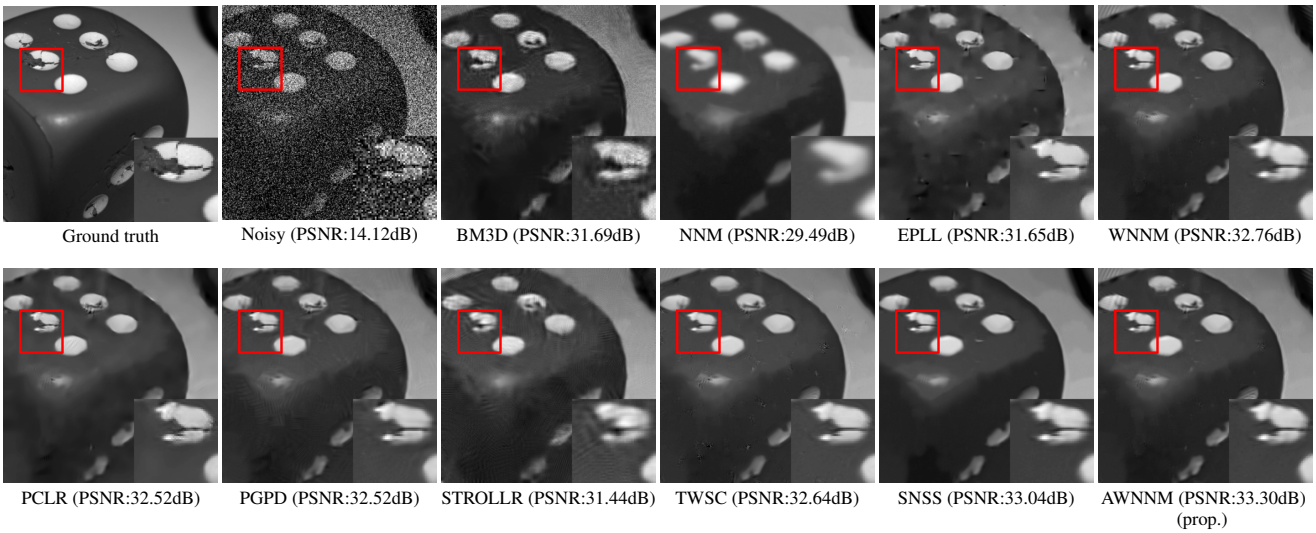


Figure 5. Denoising results on image *Dice* by different methods (noise level $\sigma_n = 50$).

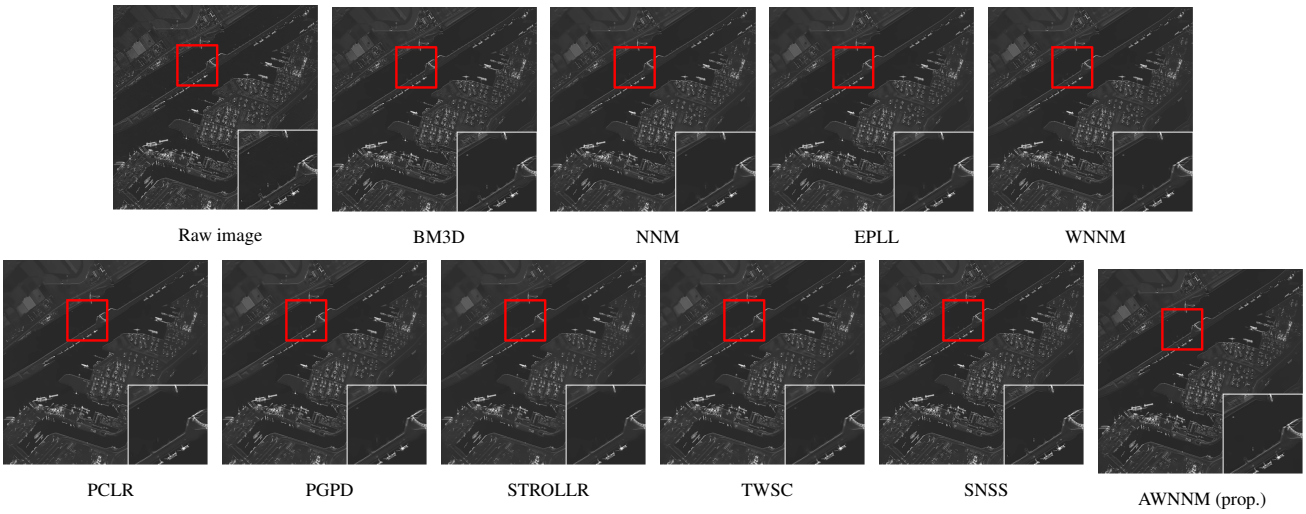


Figure 6. Denoising results on a real SAR image by all competing methods.

5.2.3. Results on Real-World Images

We also evaluate the denoising performance of the competing methods on real images[1], including a grayscale SAR image. Fig. 6 shows a comparison of denoising results obtained by the competing methods. It is evident that the proposed AWNNM method preserves local structures with clearer textures in the image and produces the fewest visual artifacts compared to other methods.

6. Conclusion

In this paper, we have studied the low-rank matrix recovery problem by introducing a non-convex penalty function on the singular values to approximate the rank function. Our proposed AWNNM method has been proven to exhibit a superior statistical rate of convergence compared to existing approaches. We then applied the proposed AWNNM method to the image-denoising task. Experimental results on widely used benchmark images demonstrate that AWNNM consistently outperforms state-of-the-art penalty-based algorithms. To the best of our knowledge, this is the first work to establish a statistical guarantee for image denoising based on low-rank matrix denoising.

[1]The SAT was downloaded at https://www.iceye.com/resources/datasets

References

[1] Aeron M Buchanan and Andrew W Fitzgibbon. Damped newton algorithms for matrix factorization with missing data. In *IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 316–322. IEEE, 2005.

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

[3] Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[4] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[5] Fei Chen, Lei Zhang, and Huimin Yu. External patch prior guided internal clustering for image denoising. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 603–611, 2015.

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

[7] Yongyong Chen, Xiaolin Xiao, and Yicong Zhou. Low-rank quaternion approximation for color image processing. *IEEE Transactions on Image Processing*, 29:1426–1439, 2020.

[8] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16(8):2080–2095, 2007.

[9] David L Donoho, Matan Gavish, and Andrea Montanari. The phase transition of matrix recovery from gaussian measurements matches the minimax mse of matrix denoising. *Proceedings of the National Academy of Sciences*, 110(21): 8405–8410, 2013.

[10] Anders Eriksson and Anton van den Hengel. Efficient computation of robust low-rank matrix approximations in the presence of missing data using the L_1 norm. In *IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 771–778. IEEE, 2010.

[11] 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.

[12] 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.*, pages 4734–4739. IEEE, 2001.

[13] Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. Weighted nuclear norm minimization with application to image denoising. In *IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 2862–2869, 2014.

[14] Huan Gui, Jiawei Han, and Quanquan Gu. Towards faster rates and oracle property for low-rank matrix estimation. In *International Conference on Machine Learning*, pages 2300–2309. PMLR, 2016.

[15] Yao Hu, Debing Zhang, Jieping Ye, Xuelong Li, and Xiaofei He. Fast and accurate matrix completion via truncated nuclear norm regularization. *IEEE transactions on pattern analysis and machine intelligence*, 35(9):2117–2130, 2013.

[16] David R Hunter and Kenneth Lange. A tutorial on mm algorithms. *The American Statistician*, 58(1):30–37, 2004.

[17] Hui Ji, Chaoqiang Liu, Zuowei Shen, and Yuhong Xu. Robust video denoising using low rank matrix completion. In *IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 1791–1798, 2010.

[18] Qifa Ke and Takeo Kanade. Robust L_1 norm factorization in the presence of outliers and missing data by alternative convex programming. In *IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 739–746. IEEE, 2005.

[19] N Kishore Kumar and Jan Schneider. Literature survey on low rank approximation of matrices. *Linear and Multilinear Algebra*, 65(11):2212–2244, 2016.

[20] Joonseok Lee, Seungyeon Kim, Guy Lebanon, Yoram Singer, and Samy Bengio. Llorma: Local low-rank matrix approximation. *Journal of Machine Learning Research*, 17 (1):442–465, 2016.

[21] Yixiao Li, Yifan Yu, Qingru Zhang, Chen Liang, Pengcheng He, Weizhu Chen, and Tuo Zhao. Losparse: Structured compression of large language models based on low-rank and sparse approximation. In *International Conference on Machine Learning*, pages 20336–20350. PMLR, 2023.

[22] Zhouchen Lin, Risheng Liu, and Zhixun Su. Linearized alternating direction method with adaptive penalty for low-rank representation. *Advances in Neural Information Processing Systems*, pages 612–620, 2011.

[23] Guangcan Liu, Zhouchen Lin, and Yong Yu. Robust subspace segmentation by low-rank representation. In *Proceedings of the 27th International Conference on Machine Learning*, pages 663–670, 2010.

[24] Risheng Liu, Zhouchen Lin, Fernando De la Torre, and Zhixun Su. Fixed-rank representation for unsupervised visual learning. In *IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 598–605. IEEE, 2012.

[25] Tianxiang Liu, Zhaosong Lu, Xiaojun Chen, and Yu-Hong Dai. An exact penalty method for semidefinite-box-constrained low-rank matrix optimization problems. *IMA Journal of Numerical Analysis*, 40(1):563–586, 2020.

[26] Xin Liu, Guoying Zhao, Jiawen Yao, and Chun Qi. Background subtraction based on low-rank and structured sparse decomposition. *IEEE Transactions on Image Processing*, 24 (8):2502–2514, 2015.

[27] Canyi Lu, Jinhui Tang, Shuicheng Yan, and Zhouchen Lin. Generalized nonconvex nonsmooth low-rank minimization. In *IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 4130–4137, 2014.

[28] Canyi Lu, Jinhui Tang, Shuicheng Yan, and Zhouchen Lin. Nonconvex nonsmooth low rank minimization via iteratively reweighted nuclear norm. *IEEE Transactions on Image Processing*, 25(2):829–839, 2015.

[29] Ivan Markovsky. Structured low-rank approximation and its applications. *Automatica*, 44(4):891–909, 2008.

[30] Ivan Markovsky. *Low rank approximation: algorithms, implementation, applications*. Springer, 2011.

[31] Bamdev Mishra, Gilles Meyer, Francis Bach, and Rodolphe Sepulchre. Low-rank optimization with trace norm penalty. *SIAM Journal on Optimization*, 23(4):2124–2149, 2013.

[32] Yuji Nakatsukasa. Fast and stable randomized low-rank matrix approximation. *arXiv preprint arXiv:2009.11392*, 2020.

[33] Feiping Nie, Hua Wang, Xiao Cai, Heng Huang, and Chris Ding. Robust matrix completion via joint schatten p-norm

and lp-norm minimization. In *2012 IEEE 12th International Conference on Data Mining*, pages 566–574. IEEE, 2012.

[34] Ankit Parekh and Ivan W Selesnick. Enhanced low-rank matrix approximation. *IEEE Signal Processing Letters*, 23(4): 493–497, 2016.

[35] Hong Peng, Shuyi Hong, Linkai Luo, Qifeng Zhou, and Xiaoqin Huang. A new approach of matrix factorization and its application in recommender systems. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 682–686, 2016.

[36] Nathan Srebro and Tommi Jaakkola. Weighted low-rank approximations. In *Proceedings of the 20th International Conference on Machine Learning*, pages 720–727, 2003.

[37] Ying Sun, Prabhu Babu, and Daniel P Palomar. Majorization-minimization algorithms in signal processing, communications, and machine learning. *IEEE Transactions on Signal Processing*, 65(3):794–816, 2017.

[38] Joel A Tropp and Robert J Webber. Randomized algorithms for low-rank matrix approximation: Design, analysis, and applications. *arXiv preprint arXiv:2306.12418*, 2023.

[39] Shenlong Wang, Lei Zhang, and Yan Liang. Nonlocal spectral prior model for low-level vision. In *Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part III 11*, pages 231–244. Springer, 2012.

[40] Zhaoran Wang, Han Liu, and Tong Zhang. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Annals of Statistics*, 42(6):2164, 2014.

[41] Bihan Wen, Yanjun Li, and Yoram Bresler. When sparsity meets low-rankness: Transform learning with non-local low-rank constraint for image restoration. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2297–2301. IEEE, 2017.

[42] John Wright, Arvind Ganesh, Shankar Rao, Yigang Peng, and Yi Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. *Advances in neural information processing systems*, 22, 2009.

[43] Jun Xu, Lei Zhang, Wangmeng Zuo, David Zhang, and Xiangchu Feng. Patch group based nonlocal self-similarity prior learning for image denoising. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 244–252, 2015.

[44] Ming Yang, Qilun Luo, Wen Li, and Mingqing Xiao. Nonconvex 3d array image data recovery and pattern recognition under tensor framework. *Pattern recognition*, 122:108311, 2022.

[45] Zhiyuan Zha, Xin Yuan, Bihan Wen, Jiantao Zhou, Jiachao Zhang, and Ce Zhu. From rank estimation to rank approximation: Rank residual constraint for image restoration. *IEEE Transactions on Image Processing*, 29:3254–3269, 2020.

[46] Zhiyuan Zha, Xin Yuan, Jiantao Zhou, Ce Zhu, and Bihan Wen. Image restoration via simultaneous nonlocal self-similarity priors. *IEEE Transactions on Image Processing*, 29:8561–8576, 2020.

[47] Lei Zhang and Wangmeng Zuo. Image restoration: From sparse and low-rank priors to deep priors [lecture notes]. *IEEE Signal Processing Magazine*, 34(5):172–179, 2017.

[48] Tong Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, 11(3), 2010.

[49] Xinfeng Zhang, Weisi Lin, Ruiqin Xiong, Xianming Liu, Siwei Ma, and Wen Gao. Low-rank decomposition-based restoration of compressed images via adaptive noise estimation. *IEEE Transactions on Image Processing*, 25(9):4158–4171, 2016.

[50] Xiaowei Zhou, Can Yang, Hongyu Zhao, and Weichuan Yu. Low-rank modeling and its applications in image analysis. *ACM Computing Surveys*, 47(2):36, 2014.

[51] Daniel Zoran and Yair Weiss. From learning models of natural image patches to whole image restoration. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 479–486. IEEE, 2011.

Achieving Oracle Statistical Property for Low-Rank Matrix Recovery with Application to Image Denoising

Supplementary Material

7. Backgrounds

Here, we first introduce essential notations. We denote the all-zero matrix and the identity matrix by $\mathbf{0}$ and $\mathbf{I}$, respectively, with dimensions inferred from context. Consider a matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$. The smallest of $\mathbf{X}$ are represented as $\lambda_{\min}(\mathbf{X})$. The p-norm of $\mathbf{X}$ is defined as $\|\mathbf{X}\|_p = \left(\sum_{i=1}^m \sum_{j=1}^n |X_{ij}|^p\right)^{\frac{1}{p}}$ for a real $p > 0$. Notably, the Frobenius norm and spectral norm are denoted by $\|\mathbf{X}\|_{\mathrm{F}}$ and $\|\mathbf{X}\|_2$, respectively, while $\|\mathbf{X}\|_1$ represents the sum of the absolute values of all entries of $\mathbf{X}$. The maximum-absolute-value norm of $\mathbf{X}$ is expressed as $\|\mathbf{X}\|_{\max}$, and the minimum-absolute-value norm as $\|\mathbf{X}\|_{\min}$. The nuclear norm of $\mathbf{X}$, defined as the sum of the singular values of $\mathbf{X}$, is expressed as $\|\mathbf{X}\|_* = \sum_{i=1}^d \sigma_i(\mathbf{X})$. The vectorization of $\mathbf{X}$, achieved by stacking its columns, is denoted as $\mathrm{vec}(\mathbf{X})$. Furthermore, the Kronecker product and Hadamard product (also referred to as the entry-wise product) of matrices $\mathbf{X}$ and $\mathbf{Y}$ are denoted by $\mathbf{X}\otimes\mathbf{Y}$ and $\mathbf{X}\odot\mathbf{Y}$, respectively. The Euclidean inner product is defined as $\langle\mathbf{X},\mathbf{Y}\rangle := \mathrm{tr}\left(\mathbf{X}\mathbf{Y}^\top\right)$, where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. For a real-valued function $f(\mathbf{X})$, the gradient $\nabla f(\mathbf{X})$ is a $d\times d$ matrix with the (i,j)-th element given by $\frac{\partial}{\partial X_{ij}} f(\mathbf{X})$, denoted by $\nabla_{ij} f(\mathbf{X})$, while $\nabla^2 f(\mathbf{X})$ represents the $d^2\times d^2$ Hessian matrix. For an index set $\mathcal{E}$, its cardinality is represented as $|\mathcal{E}|$, while its complement is denoted by $\overline{\mathcal{E}}$. Given a vector $\mathbf{w}$, $\mathbf{w}_{\mathcal{E}}$ is defined such that its i-th entry equals to w_i if $i\in\mathcal{E}$ and is 0 otherwise. For the functionals $f(n)$ and $g(n)$, we express $f(n)\gtrsim g(n)$ if $f(n)\geq cg(n)$, $f(n)\lesssim g(n)$ if $f(n)\leq Cg(n)$, and $f(n)\asymp g(n)$ if $cg(n)\leq f(n)\leq Cg(n)$ for some positive constants c and C. Additionally, $\mathcal{O}_p(\cdot)$ is utilized to indicate boundedness in probability.

We recall the problem presented in (3):

$$\widehat{\mathbf{X}}^{(k)} \in \arg\min_{\mathbf{X}} \left\{ f(\mathbf{X}) + \sum_{i=1}^d w_i^{(k-1)} \sigma_i(\mathbf{X}) \right\}.$$

This weighted nuclear norm penalized problem can be generally expressed as:

$$\widehat{\mathbf{X}}^{(k)} \in \arg\min_{\mathbf{X}} \left\{ f(\mathbf{X}) + \|\mathbf{w}\odot\boldsymbol{\sigma}(\mathbf{X})\|_1 \right\}, \tag{6}$$

where $\mathbf{w}\in\mathbb{R}^d$ is a vector of regularization parameters with $w_i\in[0,\lambda]$ for $i=1,\ldots,d$.

8. Proofs of Statistical Theory

8.1. Technical Lemmata

Lemma 8. *Consider the general problem in (6). Assume that there exists a set $\mathcal{E}$ such that*

$$\mathcal{S}^*\subseteq\mathcal{E},\ |\mathcal{E}|\leq 2r^*,\ \textit{and}\ \|\mathbf{w}_{\overline{\mathcal{E}}}\|_{\min}\geq\frac{\lambda}{2}.$$

If $\lambda\geq 2\|\nabla f(\mathbf{X}^*)\|_2$, then the optimal solution $\widehat{\mathbf{X}}$ satisfies

$$\|\widehat{\mathbf{X}}-\mathbf{X}^*\|_{\mathrm{F}}\leq\|\mathbf{w}_{\mathcal{S}^*}\|_2+\|\Pi_{\mathcal{F}_{\mathcal{E}}}(\nabla f(\mathbf{X}^*))\|_{\mathrm{F}}\leq\frac{3}{2}\lambda\sqrt{r^*}.$$

Proof. Given that $\nabla^2 f(\mathbf{X}) = \mathbf{I}\otimes\mathbf{I}$, the mean value theorem guarantees the existence of $\rho\in[0,1]$ such that

$$\begin{aligned}&\langle\nabla f(\mathbf{X})-\nabla f(\mathbf{X}^*),\boldsymbol{\Delta}\rangle\\ =&\mathrm{vec}^\top(\boldsymbol{\Delta})\nabla^2 f(\mathbf{X}^*+\rho\boldsymbol{\Delta})\mathrm{vec}(\boldsymbol{\Delta}) = \|\boldsymbol{\Delta}\|_{\mathrm{F}}^2,\end{aligned}$$

where $\boldsymbol{\Delta}=\mathbf{X}-\mathbf{X}^*$.

Hence, we obtain

$$\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2\leq\langle\nabla f(\mathbf{X})-\nabla f(\mathbf{X}^*),\boldsymbol{\Delta}\rangle. \tag{7}$$

Let $\widehat{\boldsymbol{\Delta}}=\widehat{\mathbf{X}}-\mathbf{X}^*$. Applying the inequality (7) with $\mathbf{X}=\widehat{\mathbf{X}}$ and $\boldsymbol{\Delta}=\widehat{\boldsymbol{\Delta}}$ yields

$$\left\|\widehat{\boldsymbol{\Delta}}\right\|_{\mathrm{F}}^2\leq\left\langle\nabla f\left(\widehat{\mathbf{X}}\right)-\nabla f(\mathbf{X}^*),\widehat{\boldsymbol{\Delta}}\right\rangle.$$

Since $\widehat{\mathbf{X}}$ is the minimizer of the loss $f(\mathbf{X})$ and the penalty $\sum_{i=1}^m w_i^{(k-1)}\sigma_i(\mathbf{X}) = \langle\mathrm{Diag}(\mathbf{w}),\mathrm{Diag}(\boldsymbol{\sigma}(\mathbf{X}))\rangle$, we have the optimality condition $\nabla f\left(\widehat{\mathbf{X}}\right)+\mathrm{Diag}(\mathbf{w})\cdot\boldsymbol{\Xi}=\mathbf{0}$ for the subgradient $\boldsymbol{\Xi}\in\partial\|\mathrm{Diag}(\boldsymbol{\sigma}(\mathbf{X}))\|_1$ at $\mathbf{X}=\widehat{\mathbf{X}}$. Therefore, we have

$$\left\|\widehat{\boldsymbol{\Delta}}\right\|_{\mathrm{F}}^2\leq\left\langle-\nabla f(\mathbf{X}^*)-\mathrm{Diag}(\mathbf{w})\cdot\boldsymbol{\Xi},\widehat{\boldsymbol{\Delta}}\right\rangle \tag{8}$$

$$=-\underbrace{\left\langle\nabla f(\mathbf{X}^*),\widehat{\boldsymbol{\Delta}}\right\rangle}_{\mathrm{I}}-\underbrace{\left\langle\mathrm{Diag}(\mathbf{w})\cdot\boldsymbol{\Xi},\widehat{\boldsymbol{\Delta}}\right\rangle}_{\mathrm{II}}. \tag{9}$$

It remains to bound terms I and II, respectively. For term I, projecting the support of $\nabla f(\mathbf{X}^*)$ and $\widehat{\boldsymbol{\Delta}}$ to $\mathcal{E}$ and $\overline{\mathcal{E}}$, and

then using the matrix Hölder inequality, we obtain

$$\begin{aligned}
&\left\langle \nabla f\left(\mathbf{X}^*\right), \widehat{\mathbf{\Delta}} \right\rangle \\
=&\left\langle \Pi_{\mathcal{F}_{\mathcal{E}}}\left(\nabla f\left(\mathbf{X}^*\right)\right), \Pi_{\mathcal{F}_{\mathcal{E}}}\left(\widehat{\mathbf{\Delta}}\right)\right\rangle \\
&\quad + \left\langle \Pi_{\mathcal{F}_{\overline{\mathcal{E}}}}\left(\nabla f\left(\mathbf{X}^*\right)\right), \Pi_{\mathcal{F}_{\overline{\mathcal{E}}}}\left(\widehat{\mathbf{\Delta}}\right)\right\rangle \\
\geq& -\left\|\Pi_{\mathcal{F}_{\mathcal{E}}}\left(\nabla f\left(\mathbf{X}^*\right)\right)\right\|_{\mathrm{F}} \left\|\Pi_{\mathcal{F}_{\mathcal{E}}}\left(\widehat{\mathbf{\Delta}}\right)\right\|_{\mathrm{F}} \\
&\quad - \left\|\Pi_{\mathcal{F}_{\overline{\mathcal{E}}}}\left(\nabla f\left(\mathbf{X}^*\right)\right)\right\|_2 \left\|\Pi_{\mathcal{F}_{\overline{\mathcal{E}}}}\left(\widehat{\mathbf{\Delta}}\right)\right\|_*.
\end{aligned}$$

For term II, projecting the support of $\mathrm{Diag}(\mathbf{w})\cdot\mathbf{\Xi}$ and $\widehat{\mathbf{\Delta}}$ to $\mathcal{S}^*$, $\mathcal{E}\backslash\mathcal{S}^*$, and $\overline{\mathcal{E}}$, we obtain

$$\begin{aligned}
&\left\langle \mathrm{Diag}(\mathbf{w})\cdot\mathbf{\Xi}, \widehat{\mathbf{\Delta}} \right\rangle \\
=&\left\langle \Pi_{\mathcal{F}_{\mathcal{S}^*}}\left(\mathrm{Diag}(\mathbf{w})\cdot\mathbf{\Xi}\right), \Pi_{\mathcal{F}_{\mathcal{S}^*}}\left(\widehat{\mathbf{\Delta}}\right)\right\rangle \\
&+ \left\langle \Pi_{\mathcal{F}_{\mathcal{E}\backslash\mathcal{S}^*}}\left(\mathrm{Diag}(\mathbf{w})\cdot\mathbf{\Xi}\right), \Pi_{\mathcal{F}_{\mathcal{E}\backslash\mathcal{S}^*}}\left(\widehat{\mathbf{\Delta}}\right)\right\rangle \\
&+ \left\langle \Pi_{\mathcal{F}_{\overline{\mathcal{E}}}}\left(\mathrm{Diag}(\mathbf{w})\cdot\mathbf{\Xi}\right), \Pi_{\mathcal{F}_{\overline{\mathcal{E}}}}\left(\widehat{\mathbf{\Delta}}\right)\right\rangle \\
=&\left\langle \Pi_{\mathcal{F}_{\mathcal{S}^*}}\left(\mathrm{Diag}(\mathbf{w})\cdot\mathbf{\Xi}\right), \Pi_{\mathcal{F}_{\mathcal{S}^*}}\left(\widehat{\mathbf{\Delta}}\right)\right\rangle \\
&+ \left\langle \Pi_{\mathcal{F}_{\overline{\mathcal{E}}}}\left(\mathrm{Diag}(\mathbf{w})\cdot\mathbf{\Xi}\right), \Pi_{\mathcal{F}_{\overline{\mathcal{E}}}}\left(\widehat{\mathbf{\Delta}}\right)\right\rangle \\
\geq& -\left\|\Pi_{\mathcal{F}_{\mathcal{S}^*}}\left(\mathrm{Diag}(\mathbf{w})\cdot\mathbf{\Xi}\right)\right\|_{\mathrm{F}} \left\|\Pi_{\mathcal{F}_{\mathcal{S}^*}}\left(\widehat{\mathbf{\Delta}}\right)\right\|_{\mathrm{F}} \\
&+ \left\langle \Pi_{\mathcal{F}_{\overline{\mathcal{E}}}}\left(\mathrm{Diag}\left(\mathbf{w}\right)\right), \Pi_{\mathcal{F}_{\overline{\mathcal{E}}}}\left(\widehat{\mathbf{\Delta}}\right)\right\rangle \\
\geq& -\left\|\mathbf{w}_{\mathcal{S}^*}\right\|_2 \left\|\Pi_{\mathcal{F}_{\mathcal{S}^*}}\left(\widehat{\mathbf{\Delta}}\right)\right\|_{\mathrm{F}} + \left\|\mathbf{w}_{\overline{\mathcal{E}}}\right\|_{\min} \left\|\Pi_{\mathcal{F}_{\overline{\mathcal{E}}}}\left(\widehat{\mathbf{\Delta}}\right)\right\|_*,
\end{aligned}$$

where the second equality is due to the fact that

$$\left\langle \Pi_{\mathcal{F}_{\mathcal{E}\backslash\mathcal{S}^*}}\left(\mathrm{Diag}(\mathbf{w})\cdot\mathbf{\Xi}\right), \Pi_{\mathcal{F}_{\mathcal{E}\backslash\mathcal{S}^*}}\left(\widehat{\mathbf{\Delta}}\right)\right\rangle = 0,$$

and the last inequality is due to

$$\begin{aligned}
&\left\langle \Pi_{\mathcal{F}_{\overline{\mathcal{E}}}}\left(\mathrm{Diag}(\mathbf{w})\cdot\mathbf{\Xi}\right), \Pi_{\mathcal{F}_{\overline{\mathcal{E}}}}\left(\widehat{\mathbf{\Delta}}\right)\right\rangle \\
=&\left\langle \Pi_{\mathcal{F}_{\overline{\mathcal{E}}}}\left(\mathrm{Diag}\left(\mathbf{w}\right)\right), \Pi_{\mathcal{F}_{\overline{\mathcal{E}}}}\left(\widehat{\mathbf{\Sigma}}\right)\right\rangle \\
=&\left\langle \Pi_{\mathcal{F}_{\overline{\mathcal{E}}}}\left(\mathrm{Diag}\left(\mathbf{w}\right)\right), \Pi_{\mathcal{F}_{\overline{\mathcal{E}}}}\left(\widehat{\mathbf{\Delta}}\right)\right\rangle \\
\geq& \left\|\mathbf{w}_{\overline{\mathcal{E}}}\right\|_{\min} \left\|\Pi_{\mathcal{F}_{\overline{\mathcal{E}}}}\left(\widehat{\mathbf{\Delta}}\right)\right\|_*.
\end{aligned}$$

Substituting the above results into (9) yields

$$\begin{aligned}
\left\|\widehat{\mathbf{\Delta}}\right\|_{\mathrm{F}}^2 \leq& \left\|\Pi_{\mathcal{F}_{\mathcal{E}}}\left(\nabla f\left(\mathbf{X}^*\right)\right)\right\|_{\mathrm{F}} \left\|\Pi_{\mathcal{F}_{\mathcal{E}}}\left(\widehat{\mathbf{\Delta}}\right)\right\|_{\mathrm{F}} \\
&+ \left\|\Pi_{\mathcal{F}_{\overline{\mathcal{E}}}}\left(\nabla f\left(\mathbf{X}^*\right)\right)\right\|_2 \left\|\Pi_{\mathcal{F}_{\overline{\mathcal{E}}}}\left(\widehat{\mathbf{\Delta}}\right)\right\|_* \\
&- \left\|\mathbf{w}_{\overline{\mathcal{E}}}\right\|_{\min} \left\|\Pi_{\mathcal{F}_{\overline{\mathcal{E}}}}\left(\widehat{\mathbf{\Delta}}\right)\right\|_* \\
&+ \left\|\mathbf{w}_{\mathcal{S}^*}\right\|_2 \left\|\Pi_{\mathcal{F}_{\mathcal{S}^*}}\left(\widehat{\mathbf{\Delta}}\right)\right\|_{\mathrm{F}} \\
\leq& \left(\left\|\Pi_{\mathcal{F}_{\mathcal{E}}}\left(\nabla f\left(\mathbf{X}^*\right)\right)\right\|_{\mathrm{F}} + \left\|\mathbf{w}_{\mathcal{S}^*}\right\|_2\right) \left\|\Pi_{\mathcal{F}_{\mathcal{E}}}\left(\widehat{\mathbf{\Delta}}\right)\right\|_{\mathrm{F}} \\
&+ \left\|\Pi_{\mathcal{F}_{\overline{\mathcal{E}}}}\left(\widehat{\mathbf{\Delta}}\right)\right\|_* \left(\left\|\nabla f\left(\mathbf{X}^*\right)\right\|_2 - \left\|\mathbf{w}_{\overline{\mathcal{E}}}\right\|_{\min}\right) \\
\leq& \left(\left\|\Pi_{\mathcal{F}_{\mathcal{E}}}\left(\nabla f\left(\mathbf{X}^*\right)\right)\right\|_{\mathrm{F}} + \left\|\mathbf{w}_{\mathcal{S}^*}\right\|_2\right) \left\|\Pi_{\mathcal{F}_{\mathcal{E}}}\left(\widehat{\mathbf{\Delta}}\right)\right\|_{\mathrm{F}} \\
\leq& \left(\left\|\Pi_{\mathcal{F}_{\mathcal{E}}}\left(\nabla f\left(\mathbf{X}^*\right)\right)\right\|_{\mathrm{F}} + \left\|\mathbf{w}_{\mathcal{S}^*}\right\|_2\right) \left\|\widehat{\mathbf{\Delta}}\right\|_{\mathrm{F}}, \quad (10)
\end{aligned}$$

where the third inequality is due to $\left\|\mathbf{w}_{\overline{\mathcal{E}}}\right\|_{\min} \geq \frac{\lambda}{2} \geq \left\|\nabla f\left(\mathbf{\Sigma}^*\right)\right\|_2$. Dividing by $\left\|\widehat{\mathbf{\Delta}}\right\|_{\mathrm{F}}$ on both sides of the inequality (10), we have

$$\begin{aligned}
\left\|\widehat{\mathbf{\Delta}}\right\|_{\mathrm{F}} \leq& \left\|\Pi_{\mathcal{F}_{\mathcal{E}}}\left(\nabla f\left(\mathbf{X}^*\right)\right)\right\|_{\mathrm{F}} + \left\|\mathbf{w}_{\mathcal{S}^*}\right\|_2 \\
\leq& \sqrt{r^*}\left\|\Pi_{\mathcal{F}_{\mathcal{E}}}\left(\nabla f\left(\mathbf{X}^*\right)\right)\right\|_2 + \sqrt{r^*}\left\|\mathbf{w}_{\mathcal{S}^*}\right\|_{\max} \\
\leq& \frac{3}{2}\lambda\sqrt{r^*}.
\end{aligned}$$

□

Lemma 9. *Suppose that Assumption 1 holds. Consider the problem (3). Define the set $\mathcal{E}^{(k)}$ by*

$$\mathcal{E}^{(k)} = \mathcal{S}^* \cup \mathcal{S}^{(k)}, \text{ with } \mathcal{S}^{(k)} = \left\{ i \mid w_i^{(k-1)} \leq p_\lambda'\left(u\right) \right\},$$

where $u = c\lambda$ and $c = \frac{3}{2}$ is the same to that given in Assumption 1. If $\lambda \geq 4\left\|\nabla f\left(\mathbf{\Sigma}^*\right)\right\|_2$, then for $k \geq 1$, we have $\left|\mathcal{E}^{(k)}\right| \leq 2r^*$ and the optimal solution $\widehat{\mathbf{\Sigma}}^{(k)}$ satisfies

$$\begin{aligned}
\left\|\widehat{\mathbf{X}}^{(k)} - \mathbf{X}^*\right\|_{\mathrm{F}} \leq& \left\|\mathbf{w}_{\mathcal{S}^*}^{(k-1)}\right\|_2 + \left\|\Pi_{\mathcal{F}_{\mathcal{E}^{(k)}}} \nabla f\left(\mathbf{X}^*\right)\right\|_{\mathrm{F}} \\
\leq& \frac{3}{2}\lambda\sqrt{r^*}.
\end{aligned}$$

Proof. We first prove $\left|\mathcal{E}^{(k)}\right| \leq 2r^*$ holds by induction. For $k = 1$, we have $w_i^{(0)} = \lambda \geq p_\lambda'\left(u\right)$ and thus $\mathcal{S}^{(0)} = \emptyset$ and $\mathcal{E}^{(1)} = \mathcal{S}^*$, which implies $\left|\mathcal{E}^{(1)}\right| \leq 2r^*$ holds. Assume $\left|\mathcal{E}^{(k)}\right| \leq 2r^*$ holds at $k-1$, i.e., $\left|\mathcal{E}^{(k-1)}\right| \leq 2r^*$ holds for some $k \geq 2$. Next, we will prove $\left|\mathcal{E}^{(k)}\right| \leq 2r^*$ holds at k. For any $i \in \mathcal{S}^{(k)}$, we obtain $\sigma_i\left(\widehat{\mathbf{X}}^{(k-1)}\right) \geq u$ and further

have

$$\sqrt{\left|\mathcal{E}^{(k)} \setminus \mathcal{S}^*\right|} \leq \frac{\left\|\Pi_{\mathcal{F}_{\mathcal{E}^{(k)} \setminus \mathcal{S}}}\left(\widehat{\mathbf{X}}^{(k-1)}\right)\right\|_{\mathrm{F}}}{u} \leq \frac{\left\|\widehat{\mathbf{X}}^{(k-1)} - \mathbf{X}^*\right\|_{\mathrm{F}}}{u}. \tag{11}$$

For any $i \in \overline{\mathcal{S}^{(k-1)}}$, we have

$$\mathbf{w}_i^{(k-2)} = p_\lambda'\left(\sigma_i\left(\widehat{X}^{(k-2)}\right)\right) \geq p_\lambda'(u) \geq \frac{\lambda}{2},$$

which implies

$$\left\|\mathbf{w}_{\overline{\mathcal{E}^{(k-1)}}}^{(k-2)}\right\|_{\min} \geq \left\|\mathbf{w}_{\overline{\mathcal{S}^{(k-1)}}}^{(k-2)}\right\|_{\min} \geq p_\lambda'(u) \geq \lambda/2.$$

One also has $\left|\mathcal{E}^{(k-1)}\right| \leq 2r^*$ and $\mathcal{S}^* \subseteq \mathcal{E}^{k-1}$. Applying Lemma 8 with $\widehat{\mathbf{X}} = \widehat{\mathbf{X}}^{(k-1)}$, $\mathcal{E} = \mathcal{E}^{(k-1)}$, and $\mathbf{w}_{\mathcal{S}^*} = \mathbf{w}_{\mathcal{S}^*}^{(k-2)}$ yields

$$\|\widehat{\mathbf{X}}^{(k-1)} - \mathbf{X}^*\|_{\mathrm{F}} \leq \frac{3}{2}\lambda\sqrt{r^*}.$$

Substituting the above result into the inequality (11) yields

$$\sqrt{\left|\mathcal{E}^{(k)} \setminus \mathcal{S}^*\right|} \leq \frac{3}{2u}\lambda\sqrt{r^*} = \sqrt{r^*}.$$

Thus, we have

$$\left|\mathcal{E}^{(k)}\right| = \left|\mathcal{S}^* \cup \left(\mathcal{E}^{(k)} \setminus \mathcal{S}^*\right)\right| = |\mathcal{S}^*| + \left|\mathcal{E}^{(k)} \setminus \mathcal{S}^*\right| \leq 2r^*,$$

completing the induction. Then, by the definition of $\mathcal{E}^{(k)}$, we have

$$\left\|\mathbf{w}_{\overline{\mathcal{E}^{(k)}}}\right\|_{\min} \geq p_\lambda'(u) \geq \lambda/2.$$

Applying Lemma 8 with $\widehat{\mathbf{X}} = \widehat{\mathbf{X}}^{(k)}$, $\mathcal{E} = \mathcal{E}^{(k)}$, and $\mathbf{w}_{\mathcal{S}^*} = \mathbf{w}_{\mathcal{S}^*}^{(k-1)}$, the optimal solution $\widehat{\mathbf{X}}^{(k)}$ to (3) satisfies

$$\left\|\widehat{\mathbf{X}}^{(k)} - \mathbf{X}^*\right\|_{\mathrm{F}} \leq \left\|\mathbf{w}_{\mathcal{S}^*}^{(k-1)}\right\|_2 + \left\|\Pi_{\mathcal{F}_{\mathcal{E}^{(k)}}}\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)\right\|_{\mathrm{F}} \leq \frac{3}{2}\lambda\sqrt{r^*}.$$

□

Lemma 10. *For any norm $\|\cdot\|$, the following inequality holds:*

$$\|\mathbf{w}_{\mathcal{S}^*}\| \leq \|\mathbf{w}\left(\boldsymbol{\sigma}_{\mathcal{S}^*}\left(\mathbf{X}^*\right) - u\mathbf{1}_{\mathcal{S}^*}\right)\| + \lambda u^{-1}\left\|\mathbf{X}_{\mathcal{S}^*} - \mathbf{X}_{\mathcal{S}}^*\right\|,$$

where $\mathbf{w}\left(\boldsymbol{\sigma}_{\mathcal{S}^*}\left(\mathbf{X}^*\right) - u\mathbf{1}_{\mathcal{S}^*}\right) = \left[p_\lambda'\left(\boldsymbol{\sigma}_1\left(\mathbf{X}^*\right) - u\right), \cdots, p_\lambda'\left(\boldsymbol{\sigma}_d\left(\mathbf{X}^*\right) - u\right)\right]$

Proof. We consider two cases for each index $i \in \mathcal{S}^*$.

Case I: If $|\sigma_i(\mathbf{X}) - \sigma_i(\mathbf{X}^*)| \geq u$, under Assumption 1, we have

$$p_\lambda'\left(\sigma_i(\mathbf{X})\right) \leq \lambda \leq \lambda u^{-1}\left|\sigma_i(\mathbf{X}) - \sigma_i\left(\mathbf{X}^*\right)\right|.$$

Case II: If $|\sigma_i(\mathbf{X}) - \sigma_i(\mathbf{X}^*)| < u$, then due to the non-increasing nature of $p_\lambda'(\cdot)$, it follows that

$$p_\lambda'\left(\sigma_i(\mathbf{X})\right) \leq p_\lambda'\left(\sigma_i\left(\mathbf{X}^*\right) - u\right).$$

By combining both cases, we establish that for all $i \in \mathcal{S}^*$, the following inequality holds:

$$p_\lambda'\left(\sigma_i(\mathbf{X})\right) \leq p_\lambda'\left(\sigma_i\left(\mathbf{X}^*\right) - u\right) + \lambda u^{-1}\left|\sigma_i(\mathbf{X}) - \sigma_i\left(\mathbf{X}^*\right)\right|.$$

Expressing this inequality in vector form for all indices in $\mathcal{S}^*$, applying the any norm $\|\cdot\|$ and utilizing the triangle inequality, we obtain

$$\begin{aligned}&\|\mathbf{w}_{\mathcal{S}^*}\| \\ \leq &\|\mathbf{w}\left(\sigma_{\mathcal{S}^*}\left(\mathbf{X}^*\right) - u\mathbf{1}_{\mathcal{S}^*}\right)\| \\ &+ \lambda u^{-1}\left\|\boldsymbol{\sigma}_{\mathcal{S}^*}(\mathbf{X}) - \boldsymbol{\sigma}_{\mathcal{S}^*}\left(\mathbf{X}^*\right)\right\|. \end{aligned} \tag{12}$$

Recognizing that the singular values are Lipschitz continuous functions of the matrix entries, we have

$$\left\|\boldsymbol{\sigma}_{\mathcal{S}^*}(\mathbf{X}) - \boldsymbol{\sigma}_{\mathcal{S}^*}\left(\mathbf{X}^*\right)\right\| \leq \left\|\Pi_{\mathcal{F}_{\mathcal{S}^*}}(\mathbf{X}) - \Pi_{\mathcal{F}_{\mathcal{S}^*}}\left(\mathbf{X}^*\right)\right\|.$$

Substituting this back into the inequality (12), we obtain

$$\begin{aligned}\|\mathbf{w}_{\mathcal{S}^*}\| \leq &\|\mathbf{w}\left(\boldsymbol{\sigma}_{\mathcal{S}^*}\left(\mathbf{X}^*\right) - u\mathbf{1}_{\mathcal{S}^*}\right)\| \\ &+ \lambda u^{-1}\left\|\mathbf{X}_{\mathcal{S}^*} - \mathbf{X}_{\mathcal{S}}^*\right\|. \end{aligned}$$

□

8.2. Proof of Theorem 4

Proof. By Lemma 9, we have

$$\left\|\widehat{\mathbf{X}}^{(k)} - \mathbf{X}^*\right\|_{\mathrm{F}} \leq \left\|\mathbf{w}_{\mathcal{S}^*}^{(k-1)}\right\|_2 + \left\|\Pi_{\mathcal{F}_{\mathcal{E}^{(k)}}}\left(\nabla f\left(\mathbf{X}^*\right)\right)\right\|_{\mathrm{F}}.$$

Using Lemma 10 yields that

$$\begin{aligned}\|\mathbf{w}_{\mathcal{S}^*}\| \leq &\|w\left(\boldsymbol{\sigma}_{\mathcal{S}^*}\left(\mathbf{X}^*\right) - u\mathbf{1}_{\mathcal{S}^*}\right)\| \\ &+ \lambda u^{-1}\left\|\mathbf{X}_{\mathcal{S}^*} - \mathbf{X}_{\mathcal{S}}^*\right\|. \end{aligned}$$

Plugging two inequalities above obtains that

$$\begin{aligned}\left\|\widehat{\mathbf{X}}^{(k)} - \mathbf{X}^*\right\|_{\mathrm{F}} \leq &\|\mathbf{w}\left(\boldsymbol{\sigma}_{\mathcal{S}^*}\left(\mathbf{X}^*\right) - u\mathbf{1}_{\mathcal{S}^*}\right)\|_2 \\ &+ \lambda u^{-1}\left\|\Pi_{\mathcal{F}_{\mathcal{S}^*}}\left(\widehat{\mathbf{X}}^{(k-1)}\right) - \Pi_{\mathcal{F}_{\mathcal{S}^*}}\left(\mathbf{X}^*\right)\right\|_{\mathrm{F}} \\ &+ \left\|\Pi_{\mathcal{F}_{\mathcal{E}^{(k)}}}\left(\nabla f\left(\mathbf{X}^*\right)\right)\right\|_{\mathrm{F}} \\ \leq &\lambda u^{-1}\left\|\widehat{\mathbf{X}}^{(k-1)} - \mathbf{X}^*\right\|_{\mathrm{F}} \\ &+ \left\|\Pi_{\mathcal{F}_{\mathcal{E}^{(k)}}}\left(\nabla f\left(\mathbf{X}^*\right)\right)\right\|_{\mathrm{F}}, \end{aligned}$$

where $\mathbf{w}\left(\boldsymbol{\sigma}_{\mathcal{S}^*}\left(\mathbf{X}^*\right) - u\mathbf{1}_{\mathcal{S}^*}\right) \leq \mathbf{w}\left(\gamma\lambda \cdot \mathbf{1}_{\mathcal{S}}\right) = \mathbf{0}_{\mathcal{S}^*}$ due to $\left\|\Pi_{\mathcal{F}_{\mathcal{S}^*}}\left(\mathbf{X}^*\right)\right\|_{\min} \geq (\gamma + c)\lambda \gtrsim \lambda$ in Assumption 3. In the next, we bound term $\left\|\Pi_{\mathcal{F}_{\mathcal{E}^{(k)}}}\left(\nabla f\left(\mathbf{X}^*\right)\right)\right\|_{\mathrm{F}}$. Separating the support of $\Pi_{\mathcal{F}_{\mathcal{E}^{(k)}}}\left(\nabla f\left(\mathbf{X}^*\right)\right)$ to $\mathcal{S}^*$ and $\mathcal{E}^{(k)}\backslash\mathcal{S}^*$ and then using triangle inequality, we obtain

$$
\begin{aligned}
&\left\|\Pi_{\mathcal{F}_{\mathcal{E}^{(k)}}}\left(\nabla f\left(\mathbf{X}^*\right)\right)\right\|_{\mathrm{F}} \\
\leq &\left\|\Pi_{\mathcal{F}_{\mathcal{S}^*}}\left(\nabla f\left(\mathbf{X}^*\right)\right)\right\|_{\mathrm{F}} + \left\|\Pi_{\mathcal{F}_{\mathcal{E}^{(k)}\backslash\mathcal{S}^*}}\left(\nabla f\left(\mathbf{X}^*\right)\right)\right\|_{\mathrm{F}} \\
\leq &\left\|\Pi_{\mathcal{F}_{\mathcal{S}^*}}\left(\nabla f\left(\mathbf{X}^*\right)\right)\right\|_{\mathrm{F}} + \sqrt{\left|\mathcal{E}^{(k)}\backslash\mathcal{S}^*\right|}\left\|\nabla f\left(\mathbf{X}^*\right)\right\|_2 .
\end{aligned}
$$

From the assumption, we know $\lambda \geq 4\left\|\nabla f\left(\mathbf{X}^*\right)\right\|_2$. Following the proof of Lemma 9, $\sqrt{\left|\mathcal{E}^{(k)}\backslash\mathcal{S}^*\right|}$ can be bounded by

$$
\begin{aligned}
\sqrt{\left|\mathcal{E}^{(k)}\backslash\mathcal{S}^*\right|} &\leq \frac{\left\|\Pi_{\mathcal{F}_{\mathcal{E}^{(k)}\backslash\mathcal{S}^*}}\left(\widehat{\mathbf{X}}^{(k-1)}\right)\right\|_{\mathrm{F}}}{u} \\
&\leq \frac{\left\|\widehat{\mathbf{X}}^{(k-1)} - \mathbf{X}^*\right\|_{\mathrm{F}}}{u}.
\end{aligned}
$$

Therefore, $\left\|\Pi_{\mathcal{F}_{\mathcal{E}^{(k)}}}\left(\nabla f\left(\mathbf{X}^*\right)\right)\right\|_{\mathrm{F}}$ can be simplified to

$$
\begin{aligned}
\left\|\Pi_{\mathcal{F}_{\mathcal{E}^{(k)}}}\left(\nabla f\left(\mathbf{X}^*\right)\right)\right\|_{\mathrm{F}} \leq &\left\|\Pi_{\mathcal{F}_{\mathcal{S}^*}}\left(\nabla f\left(\mathbf{X}^*\right)\right)\right\|_{\mathrm{F}} \\
&+ \frac{\lambda}{4u}\left\|\widehat{\mathbf{X}}^{(k-1)} - \mathbf{X}^*\right\|_{\mathrm{F}},
\end{aligned}
$$

which yields the contraction property

$$
\begin{aligned}
&\left\|\widehat{\mathbf{X}}^{(k)} - \mathbf{X}^*\right\|_{\mathrm{F}} \\
\leq &\left\|\Pi_{\mathcal{F}_{\mathcal{S}^*}}\left(\nabla f\left(\mathbf{X}^*\right)\right)\right\|_{\mathrm{F}} + 1.25\lambda u^{-1}\left\|\widehat{\mathbf{X}}^{(k-1)} - \mathbf{X}^*\right\|_{\mathrm{F}} \\
= &\left\|\Pi_{\mathcal{F}_{\mathcal{S}^*}}\left(\nabla f\left(\mathbf{X}^*\right)\right)\right\|_{\mathrm{F}} + \delta\left\|\widehat{\mathbf{X}}^{(k-1)} - \mathbf{X}^*\right\|_{\mathrm{F}},
\end{aligned}
$$

where $\delta = 1.25\lambda u^{-1}$. Consequently, we obtain

$$
\begin{aligned}
\left\|\widehat{\mathbf{X}}^{(k)} - \mathbf{X}^*\right\|_{\mathrm{F}} \leq &\frac{1}{1-\delta}\left\|\Pi_{\mathcal{F}_{\mathcal{S}^*}}\left(\nabla f\left(\mathbf{X}^*\right)\right)\right\|_{\mathrm{F}} \\
&+ \delta^{k-1}\left\|\widehat{\mathbf{X}}^{(1)} - \mathbf{X}^*\right\|_{\mathrm{F}}.
\end{aligned}
$$

□

8.3. Concentration Inequalities

Lemma 11. *Let* $\mathbf{E}$ *be a sub-Gaussian random matrix with* E_{ij} *are sub-Gaussian random variables with zero mean and covariance* ε*. Consider that* $\mathbf{Y} = \mathbf{X}^* + \mathbf{E}$*. There exists some constant* c_1*,* c_2*, the following tail bound holds*

$$
\mathbb{P}\left(\left|X_{ij}^* - Y_{ij}\right| > t\right) \leq c_1 \exp\left(-c_2 t^2\right).
$$

Lemma 12. *Under the same conditions in Lemma 11, if taking* $\lambda = \sqrt{\frac{3\log mn}{c_2}} \asymp \sqrt{\log mn}$*, then the following result holds*

$$
\mathbb{P}\left(\left\|\mathbf{X}^* - \mathbf{Y}\right\|_{\mathrm{F}} \leq \lambda\right) \geq 1 - \frac{c_1}{mn}.
$$

Proof. Applying Lemma 11 and union bound, for any λ such that $0 < \lambda < t_0$, we obtain

$$
\begin{aligned}
&\mathbb{P}\left(\left\|\mathbf{X}^* - \mathbf{Y}\right\|_{\mathrm{F}} > \lambda\right) \\
\leq &c_1 mn \exp\left(-c_2\lambda^2\right) \\
= &c_1 \exp\left(-c_2\lambda^2 + \log mn\right).
\end{aligned}
$$

By taking $\lambda = \sqrt{\frac{2\log mn}{c_2}} \asymp \sqrt{\log mn}$, we obtain

$$
\begin{aligned}
&\mathbb{P}\left(\left\|\mathbf{X}^* - \mathbf{Y}\right\|_{\mathrm{F}} \leq \lambda\right) \\
\geq &1 - c_1 \exp\left(-c_2\lambda^2 + \log mn\right) \\
= &1 - \frac{c_1}{mn}.
\end{aligned}
$$

□

Lemma 13. *Under the same conditions in Lemma 11, the following result holds*

$$
\mathbb{P}\left(\left\|\mathbf{X}^* - \mathbf{Y}\right\|_2 \leq \lambda\right) \geq 1 - \frac{c_1}{mn}.
$$

Proof. Recall that the spectral norm of a matrix is bounded above by its Frobenius norm

$$
\left\|\mathbf{X}^* - \mathbf{Y}\right\|_2 \leq \left\|\mathbf{X}^* - \mathbf{Y}\right\|_{\mathrm{F}}.
$$

Therefore, we have

$$
\mathbb{P}\left(\left\|\mathbf{X}^* - \mathbf{Y}\right\|_2 > \lambda\right) \leq \mathbb{P}\left(\left\|\mathbf{X}^* - \mathbf{Y}\right\|_{\mathrm{F}} > \lambda\right).
$$

Based on Lemma 12, we complete the proof. □

Lemma 14. *Under the same conditions in Lemma 11, the following result holds*

$$
\left\|\Pi_{\mathcal{F}_{\mathcal{S}^*}}\left(\mathbf{X}^* - \mathbf{Y}\right)\right\|_{\mathrm{F}} = \mathcal{O}_P\left(\sqrt{r^*}\right).
$$

Proof. Applying Lemma 11 and union bound, for any M such that $0 < M\sqrt{\frac{1}{n}} < t_0$, we obtain

$$
\begin{aligned}
&\mathbb{P}\left(\left\|\Pi_{\mathcal{F}_{\mathcal{S}^*}}\left(\mathbf{X}^* - \mathbf{Y}\right)\right\|_2 > M\right) \\
\leq &\mathbb{P}\left(\left\|\Pi_{\mathcal{F}_{\mathcal{S}^*}}\left(\mathbf{X}^* - \mathbf{Y}\right)\right\|_{\mathrm{F}} > M\right) \\
\leq &c_1 mn \exp\left(-c_2 M^2\right) \\
= &c_1 \exp\left(-c_2 M^2 + \log mn\right).
\end{aligned}
$$

By taking M such that $\sqrt{\frac{\log mn}{c_2}} < M$ and $M \to \infty$ in above inequality obtains

$$
\lim_{M\to\infty}\sup_n \mathbb{P}\left(\left\|\Pi_{\mathcal{F}_{\mathcal{S}^*}}\left(\mathbf{X}^* - \mathbf{Y}\right)\right\|_2 > M\right) = 0. \tag{13}
$$

This proof is completed by applying $\left\|\Pi_{\mathcal{F}_{\mathcal{S}^*}}\left(\mathbf{X}^* - \mathbf{Y}\right)\right\|_{\mathrm{F}} \leq \sqrt{r^*}\left\|\Pi_{\mathcal{F}_{\mathcal{S}^*}}\left(\mathbf{X}^* - \mathbf{Y}\right)\right\|_2$. □

8.4. Proof of Corollary 6

Proof. If λ satisfies $\lambda \asymp \sqrt{\log nm}$, then by Lemma 13, $\lambda \geq 4\left\|\nabla f\left(\mathbf{X}^*\right)\right\|_2$ holds with high probability. Applying Lemma 9 with $k=1$, we obtain

$$\|\widehat{\mathbf{X}}^{(1)} - \mathbf{X}^*\|_{\mathrm{F}} \leq \frac{3}{2}\lambda\sqrt{r^*}.$$

If $\lambda \asymp \sqrt{\log mn}$, then $\|\widehat{\mathbf{X}}^{(1)} - \mathbf{X}^*\|_{\mathrm{F}} \lesssim \sqrt{r^* \log mn}$. □

8.5. Proof of Corollary 7

Proof. If λ satisfies $\lambda \asymp \sqrt{\log mn}$, then by Lemma 13, $\lambda \geq 4\left\|\nabla f\left(\mathbf{\Sigma}^*\right)\right\|_2$ holds with high probability. Applying Theorem 4, we obtain

$$\begin{aligned}
&\left\|\widehat{\mathbf{X}}^{(k)} - \mathbf{X}^*\right\|_{\mathrm{F}} \\
\leq &\left\|\Pi_{\mathcal{F}_{\mathcal{S}^*}}\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)\right\|_{\mathrm{F}} + \delta\left\|\widehat{\mathbf{X}}^{(k-1)} - \mathbf{X}^*\right\|_{\mathrm{F}} \\
\leq &\frac{1}{1-\delta}\left\|\Pi_{\mathcal{F}_{\mathcal{S}^*}}\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)\right\|_{\mathrm{F}} + \delta^{k-1}\left\|\widehat{\mathbf{X}}^{(1)} - \mathbf{X}^*\right\|_{\mathrm{F}} \\
\leq &\frac{1}{1-\delta}\left\|\Pi_{\mathcal{F}_{\mathcal{S}^*}}\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)\right\|_{\mathrm{F}} + \delta^{k-1}\frac{3}{2}\lambda\sqrt{r^*},
\end{aligned}$$

where the last inequality is due to $\|\widehat{\mathbf{X}}^{(1)} - \mathbf{X}^*\|_{\mathrm{F}} \leq \frac{3}{2}\lambda\sqrt{r^*}$, which follow from Lemma 9 with $k=1$. One has

$$\left\|\Pi_{\mathcal{F}_{\mathcal{S}^*}}\left(\nabla f\left(\mathbf{X}^*\right)\right)\right\|_{\mathrm{F}} = \left\|\Pi_{\mathcal{F}_{\mathcal{S}^*}}\left(\mathbf{X}^* - \mathbf{Y}\right)\right\|_{\mathrm{F}}.$$

By Lemma 14, $\left\|\Pi_{\mathcal{F}_{\mathcal{S}^*}}\left(\mathbf{X}^* - \mathbf{Y}\right)\right\|_{\mathrm{F}} = \mathcal{O}_P\left(\sqrt{r^*}\right)$. If $K \geq 1 + \frac{\log\lambda}{\log\delta^{-1}} \gtrsim \log\lambda \gtrsim \log\log mn$, then we have

$$\delta^{k-1}\lambda\sqrt{r^*} \leq \frac{1}{\lambda}\lambda\sqrt{r^*} \leq \sqrt{r^*}.$$

Combining the above results yields that $\left\|\widehat{\mathbf{X}}^{(K)} - \mathbf{X}^*\right\|_{\mathrm{F}} = \mathcal{O}_P\left(\sqrt{r^*}\right)$.

□

9. More Experimental Setup

We also check the performance of our proposed estimator on the BSD100 datasets. The BSD100 dataset contains a diverse set of natural images with complex textures and real-world scenes, posing significant challenges for denoising algorithms. The following experiments demonstrate the superior performance of our algorithm in handling real-world image noise and enhancing the visual quality of denoised images.

- We primarily evaluate the denoising performance by comparing the visual quality of the output images. As shown in Fig 7 - Fig 10, the AWNNM algorithm outperforms competing methods in preserving fine details and effectively suppressing noise. This is particularly evident in high-texture regions, where AWNNM demonstrates exceptional robustness and fidelity. These results highlight the effectiveness of AWNNM for real-world denoising tasks.
- In addition to visual comparisons, we also compute commonly used metric PSNR to further validate the performance of the AWNNM algorithm. As evidenced by the results presented in the figures and tables, our algorithm achieves higher scores on PNSR compared to other methods. This indicates that AWNNM not only provides high-quality denoising results but also preserves structural information and texture details effectively, even under challenging noise conditions.

The experimental results on the BSD100 dataset demonstrate the outstanding performance of the proposed AWNNM algorithm for real-world image denoising. Its ability to balance texture preservation, noise suppression, and detail restoration surpasses existing methods, underscoring its potential for widespread applications in image processing.

Table 2. Denoising results (PSNR) by different methods on BSD100.

	$\sigma_n = 10$										$\sigma_n = 20$									
	BM3D	NNM	EPLL	WNNM	PCLR	PGPD	STROLLR	TWSC	SNSS	AWNNM	BM3D	NNM	EPLL	WNNM	PCLR	PGPD	STROLLR	TWSC	SNSS	AWNNM
001	35.97	28.42	36.98	37.63	37.76	37.26	37.46	37.67	37.71	**37.78**	31.97	28.00	33.03	33.39	33.58	33.16	33.27	33.49	33.51	**33.60**
002	37.01	31.23	39.51	40.41	40.60	39.73	39.75	40.59	40.55	**40.65**	32.26	30.77	35.41	36.14	36.42	35.89	36.17	36.34	36.40	**36.46**
003	38.18	31.41	39.34	40.05	40.31	39.28	40.27	40.15	40.20	**40.32**	33.92	30.83	35.42	35.86	36.31	35.52	36.27	36.06	36.11	**36.39**
004	37.27	31.50	38.50	39.03	39.19	38.53	38.56	39.12	39.16	**39.24**	33.26	30.77	34.43	34.65	34.48	34.43	34.48	34.80	34.64	**34.66**
005	36.13	29.13	37.59	38.37	**38.50**	37.90	38.24	38.47	38.36	**38.50**	31.73	28.73	33.51	33.99	34.23	33.76	34.16	34.13	34.11	**34.26**
006	39.06	34.79	40.37	40.98	41.47	40.25	41.17	41.22	41.31	**41.54**	34.36	33.73	36.26	36.83	37.35	36.57	37.86	37.11	37.52	**37.87**
007	36.05	28.22	37.09	37.76	37.88	37.32	37.28	37.88	37.79	**37.89**	31.98	27.78	33.03	33.42	33.57	33.14	33.41	33.57	33.55	**33.62**
008	35.82	27.75	36.85	37.45	37.63	37.04	37.58	37.61	37.63	**37.65**	31.77	27.39	32.76	33.01	33.23	32.79	33.29	33.17	33.22	**33.31**
009	36.77	29.14	37.91	38.66	38.78	38.09	38.36	38.77	38.76	**38.79**	32.82	28.67	33.82	34.38	34.46	33.99	34.23	34.52	34.49	**34.48**
010	37.27	30.88	38.44	38.94	39.14	38.43	38.80	39.05	39.09	**39.16**	33.19	30.25	34.41	34.59	34.88	34.36	34.80	34.78	34.81	**34.89**
011	36.56	28.14	37.28	38.28	38.30	37.76	38.26	38.34	38.27	**38.32**	32.61	27.76	33.30	34.19	34.23	33.78	34.21	34.32	34.29	**34.25**
012	36.19	28.64	37.11	37.76	37.92	37.28	37.71	37.85	37.88	**37.97**	32.16	28.16	33.05	33.37	**33.57**	33.09	33.26	33.49	33.52	**33.57**
013	36.74	30.93	37.76	38.34	38.51	37.94	38.36	38.45	38.51	**38.56**	32.72	30.25	33.72	33.99	34.22	33.81	33.99	34.11	34.16	**34.23**
014	38.39	31.84	39.60	40.35	40.50	39.53	40.55	40.46	40.52	**40.58**	34.20	31.28	35.56	36.14	**36.26**	35.76	36.15	36.24	36.25	**36.26**
015	36.45	28.53	37.44	38.04	38.23	37.59	38.26	38.12	38.20	**38.26**	32.37	28.16	33.44	33.77	33.97	33.51	33.68	33.86	33.89	**33.98**
016	34.95	25.25	35.85	36.83	36.95	36.35	36.61	36.93	36.95	**36.96**	30.96	25.04	31.85	32.55	32.71	32.19	32.30	32.65	32.71	**32.76**
017	39.63	35.47	40.68	41.24	41.30	40.33	41.25	41.36	41.32	**41.37**	35.25	34.17	36.63	36.96	37.10	36.64	37.15	37.13	37.15	**37.18**
018	39.68	35.17	40.62	41.29	41.65	40.34	41.37	41.39	41.57	**41.73**	35.44	34.03	36.69	37.11	37.60	36.78	37.64	37.34	37.54	**37.64**
019	37.49	30.08	38.17	39.04	39.13	38.50	38.67	39.12	39.14	**39.16**	33.42	29.59	34.30	34.84	**35.05**	34.66	34.96	34.94	35.02	**35.05**
020	37.66	29.37	38.65	39.52	39.71	38.78	39.56	39.63	39.68	**39.75**	33.69	29.00	32.76	35.29	35.54	34.90	35.55	35.46	35.55	**35.61**
AVE.	37.16	30.29	38.29	39.00	39.17	38.41	38.90	39.11	39.13	**39.21**	33.00	29.27	34.17	34.72	34.49	34.44	34.84	34.88	34.92	**35.00**
	$\sigma_n = 40$										$\sigma_n = 60$									
	BM3D	NNM	EPLL	WNNM	PCLR	PGPD	STROLLR	TWSC	SNSS	AWNNM	BM3D	NNM	EPLL	WNNM	PCLR	PGPD	STROLLR	TWSC	SNSS	AWNNM
001	27.62	26.14	29.15	29.45	**29.68**	29.48	29.64	29.38	29.42	**29.68**	25.01	23.01	26.99	27.28	27.53	27.35	27.46	26.99	27.14	**27.62**
002	26.97	28.59	31.28	32.20	32.40	32.24	32.37	32.22	32.31	**32.49**	23.97	24.93	28.94	30.05	30.23	29.93	29.98	29.95	30.28	**30.87**
003	29.57	28.68	31.34	31.98	32.34	31.81	32.26	31.95	32.23	**32.39**	26.94	25.55	28.92	29.83	29.85	29.36	29.84	29.49	24.76	**29.94**
004	29.04	28.57	30.50	30.73	31.01	30.76	31.01	30.72	30.87	**31.04**	26.74	25.62	28.33	28.63	**28.91**	28.51	28.86	28.50	28.67	**28.91**
005	26.83	26.77	29.47	30.00	30.28	30.08	30.16	29.94	30.17	**30.30**	23.83	23.46	27.24	27.89	28.16	27.97	28.01	27.73	28.11	**28.32**
006	28.82	30.72	32.09	32.91	33.55	33.08	33.41	32.98	33.22	**33.62**	25.63	26.63	29.77	30.78	31.52	30.74	30.36	30.62	30.98	**31.53**
007	28.03	25.94	29.08	29.41	29.53	29.30	29.28	29.39	29.44	**29.54**	25.80	23.32	26.84	27.23	27.14	27.13	27.17	27.01	27.12	**27.24**
008	27.87	25.82	28.85	29.11	29.22	29.04	29.19	29.02	29.22	**29.25**	25.49	23.55	26.68	27.02	26.96	26.96	26.85	26.73	27.01	**27.07**
009	28.74	26.75	29.77	30.27	30.28	30.11	30.21	30.25	30.28	**30.31**	25.58	24.08	27.52	27.96	27.89	27.88	27.89	27.72	27.81	**28.01**
010	28.83	28.26	30.43	30.60	30.88	30.66	30.89	30.58	30.67	**30.93**	26.67	25.45	28.17	28.45	28.69	28.45	28.62	28.23	28.35	**28.73**
011	28.08	25.75	29.25	30.20	30.21	29.86	30.16	30.15	30.19	**30.22**	24.81	22.39	26.91	**27.89**	27.79	27.43	27.77	27.69	27.77	**27.89**
012	28.04	26.33	29.04	29.27	29.49	29.23	29.35	29.19	29.21	**29.51**	25.78	23.81	26.82	27.09	27.17	27.10	26.98	26.78	26.98	**27.19**
013	28.62	28.14	29.81	30.09	**30.34**	30.20	30.28	30.02	30.23	**30.34**	26.39	25.29	27.71	28.03	28.29	28.09	27.96	27.88	28.13	**28.35**
014	29.81	29.21	31.37	32.12	32.11	31.90	32.14	32.11	32.14	**32.15**	27.12	26.28	29.02	29.83	29.67	29.53	29.52	29.54	29.67	**29.85**
015	28.13	26.44	29.52	29.78	29.96	29.77	29.74	29.72	29.87	**29.99**	25.57	23.83	27.35	27.62	27.80	27.63	27.67	27.44	27.68	**27.89**
016	26.96	23.43	27.88	28.48	28.55	28.20	28.54	28.43	28.46	**28.59**	24.66	20.87	25.59	26.20	26.16	25.87	25.99	25.99	26.15	**26.27**
017	29.47	31.49	32.63	33.10	33.35	33.12	33.30	33.13	33.24	**33.42**	25.89	28.52	30.45	31.05	31.33	30.83	31.02	30.75	31.22	**31.42**
018	30.63	31.32	32.60	33.25	33.74	33.30	33.74	33.34	33.46	**33.77**	27.80	27.56	30.20	31.10	31.49	30.83	31.26	30.81	31.23	**31.52**
019	28.99	27.54	30.61	31.20	31.46	31.21	31.45	31.18	31.27	**31.51**	26.02	24.48	28.47	29.25	29.42	29.06	29.26	29.09	29.35	**29.44**
020	29.33	27.16	30.83	31.32	31.58	31.17	31.57	31.26	31.46	**31.63**	26.08	24.41	28.57	29.08	29.30	28.90	29.17	28.31	28.46	**29.39**
AVE.	28.52	27.65	30.28	30.77	31.00	30.73	30.93	30.75	30.87	**31.03**	25.79	24.65	28.02	28.61	28.77	28.48	28.58	28.36	28.34	**28.87**
	$\sigma_n = 80$										$\sigma_n = 100$									
	BM3D	NNM	EPLL	WNNM	PCLR	PGPD	STROLLR	TWSC	SNSS	AWNNM	BM3D	NNM	EPLL	WNNM	PCLR	PGPD	STROLLR	TWSC	SNSS	AWNNM
001	22.81	24.14	25.53	25.80	26.07	25.86	25.74	25.44	25.65	**26.10**	20.80	23.64	24.41	24.68	25.00	24.67	**25.09**	24.26	24.38	24.73
002	21.83	24.57	27.27	28.41	28.65	28.26	28.26	28.30	28,44	**28.66**	20.04	23.63	25.94	27.17	**27.40**	27.03	27.33	26.90	27.12	27.29
003	24.76	24.70	27.20	28.15	**28.26**	27.60	27.83	27.66	28.13	**28.26**	22.91	23.75	25.88	26.72	26.81	26.33	26.82	26.03	26.45	**26.83**
004	24.65	22.64	26.82	27.13	27.34	26.87	27.13	26.86	26.98	**27.35**	22.92	22.20	25.66	25.91	26.11	25.69	**26.16**	25.50	25.88	26.07
005	21.60	25.75	25.70	26.35	26.63	26.44	26.40	26.22	26.43	**26.68**	19.85	25.13	24.55	25.16	25.54	25.30	**25.60**	24.99	25.24	25.47
006	22.88	22.47	28.16	29.14	**30.01**	28.93	28.85	28.91	28.56	29.89	20.97	21.76	26.86	27.86	**28.70**	27.80	28.74	27.56	27.88	28.03
007	23.68	22.80	25.25	25.72	25.61	25.57	25.18	25.31	25.45	**25.77**	21.76	22.22	24.09	24.53	24.42	24.38	24.17	24.00	24.37	**24.61**
008	23.28	23.28	25.19	25.56	25.49	25.49	25.27	25.10	25.23	**25.60**	21.47	22.67	24.10	24.41	24.31	24.37	24.26	23.95	24.21	**24.50**
009	22.83	24.66	26.00	26.37	26.43	26.35	25.98	25.95	26.12	**26.45**	20.63	23.87	24.84	25.19	25.27	25.22	25.01	24.71	25.12	**25.28**
010	24.89	21.48	26.62	26.98	27.18	26.85	26.74	26.63	26.78	**27.21**	23.37	21.01	25.43	25.79	**25.98**	25.62	25.65	25.34	25.67	25.96
011	22.45	23.09	25.25	26.20	26.16	25.73	26.21	25.87	26.12	**26.32**	20.32	22.41	23.99	24.80	24.88	24.45	24.84	24.42	24.64	**24.93**
012	24.31	24.38	25.29	25.61	25.61	25.61	25.40	25.19	25.23	**25.66**	22.67	23.58	24.16	24.44	24.46	24.44	24.23	24.00	24.32	**24.52**
013	24.76	25.29	26.28	26.63	26.86	26.56	26.89	26.41	26.56	**26.93**	23.22	24.41	25.16	25.50	**25.78**	25.45	25.19	25.16	25.67	25.73
014	25.03	23.11	27.40	28.17	27.95	27.80	27.95	27.69	27.98	**28.25**	22.99	22.64	26.18	26.85	26.71	26.60	26.35	26.21	26.34	**26.94**
015	23.34	20.15	25.90	26.08	26.27	26.12	26.33	25.84	25.99	**26.29**	21.43	19.79	24.82	24.92	**25.19**	23.07	24.99	24.67	24.78	25.11
016	22.71	23.57	23.95	24.56	24.42	24.31	24.44	24.21	24.43	**24.61**	20.86	26.59	22.66	23.25	23.09	23.12	22.86	22.81	23.15	**24.77**
017	23.27	26.48	28.88	29.59	29.80	29.09	29.73	29.07	29.47	**29.86**	21.13	25.33	27.62	28.42	28.54	27.96	28.49	27.68	27.43	**28.68**
018	25.39	23.65	28.49	29.45	29.71	28.93	29.75	28.82	28.96	**29.77**	23.25	23.19	27.16	27.99	28.28	27.60	28.23	27.17	28.34	**28.49**
019	23.75	23.66	26.90	27.74	27.86	27.43	27.58	27.50	27.67	**27.89**	21.81	23.23	25.62	26.53	26.65	26.22	26.58	26.14	26.78	**27.01**
020	23.77	21.80	26.93	27.49	27.65	27.25	27.74	27.07	27.47	**27.72**	21.71	22.47	25.65	26.25	26.40	26.07	26.37	25.78	26.12	**26.87**
AVE.	23.13	24.47	26.67	27.45	27.66	27.24	27.28	27.13	27.41	**27.67**	21.25	23.67	25.41	26.19	**26.40**	26.01	26.41	25.73	25.98	26.28

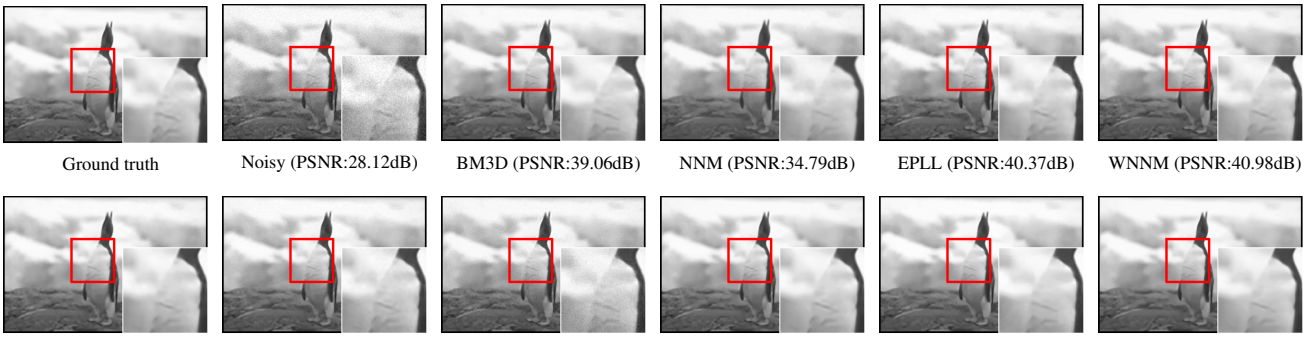


Figure 7. Denoising results on image *006* by different methods (noise level $\sigma_n = 10$).

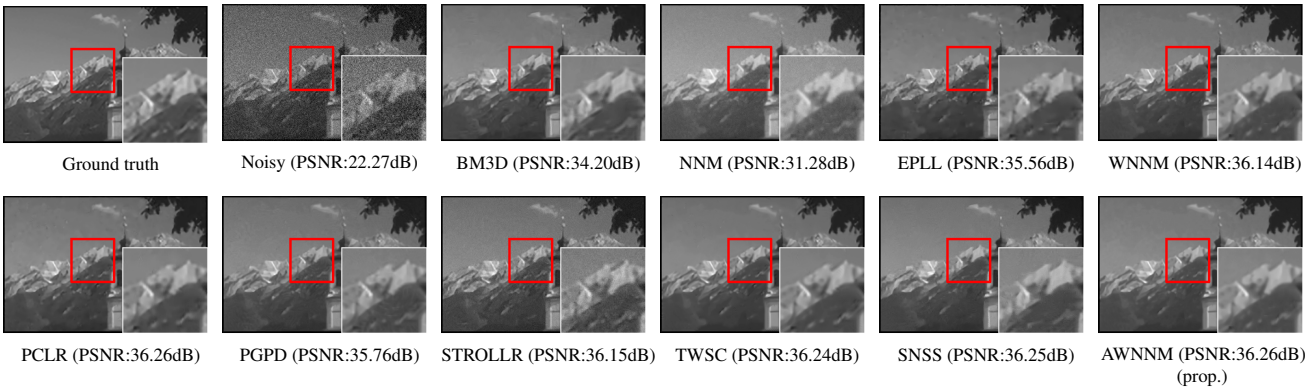


Figure 8. Denoising results on image *014* by different methods (noise level $\sigma_n = 20$).

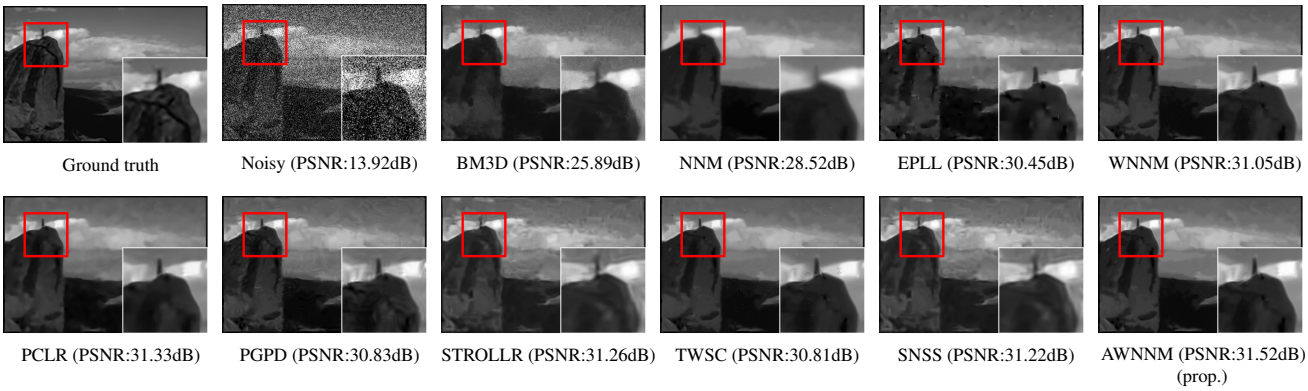


Figure 9. Denoising results on image *017* by different methods (noise level $\sigma_n = 60$).

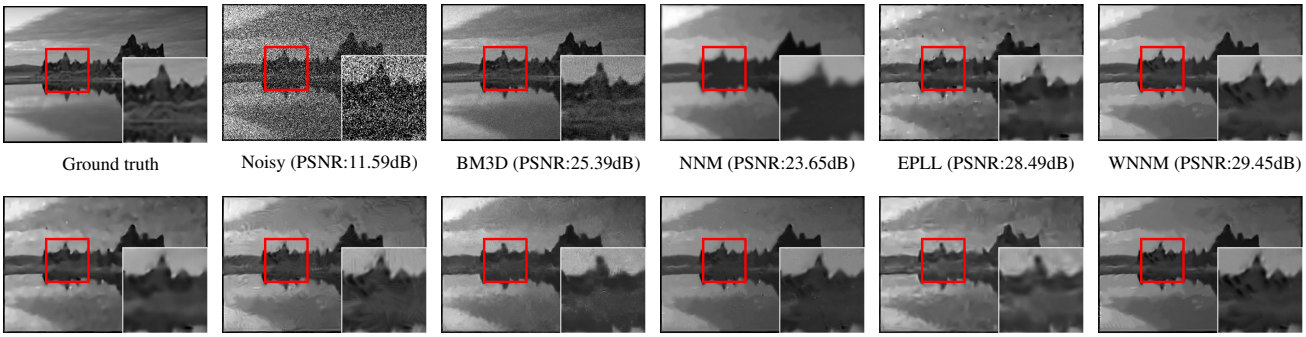


Figure 10. Denoising results on image *018* by different methods (noise level $\sigma_n = 80$).