Low-Rank Tensor Learning by Generalized Nonconvex Regularization

Sijia Xia*, Michael K. Ng†, and Xiongjun Zhang‡

October 25, 2024

Abstract

In this paper, we study the problem of low-rank tensor learning, where only a few of training samples are observed and the underlying tensor has a low-rank structure. The existing methods are based on the sum of nuclear norms of unfolding matrices of a tensor, which may be suboptimal. In order to explore the low-rankness of the underlying tensor effectively, we propose a nonconvex model based on transformed tensor nuclear norm for low-rank tensor learning. Specifically, a family of nonconvex functions are employed onto the singular values of all frontal slices of a tensor in the transformed domain to characterize the low-rankness of the underlying tensor. An error bound between the stationary point of the nonconvex model and the underlying tensor is established under restricted strong convexity on the loss function (such as least squares loss and logistic regression) and suitable regularity conditions on the nonconvex penalty function. By reformulating the nonconvex function into the difference of two convex functions, a proximal majorization-minimization (PMM) algorithm is designed to solve the resulting model. Then the global convergence and convergence rate of PMM are established under very mild conditions. Numerical experiments are conducted on tensor completion and binary classification to demonstrate the effectiveness of the proposed method over other state-of-the-art methods.

Key Words: Low-rank tensor learning, nonconvex regularization, transformed tensor SVD, proximal majorization-minimization, error bound

2020 Mathematics Subject Classification: 15A69, 90C25

1 Introduction

Tensors, which are higher-order generalization of vectors and matrices, have attracted much attention in the past decades and have a broad range of applications in various fields such as image processing [38], computer vision [48], machine learning [32, 65], and bioinformation [8]. These tensor data are generally lied in low-dimensional subspaces in realistic scenes, which substantially have low-rank

*School of Mathematics and Statistics, Central China Normal University, Wuhan 430079, China (e-mail: sijiax@mails.ccnu.edu.cn).

†Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong (e-mail: michaelng@hkbu.edu.hk). The research of this author was supported in part by the Hong Kong Research Grant Council GRF 17300021, C7004-21GF and Joint NSFC-RGC N-HKU76921.

‡School of Mathematics and Statistics, and Key Laboratory of Nonlinear Analysis & Applications (Ministry of Education), Central China Normal University, Wuhan 430079, China (e-mail: xjzhang@ccnu.edu.cn). The research of this author was supported in part by the National Natural Science Foundation of China under Grant No. 12171189 and the Fundamental Research Funds for the Central Universities under Grant No. CCNU24AI002.

structure. In turn, the low-rank tensor data lead to efficient estimation and prediction. As a result, the low-rank tensor based approaches are utilized to exploit the internal structure of a tensor efficiently. In this paper, we focus on the problem of low-rank tensor learning, where only a few of samples are given to learn the underlying tensor. Specifically, the general model of low-rank tensor learning is formulated as follows:

$$\min_{\mathcal{X}\in D} f_{n,\mathcal{Y}}(\mathcal{X}) + \beta \cdot \text{rank}(\mathcal{X}), \tag{1}$$

where $f_{n,\mathcal{Y}}(\cdot)$ is the loss function related to the number of samples n and the observed tensor $\mathcal{Y}$, D is a given constraint set, $\beta > 0$ is the regularization parameter, and $\text{rank}(\mathcal{X})$ denotes the rank function of the underlying parameter tensor $\mathcal{X} \in \mathbb{R}^{n_1\times n_2\times n_3}$.

Different tasks result in different loss functions in (1). Under the least squares loss, the model in (1) is suitable for low-rank tensor completion with noises. For example, Gandy et al. [19] proposed a model composed of least squares loss and nuclear norms of unfolding matrices of a tensor for tensor completion, where the nuclear norms of unfolding matrices were utilized to approximate the Tucker rank. Besides, Qiu et al. [53] and Zhang et al. [77] also employed the least squares loss with different low-rank approximations for noisy tensor completion, respectively. Under logistic regression loss, the model in (1) can be used for binary classification, where the low-rankness with respect to learning coefficients represents that only a subspace of feature space is utilized for classification. For high-dimensional tensor regression to classification problems, Lian [37] proposed a support tensor machine model with hinge loss and low-rank regularization, and also established the convergence rate of the estimator. Tan et al. [61] proposed a logistic tensor regression model for classification of high-dimensional data with structural information, which employed the CANDECOMP/PARAFAC (CP) decomposition [12] to measure the low-rankness of a tensor. Wimalawarne et al. [69] proposed a regularization method combined the logistic regression loss function and tensor norms based on the unfolding matrices of a tensor along each mode for binary classification. Furthermore, when the order of the tensor is second, low-rank tensor learning reduces to low-rank matrix learning, which has attracted much attention in the past few decades. For instance, when the loss function is the least squares loss, it is low-rank matrix completion with noisy observations. There exist many work for matrix completion in the literature, see [9, 10, 54] and references therein. Besides, for logistic regression, Yin et al. [73] proposed a low-rank plus sparse method to detect the intra-sample outliers via training data.

1.1 Low-Rank Tensor Learning

For low-rank tensor learning, the key issue is the definition of the rank of a tensor, which is acquired by tensor decomposition. Some widely used tensor decompositions include CP decomposition [12], Tucker decomposition [63], tensor singular value decomposition (SVD) [31], tensor train decomposition [47], tensor ring decomposition [80], and fully-connected tensor network decomposition [82]. However, computing the CP rank of a tensor is NP-hard in general [26]. Although the Tucker rank of a tensor can be computed via the SVD of the unfolding matrices easily, the Tucker rank minimization is also NP-hard due to the difficulty of matrix rank minimization [9]. Some convex approximation methods were proposed and studied via the sum of nuclear norms (SNN) of unfolding matrices of a tensor for tensor completion [19, 38]. However, the SNN is not the convex envelope of the sum of entries of Tucker rank of a tensor [57].

For tensor train rank minimization, which constitutes of ranks of matrices formed by a well-balanced matricization scheme, Bengua et al. [6] proposed two approaches for low-rank tensor completion via tensor train nuclear norm and its parallel matrix factorization, which were capable of capturing the global correlation of the tensor entries. However, the above work may destroy the intrinsic structure of a tensor since a low-order tensor should be represented into a higher-order tensor by the ket aug-

mentation technique. Based on tensor ring rank minimization, Yuan et al. [74] proposed a tensor ring nuclear norm method for noisy tensor completion to reduce the computational complexity by using the low-rank assumption on tensor factors instead of on the original tensor, where the tensor ring nuclear norm was defined as the nuclear norms of the unfolding matrices of all factor tensors in tensor ring decomposition. While the tensor train and tensor ring methods only established a connection between adjacent two factors, rather than any two factors. In order to overcome the limitation of tensor train and tensor ring decomposition, Zheng et al. [82] proposed a fully-connected tensor network decomposition and presented a novel method for tensor completion via this kind of decomposition, which established an operation between any two factor tensors. However, the fully-connected tensor network decomposition may be resulted in large computational burden due to the multiple connections of factor tensors.

1.2 Tensor SVD

The tensor SVD was proposed and studied based on tensor-tensor product for third-order tensors in [31], which was further extended to higher-order tensors [42, 49]. Moreover, the optimal representation and compression about tensor SVD were studied in [30]. For the problem of low-rank tensor minimization, the tensor nuclear norm was proposed to approximate the tensor tubal rank in [58]. And the tensor nuclear norm based methods were further presented to low-rank tensor learning problems, see [41, 66] and references therein. However, the tensor nuclear norm is based on Fourier transform, which may be challenged since the periodicity is assumed. Recently, Song et al. [59] proposed a transformed tensor SVD via any unitary transform instead of Fourier transform. Then the transformed tensor nuclear norm (TTNN) was proposed to approximate the transformed multi-rank of a tensor for robust tensor completion, which was capable of acquiring a lower rank tensor under suitable unitary transformations. Moreover, the TTNN was applied to other low-rank tensor optimization problems successfully, see [50, 51, 60, 78] and references therein. However, the TTNN is just the convex envelope of sum of each entry of transformed multi-rank of a tensor under the unit ball of tensor spectral norm, which may be suboptimal. Furthermore, the tensor nuclear norms based on non-invertible transformations [28] and nonlinear transformations [35] were proposed and studied for tensor completion.

On the other hand, some nonconvex approximations were also proposed and studied for low-rank tensor learning, which can generally approximate the tensor rank better than the convex relaxation methods. For example, based on SNN for tensor completion, Zhang proposed a nonconvex method by folding a tenor into a square matrix to approximate the Tucker rank minimization [76]. Xu et al. [71] proposed a low-rank factorization model for tensor completion, where the unfolding matrices of a tensor along each mode were factorized into the product of two smaller matrices. Besides, Yao et al. [72] proposed a nonconvex model based on SNN for low-rank tensor learning and established the statistical performance on the tensor completion problem, where the nonconvex regularization was employed onto the singular values of unfolding matrices of a tensor along each mode. However, the error bound between the stationary point of the nonconvex model and the underlying tensor was only established for low-rank tensor learning with least squares loss. Moreover, the nonconvex approximation based on SNN may be suboptimal since the SNN is not the tightest convex relaxation of Tucker rank minimization [44].

Based on tensor SVD in low-rank tensor learning, Wang et al. [66] proposed a nonconvex model via using the nonconvex function onto the singular values of all frontal slices in the Fourier domain for tensor completion without noise. Furthermore, Qiu et al. [50, 52] proposed a nonconvex approach via using nonconvex regularization for robust tensor completion, where a family of nonconvex functions were employed onto the low-rank and sparse components, respectively. Moreover, Zhao et al. [81] proposed a nonconvex tensor surrogate function for robust tensor completion via equivalent nonconvex

surrogates with difference of convex functions structures. While they only established the recovery error bound between the estimator of an approximate convex model and the underlying tensor, and did not analyzed the error bound of the nonconvex model. In addition, Gao et al. [20] proposed an ℓ_p $(0 < p < 1)$ method by employing the ℓ_p function to each singular value of the low-rank tensor in the Fourier domain and each entry of the sparse tensor for tensor robust principle component analysis, where the ℓ_p norm was used to approximate tensor fibered rank and measure sparsity, respectively. Recently, Zhang et al. [79] proposed a sparse tensor factorization method based on tensor-tensor product under general observations. In the previous two work, the error bounds between the global minimizer of the nonconvex model and the underlying tensor were established under some conditions. However, there is a significant gap between the theory and practice computation since it is difficult to achieve the global minimizers of the these nonconvex models in numerical algorithms.

1.3 The Contribution

In this paper, we propose a nonconvex approach for low-rank tensor learning. Specifically, a general loss function is utilized between given samples and the underlying tensor to fit the observed data. Moreover, in order to explore the global low-rankness of the underlying tensor, a family of nonconvex functions are employed onto the singular values of all frontal slices of a tensor in the transformed domain. Compared with TTNN, the nonconvex method is capable of acquiring a lower rank tensor, which is preferred for low-rank tensor learning. Besides, an error bound between any stationary point of the proposed nonconvex model and the underlying tensor is established under the restricted strong convexity (RSC) condition on the loss function and suitable regularity conditions on the nonconvex penalty, which is smaller than that of the Tucker based method in [72]. In particular, our model covers the least squares loss for tensor completion and logistic regression for binary classification. Moreover, by reformulating the nonconvex regularization into a difference of convex (DC) functions, a proximal majorization-minimization (PMM) algorithm is designed to solve the proposed nonconvex model, where the loss function and one convex function in the DC structure are linearized at the current point of iterations. Moreover, we show that the PMM algorithm globally converges to a stationary point of the proposed nonconvex model under the Kurdyka-Łojasiewicz (KL) assumption on the nonconvex penalty, where the convergence rate of PMM is also established. And an alternating direction method of multipliers (ADMM) is presented to solve the resulting subproblem in PMM. Numerical examples on tensor completion and binary classification are conducted to demonstrate the superiority of the proposed method compared with several state-of-the-art methods.

The remaining parts of this paper are organized as follows. Next, we give some notations and notions about tensors and transformed tensor SVD. In Section 2, we propose a nonconvex model for low-rank tensor learning based on TTNN. The error bound of any stationary point of the proposed model is established under some conditions. A PMM algorithm is designed to solve the proposed nonconvex model and the ADMM is applied to solve the resulting subproblem in Section 3, where the global convergence and convergence rate of PMM are also established. In Section 4, some numerical experiments are conducted on tensor completion and logistical regression to illustrate the advantage of our method over other existing approaches. We conclude this paper in Section 5. Finally, all the technical proofs are deferred to the Appendix.

1.4 Preliminaries

Some notations used throughout this paper are summarized in Table 1, where the size of a tensor is $n_1 \times n_2 \times n_3$.

Now we give the definition of subdifferential of a function.

Table 1: Notations

Notations	Description
a/$\mathbf{a}$/$\mathbf{A}$/$\mathcal{A}$	Scalars/Vectors/Matrices/Tensors
$\mathcal{A}_{ijk}$	The (i,j,k)-th element of $\mathcal{A}$
$\cdot^T$	The conjugate transpose operator
$\mathrm{Tr}(\cdot)$	The trace of a matrix
$\mathcal{A}^{\langle i\rangle}$	The i-th frontal slice of $\mathcal{A}$
$\mathcal{A}_{(i)}$	The mode-i unfolding of $\mathcal{A}$
$\mathrm{Fold}_i(\cdot)$	The inverse operator of mode-i unfolding, i.e., $\mathrm{Fold}_i(\mathcal{A}_{(i)}) = \mathcal{A}$
$\|\mathbf{a}\|_2$	The ℓ_2 norm of $\mathbf{a}$
$\mathrm{Diag}(\mathbf{a})$	A diagonal matrix with the i-th diagonal element being the i-th component of $\mathbf{a}$
$\sigma_j(\mathbf{A})$	The j-th largest singular value of $\mathbf{A}$
$\|\mathbf{A}\|_*$	The nuclear norm of $\mathbf{A}$ defined as $\|\mathbf{A}\|_* := \sum_{j=1}^{\min\{n_1,n_2\}} \sigma_j(\mathbf{A})$
$\|\mathbf{A}\|$	The spectral norm of $\mathbf{A}$ defined as $\|\mathbf{A}\| := \sigma_1(\mathbf{A})$
$\|\mathbf{A}\|_F$	Frobenius norm of $\mathbf{A}$ defined as $\|\mathbf{A}\|_F := \sqrt{\mathrm{Tr}(\mathbf{A}^T\mathbf{A})}$
$\mathbf{U}$	A unitary matrix satisfying $\mathbf{U}\mathbf{U}^T = \mathbf{U}^T\mathbf{U} = \mathbf{I}_{n_3}$, where $\mathbf{I}_{n_3}$ is the $n_3 \times n_3$ identity matrix
$\widehat{\mathcal{X}}_{\mathbf{U}}$ or $\mathbf{U}[\mathcal{X}]$	$\widehat{\mathcal{X}}_{\mathbf{U}} = \mathbf{U}[\mathcal{X}] := \mathrm{Fold}_3(\mathbf{U}\mathcal{X}_{(3)})$
$\langle \mathcal{A}, \mathcal{B}\rangle$	The inner product of two tensors defined as $\langle \mathcal{A}, \mathcal{B}\rangle := \sum_{i=1}^{n_3} \mathrm{Tr}((\mathcal{A}^{\langle i\rangle})^T \mathcal{B}^{\langle i\rangle})$
$\mathrm{vec}(\mathcal{A})$	Vectorizing a tensor $\mathcal{A}$ into a vector
$\|\mathcal{A}\|_\infty$	Tensor ℓ_∞ norm of $\mathcal{A}$ defined as $\|\mathcal{A}\|_\infty := \max \lvert\mathcal{A}_{ijk}\rvert$
$\|\mathcal{A}\|_F$	Tensor Frobenius norm of $\mathcal{A}$ defined as $\|\mathcal{A}\|_F := \sqrt{\langle \mathcal{A}, \mathcal{A}\rangle}$
$\delta_D(\cdot)$	The indicator function of a set D with $\delta_D(a) = 0$ if $a \in D$, otherwise $+\infty$
$\mathrm{dist}(\mathbf{a}, D)$	The distance from $\mathbf{a}$ to D and is defined as $\mathrm{dist}(\mathbf{a}, D) := \inf\{\|\mathbf{y} - \mathbf{a}\|_2, \mathbf{y} \in D\}$

Definition 1 *[56, Definition 8.3] Consider a function $f : \mathbb{R}^n \to (-\infty, +\infty]$ and a point $\mathbf{x} \in \mathbb{R}^n$ with the finite $f(\mathbf{x})$. The regular subdifferential of f at $\mathbf{x}$ is defined by*

$$\widehat{\partial} f(\mathbf{x}) := \left\{ \mathbf{y} \in \mathbb{R}^n : \liminf_{\mathbf{z}\to\mathbf{x}, \mathbf{z}\neq\mathbf{x}} \frac{f(\mathbf{z}) - f(\mathbf{x}) - \langle \mathbf{y}, \mathbf{z} - \mathbf{x}\rangle}{\|\mathbf{z} - \mathbf{x}\|_2} \geq 0 \right\}.$$

The (limiting) subdifferential of the function f at $\mathbf{x}$ is defined by

$$\partial f(\mathbf{x}) := \left\{ \mathbf{y} \in \mathbb{R}^n : \exists \mathbf{x}^k \xrightarrow{f} \mathbf{x}, \mathbf{y}^k \to \mathbf{y} \text{ with } \mathbf{y}^k \in \widehat{\partial} f(\mathbf{x}^k) \text{ for each } k \right\},$$

where $\mathbf{x}^k \xrightarrow{f} \mathbf{x}$ means $\mathbf{x}^k \to \mathbf{x}$ with $f(\mathbf{x}^k) \to f(\mathbf{x})$.

The KL function plays a vital role for the convergence analysis in our algorithm, and we list the definition of KL function in the following definition.

Definition 2 *[7, Definition 3] Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous function. We say that f has the Kurdyka-Łojasiewicz (KL) property at point $\mathbf{x}^* \in dom(\partial f)$, if there exist a neighborhood U of $\mathbf{x}^*$, $\eta \in (0, +\infty]$ and a continuous concave function $\varphi : [0, \eta) \to \mathbb{R}_+$ such that:*

(i) $\varphi(0) = 0$,

(ii) φ is C^1 on $(0, \eta)$,

(iii) for all $s \in (0, \eta)$, $\varphi'(s) > 0$,

(iv) for all $\mathbf{x}$ in $U \cap [f(\mathbf{x}^*) < f < f(\mathbf{x}^*) + \eta]$, the KL inequality holds:

$$\varphi'(f(\mathbf{x}) - f(\mathbf{x}^*))\, dist(0, \partial f(\mathbf{x})) \geq 1. \tag{2}$$

If f satisfy the KL property at each point of $dom(\partial f) := \{\mathbf{v} \in \mathbb{R}^n : \partial f(\mathbf{v}) \neq \emptyset\}$, then f is called a KL function.

A function is said to have the KL property at $\mathbf{x}^*$ with an exponent α if the function φ in Definition 2 takes the form of $\varphi(s) = \mu_1 s^{1-\alpha}$ with $\mu_1 > 0$ and $\alpha \in [0, 1)$. The proper closed semi-algebraic functions are KL functions with exponent $\alpha \in [0, 1)$ [36].

Next, we review the definition of transformed tensor SVD for third-order tensors, see [59] for more details. We denote $\overline{\mathcal{X}}$ (also denoted by $\text{bdiag}(\widehat{\mathcal{X}}_{\mathbf{U}})$) as a block diagonal matrix, where the i-th block is the matrix $\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i \rangle}$, $i = 1, \ldots, n_3$, i.e.,

$$\overline{\mathcal{X}} = \text{bdiag}(\widehat{\mathcal{X}}_{\mathbf{U}}) := \begin{bmatrix} \widehat{\mathcal{X}}_{\mathbf{U}}^{\langle 1 \rangle} & & \\ & \ddots & \\ & & \widehat{\mathcal{X}}_{\mathbf{U}}^{\langle n_3 \rangle} \end{bmatrix}.$$

The corresponding inverse operator, denoted by "$\text{fold}_3(\cdot)$", is defined as

$$\text{fold}_3(\text{bdiag}(\widehat{\mathcal{X}}_{\mathbf{U}})) = \widehat{\mathcal{X}}_{\mathbf{U}}. \tag{3}$$

Definition 3 *[59, Definition 1] The* $\mathbf{U}$*-product of arbitrary two tensors* $\mathcal{A} \in \mathbb{C}^{n_1 \times n_2 \times n_3}$ *and* $\mathcal{B} \in \mathbb{C}^{n_2 \times l \times n_3}$ *is defined as* $\mathcal{A} \diamond_{\mathbf{U}} \mathcal{B} = \mathbf{U}^T[\text{fold}_3(\text{bdiag}(\widehat{\mathcal{A}}_{\mathbf{U}}) \cdot \text{bdiag}(\widehat{\mathcal{B}}_{\mathbf{U}}))] \in \mathbb{C}^{n_1 \times l \times n_3}$.

Definition 4 *[59, Definition 7] The transformed tensor nuclear norm of* $\mathcal{X} \in \mathbb{C}^{n_1 \times n_2 \times n_3}$ *is defined as* $\|\mathcal{X}\|_{\text{TTNN}} = \sum_{i=1}^{n_3} \|\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i \rangle}\|_*$.

It can be easily verified that $\|\mathcal{X}\|_{\text{TTNN}} = \|\overline{\mathcal{X}}\|_*$.

Definition 5 *[59] The tensor spectral norm with respect to* $\mathbf{U}$*, denoted by* $\|\mathcal{X}\|_{\mathbf{U}}$*, is defined as* $\|\mathcal{X}\|_{\mathbf{U}} = \|\overline{\mathcal{X}}\|$.

Now we recall the definitions of the diagonal tensor, conjugate transpose, and unitary tensor [31, 59]. A third-order tensor is called to be diagonal if each frontal slice is a diagonal matrix. The conjugate transpose of $\mathcal{X} \in \mathbb{C}^{n_1 \times n_2 \times n_3}$ with respect to $\mathbf{U}$, denoted by $\mathcal{X}^T$, is defined as $\mathcal{X}^T = \mathbf{U}^T[\text{fold}_3((\text{bdiag}(\widehat{\mathcal{X}}_{\mathbf{U}}))^T)] \in \mathbb{C}^{n_2 \times n_1 \times n_3}$. A tensor $\mathcal{U}$ is called to be unitary if $\mathcal{U} \diamond_{\mathbf{U}} \mathcal{U}^T = \mathcal{U}^T \diamond_{\mathbf{U}} \mathcal{U} = \mathcal{I}_{\mathbf{U}}$, where $\mathcal{I}_{\mathbf{U}}$ denotes the identity tensor and is defined as each frontal slice of $\mathbf{U}[\mathcal{I}_{\mathbf{U}}]$ being the identity matrix. Next we give the definition of transformed tensor SVD of a third-order tensor.

Theorem 1 *[29, Theorem 5.1] The transformed tensor singular value decomposition of* $\mathcal{X} \in \mathbb{C}^{n_1 \times n_2 \times n_3}$ *is given by* $\mathcal{X} = \mathcal{U} \diamond_{\mathbf{U}} \Sigma \diamond_{\mathbf{U}} \mathcal{V}^T$*, where* $\Sigma \in \mathbb{C}^{n_1 \times n_2 \times n_3}$ *is a diagonal tensor,* $\mathcal{U} \in \mathbb{C}^{n_1 \times n_1 \times n_3}$ *and* $\mathcal{V} \in \mathbb{C}^{n_2 \times n_2 \times n_3}$ *are unitary tensors with respect to* $\mathbf{U}$*-product.*

Definition 6 *[59, Definition 6] The transformed multi-rank of* $\mathcal{X} \in \mathbb{C}^{n_1 \times n_2 \times n_3}$ *is a vector* $\mathbf{r} = (r_1, r_2, ..., r_{n_3})$ *with* $r_i = \text{rank}(\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i \rangle})$, $i = 1, 2, \ldots, n_3$*, where* $\text{rank}(\widehat{\mathcal{X}}_{\mathbf{U}}^{(i)})$ *denotes the rank of the matrix* $\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i \rangle}$.

2 Nonconvex Model for Low-Rank Tensor Learning

In this section, we present the framework of our approach for the problem of low-rank tensor learning. Given n samples, we consider a general loss function $f_{n,\mathcal{Y}}(\mathcal{X})$, which is used to fit the parameter tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ and the observed data $\mathcal{Y}$. Suppose that the parameter tensor $\mathcal{X}$ is low-rank, which we

aim to estimate. Now a nonconvex model based on TTNN with a general loss function is presented for low-rank tensor learning:

$$\begin{aligned} &\min_{\mathcal{X}} f_{n,\mathcal{Y}}(\mathcal{X}) + \beta G_\lambda(\mathcal{X}) \\ &\text{s.t. } \|\mathcal{X}\|_\infty \leq c, \end{aligned} \tag{4}$$

where $f_{n,\mathcal{Y}}(\mathcal{X})$ represents a differentiable loss function with n samples, $G_\lambda(\mathcal{X})$ represents the nonconvex regularization defined as

$$G_\lambda(\mathcal{X}) := \sum_{i=1}^{n_3} \sum_{j=1}^{\min\{n_1,n_2\}} g_\lambda(\sigma_j(\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i\rangle})), \tag{5}$$

$\beta > 0$ is a penalty parameter, and $c > 0$ is a given constant. Here $g_\lambda(\cdot)$ is a nonconvex function with respect to the parameter λ. In addition, the tensor ℓ_∞ norm constraint is effective to exclude over-spiky tensors.

Throughout this paper, the nonconvex function $g_\lambda(\cdot)$ should satisfy the following assumptions.

Assumption 1 *The function $g_\lambda(x) : \mathbb{R} \to \mathbb{R}_+$ is symmetric and has the following properties:*

(i) $g_\lambda(x)$ is concave and non-decreasing for $x \geq 0$ with $g_\lambda(0) = 0$.

(ii) The function $\frac{g_\lambda(x)}{x}$ is non-increasing for $x > 0$.

(iii) $g_\lambda(x)$ is differentiable for any $x \neq 0$ and $\lim_{x\to 0^+} g'_\lambda(x) = \lambda k_0$, where $k_0 > 0$ is a constant and λk_0 is an upper bound of $g'_\lambda(\cdot)$ on $(0, +\infty)$.

(iv) There exists a parameter $\mu > 0$ such that $g_\lambda(x) + \frac{\mu}{2}x^2$ is convex on $(0, +\infty)$.

A lot of nonconvex functions satisfy Assumption 1 in the literature, such as smoothly clipped absolute deviation (SCAD) [16], minimax concave penalty (MCP) [75], logarithmic function [11], which will be given in Table 2. These nonconvex function can approximate the ℓ_0 norm better than the ℓ_1 norm, which can yield a sparser solution in statistical learning [22]. In model (4), we propose to employ a family of nonconvex functions based on TTNN to explore the global low-rankness of the underlying tensor in low-rank tensor learning. Although the TTNN can get a lower rank tensor compared with tensor nuclear norm under suitable unitary transformations [59], the TTNN is just the ℓ_1 norm of all singular values vectors of the frontal slices of a tensor in the transformed domain, which also suffers from the drawback of ℓ_1 norm. Recently, some nonconvex surrogate functions have been demonstrated more efficiently than the convex relaxation in various fields such as statistical learning [16, 75], compressed sensing [11], and tensor completion [50, 76]. Moreover, the nonconvex relaxation based on TTNN can help to achieve a lower rank tensor than TTNN for robust tensor completion [50]. In low-rank tensor learning, we employ a family of nonconvex functions onto each singular value of the frontal slices of the underlying tensor in the transformed domain. Compared with the TTNN, the main advantage of the proposed method is that the nonconvex functions can approximate the sum of entries of the transformed multi-rank of a tensor better.

For the general loss function $f_{n,\mathcal{Y}}$ in model (4), we just need it to be differentiable. In various real-world applications, we can specify the loss function $f_{n,\mathcal{Y}}$. In particular, when $f_{n,\mathcal{Y}}$ is the least squares loss function, model (4) can be used for noisy tensor completion. When $f_{n,\mathcal{Y}}$ is the logistic regression loss function, model (4) can be utilized for binary classification in machine learning.

Remark 1 *When $f_{n,\mathcal{Y}}(\mathcal{X}) = \frac{1}{2}\|\mathcal{P}_\Omega(\mathcal{Y}) - \mathcal{P}_\Omega(\mathcal{X})\|_F^2$, model (4) can be used for tensor completion with noisy observations. Recently, some nonconvex surrogates were proposed and studied for tensor completion [50, 66, 76], where a family of nonconvex functions were employed onto the low-rank component of a tensor. For example, Wang et al. [66] proposed to utilize a general nonconvex surrogate of*

the tensor tubal rank for tensor completion and a least squares loss function for the data-fitting term, where the tensor nuclear norm was used in the Fourier domain. And Zhang [76] used a family of non-convex function onto the singular values of the square matrice via matricizing a tensor to approximate the Tucker rank of a tensor. However, they do not analyze the statistical performance of their proposed models.

Remark 2 *Based on the overlapped nuclear norm of a tensor, Yao et al. [72] proposed a nonconvex approach for low-rank tensor learning, where the nonconvex functions were employed onto each singular value of unfolding matrices of a tensor. However, the unfolding based methods are challenged due to its suboptimality, where the overlapped nuclear norm was not the convex envelope of the sum of Tucker rank of a tensor [57]. Moreover, they only analyzed the statistical performance of their model for the least squares loss function. We will establish the error bound between any stationary point of model (4) and the underlying tenor for a general loss function in the next subsection.*

2.1 Statistical Guarantee

In this subsection, we first introduce the definition of the RSC condition for a general loss function and then establish the error bound between any stationary point of the proposed model and the underlying tensor.

The RSC condition of a differentiable function in the tensor case plays a vital role in establishing the error bound of the proposed model. The foundational works on the RSC condition in the vector case are due to [40, 46]. For any tensor $\tilde{\mathcal{V}} \in \mathbb{R}^{n_1\times n_2\times n_3}$, we say that the function $f_{n,\mathcal{Y}}(\mathcal{X}) : \mathbb{R}^{n_1\times n_2\times n_3} \to \mathbb{R}$ satisfies the RSC condition if

$$\langle \nabla f_{n,\mathcal{Y}}(\mathcal{X}^* + \tilde{\mathcal{V}}) - \nabla f_{n,\mathcal{Y}}(\mathcal{X}^*), \tilde{\mathcal{V}}\rangle \geq \begin{cases} \alpha_1\|\tilde{\mathcal{V}}\|_F^2 - \tau_1\frac{\log d}{n}\|\tilde{\mathcal{V}}\|_{\text{TTNN}}^2, & \text{if } \|\tilde{\mathcal{V}}\|_F \leq 1, \\ \alpha_2\|\tilde{\mathcal{V}}\|_F - \tau_2\sqrt{\frac{\log d}{n}}\|\tilde{\mathcal{V}}\|_{\text{TTNN}}, & \text{otherwise}, \end{cases} \tag{6}$$

where $\mathcal{X}^*$ denotes the ground-truth tensor, $d := n_1n_2n_3$, n is the number of samples, and $\alpha_1, \alpha_2 > 0$, $\tau_1, \tau_2 \geq 0$ are given constants.

The RSC condition involves a lower bound on the remainder in the first-order Taylor expansion of $f_{n,\mathcal{Y}}$, where we only require $f_{n,\mathcal{Y}}$ to be differentiable. If $f_{n,\mathcal{Y}}$ is convex, the left hand in (6) is always nonnegative, and then (6) holds trivially for $\frac{\|\tilde{\mathcal{V}}\|_{\text{TTNN}}}{\|\tilde{\mathcal{V}}\|_F} \geq \sqrt{\frac{\alpha_1 n}{\tau_1 \log d}}$ and $\frac{\|\tilde{\mathcal{V}}\|_{\text{TTNN}}}{\|\tilde{\mathcal{V}}\|_F} \geq \frac{\alpha_2}{\tau_2}\sqrt{\frac{n}{\log d}}$. As a result, the inequality in (6) only enforces a type of strong convexity condition over the cone $\left\{\frac{\|\tilde{\mathcal{V}}\|_{\text{TTNN}}}{\|\tilde{\mathcal{V}}\|_F} \leq \tilde{c}\sqrt{\frac{n}{\log d}}\right\}$, where $\tilde{c} > 0$ is a constant. In particular, we will show the least squares loss and logistic regression loss satisfy the RSC condition in Appendix A and B. More discussions about the RSC condition in the vector case can be referred to [40].

Remark 3 *The RSC condition has been widely studied for statistical learning in the literature. For example, the least squares loss for linear regression in the matrix case satisfies the RSC condition [23]. Moreover, the loss functions of the generalized linear model and corrected linear model satisfy the RSC condition by selecting appropriate parameters [39, 40, 46], where the RSC condition of these loss functions is employed in the vector case.*

Let $\tilde{\mathcal{X}}$ be a stationary point of problem (4). In the following analysis of the statistical performance guarantee of the proposed model, we enforce the additional constraint $\|\mathcal{X}\|_{\text{TTNN}} \leq t$ for any $\mathcal{X} \in \mathbb{R}^{n_1\times n_2\times n_3}$ in model (4), where $t > 0$ is a given constant. The main result is described in the following theorem.

Theorem 2 *Suppose that the regularizer $g_\lambda(\cdot)$ satisfies Assumption 1 and the loss function $f_{n,\mathcal{Y}}$ satisfies the RSC condition (6) with $\frac{3}{4}\beta\mu < \alpha_1$, where μ is defined in Assumption 1(iv). Consider the parameter λ with*

$$\frac{4}{\beta k_0}\max\left\{\|\nabla f_{n,\mathcal{Y}}(\mathcal{X}^*)\|_{\mathbf{U}}, \alpha_2\sqrt{\frac{\log d}{n}}\right\} \le \lambda \le \frac{\alpha_2}{6t\beta k_0}, \tag{7}$$

and

$$n \ge \frac{16t^2\max\{\tau_1^2,\tau_2^2\}}{\alpha_2^2}\log d, \tag{8}$$

where $d := n_1n_2n_3$. Then any stationary point $\tilde{\mathcal{X}}$ of model (4) satisfies

$$\|\tilde{\mathcal{X}} - \mathcal{X}^*\|_F \le \frac{6\beta\lambda k_0\sqrt{\sum_{i=1}^{n_3} r_i}}{4\alpha_1 - 3\beta\mu},$$

where r_i denotes the rank of matrix $(\widehat{\mathcal{X}^*})_{\mathbf{U}}^{\langle i\rangle}, i = 1,\ldots,n_3$.

The proof of Theorem 2 is left to Appendix E. From Theorem 2, we know that the upper bound is related to the sum of each entry of the transformed multi-rank of the underlying tensor, which will be small under suitable unitary transformations [59]. More details about the choice of unitary transformations in TTNN can be referred to [59, 60]. Note that the parameters k_0, μ, λ are related to the nonconvex function g_λ, which are given in Assumption 1. Moreover, we need the loss function $f_{n,\mathcal{Y}}$ to satisfy the RSC condition (6) in Theorem 2. We will show the least squares loss for tensor completion and logistic regression for binary classification in low-rank tensor learning satisfy the RSC condition in Appendix A and B, where the two loss functions are widely used in practice.

In particular, the error bound in Theorem 2 can reduce to that of the TTNN model, where the nonconvex regularization term of model (4) is replaced by TTNN. Notice that the error bound in Theorem 2 is just the tensor Frobenius norm of the difference between any stationary point of model (4) and the underlying tensor, which is the worst case for the error bound of the nonconvex model in (4). In numerical computation, we can design an efficient algorithm to obtain a stationary point of the nonconvex model, which will be shown in the next section. Now we compare with the error bound of the model in [72], which utilized the least squares loss and the nonconvex regularization based on the nuclear norms of all unfolding matrices of a tensor for low-rank tensor learning. Note that $\frac{r_1+\cdots+r_{n_3}}{n_3} \le \max\{r_1,\ldots,r_{n_3}\} \le \text{rank}(\mathcal{X}^*_{(1)})$. This demonstrates that a tensor with low Tucker rank has low average transformed multi-rank. Compared with the error bound in [72], which used the Tucker rank in the model and was on the order of $O(\sum_{i=1}^d\sqrt{\text{rank}(\mathcal{X}^*_{(i)})})$, the error bound in Theorem 2 is smaller when n_3 is not too large or the rank of unfolding matrices of the underlying tensor is large.

3 Optimization Algorithm

In this section, we first design a proximal majorization-minimization (PMM) algorithm [62, 81] to solve problem (4), and then establish the global convergence and convergence rate of PMM under very mild conditions. Finally, an ADMM based algorithm is utilized to solve the resulting subproblem in PMM.

3.1 PMM Algorithm

Note that problem (4) is nonconvex and nonsmooth since $G_\lambda(\mathcal{X})$ is nonconvex and nonsmooth. The nonconvexity of the objective function results in great challenges to numerical computation and theoretical analysis. By the special structure of some nonconvex functions, they can be written as the

difference of two convex functions, which has a variety of applications in statistical and machine learning [1, 22, 24, 33]. In particular, we assume that the nonconvex function $g_\lambda(x)$ in our model can be written as

$$g_\lambda(x) = s_1(x) - s_2(x), \tag{9}$$

where s_1 is convex and s_2 satisfies Assumption 2.

Assumption 2 *The function $s_2 : \mathbb{R} \to \mathbb{R}$ satisfies the following assumptions.*
(a) s_2 is convex and symmetric, i.e., $s_2(-x) = s_2(x)$.
(b) s_2 is differentiable and its derivative s_2' is locally Lipschitz continuous.

A lot of nonconvex functions satisfying Assumption 1 can be expressed in the decomposition formulation (9), which are summarized in Table 2. Besides, s_2 in Table 2 also satisfies Assumption 2.

Table 2: Examples of the nonconvex function $g_\lambda(x)$ and its corresponding difference of two convex functions (i.e., $g_\lambda(x) = s_1(x) - s_2(x)$), where $\gamma > 0$ for Logarithm and MCP, and $\gamma > 1$ for SCAD.

Nonconvex function	$g_\lambda(x)(x \geq 0, \lambda > 0)$	$s_1(x)$	$s_2(x)$
Logarithm [11]	$\lambda \log(\frac{x}{\gamma} + 1)$	λx	$\lambda x - \lambda \log(\frac{x}{\gamma} + 1), x > 0$
MCP [75]	$\begin{cases} \lambda x - \frac{x^2}{2\gamma}, & x \leq \gamma\lambda, \\ \frac{1}{2}\gamma\lambda^2, & x > \gamma\lambda. \end{cases}$	λx	$\begin{cases} \frac{x^2}{2\gamma}, & x \leq \gamma\lambda, \\ \lambda x - \frac{\gamma\lambda^2}{2}, & x > \gamma\lambda. \end{cases}$
SCAD [16]	$\begin{cases} \lambda x, & x < \lambda. \\ \frac{-x^2+2\gamma\lambda x-\lambda^2}{2(\gamma-1)}, & \lambda \leq x < \gamma\lambda, \\ \frac{\lambda^2(\gamma+1)}{2}, & x \geq \lambda\gamma. \end{cases}$	λx	$\begin{cases} 0, & x < \lambda, \\ \frac{x^2-2\lambda x+\lambda^2}{2(\gamma-1)}, & \lambda \leq x < \gamma\lambda, \\ \lambda x - \frac{(\gamma+1)\lambda^2}{2}, & x \geq \gamma\lambda. \end{cases}$

Based on (9), we can rewritten $G_\lambda(\mathcal{X})$ in (4) as follows:

$$G_\lambda(\mathcal{X}) = S_1(\mathcal{X}) - S_2(\mathcal{X}), \tag{10}$$

where $S_1(\mathcal{X}) = \sum_{i=1}^{n_3} \sum_{j=1}^{\min\{n_1,n_2\}} s_1(\sigma_j(\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i\rangle}))$ and

$$S_2(\mathcal{X}) = \sum_{i=1}^{n_3} \sum_{j=1}^{\min\{n_1,n_2\}} s_2(\sigma_j(\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i\rangle})). \tag{11}$$

Consequently, problem (4) can be reformulated equivalently as follows:

$$\min_{\mathcal{X}} f_{n,\mathcal{Y}}(\mathcal{X}) + \beta S_1(\mathcal{X}) - \beta S_2(\mathcal{X}) + \delta_D(\mathcal{X}), \tag{12}$$

where $\delta_D(\mathcal{X})$ is the indicator function of the set D with $D := \{\mathcal{X} : \|\mathcal{X}\|_\infty \leq c\}$.

We adopt the PMM algorithm to solve problem (12), whose main idea is to linearize the concave function $-S_2(\mathcal{X})$ and the smooth loss function $f_{n,\mathcal{Y}}(\mathcal{X})$ of the objective function in (12) at the current iteration point $\mathcal{X}^t$. Specifically, given $\mathcal{X}^t \in \mathbb{R}^{n_1\times n_2\times n_3}$, we consider to solve the following problem:

$$\begin{aligned} \min_{\mathcal{X}} \; & f_{n,\mathcal{Y}}(\mathcal{X}^t) + \langle \nabla f_{n,\mathcal{Y}}(\mathcal{X}^t), \mathcal{X} - \mathcal{X}^t\rangle + \frac{\rho}{2}\|\mathcal{X} - \mathcal{X}^t\|_F^2 + \beta S_1(\mathcal{X}) - \beta S_2(\mathcal{X}^t) \\ & - \beta\langle \nabla S_2(\mathcal{X}^t), \mathcal{X} - \mathcal{X}^t\rangle + \delta_D(\mathcal{X}), \end{aligned} \tag{13}$$

where $\rho > 0$ is a given constant. Notice that problem (13) is equivalent to

$$\min_{\mathcal{X}} \beta S_1(\mathcal{X}) + \langle \nabla f_{n,\mathcal{Y}}(\mathcal{X}^t) - \beta \nabla S_2(\mathcal{X}^t), \mathcal{X} - \mathcal{X}^t \rangle + \frac{\rho}{2}\|\mathcal{X} - \mathcal{X}^t\|_F^2 + \delta_D(\mathcal{X}). \tag{14}$$

Denote

$$H(\mathcal{X}) := f_{n,\mathcal{Y}}(\mathcal{X}) + \beta G_\lambda(\mathcal{X}) + \delta_D(\mathcal{X}) = f_{n,\mathcal{Y}}(\mathcal{X}) + \beta S_1(\mathcal{X}) - \beta S_2(\mathcal{X}) + \delta_D(\mathcal{X}), \tag{15}$$

and

$$\begin{aligned} Q(\mathcal{X}, \mathcal{X}^t) := f_{n,\mathcal{Y}}(\mathcal{X}^t) + \langle \nabla f_{n,\mathcal{Y}}(\mathcal{X}^t), \mathcal{X} - \mathcal{X}^t \rangle + \frac{\rho}{2}\|\mathcal{X} - \mathcal{X}^t\|_F^2 + \beta S_1(\mathcal{X}) - \beta S_2(\mathcal{X}^t) \\ - \beta \langle \nabla S_2(\mathcal{X}^t), \mathcal{X} - \mathcal{X}^t \rangle + \delta_D(\mathcal{X}). \end{aligned} \tag{16}$$

It can be easily verified that $Q(\mathcal{X}^t, \mathcal{X}^t) = H(\mathcal{X}^t)$.

However, it is difficult to compute the exact solution of problem (14) in practice. We consider to solve an inexact solution at each iteration of PMM. In particular, we propose an inexact version of PMM algorithm for solving problem (14). The error criteria of each iteration should satisfy the following condition: Find $\mathcal{W}^{t+1} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ such that

$$\mathcal{W}^{t+1} \in \partial Q(\mathcal{X}^{t+1}, \mathcal{X}^t) \text{ and } \|\mathcal{W}^{t+1}\|_F \leq \xi\rho\|\mathcal{X}^{t+1} - \mathcal{X}^t\|_F, \tag{17}$$

where $\xi \in (0, \frac{1}{2})$ is a constant.

Now, we state the PMM in Algorithm 1.

Algorithm 1 A PMM Algorithm for Solving Problem (12)

1: **Initialization:** Given parameter $\rho, \lambda, \gamma, \beta > 0$ and $\mathcal{W}^0$. For $t = 0, 1, 2, \ldots$

2: **repeat**

Find $\mathcal{W}^{t+1}$ such that $\mathcal{W}^{t+1} \in \partial Q(\mathcal{X}^{t+1}, \mathcal{X}^t)$ and $\|\mathcal{W}^{t+1}\|_F \leq \xi\rho\|\mathcal{X}^{t+1} - \mathcal{X}^t\|_F$.

3: **until** A stopping condition is satisfied.

Remark 4 *In Algorithm 1, we propose an inexact PMM algorithm for solving problem (12), where the condition in (17) should be satisfied. In particular, when $\mathcal{X}^{t+1}$ is the optimal solution of (14), we just choose $\mathcal{W}^{t+1}$ to be a zero tensor.*

3.2 Convergence Analysis

In this subsection, the global convergence and convergence rate of Algorithm 1 are established. First, the convergent result of Algorithm 1 is presented in the following theorem.

Theorem 3 *Let $\{\mathcal{X}^t\}$ be the sequence generated by Algorithm 1. Suppose that Assumption 2 hold, where the locally Lipschitz constant of s_2' is set as L_0. Assume that $\nabla f_{n,\mathcal{Y}}$ is Lipschitz continuous with Lipschitz constant L, and $f_{n,\mathcal{Y}}, g_\lambda(x)$ are KL functions. Then for any $\rho > \frac{L}{1-2\xi}$ with $\xi \in (0, \frac{1}{2})$, the sequence $\{\mathcal{X}^t\}$ converges to a stationary point $\tilde{\mathcal{X}}$ of (4) as t goes to infinity.*

The proof of Theorem 3 is left to Appendix G. The assumptions in Theorem 3 are very mild. The nonconvex functions in Table 2 are KL functions [50, 68]. Moreover, when $f_{n,\mathcal{Y}}$ is the logistic loss function (e.g., see (27)) or the least squares loss function (e.g., see (25)), it is a KL function [7, 68]. Besides, when $f_{n,\mathcal{Y}}$ is the least squares loss, the Lipschitz constant of $\nabla f_{n,\mathcal{Y}}$ is 1. And when $f_{n,\mathcal{Y}}$ is the logistic loss function in (27), it follows from Lemma 4 that the Lipschitz constant of $\nabla f_{n,\mathcal{Y}}$ is $\frac{1}{4n}\sum_{i=1}^n \|\mathcal{Z}_i\|_F^2$.

Remark 5 *The assumption about the locally Lipschitz continuity of s_2' is very mild. In fact, the derivatives of these functions in Table 2 are Lipschitz continuous. In particular, it is evident from [1, 68] that for Logarithm function, the Lipschitz constant of s_2' is $\frac{\lambda}{\gamma^2}$. For SCAD and MCP, the Lipschitz constants of s_2' are $\frac{1}{\gamma-1}$ and $\frac{1}{\gamma}$, respectively.*

Theorem 3 shows that the sequence $\{\mathcal{X}^t\}$ generated by Algorithm 1 converges to $\tilde{\mathcal{X}}$ globally as t tends to infinity, i.e. $\lim_{t\to\infty}\mathcal{X}^t=\tilde{\mathcal{X}}$. Now we also give the convergence rate of Algorithm 1, which is stated in the following theorem.

Theorem 4 *Let $\{\mathcal{X}^t\}$ be the sequence generated by Algorithm 1. Suppose that the assumptions in Theorem 3 hold and $H(\mathcal{X})$ defined in (15) satisfies the KL property at $\tilde{\mathcal{X}}$ with an exponent $\alpha\in[0,1)$. Then, we have the following results:*

(i) If $\alpha=0$, then the sequence $\{\mathcal{X}^t\}$ converges in a finite number of steps.

(ii) If $0<\alpha\leq\frac{1}{2}$, then the sequence $\{\mathcal{X}^t\}$ converges R-linearly, i.e., there exist $w>0$ and $\vartheta\in[0,1)$ such that $\|\mathcal{X}^t-\tilde{\mathcal{X}}\|_F\leq w\vartheta^t$.

(iii) If $\frac{1}{2}<\alpha<1$, then the sequence $\{\mathcal{X}^t\}$ converges R-sublinearly, i.e., there exists $w>0$ such that $\|\mathcal{X}^t-\tilde{\mathcal{X}}\|_F\leq wt^{-\frac{1-\alpha}{2\alpha-1}}$.

The proof of Theorem 4 is left to Appendix H. In particular, if we know the KL exponent of $H(\mathcal{X})$, the detailed convergence rate of PMM can be determined.

3.3 ADMM for Solving the Subproblem

Now we consider to solve the subproblem (14) by applying ADMM [17, 21]. Let $\mathcal{X}=\mathcal{M}$, then problem (14) is equivalent to

$$\begin{aligned}&\min_{\mathcal{X},\mathcal{M}}\ \beta S_1(\mathcal{M})+\langle\nabla f_{n,\mathcal{Y}}(\mathcal{X}^t)-\beta\nabla S_2(\mathcal{X}^t),\mathcal{X}-\mathcal{X}^t\rangle+\frac{\rho}{2}\|\mathcal{X}-\mathcal{X}^t\|_F^2+\delta_D(\mathcal{X})\\&\text{s.t. }\mathcal{X}=\mathcal{M}.\end{aligned}\tag{18}$$

The augmented Lagrangian function associated with (18) is given by

$$\begin{aligned}L(\mathcal{X},\mathcal{M},\mathcal{Z})=&\beta S_1(\mathcal{M})+\langle\nabla f_{n,\mathcal{Y}}(\mathcal{X}^t)-\beta\nabla S_2(\mathcal{X}^t),\mathcal{X}-\mathcal{X}^t\rangle+\frac{\rho}{2}\|\mathcal{X}-\mathcal{X}^t\|_F^2+\delta_D(\mathcal{X})\\&+\langle\mathcal{Z},\mathcal{X}-\mathcal{M}\rangle+\frac{\eta}{2}\|\mathcal{X}-\mathcal{M}\|_F^2,\end{aligned}$$

where $\eta>0$ is the penalty parameter and $\mathcal{Z}\in\mathbb{R}^{n_1\times n_2\times n_3}$ is the Lagrangian multiplier. Then the iteration of ADMM is given as follows:

$$\mathcal{M}^{k+1}=\arg\min_{\mathcal{M}}L(\mathcal{X}^k,\mathcal{M},\mathcal{Z}^k),\tag{19}$$

$$\mathcal{X}^{k+1}=\arg\min_{\mathcal{X}}L(\mathcal{X},\mathcal{M}^{k+1},\mathcal{Z}^k),\tag{20}$$

$$\mathcal{Z}^{k+1}=\mathcal{Z}^k+\tau\eta(\mathcal{X}^{k+1}-\mathcal{M}^{k+1}),\tag{21}$$

where $\tau\in(0,\frac{1+\sqrt{5}}{2})$ is the step size.

Now we consider to solve the subproblems in (19) and (20) in detail. Note that problem (19) can be equivalently written as

$$\begin{aligned}\mathcal{M}^{k+1}&=\arg\min_{\mathcal{M}}\beta S_1(\mathcal{M})+\frac{\eta}{2}\left\|\mathcal{M}-\left(\mathcal{X}^k+\frac{1}{\eta}\mathcal{Z}^k\right)\right\|_F^2\\&=\text{Prox}_{\frac{\beta}{\eta}S_1}\left(\mathcal{X}^k+\frac{1}{\eta}\mathcal{Z}^k\right).\end{aligned}\tag{22}$$

Problem (20) can be rewritten as

$$\begin{aligned}\mathcal{X}^{k+1} &= \arg\min_{\mathcal{X}} \delta_D(\mathcal{X}) + \langle \nabla f_{n,\mathcal{Y}}(\mathcal{X}^t) - \beta \nabla S_2(\mathcal{X}^t), \mathcal{X} - \mathcal{X}^t \rangle + \langle \mathcal{Z}^k, \mathcal{X} - \mathcal{M}^{k+1} \rangle \\ &\quad + \frac{\eta}{2}\|\mathcal{X} - \mathcal{M}^{k+1}\|_F^2 + \frac{\rho}{2}\|\mathcal{X} - \mathcal{X}^t\|_F^2 \\ &= \arg\min_{\mathcal{X}} \delta_D(\mathcal{X}) + \frac{\rho+\eta}{2}\left\|\mathcal{X} - \frac{1}{\rho+\eta}\mathcal{H}^{k+1}\right\|_F^2,\end{aligned}$$

where $\mathcal{H}^{k+1} = \rho\mathcal{X}^t - \nabla f_{n,\mathcal{Y}}(\mathcal{X}^t) + \beta\nabla S_2(\mathcal{X}^t) + \eta\mathcal{M}^{k+1} - \mathcal{Z}^k$. A simple computation leads to

$$\mathcal{X}^{k+1} = \mathcal{P}_D\left(\frac{1}{\rho+\eta}\mathcal{H}^{k+1}\right), \tag{23}$$

where $\mathcal{P}_D(\cdot)$ is the projection operator onto the set D given by $\mathcal{P}_D(\mathcal{Y}) = \max\{\min\{\mathcal{Y}_{ijk}, c\}, -c\}$.

Then the ADMM for solving problem (18) is stated in Algorithm 2.

Algorithm 2 An ADMM Algorithm for Solving Problem (18)

1: **Initialization:** Given initial value $\mathcal{X}^0, \mathcal{M}^0, \mathcal{Z}^0, \mathcal{X}^t$, parameters $\eta > 0, \tau \in (0, \frac{1+\sqrt{5}}{2})$.
2: **repeat**
3: **Step 1**. Compute $\mathcal{M}^{k+1}$ by (22).
4: **Step 2**. Compute $\mathcal{X}^{k+1}$ by (23).
5: **Step 3**. Update $\mathcal{Z}^{k+1}$ by (21).
6: **until** A stopping condition is satisfied.

Remark 6 *In Algorithm 2, we need to compute the proximal mapping of S_1. In the experiments, the function $s_1(x)$ is taken as $s_1(x) = \lambda x$ and then $S_1(\mathcal{X}) = \lambda\|\mathcal{X}\|_{\text{TTNN}}$. As a result, problem (22) can be equivalently written as*

$$\mathcal{M}^{k+1} = \text{Prox}_{\frac{\beta\lambda}{\eta}\|\cdot\|_{\text{TTNN}}}\left(\mathcal{X}^k + \frac{1}{\eta}\mathcal{Z}^k\right).$$

It follows from [59, Theorem 3] that

$$\mathcal{M}^{k+1} = \mathcal{U} \diamond_{\mathbf{U}} \Sigma_\lambda \diamond_{\mathbf{U}} \mathcal{V}^T,$$

where $\mathcal{X}^k + \frac{1}{\eta}\mathcal{Z}^k = \mathcal{U} \diamond_{\mathbf{U}} \Sigma \diamond_{\mathbf{U}} \mathcal{V}^T$, $\Sigma_\lambda = \mathbf{U}^T[\widehat{\Sigma}_\lambda]$ and $\widehat{\Sigma}_\lambda = \max\{\widehat{\Sigma}_{\mathbf{U}} - \frac{\beta\lambda}{\eta}, 0\}$.

Remark 7 *For (23), we need to compute $\nabla S_2(\mathcal{X}^t)$ in order to get $\mathcal{H}^{k+1}$. Note that s_2 is differentiable. By utilizing the differentiable case in Lemma 12, we obtain that*

$$\nabla S_2(\mathcal{X}^t) = \mathcal{U}^t \diamond_{\mathbf{U}} \mathcal{D}^t \diamond_{\mathbf{U}} (\mathcal{V}^t)^T,$$

where $\mathcal{X}^t = \mathcal{U}^t \diamond_{\mathbf{U}} \Sigma^t \diamond_{\mathbf{U}} (\mathcal{V}^t)^T$ and $(\widehat{\mathcal{D}^t})_{\mathbf{U}}^{\langle i \rangle} = \text{Diag}(s_2'((\widehat{\Sigma^t}_{\mathbf{U}}^{\langle i \rangle})_{11}), \ldots, s_2'((\widehat{\Sigma^t}_{\mathbf{U}}^{\langle i \rangle})_{mm}))$, $i = 1, \ldots, n_3$. Here $(\widehat{\Sigma^t}_{\mathbf{U}}^{\langle i \rangle})_{jj}$ represents the (j,j)-th entry of a matrix and $m = \min\{n_1, n_2\}$.

Note that problem (18) is convex, and Algorithm 2 is just the classical two-block ADMM, whose convergence has been established in [17, 21]. For brevity, we omit the details of convergence of Algorithm 2 here.

Now we give the computational cost of ADMM for solving the subproblem (14) based on the nonconvex functions in Table 2 and the least squares loss or logistic regression, which is the main iteration of PMM. Note that $S_1(\mathcal{X}) = \lambda\|\mathcal{X}\|_{\text{TTNN}}$ in Table 2. In this case, the computational cost of $\mathcal{M}^{k+1}$ is $O(n_{(1)}n_{(2)}^2n_3 + n_1n_2n_3^2)$ [59, Section 4.1], where $n_{(1)} = \max\{n_1, n_2\}$ and $n_{(2)} = \min\{n_1, n_2\}$. The main cost of $\mathcal{X}^{k+1}$ is to compute $\nabla S_2(\mathcal{X}^t)$, which is $O(n_{(1)}n_{(2)}^2n_3 + n_1n_2n_3^2)$ and only computes one time in ADMM. Therefore, the computational cost of ADMM in each iteration is $O(n_{(1)}n_{(2)}^2n_3 + n_1n_2n_3^2)$. Furthermore, the computational complexity of PMM in each iteration is $O((n_{(1)}n_{(2)}^2n_3 + n_1n_2n_3^2)k_m)$, where k_m represents the number of iterations of ADMM.

4 Numerical Experiments

In this section, some experiments are conducted to demonstrate the effectiveness of the proposed method. Our model combines the loss function and nonconvex regularization (called LFNR for short), where the least squares loss and logistic regression loss are used for the loss function in model (4). In this case, the corresponding problems are tensor completion and binary classification, respectively. For the nonconvex function $g_\lambda(\cdot)$ in model (4), we use the MCP in all experiments for simplicity. We remark that the performance of other nonconvex functions is similar to that of MCP. All experiments are conducted in MATLAB R2020b with an Intel Core i7-10750H 2.6GHz and 16GB RAM.

For the unitary matrix in TTNN, the choice is given as follows: First, an initial estimator $\mathcal{X}_1$ is obtained by discrete cosine transform in (12). Afterwards, we unfold $\mathcal{X}_1$ into a matrix $(\mathcal{X}_1)_{(3)}$ along the third-dimension and take the SVD of $(\mathcal{X}_1)_{(3)}$ as $(\mathcal{X}_1)_{(3)} = \mathbf{U}\Sigma\mathbf{V}^T$. Then $\mathbf{U}^T$ is the desirable unitary matrix in TTNN. More details about the choice of unitary transform can be referred to [59].

4.1 Stopping Criterion

Algorithm 1 will be terminated if $\frac{\|\mathcal{X}^{t+1}-\mathcal{X}^t\|_F}{\|\mathcal{X}^t\|_F} \le 5\times 10^{-4}$ or the number of iterations reaches 100.

The Karush-Kuhn-Tucker (KKT) condition of (18) is given as follows:

$$\mathcal{Z} \in \partial(\beta S_1(\mathcal{M})),\ \mathcal{M} = \mathcal{X},\ \rho(\mathcal{X}^t - \mathcal{X}) - \nabla f_{n,\mathcal{Y}}(\mathcal{X}^t) + \beta\nabla S_2(\mathcal{X}^t) - \mathcal{Z} \in \partial\delta_D(\mathcal{X}).$$

The relative KKT residual is employed to evaluate the accuracy of Algorithm 2:

$$\eta_{res} := \max\{\eta_e, \eta_d, \eta_p\}, \tag{24}$$

where

$$\eta_e := \frac{\|\mathcal{M}-\mathcal{X}\|_F}{1+\|\mathcal{M}\|_F+\|\mathcal{X}\|_F},\ \eta_d := \frac{\|\mathcal{M}-\text{Prox}_{\beta S_1}(\mathcal{M}+\mathcal{Z})\|_F}{1+\|\mathcal{M}\|_F+\|\mathcal{Z}\|_F},$$
$$\eta_p := \frac{\|\mathcal{X}-\text{Prox}_{\rho^{-1}\delta_D(\cdot)}(\mathcal{X}^t-\rho^{-1}(\nabla f_{n,\mathcal{Y}}(\mathcal{X}^t)-\beta\nabla S_2(\mathcal{X}^t)+\mathcal{Z}))\|_F}{1+\|\rho^{-1}\mathcal{Z}\|_F+\|\mathcal{X}^t\|_F+\|\rho^{-1}\nabla f_{n,\mathcal{Y}}(\mathcal{X}^t)\|_F+\|\rho^{-1}\beta\nabla S_2(\mathcal{X}^t)\|_F}.$$

Then Algorithm 2 will be terminated if $\eta_{res} \le 3\times 10^{-3}$ or the number of iterations exceeds 100.

4.2 Tensor Completion

In this subsection, we present numerical experiments for tensor completion, where the least squares loss is utilized for $f_{n,\mathcal{Y}}$. In this case, model (4) reduces to the least squares loss function with nonconvex regularization given by

$$\begin{aligned}&\min_{\mathcal{X}} \frac{1}{2p}\|\mathcal{P}_\Omega(\mathcal{X}-\mathcal{Y})\|_F^2 + \beta G_\lambda(\mathcal{X})\\ &\text{s.t. } \|\mathcal{X}\|_\infty \le c,\end{aligned} \tag{25}$$

Table 3: PSNR and SSIM values of different methods for the Balloons dataset with different σ and SRs.

	σ	SR	SNN	TMac	NORT	TTNN	LFNR
PSNR	0.005	0.05	24.50	35.23	32.79	33.91	**35.65**
		0.10	31.13	37.74	37.22	38.75	**40.01**
		0.15	34.41	39.05	39.44	41.76	**42.80**
		0.20	36.76	40.20	40.38	43.72	**44.57**
		0.25	38.57	41.25	41.59	45.46	**46.07**
		0.30	39.82	41.88	42.48	46.43	**47.21**
	0.01	0.05	23.96	33.41	33.60	33.24	**35.15**
		0.10	30.50	35.96	37.22	37.56	**38.78**
		0.15	33.38	37.65	38.79	39.24	**41.08**
		0.20	35.36	39.17	39.26	41.10	**42.47**
		0.25	36.69	40.26	39.96	42.35	**43.51**
		0.30	37.73	40.83	40.96	43.32	**44.34**
SSIM	0.005	0.05	0.8342	0.9053	0.8989	0.9057	**0.9281**
		0.10	0.9211	0.9447	0.9417	0.9615	**0.9693**
		0.15	0.9493	0.9604	0.9655	0.9768	**0.9819**
		0.20	0.9640	0.9692	0.9715	0.9838	**0.9868**
		0.25	0.9720	0.9747	0.9776	0.9875	**0.9899**
		0.30	0.9763	0.9777	0.9810	0.9898	**0.9913**
	0.01	0.05	0.8218	0.8385	0.9015	0.8854	**0.9058**
		0.10	0.9029	0.9064	0.9417	0.9421	**0.9541**
		0.15	0.9300	0.9384	0.9577	0.9593	**0.9661**
		0.20	0.9417	0.9545	0.9612	0.9640	**0.9756**
		0.25	0.9469	0.9621	0.9615	0.9752	**0.9790**
		0.30	0.9493	0.9667	0.9721	0.9787	**0.9828**

where $\mathcal{Y} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is the observed tensor only known its entries in Ω, Ω is the index set, $p := \frac{n}{n_1 n_2 n_3}$ denotes the probability of each element to be observed, and $\mathcal{P}_\Omega$ is the projection operator onto Ω such that the entry maintaining the same for the index in Ω and zero outside Ω. By choosing the regularization term $G_\lambda(\mathcal{X})$ as TTNN, model (25) is convex and called TTNN for short. We also compare with the following three methods: sum of nuclear norms of unfolding matrices of a tensor (SNN) [19], parallel matrix factorization for tensor completion (TMac)[1] [71], nonconvex regularized tensor algorithm (NORT)[2] [72].

The observed tensor is constructed as follows: For an $n_1 \times n_2 \times n_3$ tensor $\mathcal{X}$, we first add the zero mean Gaussian noise with standard deviation σ, which is denoted by $\mathcal{Y}$. Then the index set Ω is uniformly generated at random and we get $\mathcal{P}_\Omega(\mathcal{Y})$, where the sample ratio (SR) is defined as SR $:= \frac{|\Omega|}{n_1 n_2 n_3}$. Here $|\Omega|$ denotes the cardinality of Ω.

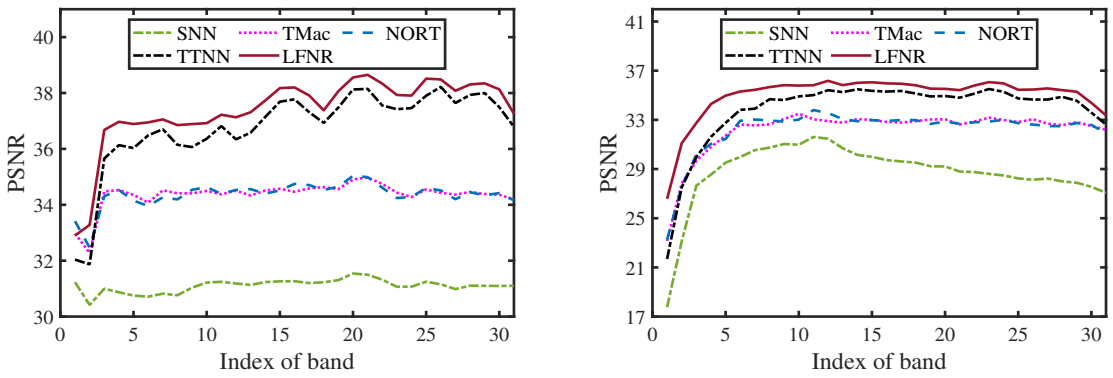
The peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) index [67] are adopted to

[1] https://xu-yangyang.github.io/TMac/

[2] https://github.com/quanmingyao/FasTer

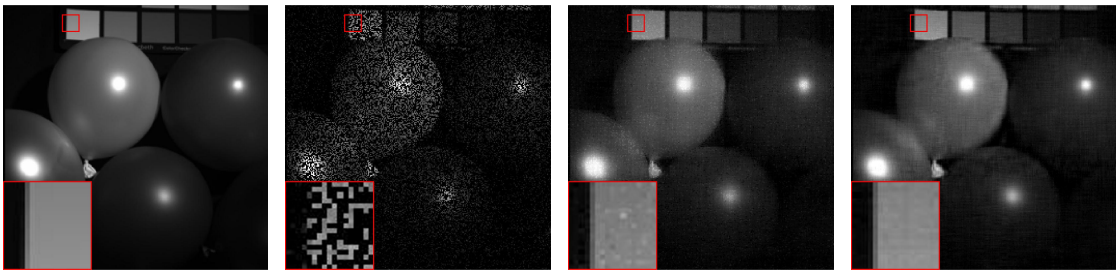
Table 4: PSNR and SSIM values of different methods for the Lemons dataset with different σ and SRs.

	σ	SR	SNN	TMac	NORT	TTNN	LFNR
PSNR	0.005	0.05	30.17	37.76	36.38	37.81	**39.98**
		0.10	35.60	40.46	39.53	42.24	**44.25**
		0.15	38.34	42.24	42.68	44.68	**46.69**
		0.20	40.19	43.56	44.24	46.37	**47.90**
		0.25	41.67	44.60	44.92	47.68	**49.15**
		0.30	42.87	45.46	45.67	48.76	**50.14**
	0.01	0.05	30.04	36.62	37.07	37.33	**38.99**
		0.10	34.54	39.93	40.57	41.25	**42.64**
		0.15	36.81	41.44	41.69	43.27	**44.44**
		0.20	38.34	42.58	42.21	44.53	**45.61**
		0.25	39.29	43.08	42.87	44.70	**46.40**
		0.30	40.02	43.45	43.66	44.90	**47.00**
SSIM	0.005	0.05	0.8775	0.9194	0.9137	0.9265	**0.9554**
		0.10	0.9410	0.9541	0.9529	0.9697	**0.9804**
		0.15	0.9608	0.9703	0.9742	0.9819	**0.9877**
		0.20	0.9699	0.9783	0.9806	0.9869	**0.9903**
		0.25	0.9756	0.9820	0.9830	0.9898	**0.9925**
		0.30	0.9787	0.9845	0.9852	0.9916	**0.9937**
	0.01	0.05	0.8760	0.9004	0.9210	0.9236	**0.9383**
		0.10	0.9231	0.9501	0.9591	0.9600	**0.9674**
		0.15	0.9390	0.9626	0.9657	0.9726	**0.9766**
		0.20	0.9464	0.9691	0.9687	0.9773	**0.9810**
		0.25	0.9490	0.9717	0.9710	0.9757	**0.9845**
		0.30	0.9498	0.9735	0.9738	0.9800	**0.9867**

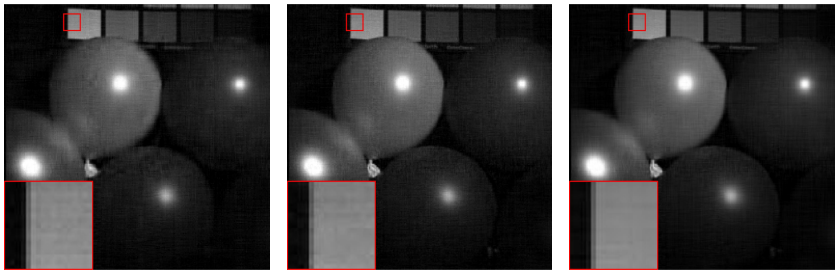


(a) Balloons (b) Lemons

Figure 1: PSNR values versus index of band of different methods for the Balloons and Lemons datasets. (a) Balloons dataset, where SR $= 0.5$ and $\sigma = 0.05$. (b) Lemons dataset, where SR $= 0.4$ and $\sigma = 0.1$.

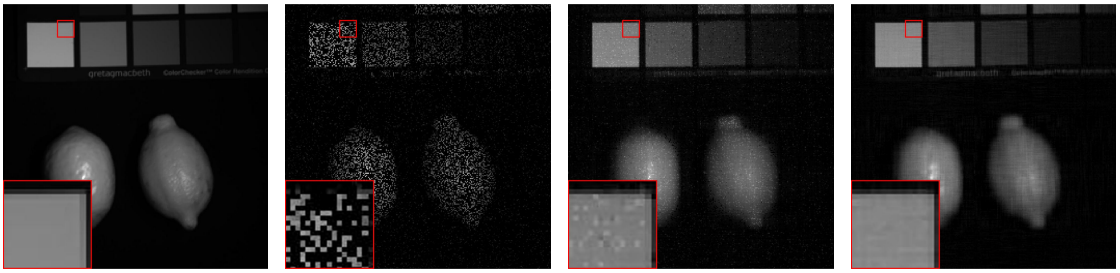


(a) Original (b) Observation (c) SNN: 30.48 (d) TMac: 34.24

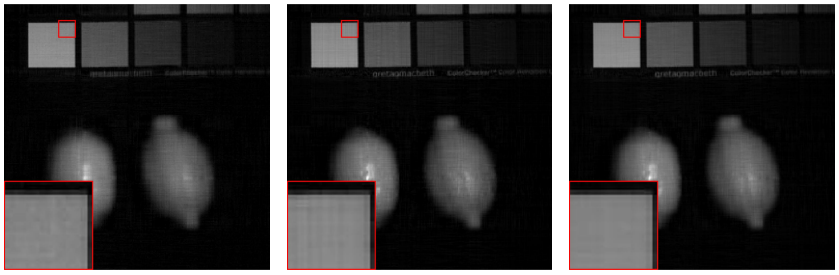


(e) NORT: 34.51 (f) TTNN: 35.57 (g) LFNR: 37.24

Figure 2: The recovered images (with PSNR values) and zoomed regions of different methods for the 30th band of the Balloons dataset, where SR $= 0.4$ and $\sigma = 0.05$.



(a) Original (b) Observation (c) SNN: 29.86 (d) TMac: 36.36



(e) NORT: 36.29 (f) TTNN: 37.58 (g) LFNR: 38.61

Figure 3: The recovered images (with PSNR values) and zoomed regions of different methods for the 30th band of the Lemons dataset, where SR $= 0.3$ and $\sigma = 0.05$.

measure the recovery performance for real-world data. Specifically, the PSNR is defined as

$$\text{PSNR} = 10\log_{10}\frac{n_1n_2n_3(\mathcal{X}_{\max}-\mathcal{X}_{\min})^2}{\|\widetilde{\mathcal{X}}-\mathcal{X}\|_F^2},$$

where $\mathcal{X}_{\max}$ and $\mathcal{X}_{\min}$ denote the maximum and minimum entries of $\mathcal{X}$, respectively, and $\widetilde{\mathcal{X}}$ and $\mathcal{X}$ are the recovered and underlying tensors, respectively. The SSIM is defined as

$$\text{SSIM} = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1) + (\sigma_x^2 + \sigma_y^2 + c_2)},$$

where μ_x and σ_x represent the mean intensity and standard deviation of the original image, respectively, μ_y and σ_y represent the mean intensity and standard deviation of the recovered image, respectively, σ_{xy} denotes the covariance between the original and recovered images, and $c_1, c_2 > 0$ are constants. For multi-dimensional images, the SSIM denotes the mean value of SSIM of all images.

4.2.1 Parameter Settings

For the parameters η and τ in Algorithm 2, we set $\eta = 10$ and $\tau = 1.618$ in the experiments for simplicity. For the parameter ρ in (14), it is set to 10. β is sensitive to the results for different cases and we select it from the set $\{0.2, 0.3, 0.4, 0.6, 0.7, 0.8, 0.9, 1.2, 1.3, 1.4, 1.5, 2\}$ to obtain the best recovery performance for tensor completion. For the two parameters γ and λ in MCP, we simply set $\gamma = 2.7$ and choose λ from the set $\{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.1, 1.2, 1.5, 1.6, 1.8, 2, 2.5\}$ to achieve the best performance in the testing cases.

4.2.2 Multispectral Images

For tensor completion, we test two multispectral images to demonstrate the effectiveness of the proposed method, including Balloons and Lemons[3], whose sizes are both $256 \times 256 \times 31$. In Tables 3 and 4, we show the PSNR and SSIM values of different methods for the Balloons and Lemons datasets, respectively, where $\sigma = 0.005, 0.01$ and SR $= 0.05, 0.10, 0.15, 0.20, 0.25, 0.30$. The best results are highlighted in bold and the second best results are highlighted in underline. We can see that the PSNR and SSIM obtained by LFNR are higher than those obtained by other methods. In particular, the LFNR outperforms NORT in terms of PSNR and SSIM values, which demonstrates that the nonconvex regularization based on TTNN is better than that based on the sum of nuclear norms of unfolding matrices of a tensor. Besides, the TTNN performs better than SNN, TMac, and NORT for most cases, especially for $\sigma = 0.01$.

Figure 1 displays the PSNR values versus index of band of different methods for the Balloons and Lemons datasets, where SR $= 0.5, \sigma = 0.05$ for the Balloons dataset, and SR $= 0.4, \sigma = 0.1$ for the Lemons dataset. As can be seen from the two figures that the LFNR performs best compared with SNN, TMac, NORT, and TTNN for almost all bands. And the TTNN outperforms SNN, TMac, and NORT in terms of PSNR values for most bands.

Figures 2 and 3 present the recovered images and zoomed regions of the 30th band of different methods for the Balloons and Lemons datasets, where SR $= 0.4, \sigma = 0.05$ for the Balloons datasets, and SR $= 0.3, \sigma = 0.05$ for the Lemons dataset. We can see that the images recovered by LFNR are better than other comparison methods in term of visual quality, especially from the zoomed regions. In fact, the LFNR can preserve the details and edges of images better than SNN, TMac, NORT, and TTNN.

4.2.3 MRI Dataset

In this subsection, we test a magnetic resonance imaging (MRI) dataset from Brainweb[4] to show the good performance of LFNR, where the size of MRI is $181 \times 217 \times 100$ and the noise level σ is set to

[3] https://www.cs.columbia.edu/CAVE/databases/multispectral/

[4] https://brainweb.bic.mni.mcgill.ca/brainweb/

Table 5: PSNR and SSIM values of different methods for the MRI dataset with different SRs.

Index	SR	SNN	TMac	NORT	TTNN	LFNR
PSNR	0.05	14.22	17.41	17.30	17.28	**17.88**
	0.10	15.58	19.85	19.51	19.13	**20.02**
	0.15	16.81	20.81	20.80	20.59	**21.77**
	0.20	17.99	21.4	22.03	21.84	**23.24**
	0.25	19.12	21.88	23.56	23.09	**24.50**
	0.30	20.21	22.31	24.26	24.17	**25.60**
	0.35	21.31	22.70	24.76	25.33	**26.71**
	0.40	22.44	23.08	25.23	26.36	**27.75**
	0.45	23.56	24.80	25.67	27.31	**28.72**
	0.50	24.71	25.31	26.11	28.29	**29.68**
SSIM	0.05	0.2485	0.3447	0.3104	0.3107	**0.3526**
	0.10	0.3346	0.5143	0.4955	0.4567	**0.5136**
	0.15	0.4282	0.5805	0.5786	0.6218	**0.6268**
	0.20	0.5174	0.6193	0.6572	0.6189	**0.6987**
	0.25	0.5966	0.6527	0.7398	0.6913	**0.7521**
	0.30	0.6651	0.6803	0.7644	0.7455	**0.7914**
	0.35	0.7239	0.7018	0.7838	0.7841	**0.8261**
	0.40	0.7740	0.7229	0.8019	0.8189	**0.8601**
	0.45	0.8155	0.8025	0.8192	0.8378	**0.8823**
	0.50	0.8500	0.8214	0.8359	0.8607	**0.9002**

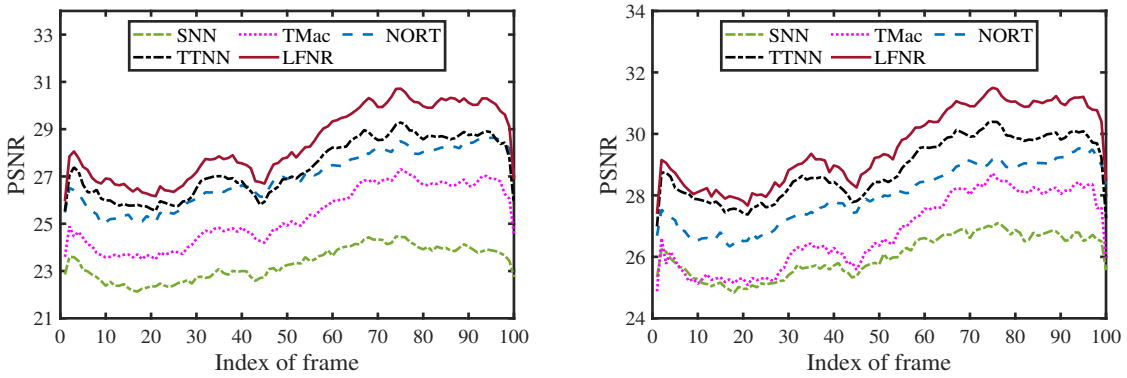
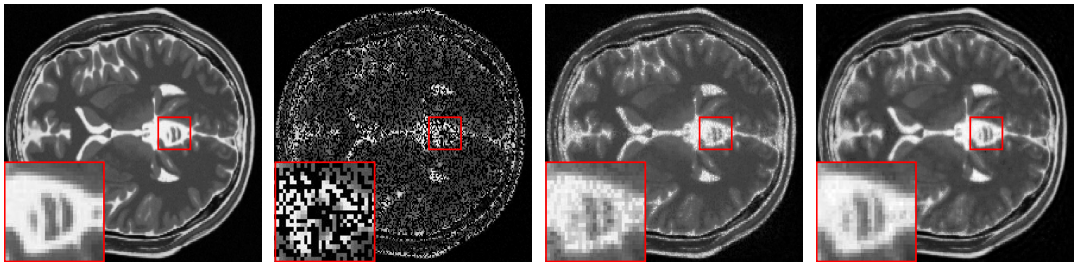
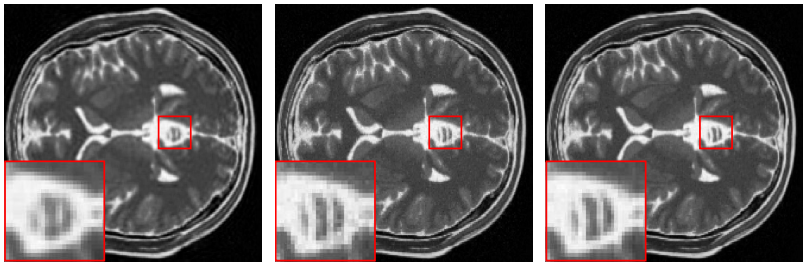


Figure 4: PSNR values versus index of frame of different methods for the MRI dataset. (a) SR $= 0.45$ and $\sigma = 0.005$. (b) SR $= 0.6$ and $\sigma = 0.02$.

be 0.01. Table 5 lists the PSNR and SSIM values of different methods for the MRI dataset at different SRs. where the best and second results are highlighted in bold and in underline, respectively. It can be observed from Table 5 that the LFNR, TMac, NORT and TTNN always reach higher PSNR and SSIM values than the SNN model. In particular, the proposed LFNR method obtains the best recovery performance on both PSNR and SSIM values compared with other approaches. And the LFNR method



(a) Original (b) Observation (c) SNN: 25.77 (d) TMac: 27.73



(e) NORT: 28.55 (f) TTNN: 30.27 (g) LFNR: 31.51

Figure 5: The recovered images (with PSNR values) and zoomed regions of different methods for the 74th frontal slice of the MRI dataset, where SR $= 0.5$ and $\sigma = 0.005$.

has at least 1dB improvement in terms of PSNR values in comparison with the TTNN model.

In Figure 4, we show the PSNR values of each frame of different methods for the MRI dataset, where SR $= 0.45$ and $\sigma = 0.005$ in Figure 4(a), and SR $= 0.6, \sigma = 0.02$ in Figure 4(b). We can see that LFNR achieves higher PSNR values than other methods for all frames. Besides, the TTNN outperforms SNN, TMac, NORT in most frames. Figure 5 exhibits the visual results of different methods on the 74th frontal slice of the MRI dataset, where SR $= 0.5, \sigma = 0.005$. Compared with SNN, TMac, NORT and TTNN, the images obtained by LFNR is visually closest to the original image, especially for the zoomed regions. In fact, the LFNR can preserve the details and texture better than other methods.

4.3 Logistic Regression for Binary Classification

In this subsection, we investigate the logistic regression loss for binary classification in problem (4). Specifically, we consider the case in which the observation pairs $\{(\mathcal{Z}_i, y_i)\}_{i=1}^n$ are drawn independent and identically distributed (i.i.d.) from a distribution of the form

$$\mathbb{P}(y_i|\mathcal{Z}_i, \mathcal{X}) = \exp\{y_i\langle \mathcal{Z}_i, \mathcal{X}\rangle - \log\left(1 + \exp(\langle \mathcal{Z}_i, \mathcal{X}\rangle)\right)\}. \tag{26}$$

Here the response y_i takes binary values $\{0, 1\}$. In addition, $\mathcal{X}$ is an unknown parameter tensor, which need to be estimated. By maximum likelihood estimate, the loss function based on the distribution of the observations in (26) can be written as

$$f_{n,y}(\mathcal{X}) := \frac{1}{n}\sum_{i=1}^{n}[\log(1 + \exp(\langle \mathcal{Z}_i, \mathcal{X}\rangle)) - y_i\langle \mathcal{Z}_i, \mathcal{X}\rangle]. \tag{27}$$

Then problem (4) reduces to

$$\widetilde{\mathcal{X}} = \underset{\|\mathcal{X}\|_\infty \le c}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} [\log\left(1 + \exp(\langle \mathcal{Z}_i, \mathcal{X} \rangle)\right) - y_i \langle \mathcal{Z}_i, \mathcal{X} \rangle] + \beta G_\lambda(\mathcal{X}) \right\}. \tag{28}$$

Model (28) combines the logistic regression loss with nonconvex regularization. In particular, when G_λ is the TTNN, model (28) is also called TTNN for brevity.

We compare our methods with the support vector machine (SVM), which comes from the **fitcsvm** function in MATLAB, the dual structure preserving kernels approach (DuSK)[5] [25], support tensor train machine by employing tensor train as the parameter model (STTM)[6] [13], logistic loss with overlapped trace norm (LOTN) [69]. The CIFAR-10 dataset[7] is tested in the experiments for binary classification, where there are 60000 color images ($32 \times 32 \times 3$) from 10 classes, including 50000 training images and 10000 testing images. In the following experiments, each case run 10 times to reduce the impact of randomness, and the final results are reported by the average value of all results of ten times. In particular, $\mathcal{Z}_i$ (with size $32 \times 32 \times 3$) are the training images and y_i are the labels of the corresponding training images in our experiments for image classification, $i = 1, \ldots, n$. Here n is the number of training samples.

Let $\tilde{m}$ be the number of the testing images. Then the response of the testing images is defined as

$$y_j^{\text{testp}} = \frac{1}{1 + \exp(-\langle \mathcal{Z}_j^{\text{test}}, \widetilde{\mathcal{X}} \rangle)}, \quad j = 1, 2, \ldots, \tilde{m},$$

where $\mathcal{Z}_j^{\text{test}}$ denotes the j-th testing image and $\widetilde{\mathcal{X}}$ is the trained parameter in (28). If $y_j^{\text{testp}} > 0.5$, set $y_j^{\text{testp}} = 1$, otherwise, $y_j^{\text{testp}} = 0$. The testing accuracy (TAc) of classification is defined as

$$\text{TAc} = 1 - \frac{1}{\tilde{m}} \sum_{j=1}^{\tilde{m}} |y_j^{\text{testp}} - y_j^{\text{test}}|,$$

where y_j^{test} denotes the true category of the j-th testing image.

4.3.1 Parameter Settings

For the experiments of image classification, we set $\rho = 100$, and choose η from the set $\{10, 100\}$ to get the best classification results. For the parameters γ and λ in the MCP function, we choose γ from the set $\{0.2, 0.3, 0.4, 0.5, 0.6, 0.8, 0.9, 1.5, 2.5, 3.4, 5, 6\}$ and λ from the set $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.8\}$ to obtain the best classification accuracy. In addition, the penalty parameter β is chosen from the set $\{0.1, 0.2, 0.3, 0.4, 0.6, 0.7\}$ to achieve the best results.

4.3.2 Image Classification

In this subsection, we demonstrate the effectiveness of the LFNR for binary image classification. First, we select two out of the ten classes in CIFAR-10. Then the first 250 images of the two classes are used in our experiments. More specifically, we randomly choose 200 images from each class for training, and employ the remaining ones as the testing images. Table 6 displays the classification accuracy of different methods for the CIFAR-10 dataset. It can be observed from this table that the LFNR

[5] https://github.com/LifangHe/SDM14_DuSK
[6] https://github.com/git2cchen/STTM
[7] https://www.cs.toronto.edu/~kriz/cifar.html

Table 6: Classification accuracy of different methods for the CIFAR-10 dataset.

Class 1	Class 2	SVM	DuSK	STTM	LOTN	TTNN	LFNR
dog	frog	0.675	0.738	0.721	0.696	0.747	**0.753**
automobile	bird	0.787	0.839	0.826	0.820	0.857	**0.861**
bird	deer	0.625	0.719	0.676	0.649	0.715	**0.737**
cat	bird	0.637	0.712	0.716	0.700	0.738	**0.742**
bird	dog	0.632	0.719	0.703	0.647	0.715	**0.730**
truck	automobile	0.598	0.668	0.666	0.624	0.680	**0.686**
bird	horse	0.685	0.727	0.695	0.714	0.723	**0.750**
deer	frog	0.665	0.735	0.705	0.712	0.747	**0.752**
frog	automobile	0.836	0.879	0.855	0.853	0.868	**0.882**
bird	airplane	0.623	0.780	0.752	0.689	0.796	**0.812**

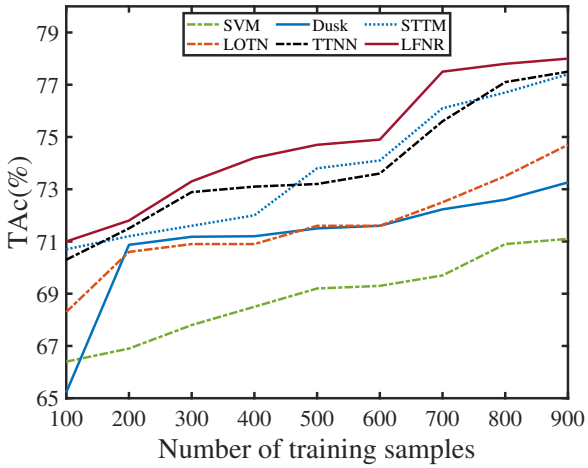


Figure 6: Classification accuracy versus number of training samples of different methods for the CIFAR-10 dataset.

performs better than SVM, DuSK, STTM, LOTN, and TTNN in terms of classification accuracy for the testing cases. Besides, for the classes 'bird' and 'deer', 'bird' and 'dog', 'bird' and 'house', 'frog' and 'automobile', the DuSK outperforms TTNN in terms of classification accuracy. However, the TTNN performs better than other methods for other testing cases. In particular, the TTNN can achieve higher classification accuracy than LOTN for these images, which implies that the TTNN can explore the low-rankness better than overlapped trace norm.

Now we test the influence of different number of training samples of different methods for image classification. Two categories of 'deer' and 'horse' are used in our experiments, where 100 images are randomly selected from each class as the testing samples and the number of training samples varies from 100 to 900 with step size 100. In Figure 6, we show the classification accuracy of different methods versus number of training samples for the 'deer' and 'horse' classes. We can see that the classification accuracy of LFNR is higher than those of SVM, DuSK, STTM, LOTN, and TTNN. Furthermore, the classification accuracy of these methods increases as the number of training samples increases.

5 Concluding Remarks

In this paper, we have studied the problem of low-rank tensor learning and proposed the LFNR model based on TTNN. Specifically, in order to explore the low-rankness of the underlying tensor, a family of nonconvex functions were employed onto the singular values of all frontal slices of a tensor in the transformed domain. Besides, a general loss function was considered in the proposed model, which could be specified to the least squares loss and logistic regression loss, respectively. And the corresponding models were suitable for tensor completion and binary classification. Then we established the error bound between the stationary point of the LFNR model and the underlying tensor under the RSC condition on the loss function and some regularity conditions on the nonconvex penalty function. Furthermore, based on the DC structure of the nonconvex regularization, a PMM algorithm was designed to solve the proposed model, where the loss function and one convex function in DC structure were linearized with a proximal term. Besides, we showed that the sequence generated by PMM converges globally to a stationary point of LFNR under very mild conditions. Numerical experiments on tensor completion and binary classification were reported to demonstrate the effectiveness of LFNR compared with other methods.

In the future work, we are going to extend our model to more general loss functions, where the RSC conditions are utilized to more general loss functions. Another possible extension is to incorporate the random method instead of using SVD directly to accelerate the PMM. It would be also of great interest to analyze the KL exponent of the objective function of the LFNR model.

Acknowledgments

The authors would like to thank Dr. Silvia Gandy for providing the code of SNN in [19].

Appendix A. Least Squares Loss for Tensor Completion

In the following, we show that the loss function in (25) satisfies the RSC condition. The least squares loss function in (25) can be also rewritten as

$$f_{n,\mathcal{Y}}(\mathcal{X}) = \frac{1}{2p}\|\mathcal{P}_\Omega(\mathcal{X}-\mathcal{Y})\|_F^2 = \frac{1}{2p}\sum_{(i,j,k)\in\Omega}(\langle \mathcal{A}^{ijk}, \mathcal{X}\rangle - \mathcal{Y}_{ijk})^2,$$

where $\mathcal{A}^{ijk} \in \mathbb{R}^{n_1\times n_2\times n_3}$ denotes a basic tensor whose (i,j,k)-th element is 1 and other elements are 0. Note that the gradient of the loss function $f_{n,\mathcal{Y}}(\mathcal{X})$ is given by

$$\nabla f_{n,\mathcal{Y}}(\mathcal{X}) = \frac{1}{p}\sum_{(i,j,k)\in\Omega}(\langle \mathcal{A}^{ijk}, \mathcal{X}\rangle - \mathcal{Y}_{ijk})\mathcal{A}^{ijk}. \tag{29}$$

Now we show that $f_{n,\mathcal{Y}}(\mathcal{X})$ satisfies the RSC condition in (6) with high probability, which is stated in the following lemma.

Lemma 1 *For any $\tilde{\mathcal{V}} \in \mathbb{R}^{n_1\times n_2\times n_3}$, there exists a constant $\eta_1 \in (0,1)$ such that*

$$\langle \nabla f_{n,\mathcal{Y}}(\mathcal{X}^* + \tilde{\mathcal{V}}) - \nabla f_{n,\mathcal{Y}}(\mathcal{X}^*), \tilde{\mathcal{V}}\rangle \geq (1-\eta_1)\|\tilde{\mathcal{V}}\|_F^2, \tag{30}$$

with probability at least $1 - \exp(-\frac{2\eta_1^2 n^2}{d^3})$.

Proof. For any $\tilde{\mathcal{V}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, we let $b := \|\tilde{\mathcal{V}}\|_\infty$. By (29), one can get

$$\begin{aligned}
&\langle \nabla f_{n,\mathcal{Y}}(\mathcal{X}^* + \tilde{\mathcal{V}}) - \nabla f_{n,\mathcal{Y}}(\mathcal{X}^*), \tilde{\mathcal{V}} \rangle \\
&= \left\langle \frac{1}{p} \sum_{(i,j,k)\in\Omega} (\langle \mathcal{A}^{ijk}, \mathcal{X}^* + \tilde{\mathcal{V}} \rangle - \mathcal{Y}_{ijk}) \mathcal{A}^{ijk} - (\langle \mathcal{A}^{ijk}, \mathcal{X}^* \rangle - \mathcal{Y}_{ijk}) \mathcal{A}^{ijk}, \tilde{\mathcal{V}} \right\rangle \\
&= \left\langle \frac{1}{p} \sum_{(i,j,k)\in\Omega} \langle \mathcal{A}^{ijk}, \tilde{\mathcal{V}} \rangle \mathcal{A}^{ijk}, \tilde{\mathcal{V}} \right\rangle \\
&= \frac{1}{p} \sum_{(i,j,k)\in\Omega} (\langle \mathcal{A}^{ijk}, \tilde{\mathcal{V}} \rangle)^2.
\end{aligned} \tag{31}$$

Denote $\xi_{ijk} = (\langle \mathcal{A}^{ijk}, \tilde{\mathcal{V}} \rangle)^2$ if $(i,j,k) \in \Omega$, 0 otherwise, $1 \le i \le n_1, 1 \le j \le n_2, 1 \le k \le n_3$. Notice that the observations are sampled at random, we know that

$$\xi_{ijk} = \begin{cases} \tilde{\mathcal{V}}_{ijk}^2, & \text{with probability } p, \\ 0, & \text{with probability } 1-p. \end{cases} \tag{32}$$

Note that $\{\xi_{ijk}\}$ are independent random variables and the expectation of ξ_{ijk} is $\mathbb{E}(\xi_{ijk}) = p\tilde{\mathcal{V}}_{ijk}^2$. Let

$$\alpha = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} \xi_{ijk},$$

then $\mathbb{E}(\alpha) = p \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} \tilde{\mathcal{V}}_{ijk}^2 = p\|\tilde{\mathcal{V}}\|_F^2$. Notice that $0 \le \xi_{ijk} \le b^2$. For any $\epsilon > 0$, it follows from the Hoeffding's inequality [64, Proposition 2.5] that

$$\mathbb{P}(\alpha - \mathbb{E}(\alpha) < -\epsilon) < \exp\left(-\frac{2\epsilon^2}{db^4}\right).$$

By taking $\epsilon := \eta_1 p \|\tilde{\mathcal{V}}\|_F^2$, where $\eta_1 \in (0,1)$ is a constant, we deduce that

$$\begin{aligned}
\mathbb{P}\left(\alpha < \mathbb{E}(\alpha) - \eta_1 p \|\tilde{\mathcal{V}}\|_F^2\right) &= \mathbb{P}\left(\frac{1}{p}\alpha < (1-\eta_1)\|\tilde{\mathcal{V}}\|_F^2\right) \\
&< \exp\left(-\frac{2\eta_1^2 p^2 \|\tilde{\mathcal{V}}\|_F^4}{db^4}\right),
\end{aligned} \tag{33}$$

which implies

$$\mathbb{P}\left(\frac{1}{p}\alpha \ge (1-\eta_1)\|\tilde{\mathcal{V}}\|_F^2\right) \ge 1 - \exp\left(-\frac{2\eta_1^2 p^2 \|\tilde{\mathcal{V}}\|_F^4}{db^4}\right). \tag{34}$$

Since $1 \le \sqrt{d}\frac{\|\tilde{\mathcal{V}}\|_\infty}{\|\tilde{\mathcal{V}}\|_F} \le \sqrt{d}$, we obtain that $\frac{1}{d^2} \le \frac{\|\tilde{\mathcal{V}}\|_F^4}{d^2\|\tilde{\mathcal{V}}\|_\infty^4} \le 1$, which implies that

$$0 < \frac{1}{d} \le \frac{\|\tilde{\mathcal{V}}\|_F^4}{d\|\tilde{\mathcal{V}}\|_\infty^4} = \frac{\|\tilde{\mathcal{V}}\|_F^4}{db^4} \le d.$$

As a consequence, (34) further implies that

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{p}\alpha \ge (1-\eta_1)\|\tilde{\mathcal{V}}\|_F^2\right) &= \mathbb{P}\left(\frac{1}{p}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\xi_{ijk} \ge (1-\eta_1)\|\tilde{\mathcal{V}}\|_F^2\right) \\
&\ge 1 - \exp\left(-\frac{2\eta_1^2 n^2}{d^3}\right),
\end{aligned} \tag{35}$$

Combining (35) with (31) yields

$$\mathbb{P}\left(\langle \nabla f_{n,\mathcal{Y}}(\mathcal{X}^* + \tilde{\mathcal{V}}) - \nabla f_{n,\mathcal{Y}}(\mathcal{X}^*), \tilde{\mathcal{V}} \rangle \geq (1-\eta_1)\|\tilde{\mathcal{V}}\|_F^2\right) \geq 1 - \exp\left(-\frac{2\eta_1^2 n^2}{d^3}\right).$$

This concludes the proof. □

From Lemma 1, we know that $f_{n,\mathcal{Y}}(\mathcal{X})$ defined in (25) satisfies the RSC condition in (6) by taking $\alpha_1 = \alpha_2 = 1 - \eta_1$, $\tau_1 = \tau_2 = 0$.

Appendix B. Logistic Regression Loss

First, we give the definition and property of a sub-Gaussian random variable, which plays a vital role in the proof of Lemma 2.

Definition 7 *[18] A random variable x is called sub-Gaussian if there exist constants $a_1, \kappa > 0$ such that $\mathbb{P}(|x| \geq t) \leq a_1 e^{-\kappa t^2}$ for any $t > 0$.*

Proposition 1 *[18, Proposition 7.24] If x is sub-Gaussian with $\mathbb{E}(x) = 0$, then there exists a constant $\tilde{b}$ (depending only on a_1 and κ) such that*

$$\mathbb{E}[\exp(\theta x)] \leq \exp\left(\tilde{b}\theta^2\right) \quad \text{for all } \theta \in \mathbb{R}, \tag{36}$$

where the constant $\tilde{b}$ is called a sub-Gaussian parameter of x.

Now we show that the logistic loss function $f_{n,y}$ in (27) satisfies the RSC condition (6) with high probability in the following lemma.

Lemma 2 *Assume that $\{(\mathcal{Z}_i, y_i)\}_{i=1}^n$ are drawn i.i.d., where $\mathcal{Z}_i, i = 1, 2, \ldots, n$, are sub-Gaussian with independent, mean-zero, and sub-Gaussian entries with the same sub-Gaussian parameter $\tilde{b}$ in (36), and the response variables $y_i \in \{0, 1\}$. Suppose that the number of samples $n \geq 4c_2^4 t^2 \log d$, where $c_2 > 0$ is a given constant and t is defined in Section 2.1. Then there exists some positive constant c_3 such that the loss function $f_{n,y}$ in (27) satisfies the RSC condition (6) with probability at least $1 - \exp(-c_3 \log d)$.*

Proof. Note that the gradient of $f_{n,y}(\mathcal{X})$ in (27) is

$$\nabla f_{n,y}(\mathcal{X}) = \frac{1}{n}\sum_{i=1}^n \left[\left(\frac{\exp(\langle \mathcal{Z}_i, \mathcal{X}\rangle)}{1+\exp(\langle \mathcal{Z}_i, \mathcal{X}\rangle)} - y_i\right)\mathcal{Z}_i\right]. \tag{37}$$

For any third-order tensor $\tilde{\mathcal{V}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, we have

$$\begin{aligned}
&\nabla f_{n,y}(\mathcal{X}^* + \tilde{\mathcal{V}}) - \nabla f_{n,y}(\mathcal{X}^*) \\
&= \frac{1}{n}\sum_{i=1}^n \left[\left(\frac{\exp(\langle \mathcal{Z}_i, \mathcal{X}^* + \tilde{\mathcal{V}}\rangle)}{1+\exp(\langle \mathcal{Z}_i, \mathcal{X}^* + \tilde{\mathcal{V}}\rangle)} - y_i\right)\mathcal{Z}_i\right] - \frac{1}{n}\sum_{i=1}^n \left[\left(\frac{\exp(\langle \mathcal{Z}_i, \mathcal{X}^*\rangle)}{1+\exp(\langle \mathcal{Z}_i, \mathcal{X}^*\rangle)} - y_i\right)\mathcal{Z}_i\right] \\
&= \frac{1}{n}\sum_{i=1}^n \left[\left(\frac{\exp(\langle \mathcal{Z}_i, \mathcal{X}^* + \tilde{\mathcal{V}}\rangle)}{1+\exp(\langle \mathcal{Z}_i, \mathcal{X}^* + \tilde{\mathcal{V}}\rangle)} - \frac{\exp(\langle \mathcal{Z}_i, \mathcal{X}^*\rangle)}{1+\exp(\langle \mathcal{Z}_i, \mathcal{X}^*\rangle)}\right)\mathcal{Z}_i\right].
\end{aligned}$$

Therefore, one can deduce that

$$\begin{aligned}&\langle \nabla f_{n,y}(\mathcal{X}^*+\tilde{\mathcal{V}})-\nabla f_{n,y}(\mathcal{X}^*),\tilde{\mathcal{V}}\rangle\\&=\frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{\exp(\langle \mathcal{Z}_i,\mathcal{X}^*+\tilde{\mathcal{V}}\rangle)}{1+\exp(\langle \mathcal{Z}_i,\mathcal{X}^*+\tilde{\mathcal{V}}\rangle)}-\frac{\exp(\langle \mathcal{Z}_i,\mathcal{X}^*\rangle)}{1+\exp\left(\langle \mathcal{Z}_i,\mathcal{X}^*\rangle\right)}\right)\right]\langle \mathcal{Z}_i,\tilde{\mathcal{V}}\rangle.\end{aligned}\tag{38}$$

Define the function $\psi(u):=\log\left(1+\exp(u)\right)$. Note that the first-order derivative of $\psi(u)$ is $\psi'(u)=\frac{\exp(u)}{1+\exp(u)}$. Then (38) can be rewritten as

$$\begin{aligned}\langle \nabla f_{n,y}(\mathcal{X}^*+\tilde{\mathcal{V}})-\nabla f_{n,y}(\mathcal{X}^*),\tilde{\mathcal{V}}\rangle&=\frac{1}{n}\sum_{i=1}^{n}\left(\psi'(\langle \mathcal{Z}_i,\mathcal{X}^*+\tilde{\mathcal{V}}\rangle)-\psi'(\langle \mathcal{Z}_i,\mathcal{X}^*\rangle)\right)\langle \mathcal{Z}_i,\tilde{\mathcal{V}}\rangle\\&=\frac{1}{n}\sum_{i=1}^{n}\psi''(\langle \mathcal{Z}_i,\mathcal{X}^*\rangle+t_i\langle \mathcal{Z}_i,\tilde{\mathcal{V}}\rangle)\langle \mathcal{Z}_i,\tilde{\mathcal{V}}\rangle^2,\end{aligned}\tag{39}$$

where the second equality holds by the mean value theorem and $t_i\in[0,1]$. Now we consider two cases $\|\tilde{\mathcal{V}}\|_F\le 1$ and $\|\tilde{\mathcal{V}}\|_F>1$.

Case I. Suppose that $\|\tilde{\mathcal{V}}\|_F=\delta\in(0,1]$. First, similar to [64, Theorem 9.36], we introduce the following assumptions

$$\mathbb{E}(\langle \mathcal{Z}_i,\tilde{\mathcal{V}}\rangle^2)\ge\tau_u \text{ and } \mathbb{E}(\langle \mathcal{Z}_i,\tilde{\mathcal{V}}\rangle^4)\le\tau_l,\quad \text{for any } \tilde{\mathcal{V}}\in\mathbb{R}^{n_1\times n_2\times n_3} \text{ with } \|\tilde{\mathcal{V}}\|_F=1,\tag{40}$$

where $\tau_u,\tau_l>0$ are given constants. In (40), the first inequality is used to control the lower bound of covariance of the entries of $\mathcal{Z}_i$ and the second inequality is used to simplify the upper bound of $\mathbb{E}(\langle \mathcal{Z}_i,\tilde{\mathcal{V}}\rangle^4)$. Indeed, by simple calculations, we can get $\mathbb{E}(\langle \mathcal{Z}_i,\tilde{\mathcal{V}}\rangle^4)\le(4\tilde{b})^2\times 2=64\tilde{b}^2$ [55, Proposition 3.2]. Similar to [4, 40], we use a truncated version to the right-hand side of (39). Denote a truncation level $\tau:=K\delta$ with some constant $K>0$. For any tensor $\tilde{\mathcal{V}}\in\mathbb{R}^{n_1\times n_2\times n_3}$ with $\|\tilde{\mathcal{V}}\|_F=\delta\in(0,1]$, we define

$$\mathbb{I}[|\langle \mathcal{Z}_i,\tilde{\mathcal{V}}\rangle|\le 2\tau]:=\begin{cases}1, & \text{if } |\langle \mathcal{Z}_i,\tilde{\mathcal{V}}\rangle|\le 2\tau,\\0, & \text{otherwise,}\end{cases}\qquad \mathbb{I}[|\langle \mathcal{Z}_i,\tilde{\mathcal{V}}\rangle|> 2\tau]:=\begin{cases}1, & \text{if } |\langle \mathcal{Z}_i,\tilde{\mathcal{V}}\rangle|> 2\tau,\\0, & \text{otherwise.}\end{cases}$$

In addition, for any given truncation level $T>0$, define the functions

$$\mathbb{I}[|\langle \mathcal{Z}_i,\mathcal{X}^*\rangle|\le T]:=\begin{cases}1, & \text{if } |\langle \mathcal{Z}_i,\mathcal{X}^*\rangle|\le T,\\0, & \text{otherwise,}\end{cases}\qquad \mathbb{I}[|\langle \mathcal{Z}_i,\mathcal{X}^*\rangle|> T]:=\begin{cases}1, & \text{if } |\langle \mathcal{Z}_i,\mathcal{X}^*\rangle|> T,\\0, & \text{otherwise.}\end{cases}$$

Now we can represent $\psi''(\langle \mathcal{Z}_i,\mathcal{X}^*\rangle+t_i\langle \mathcal{Z}_i,\tilde{\mathcal{V}}\rangle)\langle \mathcal{Z}_i,\tilde{\mathcal{V}}\rangle^2$ in (39) via above functions as follows

$$\begin{aligned}&\langle \nabla f_{n,y}(\mathcal{X}^*+\tilde{\mathcal{V}})-\nabla f_{n,y}(\mathcal{X}^*),\tilde{\mathcal{V}}\rangle\\&=\frac{1}{n}\sum_{i=1}^{n}\psi''(\langle \mathcal{Z}_i,\mathcal{X}^*\rangle+t_i\langle \mathcal{Z}_i,\tilde{\mathcal{V}}\rangle)\langle \mathcal{Z}_i,\tilde{\mathcal{V}}\rangle^2\mathbb{I}[|\langle \mathcal{Z}_i,\tilde{\mathcal{V}}\rangle|\le 2\tau]\mathbb{I}\left[|\langle \mathcal{Z}_i,\mathcal{X}^*\rangle|\le T\right]\\&\quad+\frac{1}{n}\sum_{i=1}^{n}\psi''(\langle \mathcal{Z}_i,\mathcal{X}^*\rangle+t_i\langle \mathcal{Z}_i,\tilde{\mathcal{V}}\rangle)\langle \mathcal{Z}_i,\tilde{\mathcal{V}}\rangle^2\mathbb{I}[|\langle \mathcal{Z}_i,\tilde{\mathcal{V}}\rangle|\le 2\tau]\mathbb{I}\left[|\langle \mathcal{Z}_i,\mathcal{X}^*\rangle|> T\right]\\&\quad+\frac{1}{n}\sum_{i=1}^{n}\psi''(\langle \mathcal{Z}_i,\mathcal{X}^*\rangle+t_i\langle \mathcal{Z}_i,\tilde{\mathcal{V}}\rangle)\langle \mathcal{Z}_i,\tilde{\mathcal{V}}\rangle^2\mathbb{I}[|\langle \mathcal{Z}_i,\tilde{\mathcal{V}}\rangle|> 2\tau]\mathbb{I}\left[|\langle \mathcal{Z}_i,\mathcal{X}^*\rangle|\le T\right]\\&\quad+\frac{1}{n}\sum_{i=1}^{n}\psi''(\langle \mathcal{Z}_i,\mathcal{X}^*\rangle+t_i\langle \mathcal{Z}_i,\tilde{\mathcal{V}}\rangle)\langle \mathcal{Z}_i,\tilde{\mathcal{V}}\rangle^2\mathbb{I}[|\langle \mathcal{Z}_i,\tilde{\mathcal{V}}\rangle|> 2\tau]\mathbb{I}\left[|\langle \mathcal{Z}_i,\mathcal{X}^*\rangle|> T\right]\\&\ge\frac{1}{n}\sum_{i=1}^{n}\psi''(\langle \mathcal{Z}_i,\mathcal{X}^*\rangle+t_i\langle \mathcal{Z}_i,\tilde{\mathcal{V}}\rangle)\langle \mathcal{Z}_i,\tilde{\mathcal{V}}\rangle^2\mathbb{I}[|\langle \mathcal{Z}_i,\tilde{\mathcal{V}}\rangle|\le 2\tau]\mathbb{I}\left[|\langle \mathcal{Z}_i,\mathcal{X}^*\rangle|\le T\right],\end{aligned}\tag{41}$$

where the inequality holds by the fact that the second derivative ψ'' is always positive.

Note that $0 < \tau = K\delta \leq K$. Define $\ell_\psi(T) := \min_{|u|\leq(T+2K)} \psi''(u)$. An immediate consequence is that $\psi''(\langle \mathcal{Z}_i, \mathcal{X}^*\rangle + t_i\langle \mathcal{Z}_i, \tilde{\mathcal{V}}\rangle)\mathbb{I}[|\langle \mathcal{Z}_i, \tilde{\mathcal{V}}\rangle| \leq 2\tau]\mathbb{I}\left[|\langle \mathcal{Z}_i, \mathcal{X}^*\rangle| \leq T\right] \geq \ell_\psi(T)\mathbb{I}[|\langle \mathcal{Z}_i, \tilde{\mathcal{V}}\rangle| \leq 2\tau]\mathbb{I}\left[|\langle \mathcal{Z}_i, \mathcal{X}^*\rangle| \leq T\right]$. This together with (41) leads to

$$\langle \nabla f_{n,y}(\mathcal{X}^* + \tilde{\mathcal{V}}) - \nabla f_{n,y}(\mathcal{X}^*), \tilde{\mathcal{V}}\rangle \geq \ell_\psi(T)\frac{1}{n}\sum_{i=1}^n \langle \mathcal{Z}_i, \tilde{\mathcal{V}}\rangle^2 \mathbb{I}[|\langle \mathcal{Z}_i, \tilde{\mathcal{V}}\rangle| \leq 2\tau]\mathbb{I}\left[|\langle \mathcal{Z}_i, \mathcal{X}^*\rangle| \leq T\right]. \quad (42)$$

Next, we will demonstrate that, for $\varepsilon_0 \geq 0$,

$$\frac{1}{n}\sum_{i=1}^n \langle \mathcal{Z}_i, \tilde{\mathcal{V}}\rangle^2 \mathbb{I}[|\langle \mathcal{Z}_i, \tilde{\mathcal{V}}\rangle| \leq 2\tau]\mathbb{I}\left[|\langle \mathcal{Z}_i, \mathcal{X}^*\rangle| \leq T\right] \geq \frac{\tau_u}{2}\|\tilde{\mathcal{V}}\|_F^2 - \varepsilon_0 \quad (43)$$

with high probability, where τ_u is defined in (40). Note that $\tau = K\delta$. For notational convenience, we define

$$\varpi_{\tau(\delta)}(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}\rangle) := \langle \mathcal{Z}_i, \tilde{\mathcal{V}}\rangle^2 \mathbb{I}[|\langle \mathcal{Z}_i, \tilde{\mathcal{V}}\rangle| \leq 2\tau]. \quad (44)$$

In particular, if $\delta = 1$, then $\varpi_{\tau(1)}(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}\rangle) = \langle \mathcal{Z}_i, \tilde{\mathcal{V}}\rangle^2 \mathbb{I}[|\langle \mathcal{Z}_i, \tilde{\mathcal{V}}\rangle| \leq 2K]$.

By using the notation in (44), (43) can be rewritten as

$$\frac{1}{n}\sum_{i=1}^n \varpi_{\tau(\delta)}(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}\rangle)\mathbb{I}\left[|\langle \mathcal{Z}_i, \mathcal{X}^*\rangle| \leq T\right] \geq \frac{\tau_u}{2}\|\tilde{\mathcal{V}}\|_F^2 - \varepsilon_0, \quad \text{for any } \varepsilon_0 \geq 0. \quad (45)$$

Furthermore, we claim that the inequality in (45) can be reduced to the case for $\delta = 1$. Let $\tilde{\mathcal{V}}_1 = \frac{\tilde{\mathcal{V}}}{\|\tilde{\mathcal{V}}\|_F} = \frac{\tilde{\mathcal{V}}}{\delta}$, then $\|\tilde{\mathcal{V}}_1\|_F = 1$. Note that

$$\varpi_{\tau(\delta)}(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}\rangle) = \langle \mathcal{Z}_i, \tilde{\mathcal{V}}\rangle^2 \mathbb{I}[|\langle \mathcal{Z}_i, \tilde{\mathcal{V}}\rangle| \leq 2K\delta] = \delta^2\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle^2 \mathbb{I}[|\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle| \leq 2K] = \delta^2 \varpi_{\tau(1)}(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle). \quad (46)$$

Then (45) is equivalent to the following form

$$\frac{1}{n}\sum_{i=1}^n \varpi_{\tau(1)}(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle)\mathbb{I}\left[|\langle \mathcal{Z}_i, \mathcal{X}^*\rangle| \leq T\right] \geq \frac{\tau_u}{2} - \frac{\varepsilon_0}{\delta^2}, \quad \text{for any } \varepsilon_0 \geq 0. \quad (47)$$

In the following, we begin to prove that (47). Define $x_i := \varpi_{\tau(\delta)}(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}\rangle)\mathbb{I}\left[|\langle \mathcal{Z}_i, \mathcal{X}^*\rangle| \leq T\right]$. For any i, one can easily check the random variable $x_i \in [0, 4K^2]$. Note that the bounded random variables are sub-Gaussian [64, Example 2.4], which implies that x_i is sub-Gaussian. Furthermore, we can obtain that x_i is also sub-exponential [64, Definition 2.7]. For any $\varepsilon_0 \geq 0$,

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \varpi_{\tau(1)}(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle)\mathbb{I}\left[|\langle \mathcal{Z}_i, \mathcal{X}^*\rangle| \leq T\right] \geq \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(\varpi_{\tau(1)}(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle)\mathbb{I}\left[|\langle \mathcal{Z}_i, \mathcal{X}^*\rangle| \leq T\right]\right) - \frac{\varepsilon_0}{\delta^2}\right) \\
&= \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \varpi_{\tau(\delta)}(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}\rangle)\mathbb{I}\left[|\langle \mathcal{Z}_i, \mathcal{X}^*\rangle| \leq T\right] \geq \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(\varpi_{\tau(\delta)}(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}\rangle)\mathbb{I}\left[|\langle \mathcal{Z}_i, \mathcal{X}^*\rangle| \leq T\right]\right) - \varepsilon_0\right) \\
&= \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n x_i \geq \frac{1}{n}\sum_{i=1}^n \mathbb{E}(x_i) - \varepsilon_0\right) \\
&\geq 1 - \exp\left(-\frac{n\varepsilon_0^2}{\frac{2}{n}\sum_{i=1}^n \mathbb{E}(x_i^2)}\right),
\end{aligned} \quad (48)$$

where the first equality holds by (46), the second equality holds by the definition of x_i, and the inequality holds by the Bernstein's inequality [64, Propostion 2.14]. Suppose that we can prove that

$$\mathbb{E}\left(x_i\right) = \mathbb{E}\left(\varpi_{\tau(\delta)}(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}\rangle)\mathbb{I}\left[|\langle \mathcal{Z}_i, \mathcal{X}^*\rangle| \leq T\right]\right) \geq \frac{\tau_u}{2}\|\tilde{\mathcal{V}}\|_F^2, \tag{49}$$

which taken collectively with (44), (45), (48) yields the desired claim (43).

Note that (49) can be written equivalently in the following form

$$\mathbb{E}\left(\varpi_{\tau(1)}(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle)\mathbb{I}\left[|\langle \mathcal{Z}_i, \mathcal{X}^*\rangle| \leq T\right]\right) \geq \frac{\tau_u}{2}. \tag{50}$$

Observe that

$$\begin{aligned}
&\mathbb{E}\left(\varpi_{\tau(1)}(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle)\mathbb{I}\left[|\langle \mathcal{Z}_i, \mathcal{X}^*\rangle| \leq T\right]\right) \\
&= \mathbb{E}\left(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle^2\mathbb{I}[|\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle| \leq 2K]\mathbb{I}\left[|\langle \mathcal{Z}_i, \mathcal{X}^*\rangle| \leq T\right]\right) \\
&= \mathbb{E}\left(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle^2\mathbb{I}[|\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle| \leq 2K]\right) - \mathbb{E}\left(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle^2\mathbb{I}[|\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle| \leq 2K]\mathbb{I}\left[|\langle \mathcal{Z}_i, \mathcal{X}^*\rangle| > T\right]\right) \\
&= \mathbb{E}\left(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle^2\right) - \mathbb{E}\left(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle^2\mathbb{I}[|\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle| > 2K]\right) \\
&\quad - \mathbb{E}\left(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle^2\mathbb{I}[|\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle| \leq 2K]\mathbb{I}\left[|\langle \mathcal{Z}_i, \mathcal{X}^*\rangle| > T\right]\right) \\
&\geq \tau_u - \mathbb{E}\left(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle^2\mathbb{I}[|\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle| > 2K]\right) \\
&\quad - \mathbb{E}\left(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle^2\mathbb{I}[|\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle| \leq 2K]\mathbb{I}\left[|\langle \mathcal{Z}_i, \mathcal{X}^*\rangle| > T\right]\right),
\end{aligned} \tag{51}$$

where the first equality follows from (46) and the inequality holds by the condition $\mathbb{E}\left(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle^2\right) \geq \tau_u$. By Cauchy-Schwarz inequality, we get that

$$\begin{aligned}
\mathbb{E}\left(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle^2\mathbb{I}[|\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle| > 2K]\right) &\leq \sqrt{\mathbb{E}\left(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle^4\right)}\sqrt{\mathbb{P}\left(|\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle| > 2K\right)} \\
&\leq \sqrt{\tau_l}\sqrt{\mathbb{P}\left(|\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle| > 2K\right)},
\end{aligned} \tag{52}$$

where the last inequality follows from (40). On the other hand, we can deduce

$$\mathbb{P}\left(|\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle| > 2K\right) = \mathbb{P}\left(|\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle|^4 > 16K^4\right) \leq \frac{\mathbb{E}(\langle \mathcal{Z}, \tilde{\mathcal{V}}_1\rangle^4)}{16K^4} \leq \frac{\tau_l}{16K^4}, \tag{53}$$

where the first inequality follows from Markov inequality [18, Theorem 7.3] and the second inequality follows from (40). Set $K^2 \geq \frac{\tau_l}{\tau_u}$. Combining (53) with (52) yields

$$\mathbb{E}\left(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle^2\mathbb{I}[|\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle| > 2K]\right) \leq \frac{\tau_l}{4K^2} \leq \frac{\tau_u}{4}. \tag{54}$$

By a similar discussion, we obtain

$$\mathbb{P}\left(|\langle \mathcal{Z}_i, \mathcal{X}^*\rangle| > T\right) \leq \frac{\mathbb{E}\left(\langle \mathcal{Z}_i, \mathcal{X}^*\rangle^4\right)}{T^4} = \frac{\|\mathcal{X}^*\|_F^4\mathbb{E}\left(\langle \mathcal{Z}_i, \frac{\mathcal{X}^*}{\|\mathcal{X}^*\|_F}\rangle^4\right)}{T^4} \leq \frac{\|\mathcal{X}^*\|_F^4\tau_l}{T^4}. \tag{55}$$

By Cauchy-Schwarz inequality, one can easily get

$$\begin{aligned}
&\mathbb{E}\left(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle^2 \mathbb{I}[|\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle| \le 2K]\mathbb{I}\left[|\langle \mathcal{Z}_i, \mathcal{X}^*\rangle| > T\right]\right) \\
&\le \sqrt{\mathbb{E}\left(\left(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle\right)^2 \mathbb{I}[|\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle| \le 2K]\right)^2} \sqrt{\mathbb{P}\left(|\langle \mathcal{Z}_i, \mathcal{X}^*\rangle| > T\right)} \\
&\le \sqrt{\mathbb{E}\left(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle\right)^4} \sqrt{\mathbb{P}\left(|\langle \mathcal{Z}_i, \mathcal{X}^*\rangle| > T\right)} \le \frac{\tau_l \|\mathcal{X}^*\|_F^2}{T^2},
\end{aligned}$$

where the last inequality follows from (55) and (40). Taking $T^2 \ge \frac{4\|\mathcal{X}^*\|_F^2 \tau_l}{\tau_u}$ yields

$$\mathbb{E}\left(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle^2 \mathbb{I}[|\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle| \le 2K]\mathbb{I}\left[|\langle \mathcal{Z}_i, \mathcal{X}^*\rangle| > T\right]\right) \le \frac{\tau_u}{4}. \tag{56}$$

Combining (51), (54) and (56) then gives

$$\mathbb{E}\left(\varpi_{\tau(1)}(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}_1\rangle)\mathbb{I}\left[|\langle \mathcal{Z}_i, \mathcal{X}^*\rangle| \le T\right]\right) \ge \tau_u - \frac{\tau_u}{4} - \frac{\tau_u}{4} = \frac{\tau_u}{2}. \tag{57}$$

With the above analysis, we can establish the desired result (50). From (48), (49) and (50), for any $\|\tilde{\mathcal{V}}\|_F \in (0, 1]$, we get

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \varpi_{\tau(\delta)}(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}\rangle)\mathbb{I}\left[|\langle \mathcal{Z}_i, \mathcal{X}^*\rangle| \le T\right] \ge \frac{\tau_u}{2}\|\tilde{\mathcal{V}}\|_F^2 - \varepsilon_0\right) \ge 1 - \exp\left(-\frac{n\varepsilon_0^2}{\frac{2}{n}\sum_{i=1}^n \mathbb{E}(x_i^2)}\right). \tag{58}$$

Combining this with (44) yields the claim (43). Taking this collectively with (42), we arrive at

$$\begin{aligned}
&\mathbb{P}\left(\langle \nabla f_{n,y}(\mathcal{X}^* + \tilde{\mathcal{V}}) - \nabla f_{n,y}(\mathcal{X}^*), \tilde{\mathcal{V}}\rangle \ge \frac{\ell_\psi(T)\tau_u}{2}\|\tilde{\mathcal{V}}\|_F^2 - \ell_\psi(T)\varepsilon_0\right) \\
&\ge 1 - \exp\left(-\frac{n\varepsilon_0^2}{\frac{2}{n}\sum_{i=1}^n \mathbb{E}(x_i^2)}\right).
\end{aligned} \tag{59}$$

Note that

$$x_i^2 = \langle \mathcal{Z}_i, \tilde{\mathcal{V}}\rangle^4 \mathbb{I}[|\langle \mathcal{Z}_i, \tilde{\mathcal{V}}\rangle| \le 2\tau]\mathbb{I}\left[|\langle \mathcal{Z}_i, \mathcal{X}^*\rangle| \le T\right].$$

Denote $z_i := \langle \mathcal{Z}_i, \tilde{\mathcal{V}}\rangle = \langle \mathrm{vec}(\mathcal{Z}_i), \mathrm{vec}(\tilde{\mathcal{V}})\rangle, i = 1, 2, \ldots, n$. Note that $\mathcal{Z}_i$ are sub-Gaussian with i.i.d. zero-mean entries with same sub-Gaussian parameter $\tilde{b}$ in (36), which implies that the random variable $z_i = \langle \mathrm{vec}(\mathcal{Z}_i), \mathrm{vec}(\tilde{\mathcal{V}})\rangle$ is sub-Gaussian with parameter $\tilde{b}\|\tilde{\mathcal{V}}\|_F^2$ [18, Theorem 7.27]. Then we have

$$\begin{aligned}
\mathbb{E}(x_i^2) &= \mathbb{E}\left(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}\rangle^4 \mathbb{I}[|\langle \mathcal{Z}_i, \tilde{\mathcal{V}}\rangle| \le 2\tau]\mathbb{I}\left[|\langle \mathcal{Z}_i, \mathcal{X}^*\rangle| \le T\right]\right) \\
&\le \sqrt{\mathbb{E}\left(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}\rangle^4 \mathbb{I}[|\langle \mathcal{Z}_i, \tilde{\mathcal{V}}\rangle| \le 2\tau]\right)^2} \sqrt{\mathbb{P}(|\langle \mathcal{Z}_i, \mathcal{X}^*\rangle| \le T)} \\
&\le \sqrt{\mathbb{E}\left(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}\rangle^4 \mathbb{I}[|\langle \mathcal{Z}_i, \tilde{\mathcal{V}}\rangle| \le 2\tau]\right)^2} \\
&\le \sqrt{\mathbb{E}(\langle \mathcal{Z}_i, \tilde{\mathcal{V}}\rangle^8)} \le 256\tilde{b}^2\|\tilde{\mathcal{V}}\|_F^4,
\end{aligned}$$

where the first inequality follows from the Cauchy-Schwarz inequality and the last inequality follows from [55, Proposition 3.2]. Consequently,

$$-\frac{n\varepsilon_0^2}{\frac{2}{n}\sum_{i=1}^n \mathbb{E}(x_i^2)} \le -\frac{n\varepsilon_0^2}{\frac{2}{n}\sum_{i=1}^n 256\tilde{b}^2\|\tilde{\mathcal{V}}\|_F^4} = -\frac{n\varepsilon_0^2}{512\tilde{b}^2\|\tilde{\mathcal{V}}\|_F^4} = -\frac{n\varepsilon_0^2}{\kappa\|\tilde{\mathcal{V}}\|_F^4}, \tag{60}$$

where $\kappa := 512\tilde{b}^2$. Let

$$\varepsilon_0 = c_2\sqrt{\frac{\log d}{n}}\|\tilde{\mathcal{V}}\|_{\text{TTNN}}\|\tilde{\mathcal{V}}\|_F,$$

where $c_2 > 0$ is a given constant. By using the arithmetic mean-geometric mean inequality, we have

$$\varepsilon_0 = c_2\sqrt{\frac{\log d}{n}}\|\tilde{\mathcal{V}}\|_{\text{TTNN}}\|\tilde{\mathcal{V}}\|_F = \sqrt{\frac{2c_2^2}{\tau_u}\frac{\log d}{n}\|\tilde{\mathcal{V}}\|_{\text{TTNN}}^2 \cdot \frac{\tau_u}{2}\|\tilde{\mathcal{V}}\|_F^2} \le \frac{c_2^2}{\tau_u}\frac{\log d}{n}\|\tilde{\mathcal{V}}\|_{\text{TTNN}}^2 + \frac{\tau_u}{4}\|\tilde{\mathcal{V}}\|_F^2.$$

Consequently, we deduce

$$\begin{aligned}\frac{\ell_\psi(T)\tau_u}{2}\|\tilde{\mathcal{V}}\|_F^2 - \ell_\psi(T)\varepsilon_0 &\ge \frac{\ell_\psi(T)\tau_u}{2}\|\tilde{\mathcal{V}}\|_F^2 - \frac{c_2^2\ell_\psi(T)}{\tau_u}\frac{\log d}{n}\|\tilde{\mathcal{V}}\|_{\text{TTNN}}^2 - \frac{\ell_\psi(T)\tau_u}{4}\|\tilde{\mathcal{V}}\|_F^2 \\ &= \frac{\ell_\psi(T)\tau_u}{4}\|\tilde{\mathcal{V}}\|_F^2 - \frac{c_2^2\ell_\psi(T)}{\tau_u}\frac{\log d}{n}\|\tilde{\mathcal{V}}\|_{\text{TTNN}}^2.\end{aligned} \tag{61}$$

Combining (61), (60), and (59) yields

$$\begin{aligned}&\mathbb{P}\left(\langle\nabla f_{n,y}(\mathcal{X}^* + \tilde{\mathcal{V}}) - \nabla f_{n,y}(\mathcal{X}^*), \tilde{\mathcal{V}}\rangle \ge \frac{\ell_\psi(T)\tau_u}{4}\|\tilde{\mathcal{V}}\|_F^2 - \frac{c_2^2\ell_\psi(T)}{\tau_u}\frac{\log d}{n}\|\tilde{\mathcal{V}}\|_{\text{TTNN}}^2\right) \\ &\ge 1 - \exp\left(-\frac{n\varepsilon_0^2}{\kappa\|\tilde{\mathcal{V}}\|_F^4}\right) \\ &= 1 - \exp\left(-\frac{c_2^2\log d\|\tilde{\mathcal{V}}\|_{\text{TTNN}}^2}{\kappa\|\tilde{\mathcal{V}}\|_F^2}\right) \\ &\ge 1 - \exp\left(-\frac{c_2^2\log d}{\kappa}\right) = 1 - \exp\left(-c_3\log d\right),\end{aligned} \tag{62}$$

where the second inequality holds by $\|\tilde{\mathcal{V}}\|_{\text{TTNN}}^2 \ge \|\tilde{\mathcal{V}}\|_F^2$ and the last equality holds by letting $c_3 := \frac{c_2^2}{\kappa}$.

Case II. Suppose that $\|\tilde{\mathcal{V}}\|_F > 1$. Similar to [40, Lemma 8], we assume $\|\tilde{\mathcal{V}}\|_{\text{TTNN}} \le 2t$. Define $L(\gamma) : [0,1] \to \mathbb{R}$ as $L(\gamma) := f(\mathcal{X}^* + \gamma\tilde{\mathcal{V}})$. Notice that $L(\gamma)$ is convex by the convexity of f, which implies that L' is monotonously non-decreasing. Consequently, for any $\gamma \in [0,1]$, one has $L'(1) - L'(0) \ge L'(\gamma) - L'(0)$. Then we have

$$\langle\nabla f_{n,y}(\mathcal{X}^* + \tilde{\mathcal{V}}) - \nabla f_{n,y}(\mathcal{X}^*), \tilde{\mathcal{V}}\rangle \ge \frac{1}{\gamma}\langle\nabla f_{n,y}(\mathcal{X}^* + \gamma\tilde{\mathcal{V}}) - \nabla f_{n,y}(\mathcal{X}^*), \gamma\tilde{\mathcal{V}}\rangle.$$

Taking $\gamma = \frac{1}{\|\tilde{\mathcal{V}}\|_F} \in (0,1]$, we conclude that

$$\begin{aligned}\langle\nabla f_{n,y}(\mathcal{X}^* + \tilde{\mathcal{V}}) - \nabla f_{n,y}(\mathcal{X}^*), \tilde{\mathcal{V}}\rangle &\ge \|\tilde{\mathcal{V}}\|_F\left(\frac{\ell_\psi(T)\tau_u}{4} - \frac{c_2^2\ell_\psi(T)}{\tau_u}\frac{\log d}{n}\frac{\|\tilde{\mathcal{V}}\|_{\text{TTNN}}^2}{\|\tilde{\mathcal{V}}\|_F^2}\right) \\ &\ge \|\tilde{\mathcal{V}}\|_F\left(\frac{\ell_\psi(T)\tau_u}{4} - \frac{2tc_2^2\ell_\psi(T)}{\tau_u}\frac{\log d}{n}\frac{\|\tilde{\mathcal{V}}\|_{\text{TTNN}}}{\|\tilde{\mathcal{V}}\|_F^2}\right) \\ &\ge \|\tilde{\mathcal{V}}\|_F\left(\frac{\ell_\psi(T)\tau_u}{4} - \frac{\ell_\psi(T)}{\tau_u}\sqrt{\frac{\log d}{n}}\frac{\|\tilde{\mathcal{V}}\|_{\text{TTNN}}}{\|\tilde{\mathcal{V}}\|_F^2}\right) \\ &\ge \|\tilde{\mathcal{V}}\|_F\left(\frac{\ell_\psi(T)\tau_u}{4} - \frac{\ell_\psi(T)}{\tau_u}\sqrt{\frac{\log d}{n}}\frac{\|\tilde{\mathcal{V}}\|_{\text{TTNN}}}{\|\tilde{\mathcal{V}}\|_F}\right) \\ &= \frac{\ell_\psi(T)\tau_u}{4}\|\tilde{\mathcal{V}}\|_F - \frac{\ell_\psi(T)}{\tau_u}\sqrt{\frac{\log d}{n}}\|\tilde{\mathcal{V}}\|_{\text{TTNN}},\end{aligned} \tag{63}$$

where the second inequality holds since the assumption $\|\tilde{\mathcal{V}}\|_{\text{TTNN}} \leq 2t$, the third inequality follows from $n \geq 4t^2 c_2^4 \log d$, and the last inequality holds since $\|\tilde{\mathcal{V}}\|_F^2 \geq \|\tilde{\mathcal{V}}\|_F > 1$.

Therefore, by combining (62) with (63), and taking $\alpha_1 = \alpha_2 = \frac{\ell_\psi(T)\tau_u}{4}, \tau_1 = \frac{c_2^2 \ell_\psi(T)}{\tau_u}, \tau_2 = \frac{\ell_\psi(T)}{\tau_u}$, we get that the loss function (27) satisfies the RSC condition (6) with probability at least $1 - \exp(-c_3 \log d)$. □

Appendix C. Auxiliary Lemmas

Lemma 3 *Denote the function* $h(x) := \frac{1}{1+\exp(-x)}$, *where* $x \in \mathbb{R}$. *Then*

$$|h(x_1) - h(x_2)| \leq \frac{1}{4}|x_1 - x_2|, \quad \forall x_1, x_2 \in \mathbb{R}.$$

Proof. For simplicity, we assume $x_1 < x_2$. Then there exists $x_3 \in (x_1, x_2)$ such that

$$|h(x_1) - h(x_2)| = |h'(x_3)(x_1 - x_2)|. \tag{64}$$

Note that for any $x \in \mathbb{R}$, we have

$$h'(x) = \frac{\exp(x)}{(1+\exp(x))^2} = \frac{1}{2 + \exp(x) + \exp(-x)} \leq \frac{1}{4}, \tag{65}$$

where the inequality follows from the fact that $\exp(x) + \exp(-x) \geq 2\sqrt{\exp(x) \cdot \exp(-x)} = 2$. In addition, it is worth noting that $h'(x) > 0$. Taking (64) collectively with (65) shows that

$$|h(x_1) - h(x_2)| \leq \frac{1}{4}|x_1 - x_2|.$$

This completes the proof. □

Note that the gradient of $f_{n,y}(\mathcal{X})$ in (27) is given by

$$\nabla f_{n,y}(\mathcal{X}) = \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{\exp(\langle \mathcal{Z}_i, \mathcal{X}\rangle)}{1+\exp\left(\langle \mathcal{Z}_i, \mathcal{X}\rangle\right)} - y_i\right)\mathcal{Z}_i\right]. \tag{66}$$

Now we show that $\nabla f_{n,y}$ is Lipschitz continuous in the following lemma.

Lemma 4 *The gradient* $\nabla f_{n,y}$ *in (66) is Lipschitz continuous with Lipschitz constant* $\frac{1}{4n}\sum_{i=1}^{n}\|\mathcal{Z}_i\|_F^2$, *i.e.,*

$$\|\nabla f_{n,y}(\mathcal{X}_1) - \nabla f_{n,y}(\mathcal{X}_2)\|_F \leq \frac{1}{4n}\sum_{i=1}^{n}\|\mathcal{Z}_i\|_F^2\|\mathcal{X}_1 - \mathcal{X}_2\|_F, \ \forall\, \mathcal{X}_1, \mathcal{X}_2 \in \mathbb{R}^{n_1 \times n_2 \times n_3}.$$

Proof. For any tensor $\mathcal{X}_1, \mathcal{X}_2 \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, we obtain

$$\nabla f_{n,y}(\mathcal{X}_1) - \nabla f_{n,y}(\mathcal{X}_2) = \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{\exp(\langle \mathcal{Z}_i, \mathcal{X}_1\rangle)}{1+\exp(\langle \mathcal{Z}_i, \mathcal{X}_1\rangle)} - \frac{\exp(\langle \mathcal{Z}_i, \mathcal{X}_2\rangle)}{1+\exp\left(\langle \mathcal{Z}_i, \mathcal{X}_2\rangle\right)}\right)\mathcal{Z}_i\right]. \tag{67}$$

Let $h(\langle \mathcal{Z}_i, \mathcal{X}\rangle) = \frac{1}{1+\exp(-\langle \mathcal{Z}_i, \mathcal{X}\rangle)}$. Then (67) can be equivalently expressed as

$$\nabla f_{n,y}(\mathcal{X}_1) - \nabla f_{n,y}(\mathcal{X}_2) = \frac{1}{n}\sum_{i=1}^{n}\left[\left(h(\langle \mathcal{Z}_i, \mathcal{X}_1\rangle) - h(\langle \mathcal{Z}_i, \mathcal{X}_2\rangle)\right)\mathcal{Z}_i\right].$$

Thus, we get

$$\begin{aligned}\|\nabla f_{n,y}(\mathcal{X}_1)-\nabla f_{n,y}(\mathcal{X}_2)\|_F &= \frac{1}{n}\left\|\sum_{i=1}^{n}\left[\left(h(\langle\mathcal{Z}_i,\mathcal{X}_1\rangle)-h(\langle\mathcal{Z}_i,\mathcal{X}_2\rangle)\right)\mathcal{Z}_i\right]\right\|_F \\ &\le \frac{1}{n}\sum_{i=1}^{n}\left\|\left(h(\langle\mathcal{Z}_i,\mathcal{X}_1\rangle)-h(\langle\mathcal{Z}_i,\mathcal{X}_2\rangle)\right)\mathcal{Z}_i\right\|_F \\ &= \frac{1}{n}\sum_{i=1}^{n}|h(\langle\mathcal{Z}_i,\mathcal{X}_1\rangle)-h(\langle\mathcal{Z}_i,\mathcal{X}_2\rangle)|\|\mathcal{Z}_i\|_F,\end{aligned} \tag{68}$$

where the inequality holds by the Minkowski's inequality. Using Lemma 3 further leads to

$$\begin{aligned}|h(\langle\mathcal{Z}_i,\mathcal{X}_1\rangle)-h(\langle\mathcal{Z}_i,\mathcal{X}_2\rangle)|\|\mathcal{Z}_i\|_F &\le \frac{1}{4}|\langle\mathcal{Z}_i,\mathcal{X}_1\rangle-\langle\mathcal{Z}_i,\mathcal{X}_2\rangle|\|\mathcal{Z}_i\|_F \\ &= \frac{1}{4}|\langle\mathcal{Z}_i,\mathcal{X}_1-\mathcal{X}_2\rangle|\|\mathcal{Z}_i\|_F \\ &\le \frac{1}{4}\|\mathcal{Z}_i\|_F\|\mathcal{X}_1-\mathcal{X}_2\|_F\|\mathcal{Z}_i\|_F \\ &= \frac{1}{4}\|\mathcal{Z}_i\|_F^2\|\mathcal{X}_1-\mathcal{X}_2\|_F,\end{aligned} \tag{69}$$

where the second inequality follows from the Cauchy-Schwarz inequality. Plugging (69) into (68) immediately yields

$$\|\nabla f_{n,y}(\mathcal{X}_1)-\nabla f_{n,y}(\mathcal{X}_2)\|_F \le \frac{1}{4n}\sum_{i=1}^{n}\|\mathcal{Z}_i\|_F^2\|\mathcal{X}_1-\mathcal{X}_2\|_F.$$

This concludes the proof. □

Lemma 5 *For any matrix $\mathbf{X}\in\mathbb{R}^{n_1\times n_2}$, define $h(\mathbf{X}):=\sum_{j=1}^{\min\{n_1,n_2\}}g_\lambda(\sigma_j(\mathbf{X}))$, where $g_\lambda(\cdot)$ satisfies Assumption 1. Then*

$$\|\mathbf{X}\|_* \le \frac{h(\mathbf{X})}{\lambda k_0}+\frac{\mu}{2\lambda k_0}\|\mathbf{X}\|_F^2.$$

Proof. Firstly, we prove

$$\lambda k_0 x \le g_\lambda(x)+\frac{\mu}{2}x^2, \text{ for any } x\ge 0, \tag{70}$$

where k_0,μ are the parameters in Assumption 1. The above inequality is trivial for $x=0$. For $x>0$, it follows from Assumption 1 that there exists $\mu>0$ such that $g_\lambda(x)+\frac{\mu}{2}x^2$ is convex, which yields

$$g_\lambda(x)+\frac{\mu}{2}x^2 \ge g_\lambda(x')+\frac{\mu}{2}(x')^2+\langle g_\lambda'(x')+\mu x', x-x'\rangle, \text{ for any } x,x'>0. \tag{71}$$

Here $g_\lambda'(\cdot)$ represents the derivative of $g_\lambda(\cdot)$. By taking $x'\to 0^+$ in (71), under Assumption 1, we can get that (70) holds.

Secondly, one has

$$\begin{aligned}\lambda k_0\|\mathbf{X}\|_* = \sum_{j=1}^{\min\{n_1,n_2\}}\lambda k_0\sigma_j(\mathbf{X}) &\le \sum_{j=1}^{\min\{n_1,n_2\}}\left(g_\lambda(\sigma_j(\mathbf{X}))+\frac{\mu}{2}(\sigma_j(\mathbf{X}))^2\right) \\ &= h(\mathbf{X})+\frac{\mu}{2}\|\mathbf{X}\|_F^2,\end{aligned} \tag{72}$$

where the last equation follows from the fact that $\|\mathbf{X}\|_F^2=\sum_{j=1}^{\min\{n_1,n_2\}}(\sigma_j(\mathbf{X}))^2$. This completes the proof. □

Lemma 6 *For any $\mathcal{X} \in \mathbb{R}^{n_1\times n_2\times n_3}$, define $\gamma(\mathcal{X}) := G_\lambda(\mathcal{X}) + \frac{\mu}{2}\|\mathcal{X}\|_F^2$, where $G_\lambda(\mathcal{X})$ is defined in (4) and the parameter μ is the same in Assumption 1(iv). Then $\gamma(\mathcal{X})$ is convex.*

Proof. Let $\rho(x) := g_\lambda(x) + \frac{\mu}{2}x^2$ and $m := \min\{n_1, n_2\}$. Under Assumption 1(iv), for any $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{n_1\times n_2\times n_3}$, $i = 1, \ldots, n_3, j = 1, \ldots, m$, the convexity of $\rho(x)$ leads to

$$\begin{aligned}
&g_\lambda\left(\theta\sigma_j(\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i\rangle}) + (1-\theta)\sigma_j(\widehat{\mathcal{Y}}_{\mathbf{U}}^{\langle i\rangle})\right) + \frac{\mu}{2}\left(\theta\sigma_j(\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i\rangle}) + (1-\theta)\sigma_j(\widehat{\mathcal{Y}}_{\mathbf{U}}^{\langle i\rangle})\right)^2 \\
&\le \theta g_\lambda\left(\sigma_j(\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i\rangle})\right) + \frac{\theta\mu}{2}\left(\sigma_j(\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i\rangle})\right)^2 + (1-\theta)g_\lambda\left(\sigma_j(\widehat{\mathcal{Y}}_{\mathbf{U}}^{\langle i\rangle})\right) + \frac{(1-\theta)\mu}{2}\left(\sigma_j(\widehat{\mathcal{Y}}_{\mathbf{U}}^{\langle i\rangle})\right)^2, \forall\theta\in[0,1].
\end{aligned}$$

Summing up the above inequalities from $i = 1, \ldots, n_3, j = 1, \ldots, m$ yields

$$\begin{aligned}
&\sum_{i=1}^{n_3}\sum_{j=1}^{m} g_\lambda\left(\theta\sigma_j(\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i\rangle}) + (1-\theta)\sigma_j(\widehat{\mathcal{Y}}_{\mathbf{U}}^{\langle i\rangle})\right) + \frac{\mu}{2}\left(\theta\sigma_j(\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i\rangle}) + (1-\theta)\sigma_j(\widehat{\mathcal{Y}}_{\mathbf{U}}^{\langle i\rangle})\right)^2 \\
&\le \sum_{i=1}^{n_3}\sum_{j=1}^{m} \theta g_\lambda\left(\sigma_j(\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i\rangle})\right) + \frac{\theta\mu}{2}\left(\sigma_j(\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i\rangle})\right)^2 + (1-\theta)g_\lambda\left(\sigma_j(\widehat{\mathcal{Y}}_{\mathbf{U}}^{\langle i\rangle})\right) + \frac{(1-\theta)\mu}{2}\left(\sigma_j(\widehat{\mathcal{Y}}_{\mathbf{U}}^{\langle i\rangle})\right)^2.
\end{aligned} \tag{73}$$

Moreover, by [27, Corollary 3.4.3], we have

$$\sum_{j=1}^{m}\sigma_j\left(\theta\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i\rangle} + (1-\theta)\widehat{\mathcal{Y}}_{\mathbf{U}}^{\langle i\rangle}\right) \le \sum_{j=1}^{m}\sigma_j\left(\theta\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i\rangle}\right) + \sigma_j\left((1-\theta)\widehat{\mathcal{Y}}_{\mathbf{U}}^{\langle i\rangle}\right),\ i = 1, 2, \ldots, n_3.$$

Note that $\rho(x) = g_\lambda(x) + \frac{\mu}{2}x^2$ is a real-valued convex function on $[0, +\infty)$. Then by [27, Lemma 3.3.8], we obtain

$$\sum_{i=1}^{n_3}\sum_{j=1}^{m}\rho\left(\sigma_j\left(\theta\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i\rangle} + (1-\theta)\widehat{\mathcal{Y}}_{\mathbf{U}}^{\langle i\rangle}\right)\right) \le \sum_{i=1}^{n_3}\sum_{j=1}^{m}\rho\left(\sigma_j\left(\theta\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i\rangle}\right) + \sigma_j\left((1-\theta)\widehat{\mathcal{Y}}_{\mathbf{U}}^{\langle i\rangle}\right)\right),$$

which yields

$$\begin{aligned}
&\sum_{i=1}^{n_3}\sum_{j=1}^{m} g_\lambda\left(\sigma_j\left(\theta\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i\rangle} + (1-\theta)\widehat{\mathcal{Y}}_{\mathbf{U}}^{\langle i\rangle}\right)\right) + \frac{\mu}{2}\left(\sigma_j\left(\theta\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i\rangle} + (1-\theta)\widehat{\mathcal{Y}}_{\mathbf{U}}^{\langle i\rangle}\right)\right)^2 \\
&\le \sum_{i=1}^{n_3}\sum_{j=1}^{m} g_\lambda\left(\sigma_j\left(\theta\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i\rangle}\right) + \sigma_j\left((1-\theta)\widehat{\mathcal{Y}}_{\mathbf{U}}^{\langle i\rangle}\right)\right) + \frac{\mu}{2}\left(\sigma_j\left(\theta\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i\rangle}\right) + \sigma_j\left((1-\theta)\widehat{\mathcal{Y}}_{\mathbf{U}}^{\langle i\rangle}\right)\right)^2 \\
&= \sum_{i=1}^{n_3}\sum_{j=1}^{m} g_\lambda\left(\theta\sigma_j(\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i\rangle}) + (1-\theta)\sigma_j(\widehat{\mathcal{Y}}_{\mathbf{U}}^{\langle i\rangle})\right) + \frac{\mu}{2}\left(\theta\sigma_j(\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i\rangle}) + (1-\theta)\sigma_j(\widehat{\mathcal{Y}}_{\mathbf{U}}^{\langle i\rangle})\right)^2.
\end{aligned} \tag{74}$$

Plugging (74) into (73) yields

$$\begin{aligned}
&\sum_{i=1}^{n_3}\sum_{j=1}^{m} g_\lambda\left(\sigma_j\left(\theta\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i\rangle} + (1-\theta)\widehat{\mathcal{Y}}_{\mathbf{U}}^{\langle i\rangle}\right)\right) + \frac{\mu}{2}\left(\sigma_j\left(\theta\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i\rangle} + (1-\theta)\widehat{\mathcal{Y}}_{\mathbf{U}}^{\langle i\rangle}\right)\right)^2 \\
&\le \sum_{i=1}^{n_3}\sum_{j=1}^{m} \theta g_\lambda\left(\sigma_j(\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i\rangle})\right) + \frac{\theta\mu}{2}\left(\sigma_j(\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i\rangle})\right)^2 + (1-\theta)g_\lambda\left(\sigma_j(\widehat{\mathcal{Y}}_{\mathbf{U}}^{\langle i\rangle})\right) + \frac{(1-\theta)\mu}{2}\left(\sigma_j(\widehat{\mathcal{Y}}_{\mathbf{U}}^{\langle i\rangle})\right)^2.
\end{aligned} \tag{75}$$

By the definition of G_λ and the fact that $\|\mathcal{X}\|_F^2 = \sum_{i=1}^{n_3}\sum_{j=1}^{m}(\sigma_j(\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i\rangle}))^2$, (75) implies that

$$\begin{aligned}&G_\lambda(\theta\mathcal{X}+(1-\theta)\mathcal{Y})+\frac{\mu}{2}\|\theta\mathcal{X}+(1-\theta)\mathcal{Y}\|_F^2\\ \le\ &\theta G_\lambda(\mathcal{X})+\frac{\theta\mu}{2}\|\mathcal{X}\|_F^2+(1-\theta)G_\lambda(\mathcal{Y})+\frac{(1-\theta)\mu}{2}\|\mathcal{Y}\|_F^2,\end{aligned}$$

which is equivalent to

$$\gamma(\theta\mathcal{X}+(1-\theta)\mathcal{Y})\le\theta\gamma(\mathcal{X})+(1-\theta)\gamma(\mathcal{Y}),\quad\forall\theta\in[0,1].$$

Therefore, $\gamma(\mathcal{X})$ is convex. □

Lemma 7 *For any tensors $\mathcal{X}_1,\mathcal{X}_2\in\mathbb{R}^{n_1\times n_2\times n_3}$, the following inequality holds:*

$$\langle\tilde{\mathcal{Z}},\mathcal{X}_1-\mathcal{X}_2\rangle\le\beta G_\lambda(\mathcal{X}_1)-\beta G_\lambda(\mathcal{X}_2)+\frac{\beta\mu}{2}\|\mathcal{X}_1-\mathcal{X}_2\|_F^2,\ \ \forall\tilde{\mathcal{Z}}\in\partial(\beta G_\lambda(\mathcal{X}_2)),$$

where β is defined in (4) and μ is defined in Assumption 1(iv).

Proof. As a consequence of Lemma 6, the function $\beta G_\lambda(\mathcal{X})+\frac{\beta\mu}{2}\|\mathcal{X}\|_F^2$ is convex with $\beta>0$, which yields

$$\beta G_\lambda(\mathcal{X}_1)+\frac{\beta\mu}{2}\|\mathcal{X}_1\|_F^2\ge\beta G_\lambda(\mathcal{X}_2)+\frac{\beta\mu}{2}\|\mathcal{X}_2\|_F^2+\langle\tilde{\mathcal{Z}}+\beta\mu\mathcal{X}_2,\mathcal{X}_1-\mathcal{X}_2\rangle,$$

for any $\tilde{\mathcal{Z}}\in\partial(\beta G_\lambda(\mathcal{X}_2))$. The above inequality can be rewritten as

$$\begin{aligned}\langle\tilde{\mathcal{Z}},\mathcal{X}_1-\mathcal{X}_2\rangle&\le\beta G_\lambda(\mathcal{X}_1)-\beta G_\lambda(\mathcal{X}_2)+\frac{\beta\mu}{2}\|\mathcal{X}_1\|_F^2-\frac{\beta\mu}{2}\|\mathcal{X}_2\|_F^2-\beta\mu\langle\mathcal{X}_2,\mathcal{X}_1\rangle+\beta\mu\|\mathcal{X}_2\|_F^2\\&=\beta G_\lambda(\mathcal{X}_1)-\beta G_\lambda(\mathcal{X}_2)+\frac{\beta\mu}{2}\|\mathcal{X}_1-\mathcal{X}_2\|_F^2.\end{aligned}$$

The proof is completed. □

For any matrix $\mathbf{A}\in\mathbb{R}^{n_1\times n_2}$, we define $\Phi_r(\mathbf{A})$ to be the best rank r approximation of $\mathbf{A}$. Denote

$$\Psi_r(\mathbf{A})=\mathbf{A}-\Phi_r(\mathbf{A})\tag{76}$$

and

$$\phi(\mathbf{A})=\sum_{j=1}^{m}g_\lambda(\sigma_j(\mathbf{A})),\tag{77}$$

where $g_\lambda(\cdot)$ satisfies Assumption 1.

Lemma 8 *For any $\mathbf{X},\mathbf{Y}\in\mathbb{R}^{n_1\times n_2}$ with $\mathrm{rank}(\mathbf{X})=r$, one has*

$$\phi(\Psi_r(\mathbf{Y}))\ge\phi(\Psi_r(\mathbf{X}-\mathbf{Y})),\tag{78}$$

where $\Psi_r(\cdot)$ and $\phi(\cdot)$ are defined as in (76) and (77).

Proof. For any $\mathbf{X},\mathbf{Y}\in\mathbb{R}^{n_1\times n_2}$, it follows from [27, Theorem 3.3.16] that

$$\sigma_j(\mathbf{Y})\le\sigma_1(\mathbf{X})+\sigma_j(\mathbf{Y}-\mathbf{X}),\quad j=r+1,\dots,m,$$

where $m = \min\{n_1, n_2\}$. Summing up the above inequality over $j = r+1, \ldots, m$, we obtain

$$\sum_{j=r+1}^{m} \sigma_j(\mathbf{Y}) \leq \sum_{j=r+1}^{m} (\sigma_j(\mathbf{Y}-\mathbf{X}) + \sigma_1(\mathbf{X})).$$

Note that $-g_\lambda$ is convex and non-increasing on $[0, +\infty)$ by Assumption 1(i), which together with [27, Lemma 3.3.8] leads to

$$\begin{aligned} \sum_{j=r+1}^{m} -g_\lambda(\sigma_j(\mathbf{Y})) &\leq \sum_{j=r+1}^{m} -g_\lambda(\sigma_j(\mathbf{Y}-\mathbf{X}) + \sigma_1(\mathbf{X})) \\ &\leq \sum_{j=r+1}^{m} -g_\lambda(\sigma_j(\mathbf{Y}-\mathbf{X})) = \sum_{j=r+1}^{m} -g_\lambda(\sigma_j(\mathbf{X}-\mathbf{Y})), \end{aligned} \tag{79}$$

where the second inequality holds since $\sigma_j(\mathbf{Y}-\mathbf{X}) + \sigma_1(\mathbf{X}) \geq \sigma_j(\mathbf{Y}-\mathbf{X}) \geq 0$ for any $j = r+1, \ldots, m$ and the fact $-g_\lambda$ is non-increasing. By definitions of $\Psi_r(\cdot)$ and $\phi(\cdot)$ in (76) and (77), we know that (79) is equivalent to

$$\phi(\Psi_r(\mathbf{Y})) \geq \phi(\Psi_r(\mathbf{X}-\mathbf{Y})).$$

This completes the proof. □

Lemma 9 *For any $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n_1 \times n_2}$ with $\mathrm{rank}(\mathbf{X}) = r$, the following inequality holds*

$$\phi(\mathbf{X}) - \phi(\mathbf{Y}) \leq \phi(\Phi_r(\mathbf{E})) - \phi(\Psi_r(\mathbf{E})),$$

where $\mathbf{E} := \mathbf{X} - \mathbf{Y}$, $\Phi_r(\cdot)$, $\Psi_r(\cdot)$, and $\phi(\cdot)$ are defined as in (76) and (77).

Proof. By the definitions of ϕ, Φ_r and Ψ_r in (76) and (77), we can obtain that

$$\begin{aligned} \phi(\mathbf{X}) - \phi(\mathbf{Y}) &= \phi(\Phi_r(\mathbf{X})) + \phi(\Psi_r(\mathbf{X})) - \phi(\Phi_r(\mathbf{Y})) - \phi(\Psi_r(\mathbf{Y})) \\ &= \phi(\Phi_r(\mathbf{X})) - \phi(\Phi_r(\mathbf{Y})) - \phi(\Psi_r(\mathbf{Y})) \\ &= \sum_{j=1}^{r} g_\lambda(\sigma_j(\mathbf{X})) - \sum_{j=1}^{r} g_\lambda(\sigma_j(\mathbf{Y})) - \phi(\Psi_r(\mathbf{Y})) \\ &= \sum_{j=1}^{r} g_\lambda(\sigma_j(\mathbf{X}-\mathbf{Y}+\mathbf{Y})) - \sum_{j=1}^{r} g_\lambda(\sigma_j(\mathbf{Y})) - \phi(\Psi_r(\mathbf{Y})) \\ &\leq \sum_{j=1}^{r} g_\lambda(\sigma_j(\mathbf{X}-\mathbf{Y})) - \phi(\Psi_r(\mathbf{Y})) \\ &= \phi(\Phi_r(\mathbf{E})) - \phi(\Psi_r(\mathbf{Y})) \\ &\leq \phi(\Phi_r(\mathbf{E})) - \phi(\Psi_r(\mathbf{E})), \end{aligned} \tag{80}$$

where the first inequality follows from [43, Theorem 2.6] and the second inequality follows from Lemma 8. □

Appendix D. Subdifferential of $G_\lambda(\mathcal{X})$ in (5)

Lemma 10 *For any matrix $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$ and any tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, define $(g_\lambda \circ \sigma)(\mathbf{X}) := \sum_{j=1}^{\min\{n_1,n_2\}} g_\lambda(\sigma_j(\mathbf{X}))$ and $H_\lambda(\widehat{\mathcal{X}}_{\mathbf{U}}) := \sum_{i=1}^{n_3} \sum_{j=1}^{\min\{n_1,n_2\}} g_\lambda(\sigma_j(\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i \rangle}))$. Then*

$$\overline{\mathcal{Z}} \in \partial(g_\lambda \circ \sigma)(\overline{\mathcal{X}}) \iff \widehat{\mathcal{Z}}_{\mathbf{U}} \in \partial H_\lambda(\widehat{\mathcal{X}}_{\mathbf{U}}).$$

Proof. By the definition of the subdifferential of a function, we know that for any $\overline{\mathcal{Z}} \in \partial(g_\lambda \circ \sigma)(\overline{\mathcal{X}})$, there exist a sequence $\{\overline{\mathcal{X}}^k\}$ such that $\overline{\mathcal{X}}^k$ converges to $\overline{\mathcal{X}}$ with $(g_\lambda \circ \sigma)(\overline{\mathcal{X}}^k) \to (g_\lambda \circ \sigma)(\overline{\mathcal{X}})$ and a sequence of regular subdifferential $\overline{\mathcal{Z}}^k \in \widehat{\partial}(g_\lambda \circ \sigma)(\overline{\mathcal{X}}^k)$ such that $\overline{\mathcal{Z}}^k \to \overline{\mathcal{Z}}$. Notice that $(g_\lambda \circ \sigma)(\overline{\mathcal{X}}^k) = H_\lambda(\widehat{\mathcal{X}}^k_{\mathbf{U}})$. For any $\widehat{\mathcal{Y}}^k_{\mathbf{U}} \in \mathbb{C}^{n_1 \times n_2 \times n_3}$, where $\|\widehat{\mathcal{Y}}^k_{\mathbf{U}}\|_F \to 0$ as $k \to +\infty$, the following inequality holds

$$\begin{aligned} H_\lambda(\widehat{\mathcal{X}}^k_{\mathbf{U}} + \widehat{\mathcal{Y}}^k_{\mathbf{U}}) &= (g_\lambda \circ \sigma)(\overline{\mathcal{X}}^k + \overline{\mathcal{Y}}^k) \\ &\geq (g_\lambda \circ \sigma)(\overline{\mathcal{X}}^k) + \langle \overline{\mathcal{Z}}^k, \overline{\mathcal{Y}}^k \rangle + o(\|\overline{\mathcal{Y}}^k\|_F) \\ &= H_\lambda(\widehat{\mathcal{X}}^k_{\mathbf{U}}) + \langle \widehat{\mathcal{Z}}^k_{\mathbf{U}}, \widehat{\mathcal{Y}}^k_{\mathbf{U}} \rangle + o(\|\widehat{\mathcal{Y}}^k_{\mathbf{U}}\|_F), \end{aligned}$$

where the first inequality follows from the definition of the regular subdifferential of $g_\lambda \circ \sigma$ at $\overline{\mathcal{X}}^k$ and the second equation holds since $\langle \overline{\mathcal{Z}}^k, \overline{\mathcal{Y}}^k \rangle = \langle \widehat{\mathcal{Z}}^k_{\mathbf{U}}, \widehat{\mathcal{Y}}^k_{\mathbf{U}} \rangle$ and $\|\overline{\mathcal{Y}}^k\|_F = \|\widehat{\mathcal{Y}}^k_{\mathbf{U}}\|_F$ [59, Definition 5]. Here $o(\cdot)$ denotes a higher-order infinitesimal. As a consequence, we can deduce that $\widehat{\mathcal{Z}}^k_{\mathbf{U}} \in \widehat{\partial} H_\lambda(\widehat{\mathcal{X}}^k_{\mathbf{U}})$ for each k.

Since $\overline{\mathcal{Z}}^k \to \overline{\mathcal{Z}}$, we get that $\widehat{\mathcal{Z}}^k_{\mathbf{U}} \to \widehat{\mathcal{Z}}_{\mathbf{U}}$. It can be easily shown that the sequence $\{\widehat{\mathcal{X}}^k_{\mathbf{U}}\}$ converges to $\widehat{\mathcal{X}}_{\mathbf{U}}$ and $\{H_\lambda(\widehat{\mathcal{X}}^k_{\mathbf{U}})\}$ converges to $H_\lambda(\widehat{\mathcal{X}}_{\mathbf{U}})$. Since $g_\lambda(\cdot)$ is continuous, we know that $\widehat{\partial} H_\lambda(\widehat{\mathcal{X}}^k_{\mathbf{U}})$ and $\partial H_\lambda(\widehat{\mathcal{X}}^k_{\mathbf{U}})$ are closed with $\widehat{\partial} H_\lambda(\widehat{\mathcal{X}}^k_{\mathbf{U}}) \subset \partial H_\lambda(\widehat{\mathcal{X}}^k_{\mathbf{U}})$ [56, Theorem 8.6], which yields $\widehat{\mathcal{Z}}_{\mathbf{U}} \in \widehat{\partial} H_\lambda(\widehat{\mathcal{X}}_{\mathbf{U}}) \subset \partial H_\lambda(\widehat{\mathcal{X}}_{\mathbf{U}})$.

By using similar arguments to the proof of the opposite inclusion, we can easily obtain that for any $\widehat{\mathcal{Z}}_{\mathbf{U}} \in \partial H_\lambda(\widehat{\mathcal{X}}_{\mathbf{U}})$, $\overline{\mathcal{Z}} \in \partial(g_\lambda \circ \sigma)(\overline{\mathcal{X}})$ holds. □

Lemma 11 *For any tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, one has*

$$\partial H_\lambda(\widehat{\mathcal{X}}_{\mathbf{U}}) = \mathrm{fold}_3(\partial(g_\lambda \circ \sigma)(\overline{\mathcal{X}})),$$

where H_λ and $g_\lambda \circ \sigma$ are defined as the same in Lemma 10, and $\mathrm{fold}_3(\cdot)$ operating on a set means to operate each element of the set.

Proof. For any $\widehat{\mathcal{Y}}_{\mathbf{U}} \in \partial H_\lambda(\widehat{\mathcal{X}}_{\mathbf{U}})$, it follows from Lemma 10 that $\mathrm{bdiag}(\widehat{\mathcal{Y}}_{\mathbf{U}}) = \overline{\mathcal{Y}} \in \partial(g_\lambda \circ \sigma)(\overline{\mathcal{X}})$. By the definition of $\mathrm{fold}_3(\cdot)$ in (3), we get that $\widehat{\mathcal{Y}}_{\mathbf{U}} = \mathrm{fold}_3(\overline{\mathcal{Y}}) \in \mathrm{fold}_3(\partial(g_\lambda \circ \sigma)(\overline{\mathcal{X}}))$, which implies that $\partial H_\lambda(\widehat{\mathcal{X}}_{\mathbf{U}}) \subseteq \mathrm{fold}_3(\partial(g_\lambda \circ \sigma)(\overline{\mathcal{X}}))$.

A similar argument leads to the opposite direction, i.e., $\mathrm{fold}_3(\partial(g_\lambda \circ \sigma)(\overline{\mathcal{X}})) \subseteq \partial H_\lambda(\widehat{\mathcal{X}}_{\mathbf{U}})$. This completes the proof. □

Lemma 12 *For any $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, the transformed tensor SVD of $\mathcal{X}$ is denoted by $\mathcal{X} = \mathcal{U} \diamond_{\mathbf{U}} \Sigma \diamond_{\mathbf{U}} \mathcal{V}^T$. Let $G_\lambda(\mathcal{X})$ be defined in (4). Then the subdifferential of $G_\lambda(\mathcal{X})$ at $\mathcal{X}$ is given by*

$$\partial G_\lambda(\mathcal{X}) = \left\{ \mathcal{U} \diamond_{\mathbf{U}} \mathcal{D} \diamond_{\mathbf{U}} \mathcal{V}^T \mid \widehat{\mathcal{D}}^{\langle i \rangle}_{\mathbf{U}} \in \mathfrak{B}_i, i = 1, 2, \ldots, n_3 \right\},$$

where $\mathfrak{B}_i$ is defined as

$$\mathfrak{B}_i = \left\{ \mathrm{Diag}(\mathbf{d}^i) \mid \mathbf{d}^i = (d^i_1, d^i_2, \ldots, d^i_m)^T \in \mathbb{R}^m, d^i_j \in \partial g_\lambda(\sigma_j(\widehat{\mathcal{X}}^{\langle i \rangle}_{\mathbf{U}})), j = 1, 2, \ldots, m \right\} \tag{81}$$

and $m := \min\{n_1, n_2\}$.

Proof. Note that $(g_\lambda \circ \sigma)(\mathbf{X}) := \sum_{j=1}^m g_\lambda(\sigma_j(\mathbf{X}))$ for any matrix $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$. For any $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, by [34, Theorem 7.1], we get that

$$\partial(g_\lambda \circ \sigma)(\overline{\mathcal{X}}) = \left\{ \mathrm{bdiag}(\widehat{\mathcal{U}}_{\mathbf{U}}) \cdot \mathrm{bdiag}(\widehat{\mathcal{D}}_{\mathbf{U}}) \cdot \mathrm{bdiag}(\widehat{\mathcal{V}}_{\mathbf{U}}) \mid \widehat{\mathcal{D}}^{\langle i \rangle}_{\mathbf{U}} \in \mathfrak{B}_i, i = 1, 2, \ldots, n_3 \right\}, \tag{82}$$

where $\mathfrak{B}_i$ is defined in (81). It follows from Lemma 11 that

$$\partial(g_\lambda \circ \sigma)(\overline{\mathcal{X}}) = \text{bdiag}(\partial H_\lambda(\widehat{\mathcal{X}}_{\mathbf{U}})), \tag{83}$$

where H_λ is defined as the same in Lemma 10 and bdiag$(\cdot)$ operates a set means that bdiag$(\cdot)$ operates each element of the corresponding set. In light of the definitions of G_λ and H_λ, it is obvious that

$$G_\lambda(\mathcal{X}) = H_\lambda(\widehat{\mathcal{X}}_{\mathbf{U}}) = H_\lambda(\mathbf{U}[\mathcal{X}]). \tag{84}$$

Under Assumptions 1, one can easily get that $H_\lambda(\cdot)$ is a proper and lower semicontinuous function. On the other hand, the tensor $\mathbf{U}[\mathcal{X}]$ is defined as multiplying by a unitary matrix $\mathbf{U}$ onto each tube of $\mathcal{X}$, which is an invertible linear transformation. Applying [56, Exercise 10.7] to (84) yields

$$\partial G_\lambda(\mathcal{X}) = \mathbf{U}^T[\partial H_\lambda(\mathbf{U}[\mathcal{X}])] = \mathbf{U}^T[\partial H_\lambda(\widehat{\mathcal{X}}_{\mathbf{U}})]. \tag{85}$$

Here for abuse of notation, we use $\mathbf{U}[E]$ to denote a set with each element of E being operated by $\mathbf{U}$. Taking (85) with (83), we obtain that

$$\text{bdiag}(\mathbf{U}[\partial G_\lambda(\mathcal{X})]) = \text{bdiag}(\partial H_\lambda(\widehat{\mathcal{X}}_{\mathbf{U}})) = \partial(g_\lambda \circ \sigma)(\overline{\mathcal{X}}),$$

which together with (82) and the definition of $\mathbf{U}$-product further yields

$$\begin{aligned}\partial G_\lambda(\mathcal{X}) &= \mathbf{U}^T[\text{fold}_3(\partial(g_\lambda \circ \sigma)(\overline{\mathcal{X}}))] \\ &= \left\{\mathcal{U} \diamond_{\mathbf{U}} \mathcal{D} \diamond_{\mathbf{U}} \mathcal{V}^T \mid \widehat{\mathcal{D}}_{\mathbf{U}}^{\langle i\rangle} \in \mathfrak{B}_i, i = 1, 2, \ldots, n_3\right\},\end{aligned}$$

where $\mathfrak{B}_i$ is defined in (81). The proof is completed. □

Appendix E. Proof of Theorem 2

Step 1. Let $\Delta = \tilde{\mathcal{X}} - \mathcal{X}^*$. Suppose that $\|\Delta\|_F > 1$, it can be easily seen from (6) that

$$\begin{aligned}\alpha_2\|\Delta\|_F - \tau_2\sqrt{\log(d)/n}\|\Delta\|_{\text{TTNN}} &\le \langle \nabla f_{n,\mathcal{Y}}\left(\mathcal{X}^* + \Delta\right) - \nabla f_{n,\mathcal{Y}}(\mathcal{X}^*), \Delta\rangle \\ &= \langle \nabla f_{n,\mathcal{Y}}(\tilde{\mathcal{X}}) - \nabla f_{n,\mathcal{Y}}\left(\mathcal{X}^*\right), \tilde{\mathcal{X}} - \mathcal{X}^*\rangle \\ &= \langle \nabla f_{n,\mathcal{Y}}(\tilde{\mathcal{X}}) + \mathcal{Z} - \mathcal{Z} - \nabla f_{n,\mathcal{Y}}(\mathcal{X}^*), \tilde{\mathcal{X}} - \mathcal{X}^*\rangle \\ &= \langle \nabla f_{n,\mathcal{Y}}(\tilde{\mathcal{X}}) + \mathcal{Z}, \tilde{\mathcal{X}} - \mathcal{X}^*\rangle + \langle -\mathcal{Z} - \nabla f_{n,\mathcal{Y}}(\mathcal{X}^*), \tilde{\mathcal{X}} - \mathcal{X}^*\rangle,\end{aligned} \tag{86}$$

where $\mathcal{Z} \in \partial(\beta G_\lambda(\tilde{\mathcal{X}}))$ will be given in detail later. Since $\tilde{\mathcal{X}}$ is a stationary point of (4), we get that

$$0 \in \nabla f_{n,\mathcal{Y}}(\tilde{\mathcal{X}}) + \partial(\beta G_\lambda(\tilde{\mathcal{X}})) + \partial\delta_D(\tilde{\mathcal{X}}),$$

where $\delta_D(\mathcal{X})$ is the indicator function of the set $D := \{\mathcal{X} : \|\mathcal{X}\|_\infty \le c\}$. Then there exists a tensor $\mathcal{Z} \in \partial(\beta G_\lambda(\tilde{\mathcal{X}}))$ such that $-\nabla f_{n,\mathcal{Y}}(\tilde{\mathcal{X}}) - \mathcal{Z} \in \partial\delta_D(\tilde{\mathcal{X}}) = N_D(\tilde{\mathcal{X}})$, where $N_D(\tilde{\mathcal{X}})$ denotes the normal cone of D at $\tilde{\mathcal{X}}$ and the equality holds by [5, Example 3.5]. Since $\mathcal{X}^* \in D$ is feasible, by the definition of normal cone (e.g., see [5]), we have

$$\langle \nabla f_{n,\mathcal{Y}}(\tilde{\mathcal{X}}) + \mathcal{Z}, \tilde{\mathcal{X}} - \mathcal{X}^*\rangle \le 0. \tag{87}$$

Taking (87) together with (86) yields

$$\begin{aligned}\alpha_2\|\Delta\|_F - \tau_2\sqrt{\frac{\log d}{n}}\|\Delta\|_{\text{TTNN}} &\le \langle -\mathcal{Z} - \nabla f_{n,\mathcal{Y}}(\mathcal{X}^*), \tilde{\mathcal{X}} - \mathcal{X}^*\rangle \\ &= \underbrace{\langle -\mathcal{Z}, \tilde{\mathcal{X}} - \mathcal{X}^*\rangle}_{I_1} + \underbrace{\langle -\nabla f_{n,\mathcal{Y}}(\mathcal{X}^*), \tilde{\mathcal{X}} - \mathcal{X}^*\rangle}_{I_2}.\end{aligned} \tag{88}$$

Upper bound of I_1. By virtue of the definition of the inner product of two tensors [59, Definition 5], we immediately obtain

$$\langle -\mathcal{Z}, \tilde{\mathcal{X}} - \mathcal{X}^* \rangle = \langle \overline{-\mathcal{Z}}, \overline{\Delta} \rangle \leq \left\| \overline{\mathcal{Z}} \right\| \left\| \overline{\Delta} \right\|_* = \|\mathcal{Z}\|_{\mathbf{U}} \|\Delta\|_{\text{TTNN}}, \tag{89}$$

where the first inequality holds by the Hölder's inequality. In addition, by Lemma 12 and Assumption 1(iii), we obtain $\|\mathcal{Z}\|_{\mathbf{U}} \leq \beta\lambda k_0$, which together with (89) yields

$$\langle -\mathcal{Z}, \tilde{\mathcal{X}} - \mathcal{X}^* \rangle \leq \beta\lambda k_0 \|\Delta\|_{\text{TTNN}}. \tag{90}$$

Upper bound of I_2. Similar to the analysis of I_1, we have

$$\begin{aligned} \langle -\nabla f_{n,\mathcal{Y}}(\mathcal{X}^*), \tilde{\mathcal{X}} - \mathcal{X}^* \rangle &= \langle \overline{-\nabla f_{n,\mathcal{Y}}(\mathcal{X}^*)}, \overline{\Delta} \rangle \\ &\leq \left\| \overline{\nabla f_{n,\mathcal{Y}}(\mathcal{X}^*)} \right\| \left\| \overline{\Delta} \right\|_* \\ &= \|\nabla f_{n,\mathcal{Y}}(\mathcal{X}^*)\|_{\mathbf{U}} \|\Delta\|_{\text{TTNN}} \\ &\leq \frac{\beta\lambda k_0}{4} \|\Delta\|_{\text{TTNN}}, \end{aligned} \tag{91}$$

where the last inequality follows from (7).

Substituting (91) and (90) into (88) yields

$$\alpha_2\|\Delta\|_F - \tau_2\sqrt{\frac{\log d}{n}}\|\Delta\|_{\text{TTNN}} \leq \left(\beta\lambda k_0 + \frac{\beta\lambda k_0}{4}\right)\|\Delta\|_{\text{TTNN}} = \frac{5\beta\lambda k_0}{4}\|\Delta\|_{\text{TTNN}}. \tag{92}$$

Since $\|\mathcal{X}\|_{\text{TTNN}} \leq t$, we can deduce that

$$\|\Delta\|_{\text{TTNN}} = \|\tilde{\mathcal{X}} - \mathcal{X}^*\|_{\text{TTNN}} \leq \|\tilde{\mathcal{X}}\|_{\text{TTNN}} + \|\mathcal{X}^*\|_{\text{TTNN}} \leq 2t, \tag{93}$$

where the first inequality follows from [51, Remark 2]. This taken together with (92) indicates that

$$\alpha_2\|\Delta\|_F - \tau_2\sqrt{\frac{\log d}{n}}\|\Delta\|_{\text{TTNN}} \leq \frac{5\beta\lambda k_0 t}{2}. \tag{94}$$

Note that

$$\begin{aligned} \tau_2\sqrt{\frac{\log d}{n}}\|\Delta\|_{\text{TTNN}} &\leq \tau_2\sqrt{\log d\frac{\alpha_2^2}{16t^2\tau_2^2\log d}}\|\Delta\|_{\text{TTNN}} \\ &= \frac{\alpha_2}{4t}\|\Delta\|_{\text{TTNN}} \\ &\leq \frac{\alpha_2}{4t}\cdot 2t = \frac{\alpha_2}{2}. \end{aligned} \tag{95}$$

where the first inequality holds by (8) and the second inequality holds by (93). By taking (95) with (94), we obtain that

$$\begin{aligned} \alpha_2\|\Delta\|_F &\leq \frac{\alpha_2}{2} + \frac{5\beta\lambda k_0 t}{2} \\ &\leq \frac{\alpha_2}{2} + \frac{5\alpha_2}{12} = \frac{11\alpha_2}{12}, \end{aligned} \tag{96}$$

where the second inequality holds by the assumption $\lambda \leq \frac{\alpha_2}{6t\beta k_0}$ in (7). Then we can derive that $\|\Delta\|_F \leq \frac{11}{12} < 1$, which leads to a contradiction with the assumption $\|\Delta\|_F > 1$.

Step 2. We know that $\|\Delta\|_F \leq 1$ from Step 1. Invoking the RSC condition in (6) gives

$$\alpha_1\|\Delta\|_F^2 - \tau_1\frac{\log d}{n}\|\Delta\|_{\text{TTNN}}^2 \leq \langle \nabla f_{n,\mathcal{Y}}(\mathcal{X}^* + \Delta) - \nabla f_{n,\mathcal{Y}}(\mathcal{X}^*), \Delta \rangle, \tag{97}$$

which together with (86) and (87) yields that

$$\begin{aligned}\alpha_1\|\Delta\|_F^2-\tau_1\frac{\log d}{n}\|\Delta\|_{\text{TTNN}}^2 &\le \langle -\mathcal{Z}-\nabla f_{n,\mathcal{Y}}(\mathcal{X}^*),\tilde{\mathcal{X}}-\mathcal{X}^*\rangle\\ &= \langle -\mathcal{Z},\tilde{\mathcal{X}}-\mathcal{X}^*\rangle+\langle -\nabla f_{n,\mathcal{Y}}(\mathcal{X}^*),\tilde{\mathcal{X}}-\mathcal{X}^*\rangle.\end{aligned} \tag{98}$$

It follows from Lemma 7 that

$$\langle -\mathcal{Z},\tilde{\mathcal{X}}-\mathcal{X}^*\rangle=\langle \mathcal{Z},\mathcal{X}^*-\tilde{\mathcal{X}}\rangle\le \beta G_\lambda(\mathcal{X}^*)-\beta G_\lambda(\tilde{\mathcal{X}})+\frac{\beta\mu}{2}\|\mathcal{X}^*-\tilde{\mathcal{X}}\|_F^2, \tag{99}$$

where G_λ is defined in (4). This together with (91) and (98) yields

$$\begin{aligned}\alpha_1\|\Delta\|_F^2-\tau_1\frac{\log d}{n}\|\Delta\|_{\text{TTNN}}^2 &\le \beta G_\lambda(\mathcal{X}^*)-\beta G_\lambda(\tilde{\mathcal{X}})+\frac{\beta\mu}{2}\|\mathcal{X}^*-\tilde{\mathcal{X}}\|_F^2+\frac{\beta\lambda k_0}{4}\|\Delta\|_{\text{TTNN}}\\ &= \beta G_\lambda(\mathcal{X}^*)-\beta G_\lambda(\tilde{\mathcal{X}})+\frac{\beta\mu}{2}\|\Delta\|_F^2+\frac{\beta\lambda k_0}{4}\|\Delta\|_{\text{TTNN}}.\end{aligned}$$

Rearranging the terms of the above inequality yields

$$\left(\alpha_1-\frac{\beta\mu}{2}\right)\|\Delta\|_F^2\le \beta G_\lambda\left(\mathcal{X}^*\right)-\beta G_\lambda(\tilde{\mathcal{X}})+\frac{\beta\lambda k_0}{4}\|\Delta\|_{\text{TTNN}}+\tau_1\frac{\log d}{n}\|\Delta\|_{\text{TTNN}}^2. \tag{100}$$

Notice that

$$\begin{aligned}\tau_1\frac{\log d}{n}\|\Delta\|_{\text{TTNN}}^2 &= \frac{\tau_1}{\alpha_2}\sqrt{\frac{\log d}{n}}\cdot\alpha_2\sqrt{\frac{\log d}{n}}\cdot\|\Delta\|_{\text{TTNN}}\cdot\|\Delta\|_{\text{TTNN}}\\ &\le \frac{\tau_1}{\alpha_2}\sqrt{\frac{\log d}{n}}\cdot\alpha_2\sqrt{\frac{\log d}{n}}\cdot 2t\cdot\|\Delta\|_{\text{TTNN}},\end{aligned} \tag{101}$$

where the inequality follows from (93). In view of (7) and (8), one has

$$\frac{\tau_1}{\alpha_2}\sqrt{\frac{\log d}{n}}\le\frac{\tau_1}{\alpha_2}\sqrt{\log d\frac{\alpha_2^2}{16t^2\tau_1^2\log d}}=\frac{1}{4t},\quad \alpha_2\sqrt{\frac{\log d}{n}}\le\frac{\beta\lambda k_0}{4}. \tag{102}$$

Plugging (102) into (101) then gives

$$\tau_1\frac{\log d}{n}\|\Delta\|_{\text{TTNN}}^2\le\frac{1}{4t}\cdot\frac{\beta\lambda k_0}{4}\cdot 2t\cdot\|\Delta\|_{\text{TTNN}}=\frac{\beta\lambda k_0}{8}\|\Delta\|_{\text{TTNN}},$$

which taken together with (100) gives

$$\begin{aligned}\left(\alpha_1-\frac{\beta\mu}{2}\right)\|\Delta\|_F^2 &\le \beta G_\lambda(\mathcal{X}^*)-\beta G_\lambda(\tilde{\mathcal{X}})+\frac{\beta\lambda k_0}{4}\|\Delta\|_{\text{TTNN}}+\frac{\beta\lambda k_0}{8}\|\Delta\|_{\text{TTNN}}\\ &\le \beta G_\lambda(\mathcal{X}^*)-\beta G_\lambda(\tilde{\mathcal{X}})+\frac{\beta\lambda k_0}{2}\|\Delta\|_{\text{TTNN}}.\end{aligned} \tag{103}$$

It follows from Lemma 5 and Definition 4 that

$$\begin{aligned}\frac{\beta\lambda k_0}{2}\|\Delta\|_{\text{TTNN}}=\frac{\beta\lambda k_0}{2}\sum_{i=1}^{n_3}\left\|\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle}\right\|_* &\le \frac{\beta\lambda k_0}{2}\sum_{i=1}^{n_3}\left(\frac{\phi(\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle})}{\lambda k_0}+\frac{\mu}{2\lambda k_0}\left\|\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle}\right\|_F^2\right)\\ &= \sum_{i=1}^{n_3}\left(\frac{\beta\phi(\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle})}{2}+\frac{\beta\mu}{4}\left\|\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle}\right\|_F^2\right),\end{aligned} \tag{104}$$

where $\phi(\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle}) = \sum_{j=1}^{m} g_\lambda(\sigma_j(\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle}))$.

For any matrix $\mathbf{A} \in \mathbb{R}^{n_1\times n_2}$, we define $\Phi_r(\mathbf{A})$ to be the best rank r approximation of $\mathbf{A}$ and $\Psi_r(\mathbf{A}) = \mathbf{A} - \Phi_r(\mathbf{A})$. Thus, one can rewrite (104) equivalently as follows:

$$\begin{aligned}\frac{\beta\lambda k_0}{2}\|\Delta\|_{\text{TTNN}} &\le \sum_{i=1}^{n_3}\frac{\beta\phi(\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle})}{2} + \sum_{i=1}^{n_3}\frac{\beta\mu}{4}\left\|\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle}\right\|_F^2 \\ &= \sum_{i=1}^{n_3}\frac{1}{2}\left(\beta\phi\big(\Phi_{r_i}(\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle})\big) + \beta\phi\big(\Psi_{r_i}(\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle})\big)\right) + \frac{\beta\mu}{4}\|\widehat{\Delta}_{\mathbf{U}}\|_F^2 \\ &= \sum_{i=1}^{n_3}\frac{1}{2}\left(\beta\phi\big(\Phi_{r_i}(\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle})\big) + \beta\phi\big(\Psi_{r_i}(\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle})\big)\right) + \frac{\beta\mu}{4}\|\Delta\|_F^2,\end{aligned} \tag{105}$$

where the last equation follows from the fact that $\|\widehat{\Delta}_{\mathbf{U}}\|_F = \|\Delta\|_F$. Recalling the definition of G_λ in (4), we know that $G_\lambda(\mathcal{X}) = \sum_{i=1}^{n_3}\phi(\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i\rangle})$ for any $\mathcal{X}\in\mathbb{R}^{n_1\times n_2\times n_3}$. Suppose r_i is the rank of matrix $(\widehat{\mathcal{X}^*})_{\mathbf{U}}^{\langle i\rangle}, i = 1,\dots,n_3$. Consequently, we obtain

$$\begin{aligned}\beta G_\lambda(\mathcal{X}^*) - \beta G_\lambda(\tilde{\mathcal{X}}) &= \sum_{i=1}^{n_3}\beta\big(\phi\big((\widehat{\mathcal{X}^*})_{\mathbf{U}}^{\langle i\rangle}\big) - \phi\big(\widehat{\tilde{\mathcal{X}}}_{\mathbf{U}}^{\langle i\rangle}\big)\big) \\ &\le \sum_{i=1}^{n_3}\beta\big(\phi\big(\Phi_{r_i}(\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle})\big) - \phi\big(\Psi_{r_i}(\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle})\big)\big),\end{aligned} \tag{106}$$

where the inequality holds arises from Lemma 9. Taking this together with (103) and (105), one has

$$\begin{aligned}&\left(\alpha_1 - \frac{\beta\mu}{2}\right)\|\Delta\|_F^2 \\ &\le \sum_{i=1}^{n_3}\left(\beta\phi\big(\Phi_{r_i}(\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle})\big) - \beta\phi\big(\Psi_{r_i}(\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle})\big)\right) + \sum_{i=1}^{n_3}\frac{1}{2}\left(\beta\phi\big(\Phi_{r_i}(\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle})\big) + \beta\phi\big(\Psi_{r_i}(\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle})\big)\right) + \frac{\beta\mu}{4}\|\Delta\|_F^2 \\ &= \sum_{i=1}^{n_3}\frac{3\beta}{2}\phi\big(\Phi_{r_i}(\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle})\big) - \sum_{i=1}^{n_3}\frac{\beta}{2}\phi\big(\Psi_{r_i}(\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle})\big) + \frac{\beta\mu}{4}\|\Delta\|_F^2.\end{aligned} \tag{107}$$

Rearranging the above terms yields

$$\begin{aligned}\left(\alpha_1 - \frac{3\beta\mu}{4}\right)\|\Delta\|_F^2 &\le \sum_{i=1}^{n_3}\frac{3\beta}{2}\phi\big(\Phi_{r_i}(\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle})\big) - \sum_{i=1}^{n_3}\frac{\beta}{2}\phi\big(\Psi_{r_i}(\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle})\big) \\ &\le \sum_{i=1}^{n_3}\frac{3\beta}{2}\phi\big(\Phi_{r_i}(\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle})\big).\end{aligned} \tag{108}$$

Step 3. It follows from [40, Lemma 4] that

$$\sum_{i=1}^{n_3}\phi\big(\Phi_{r_i}(\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle})\big) = \sum_{i=1}^{n_3}\sum_{j=1}^{r_i} g_\lambda(\sigma_j(\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle})) \le \sum_{i=1}^{n_3}\sum_{j=1}^{r_i}\lambda k_0\sigma_j(\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle}) = \sum_{i=1}^{n_3}\lambda k_0\left\|\Phi_{r_i}(\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle})\right\|_*. \tag{109}$$

Combining (108) and (109), one gets that

$$
\begin{aligned}
2\left(\alpha_1-\frac{3\beta\mu}{4}\right)\|\Delta\|_F^2 &\le \sum_{i=1}^{n_3} 3\beta\lambda k_0 \left\|\Phi_{r_i}\left(\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle}\right)\right\|_* \le \sum_{i=1}^{n_3} 3\beta\lambda k_0\sqrt{r_i}\left\|\Phi_{r_i}\left(\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle}\right)\right\|_F \\
&\le \sum_{i=1}^{n_3} 3\beta\lambda k_0\sqrt{r_i}\left\|\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle}\right\|_F \\
&\le 3\beta\lambda k_0\sqrt{\sum_{i=1}^{n_3} r_i}\left\|\widehat{\Delta}_{\mathbf{U}}\right\|_F = 3\beta\lambda k_0\sqrt{\sum_{i=1}^{n_3} r_i}\,\|\Delta\|_F .
\end{aligned} \tag{110}
$$

where the second inequality holds by the fact that $\|\mathbf{X}\|_* \le \sqrt{r}\|\mathbf{X}\|_F$ for any $\mathbf{X}$ with rank at most r, the third inequality holds since $\|\Phi_{r_i}(\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle})\|_F^2 = \sum_{j=1}^{r_i}(\sigma_j(\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle}))^2 \le \sum_{j=1}^{\min\{n_1,n_2\}}(\sigma_j(\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle}))^2 = \|\widehat{\Delta}_{\mathbf{U}}^{\langle i\rangle}\|_F^2$, and the fourth inequality holds by the Hölder's inequality [45]. Consequently, one can deduce that

$$
2\left(\alpha_1-\frac{3\beta\mu}{4}\right)\|\Delta\|_F \le 3\beta\lambda k_0\sqrt{\sum_{i=1}^{n_3} r_i}, \tag{111}
$$

which together with $\alpha_1 > \frac{3\beta\mu}{4}$ implies that

$$
\|\tilde{\mathcal{X}}-\mathcal{X}^*\|_F \le \frac{6\beta\lambda k_0\sqrt{\sum_{i=1}^{n_3} r_i}}{4\alpha_1-3\beta\mu}. \tag{112}
$$

This completes the proof.

Appendix F. Locally Lipschitz Continuity of S_2 in (11)

First we give the locally Lipschitz continuity for the composition function between the singular value vector and a differentiable function.

Lemma 13 *Suppose that s_2 satisfies Assumption 2. Define $f:\mathbb{R}^m\to\mathbb{R}$ as $f(\mathbf{x}) := \sum_{i=1}^m s_2(x_i)$. For any matrix $\mathbf{X}\in\mathbb{C}^{n_1\times n_2}$ with the singular value vector $\sigma(\mathbf{X}) := (\sigma_1(\mathbf{X}),\sigma_2(\mathbf{X}),\ldots,\sigma_m(\mathbf{X}))^T\in\mathbb{R}^m$, where $m=\min\{n_1,n_2\}$, define $(f\circ\sigma)(\mathbf{X}):\mathbb{C}^{n_1\times n_2}\to\mathbb{R}$ as $(f\circ\sigma)(\mathbf{X}) := f(\sigma(\mathbf{X}))$. Then $f\circ\sigma$ is differentiable. Moreover, $\nabla(f\circ\sigma)$ is locally Lipschitz continuous, i.e., for any given matrix $\tilde{\mathbf{X}}\in\mathbb{R}^{n_1\times n_2}$, there exist constants $\tilde{\delta}>0, L_1>0$ such that*

$$
\|\nabla(f\circ\sigma)(\mathbf{X}_1)-\nabla(f\circ\sigma)(\mathbf{X}_2)\|_F \le L_1\|\mathbf{X}_1-\mathbf{X}_2\|_F,\ \forall\,\mathbf{X}_1,\mathbf{X}_2\in B(\tilde{\mathbf{X}},\tilde{\delta}/(2\sqrt{m})).
$$

Proof. Since s_2 is convex, we know that $f(\mathbf{x})$ is convex. Recall that a matrix is a signed permutation matrix if there is just only one nonzero entry taken 1 or -1 in each row and column. Since s_2 is symmetric, we get that $f(\mathbf{Px}) = f(\mathbf{x})$ for any $\mathbf{x}\in\mathbb{R}$, which implies that f is absolutely symmetric. Since s_2 is differentiable and $f(\mathbf{x}) = \sum_{i=1}^m s_2(x_i)$, we can deduce that f is differentiable. As a consequence, we get that $f\circ\sigma$ is differentiable at $\mathbf{X}$ [34, Proposition 6.2]. Let the singular value decomposition of $\mathbf{X}\in\mathbb{C}^{n_1\times n_2}$ be $\mathbf{X}=\mathbf{U}\Sigma\mathbf{V}^T$, where $\mathbf{U}\in\mathbb{C}^{n_1\times n_1},\mathbf{V}\in\mathbb{C}^{n_2\times n_2}$ are unitary matrices. It follows from [34, Proposition 6.2] that

$$
\nabla(f\circ\sigma)(\mathbf{X}) = \mathbf{U}\,\mathrm{Diag}(\nabla f(\sigma(\mathbf{X})))\mathbf{V}^T.
$$

By the definition of f, we get that

$$g(\mathbf{x}) := \nabla f(\mathbf{x}) = (s_2'(x_1), s_2'(x_2), \ldots, s_2'(x_m))^T \in \mathbb{R}^m, \tag{113}$$

where x_i is the i-th component of $\mathbf{x}$. Denote

$$M(\mathbf{X}) := \nabla(f \circ \sigma)(\mathbf{X}) = \mathbf{U}\,\mathrm{Diag}(g(\sigma(\mathbf{X})))\mathbf{V}^T.$$

Since s_2 is symmetric and differentiable, we know that $s_2'(-x) = -s_2'(x)$. Therefore, for any signed permutation matrix $\mathbf{P} \in \mathbb{R}^{m\times m}$, one has $g(\mathbf{P}\mathbf{x}) = \mathbf{P}g(\mathbf{x})$, which shows that g is mixed symmetric at $\mathbf{x}$ [15, Definition 2.1].

Since the derivative s_2' of s_2 is locally Lipschitz continuous, we obtain that for any given x_i, there exist $\delta_i, L_0 > 0$ such that $|s_2'(y_i) - s_2'(z_i)| \le L_0|y_i - z_i|, \forall y_i, z_i \in B(x_i, \delta_i) := \{w : |w - x_i| \le \delta_i\}$. Therefore, for any given $\mathbf{x} = (x_1, x_2, \ldots, x_m)^T \in \mathbb{R}^m$,

$$\|g(\mathbf{y}) - g(\mathbf{z})\| \le L_0\|\mathbf{y} - \mathbf{z}\|, \ \ \forall \mathbf{y}, \mathbf{z} \in B(\mathbf{x}, \tilde{\delta}) := \{\mathbf{w} : \|\mathbf{w} - \mathbf{x}\| \le \tilde{\delta}\},$$

where $\tilde{\delta} = \sqrt{\sum_{i=1}^m \delta_i^2}$. Consequently, g is locally Lipschitz continuous near $\mathbf{x}$. For any given matrix $\tilde{\mathbf{X}} \in \mathbb{C}^{n_1\times n_2}$, it follows from [15, Theorem 3.3] that $M(\cdot)$ is locally Lipschitz continuous on $B(\tilde{\mathbf{X}}, \tilde{\delta}/(2\sqrt{m}))$ with modulus

$$L_1 = \max\{(2L_0\tilde{\delta} + \tau_0)/\tilde{\delta}, \sqrt{2}L_0\}, \tag{114}$$

where $\tau_0 := \max_{i,j}\{|s_2'(\sigma_i(\tilde{\mathbf{X}})) - s_2'(\sigma_j(\tilde{\mathbf{X}}))|, |s_2'(\sigma_i(\tilde{\mathbf{X}})) + s_2'(\sigma_j(\tilde{\mathbf{X}}))|\}$. □

Next we show the locally Lipschitz continuity of S_2 in (11), which is stated in the following lemma.

Lemma 14 *Define $S_2(\mathcal{X}) := \sum_{i=1}^{n_3}\sum_{j=1}^{\min\{n_1,n_2\}} s_2(\sigma_j(\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i\rangle}))$. Let $\mathcal{X} \in \mathbb{R}^{n_1\times n_2\times n_3}$ be a given tensor. Then there exists a constant $\tilde{\delta} > 0$ such that*

$$\|\nabla S_2(\mathcal{Y}) - \nabla S_2(\mathcal{Z})\|_F \le L_1\|\mathcal{Y} - \mathcal{Z}\|_F, \ \forall\, \mathcal{Y}, \mathcal{Z} \in B(\mathcal{X}, \tilde{\delta}/(2\sqrt{mn_3}))$$

where L_1 is the same constant defined in (114).

Proof. Define $f_1 : \mathbb{R}^{mn_3} \to \mathbb{R}$ as $f_1(\mathbf{x}) = \sum_{i=1}^{mn_3} s_2(x_i)$, where $m = \min\{n_1, n_2\}$. For any given tensor $\mathcal{X} \in \mathbb{R}^{n_1\times n_2\times n_3}$. it follows from Lemma 13 that

$$\|\nabla(f_1 \circ \sigma)(\overline{\mathcal{Y}}) - \nabla(f_1 \circ \sigma)(\overline{\mathcal{Z}})\|_F \le L_1\|\overline{\mathcal{Y}} - \overline{\mathcal{Z}}\|_F, \ \forall\, \overline{\mathcal{Y}}, \overline{\mathcal{Z}} \in B(\overline{\mathcal{X}}, \tilde{\delta}/(2\sqrt{mn_3})). \tag{115}$$

Define $P : \mathbb{C}^{n_1\times n_2\times n_3} \to \mathbb{R}$ as $P(\widehat{\mathcal{X}}_{\mathbf{U}}) := \sum_{i=1}^{n_3}\sum_{j=1}^{\min\{n_1,n_2\}} s_2(\sigma_j(\widehat{\mathcal{X}}_{\mathbf{U}}^{\langle i\rangle}))$. By a similar argument as that in Lemma 10, we get that $\overline{\mathcal{A}} = \nabla(f_1 \circ \sigma)(\overline{\mathcal{X}})$ is equivalent to $\widehat{\mathcal{A}}_{\mathbf{U}} = \nabla P(\widehat{\mathcal{X}}_{\mathbf{U}})$, which implies that $\nabla P(\widehat{\mathcal{X}}_{\mathbf{U}}) = \mathrm{fold}_3(\nabla(f_1 \circ \sigma)(\overline{\mathcal{X}}))$. Note that $\|\overline{\mathcal{Y}} - \overline{\mathcal{X}}\|_F = \|\widehat{\mathcal{Y}}_{\mathbf{U}} - \widehat{\mathcal{X}}_{\mathbf{U}}\|_F$, which yields $\widehat{\mathcal{Y}}_{\mathbf{U}} \in B(\widehat{\mathcal{X}}_{\mathbf{U}}, \tilde{\delta}/(2\sqrt{mn_3}))$. Similarly, we get that $\widehat{\mathcal{Z}}_{\mathbf{U}} \in B(\widehat{\mathcal{X}}_{\mathbf{U}}, \tilde{\delta}/(2\sqrt{mn_3}))$. Consequently, for any tensors $\widehat{\mathcal{Y}}_{\mathbf{U}}, \widehat{\mathcal{Z}}_{\mathbf{U}} \in B(\widehat{\mathcal{X}}_{\mathbf{U}}, \tilde{\delta}/(2\sqrt{mn_3}))$, we have

$$\begin{aligned}
\|\nabla P(\widehat{\mathcal{Y}}_{\mathbf{U}}) - \nabla P(\widehat{\mathcal{Z}}_{\mathbf{U}})\|_F &= \|\mathrm{fold}_3(\nabla(f_1 \circ \sigma)(\overline{\mathcal{Y}})) - \mathrm{fold}_3(\nabla(f_1 \circ \sigma)(\overline{\mathcal{Z}}))\|_F \\
&= \|\nabla(f_1 \circ \sigma)(\overline{\mathcal{Y}}) - \nabla(f_1 \circ \sigma)(\overline{\mathcal{Z}})\|_F \\
&\le L_1\|\overline{\mathcal{Y}} - \overline{\mathcal{Z}}\|_F \\
&= L_1\|\widehat{\mathcal{Y}}_{\mathbf{U}} - \widehat{\mathcal{Z}}_{\mathbf{U}}\|_F,
\end{aligned} \tag{116}$$

where the first inequality holds by (115).

By the definitions of S_2 and P, we get $S_2(\mathcal{X}) = P(\widehat{\mathcal{X}}_{\mathbf{U}}) = P(\mathbf{U}[\mathcal{X}])$. By using a similar discussion in [56, Exercise 10.7], we can easily obtain that

$$\nabla S_2(\mathcal{X}) = \mathbf{U}^T[\nabla P(\mathbf{U}[\mathcal{X}])] = \mathbf{U}^T[\nabla P(\widehat{\mathcal{X}}_{\mathbf{U}})]. \tag{117}$$

Additionally, we can easily obtain from $\widehat{\mathcal{Y}}_{\mathbf{U}}, \widehat{\mathcal{Z}}_{\mathbf{U}} \in B(\widehat{\mathcal{X}}_{\mathbf{U}}, \tilde{\delta}/(2\sqrt{mn_3}))$ that $\mathcal{Y}, \mathcal{Z} \in B(\mathcal{X}, \tilde{\delta}/(2\sqrt{mn_3}))$. By (116) and (117), we obtain

$$\begin{aligned}
\|\nabla S_2(\mathcal{Y}) - \nabla S_2(\mathcal{Z})\|_F &= \|\mathbf{U}^T[\nabla P(\widehat{\mathcal{Y}}_{\mathbf{U}})] - \mathbf{U}^T[\nabla P(\widehat{\mathcal{Z}}_{\mathbf{U}})]\|_F \\
&= \|\mathbf{U}^T[\nabla P(\widehat{\mathcal{Y}}_{\mathbf{U}}) - \nabla P(\widehat{\mathcal{Z}}_{\mathbf{U}})]\|_F \\
&= \|\nabla P(\widehat{\mathcal{Y}}_{\mathbf{U}}) - \nabla P(\widehat{\mathcal{Z}}_{\mathbf{U}})\|_F \\
&\leq L_1\|\widehat{\mathcal{Y}}_{\mathbf{U}} - \widehat{\mathcal{Z}}_{\mathbf{U}}\|_F \\
&= L_1\|\mathcal{Y} - \mathcal{Z}\|_F,
\end{aligned} \tag{118}$$

where the third equality holds since the tensor Frobenius norm is unitarily invariant [14, Definition 2.1]. This concludes the proof. □

Appendix G. Proof of Theorem 3

First, we give the sufficiently descent property of the objective function $H(\mathcal{X})$ in (15).

Lemma 15 *Let $\{\mathcal{X}^t\}$ be the sequence generated by Algorithm 1. Suppose that $\nabla f_{n,\mathcal{Y}}$ is Lipschitz continuous with Lipschitz constant L. Then for any $\rho > \frac{L}{1-2\xi}$ with $\xi \in (0, \frac{1}{2})$,*

$$H(\mathcal{X}^{t+1}) + a\|\mathcal{X}^{t+1} - \mathcal{X}^t\|_F^2 \leq H(\mathcal{X}^t), \tag{119}$$

where $a := \frac{1-2\xi}{2}\rho - \frac{L}{2} > 0$.

Proof. By virtue of the definitions (15) and (16), we immediately obtain

$$\begin{aligned}
&H(\mathcal{X}^{t+1}) - Q(\mathcal{X}^{t+1}, \mathcal{X}^t) \\
&= f_{n,\mathcal{Y}}(\mathcal{X}^{t+1}) + \beta S_1(\mathcal{X}^{t+1}) - \beta S_2(\mathcal{X}^{t+1}) + \delta_D(\mathcal{X}^{t+1}) - f_{n,\mathcal{Y}}(\mathcal{X}^t) - \langle \nabla f_{n,\mathcal{Y}}(\mathcal{X}^t), \mathcal{X}^{t+1} - \mathcal{X}^t\rangle \\
&\quad - \frac{\rho}{2}\|\mathcal{X}^{t+1} - \mathcal{X}^t\|_F^2 - \beta S_1(\mathcal{X}^{t+1}) + \beta S_2(\mathcal{X}^t) + \beta\langle \nabla S_2(\mathcal{X}^t), \mathcal{X}^{t+1} - \mathcal{X}^t\rangle - \delta_D(\mathcal{X}^{t+1}) \\
&= f_{n,\mathcal{Y}}(\mathcal{X}^{t+1}) - f_{n,\mathcal{Y}}f(\mathcal{X}^t) - \langle \nabla f_{n,\mathcal{Y}}(\mathcal{X}^t), \mathcal{X}^{t+1} - \mathcal{X}^t\rangle - \frac{\rho}{2}\|\mathcal{X}^{t+1} - \mathcal{X}^t\|_F^2 \\
&\quad - \beta(S_2(\mathcal{X}^{t+1}) - S_2(\mathcal{X}^t) - \langle \nabla S_2(\mathcal{X}^t), \mathcal{X}^{t+1} - \mathcal{X}^t\rangle).
\end{aligned} \tag{120}$$

Since $\nabla f_{n,\mathcal{Y}}$ is Lipschitz continuous with Lipschitz constant L, we can deduce from [5, Lemma 5.7] that

$$f_{n,\mathcal{Y}}(\mathcal{X}^{t+1}) - f_{n,\mathcal{Y}}(\mathcal{X}^t) - \langle \nabla f_{n,\mathcal{Y}}(\mathcal{X}^t), \mathcal{X}^{t+1} - \mathcal{X}^t\rangle \leq \frac{L}{2}\|\mathcal{X}^{t+1} - \mathcal{X}^t\|_F^2. \tag{121}$$

Additionally, it follows from the convexity of S_2 that

$$S_2(\mathcal{X}^{t+1}) \geq S_2(\mathcal{X}^t) + \langle \nabla S_2(\mathcal{X}^t), \mathcal{X}^{t+1} - \mathcal{X}^t\rangle. \tag{122}$$

Combining (120), (121) and (122), we get that

$$\begin{aligned} H(\mathcal{X}^{t+1}) - Q(\mathcal{X}^{t+1}, \mathcal{X}^t) &\leq \frac{L}{2}\|\mathcal{X}^{t+1} - \mathcal{X}^t\|_F^2 - \frac{\rho}{2}\|\mathcal{X}^{t+1} - \mathcal{X}^t\|_F^2 \\ &= \frac{L-\rho}{2}\|\mathcal{X}^{t+1} - \mathcal{X}^t\|_F^2. \end{aligned} \tag{123}$$

Recalling the definition of $Q(\mathcal{X}, \mathcal{X}^t)$ in (16), it can be easily verified that $Q(\mathcal{X}, \mathcal{X}^t)$ is convex. This together with the definition of $\mathcal{W}^{t+1}$ in (17) leads to

$$\begin{aligned} Q(\mathcal{X}^t, \mathcal{X}^t) &\geq Q(\mathcal{X}^{t+1}, \mathcal{X}^t) + \langle \mathcal{W}^{t+1}, \mathcal{X}^t - \mathcal{X}^{t+1} \rangle \\ &\geq Q(\mathcal{X}^{t+1}, \mathcal{X}^t) - \|\mathcal{W}^{t+1}\|_F \|\mathcal{X}^{t+1} - \mathcal{X}^t\|_F \\ &\geq Q(\mathcal{X}^{t+1}, \mathcal{X}^t) - \xi\rho\|\mathcal{X}^{t+1} - \mathcal{X}^t\|_F^2. \end{aligned} \tag{124}$$

Taking (124) together with (123) yields

$$\begin{aligned} H(\mathcal{X}^{t+1}) &\leq Q(\mathcal{X}^{t+1}, \mathcal{X}^t) + \frac{L-\rho}{2}\|\mathcal{X}^{t+1} - \mathcal{X}^t\|_F^2 \\ &\leq Q(\mathcal{X}^t, \mathcal{X}^t) + \xi\rho\|\mathcal{X}^{t+1} - \mathcal{X}^t\|_F^2 + \frac{L-\rho}{2}\|\mathcal{X}^{t+1} - \mathcal{X}^t\|_F^2 \\ &= Q(\mathcal{X}^t, \mathcal{X}^t) + \left(\frac{L}{2} - \frac{1-2\xi}{2}\rho\right)\|\mathcal{X}^{t+1} - \mathcal{X}^t\|_F^2. \end{aligned} \tag{125}$$

Therefore, we get that

$$H(\mathcal{X}^{t+1}) + \left(\frac{1-2\xi}{2}\rho - \frac{L}{2}\right)\|\mathcal{X}^{t+1} - \mathcal{X}^t\|_F^2 \leq Q(\mathcal{X}^t, \mathcal{X}^t) = H(\mathcal{X}^t), \tag{126}$$

where $\frac{1-2\xi}{2}\rho - \frac{L}{2} > 0$. This completes the proof. □

Next, we give the relative error property of the iterations in Algorithm 1, which is stated in the following lemma.

Lemma 16 *Let $\{\mathcal{X}^t\}$ be the sequence generated by Algorithm 1. Suppose that $\nabla f_{n,\mathcal{Y}}$ and s_2' are Lipschitz continuous and locally Lipschitz continuous with Lipschitz constants L and L_0, respectively. Then there exist $\mathcal{N}^{t+1} \in \partial H(\mathcal{X}^{t+1})$ and a positive integer K such that*

$$\|\mathcal{N}^{t+1}\|_F \leq \nu\|\mathcal{X}^{t+1} - \mathcal{X}^t\|_F, \ \forall\, t \geq K,$$

where $\nu := L + \beta L_1 + (\xi+1)\rho > 0$ and L_1 is defined in (114).

Proof. By (17), we obtain that there exists $\mathcal{Y}^{t+1} \in \partial[\beta S_1(\mathcal{X}^{t+1}) + \delta_D(\mathcal{X}^{t+1})]$ such that

$$\mathcal{W}^{t+1} = \nabla f_{n,\mathcal{Y}}(\mathcal{X}^t) + \rho(\mathcal{X}^{t+1} - \mathcal{X}^t) - \beta\nabla S_2(\mathcal{X}^t) + \mathcal{Y}^{t+1} \text{ and } \|\mathcal{W}^{t+1}\|_F \leq \xi\rho\|\mathcal{X}^{t+1} - \mathcal{X}^t\|_F.$$

Denote

$$\mathcal{N}^{t+1} := \nabla f_{n,\mathcal{Y}}(\mathcal{X}^{t+1}) - \beta\nabla S_2(\mathcal{X}^{t+1}) + \mathcal{Y}^{t+1}.$$

Note that

$$\partial H(\mathcal{X}) = \nabla f_{n,\mathcal{Y}}(\mathcal{X}) - \beta\nabla S_2(\mathcal{X}) + \partial[\beta S_1(\mathcal{X}) + \delta_D(\mathcal{X})].$$

Consequently, we deduce that $\mathcal{N}^{t+1} \in \partial H(\mathcal{X}^{t+1})$.

By (126), we know that $\|\mathcal{X}^{t+1} - \mathcal{X}^t\|_F$ tends to 0 as t tends to infinity, which implies that there exists a constant $\bar{\delta} > 0$ and positive integer K such that $\|\mathcal{X}^{t+1} - \mathcal{X}^t\|_F \leq \bar{\delta}$, for any $t \geq K$. Note that

the sequence $\{\mathcal{X}^t\}$ is an approximate solution of $Q(\mathcal{X}, \mathcal{X}^t)$ in (16) at each iteration and $\delta_D(\cdot)$ is the indicator function of D. Therefore, $\mathcal{X}^t \in D$, which implies that $\{\mathcal{X}^t\}$ is bounded. Assume that $\tilde{\mathcal{X}}$ is a cluster point of $\{\mathcal{X}^t\}$, there exists $\delta_0 > 0$ such that $\mathcal{X}^t \in B(\tilde{\mathcal{X}}, \delta_0)$ holds for $t \geq K$. We further obtain that

$$\begin{aligned}\|\mathcal{X}^{t+1} - \tilde{\mathcal{X}}\|_F &= \|\mathcal{X}^{t+1} - \mathcal{X}^t + \mathcal{X}^t - \tilde{\mathcal{X}}\|_F \\ &\leq \|\mathcal{X}^{t+1} - \mathcal{X}^t\|_F + \|\mathcal{X}^t - \tilde{\mathcal{X}}\|_F \\ &\leq \bar{\delta} + \delta_0.\end{aligned}$$

Denote $\tilde{\delta} := \bar{\delta} + \delta_0$. Then we have $\mathcal{X}^{t+1} \in B(\tilde{\mathcal{X}}, \tilde{\delta})$. Note that $\|\mathcal{X}^t - \tilde{\mathcal{X}}\|_F \leq \delta_0 < \tilde{\delta}$, which immediately yields $\mathcal{X}^t, \mathcal{X}^{t+1} \in B(\tilde{\mathcal{X}}, \tilde{\delta})$. It follows from Lemma 14 that

$$\|\nabla S_2(\mathcal{X}^{t+1}) - \nabla S_2(\mathcal{X}^t)\|_F \leq L_1 \|\mathcal{X}^{t+1} - \mathcal{X}^t\|_F. \tag{127}$$

for any $t \geq K$.

Notice that

$$\mathcal{N}^{t+1} = \nabla f_{n,\mathcal{Y}}(\mathcal{X}^{t+1}) - \beta \nabla S_2(\mathcal{X}^{t+1}) + \mathcal{W}^{t+1} - \nabla f_{n,\mathcal{Y}}(\mathcal{X}^t) - \rho(\mathcal{X}^{t+1} - \mathcal{X}^t) + \beta \nabla S_2(\mathcal{X}^t).$$

Then we get that

$$\begin{aligned}\|\mathcal{N}^{t+1}\|_F &\leq \|\nabla f_{n,\mathcal{Y}}(\mathcal{X}^{t+1}) - \nabla f_{n,\mathcal{Y}}(\mathcal{X}^t)\|_F + \|\mathcal{W}^{t+1}\|_F + \rho\|\mathcal{X}^{t+1} - \mathcal{X}^t\|_F \\ &\quad + \beta\|\nabla S_2(\mathcal{X}^{t+1}) - \nabla S_2(\mathcal{X}^t)\|_F \\ &\leq L\|\mathcal{X}^{t+1} - \mathcal{X}^t\|_F + (\xi+1)\rho\|\mathcal{X}^{t+1} - \mathcal{X}^t\|_F + \beta L_1\|\mathcal{X}^{t+1} - \mathcal{X}^t\|_F \\ &= (L + \beta L_1 + (\xi+1)\rho)\|\mathcal{X}^{t+1} - \mathcal{X}^t\|_F,\end{aligned} \tag{128}$$

where the second inequality holds by (127) and the Lipschitz continuity of $\nabla f_{n,\mathcal{Y}}$. This completes the proof. □

Now we return to prove Theorem 3 in detail. By the proof of Lemma 16, we know that $\{\mathcal{X}^t\}$ is bounded. Consequently, there exists a subsequence $\{\mathcal{X}^{t_i}\}$ such that $\mathcal{X}^{t_i}$ converges to $\tilde{\mathcal{X}}$ as i tends to ∞, where $\tilde{\mathcal{X}}$ is a cluster point of the sequence $\{\mathcal{X}^t\}$. Note that D is a closed set and $\mathcal{X}^{t_i} \in D$, then $\tilde{\mathcal{X}} \in D$. Therefore, $\delta_D(\mathcal{X}^{t_i})$ converges to $\delta_D(\tilde{\mathcal{X}})$ as i tends to ∞. Since $f_{n,\mathcal{Y}}$ and G_λ are continuous, we deduce that $f_{n,\mathcal{Y}}(\mathcal{X}^{t_i}) + \beta G_\lambda(\mathcal{X}^{t_i})$ converges to $f_{n,\mathcal{Y}}(\tilde{\mathcal{X}}) + \beta G_\lambda(\tilde{\mathcal{X}})$ as i tends to infinity, which implies that $H(\mathcal{X}^{t_i})$ converges to $H(\tilde{\mathcal{X}})$ as i tends to ∞.

Since D is a closed convex set, we know that $\delta_D(\cdot)$ is closed [5, Proposition 2.3], which implies that $\delta_D(\cdot)$ is lower semicontinouns [5, Theorem 2.6]. Since $f_{n,\mathcal{Y}}$ and G_λ are continuous, we obtain that $H(\mathcal{X})$ is lower semicontinouns. It is known that $\delta_D(\mathcal{X})$ is semialgebraic since D is a semialgebraic set [7]. Then $\delta_D(\mathcal{X})$ is a KL function [7, Theorem 3]. Since g_λ is a KL function, $G_\lambda(\mathcal{X})$ is a KL function [7, 50]. Note that $f_{n,\mathcal{Y}}$ is a KL function. As a result, $H(\mathcal{X}) = f_{n,\mathcal{Y}}(\mathcal{X}) + \beta G_\lambda(\mathcal{X}) + \delta_D(\mathcal{X})$ is a KL function. According to [3, Theorem 2.9], we conclude that the sequence $\{\mathcal{X}^t\}$ converges to $\tilde{\mathcal{X}}$ as t goes to infinity, and $\tilde{\mathcal{X}}$ is a stationary point of H.

Appendix H. Proof of Theorem 4

First, we give a lemma about the sequence $\{\mathcal{X}^t\}$ generated by Algorithm 1.

Lemma 17 *Suppose that the assumptions in Theorem 3 hold and $H(\mathcal{X})$ defined in (15) satisfies the KL property at $\tilde{\mathcal{X}}$ with an exponent $\alpha \in [0, 1)$. Then the following statements hold:*

(i) There exist constants $\rho_1, \mu_1 > 0$ and a KL exponent $\alpha \in (0,1)$, such that

$$\sum_{j=t}^{\infty} \|\mathcal{X}^{j+1} - \mathcal{X}^j\|_F \leq \|\mathcal{X}^t - \mathcal{X}^{t-1}\|_F + \frac{\mu_1}{\rho_1(1-\alpha)}(H(\mathcal{X}^t) - H(\tilde{\mathcal{X}}))^{1-\alpha}. \tag{129}$$

(ii) For any positive integer t, the following inequality holds:

$$\|\mathcal{X}^t - \tilde{\mathcal{X}}\|_F \leq \sum_{j=t}^{\infty} \|\mathcal{X}^{j+1} - \mathcal{X}^j\|_F. \tag{130}$$

Proof. (i) Define a function $p : \mathbb{R} \to \mathbb{R}_+$ as

$$p(s) := \frac{\mu_1}{1-\alpha}(s - H(\tilde{\mathcal{X}}))^{1-\alpha}, \ \forall s \geq H(\tilde{\mathcal{X}}),$$

where $\mu_1 > 0$ is a constant and $\alpha \in (0,1)$. It can be easily seen that p is a concave function, whose derivative function is given by $p'(s) = \frac{\mu_1}{(s-H(\tilde{\mathcal{X}}))^{\alpha}}$ for any $s > H(\tilde{\mathcal{X}})$. From the concavity of $p(s)$, one has

$$\begin{aligned} p(H(\mathcal{X}^t)) - p(H(\mathcal{X}^{t+1})) &\geq p'(H(\mathcal{X}^t))(H(\mathcal{X}^t) - H(\mathcal{X}^{t+1})) \\ &= \frac{\mu_1}{(H(\mathcal{X}^t) - H(\tilde{\mathcal{X}}))^{\alpha}}(H(\mathcal{X}^t) - H(\mathcal{X}^{t+1})). \end{aligned} \tag{131}$$

Note that $H(\mathcal{X})$ satisfies the KL property at $\tilde{\mathcal{X}}$ with an exponent $\alpha \in [0,1)$. Then the KL inequality (2) yields

$$(H(\mathcal{X}^t) - H(\tilde{\mathcal{X}}))^{\alpha} \leq \mu_1(1-\alpha)\,\mathrm{dist}(0, \partial H(\mathcal{X}^t)) \leq \mu_1\nu\|\mathcal{X}^t - \mathcal{X}^{t-1}\|_F, \tag{132}$$

where the last inequality follows from $\alpha \in (0,1)$ and Lemma 16. Here ν is defined in Lemma 16. This taken together with (131) and Lemma 15 gives

$$p(H(\mathcal{X}^t)) - p(H(\mathcal{X}^{t+1})) \geq \frac{a\|\mathcal{X}^{t+1} - \mathcal{X}^t\|_F^2}{\nu\|\mathcal{X}^t - \mathcal{X}^{t-1}\|_F},$$

where a is defined in Lemma 15. Furthermore, we can obtain that

$$\begin{aligned} 2\|\mathcal{X}^{t+1} - \mathcal{X}^t\|_F &\leq 2\Big(\frac{\nu}{a}\Big)^{\frac{1}{2}}\Big(p(H(\mathcal{X}^t)) - p(H(\mathcal{X}^{t+1}))\Big)^{\frac{1}{2}}\|\mathcal{X}^t - \mathcal{X}^{t-1}\|_F^{\frac{1}{2}} \\ &\leq \frac{\nu}{a}\Big(p(H(\mathcal{X}^t)) - p(H(\mathcal{X}^{t+1}))\Big) + \|\mathcal{X}^t - \mathcal{X}^{t-1}\|_F, \end{aligned}$$

where the second inequality follows from the fact that $2xy \leq x^2 + y^2$. Summing the aforementioned inequality from t to infinity yields

$$2\sum_{j=t}^{\infty}\|\mathcal{X}^{j+1} - \mathcal{X}^j\|_F \leq \sum_{j=t}^{\infty}\frac{\nu}{a}\Big(p(H(\mathcal{X}^j)) - p(H(\mathcal{X}^{j+1}))\Big) + \sum_{j=t}^{\infty}\|\mathcal{X}^j - \mathcal{X}^{j-1}\|_F.$$

Consequently,

$$\begin{aligned} \sum_{j=t}^{\infty}\big\|\mathcal{X}^{j+1} - \mathcal{X}^j\big\|_F &\leq \|\mathcal{X}^t - \mathcal{X}^{t-1}\|_F + \frac{\nu}{a}\sum_{j=t}^{\infty}(p(H(\mathcal{X}^j)) - p(H(\mathcal{X}^{j+1}))) \\ &= \|\mathcal{X}^t - \mathcal{X}^{t-1}\|_F + \frac{\nu}{a}\Big(p(H(\mathcal{X}^t)) - \lim_{j\to\infty} p(H(\mathcal{X}^j))\Big) \\ &\leq \|\mathcal{X}^t - \mathcal{X}^{t-1}\|_F + \frac{\nu}{a}p(H(\mathcal{X}^t)) \\ &= \|\mathcal{X}^t - \mathcal{X}^{t-1}\|_F + \frac{\nu\mu_1}{a(1-\alpha)}(H(\mathcal{X}^t) - H(\tilde{\mathcal{X}}))^{1-\alpha} \\ &= \|\mathcal{X}^t - \mathcal{X}^{t-1}\|_F + \frac{\mu_1}{\rho_1(1-\alpha)}(H(\mathcal{X}^t) - H(\tilde{\mathcal{X}}))^{1-\alpha}, \end{aligned}$$

where $\rho_1 := \frac{a}{\nu}$.

(ii) Define $M_t(\mathcal{X}) := \mathcal{X}^{t+1} - \mathcal{X}^t$. It follows from Theorem 3 that $\{\mathcal{X}^{t+1}\}_{t\in\mathbb{N}}$ is bounded and $\|M_t(\mathcal{X})\|_F = \|\mathcal{X}^{t+1} - \mathcal{X}^t\|_F < \infty$ for any $t = 1, 2, \ldots$ Additionally, we denote a bounded measurable set by $\mathcal{G}$, which obeys $\{\mathcal{X}^{t+1}\}_{t\in\mathbb{N}} \subset \mathcal{G}$. One can obtain that $M_t(\mathcal{X})$ is a measurable function on $\mathcal{G}$. Note that, the sequence $\{\mathcal{X}^t\}_{t\in\mathbb{N}}$ converges to $\tilde{\mathcal{X}}$, which implies that $\sum_{t=1}^n M_t(\mathcal{X}) = \sum_{t=1}^n \left(\mathcal{X}^{t+1} - \mathcal{X}^t\right) = \mathcal{X}^{n+1} - \mathcal{X}^1$ converges to $\tilde{\mathcal{X}} - \mathcal{X}^1$ as n tends to infinity. Similar to the argument in [70, Lemma 1(iv)], we can easily prove $\|\mathcal{X}^t - \tilde{\mathcal{X}}\|_F \le \sum_{j=t}^\infty \|\mathcal{X}^{j+1} - \mathcal{X}^j\|_F$. This completes the proof. □

We now provide the details of the proof of Theorem 4, which follows a similar argument of [2, 70].

(i) We must have $H(\mathcal{X}^{t_0}) = H(\tilde{\mathcal{X}})$ for some t_0 when $\alpha = 0$. Otherwise, for sufficiently large t, one has $H(\mathcal{X}^t) > H(\tilde{\mathcal{X}})$. Apply the KL inequality to obtain $\mu_1 \operatorname{dist}(0, \partial H(\mathcal{X}^t)) \ge 1$ for all t, it is impossible since $\mathcal{X}^t \to \tilde{\mathcal{X}}$ and $0 \in \partial H(\tilde{\mathcal{X}})$. Then there exists some t_0 such that $H(\mathcal{X}^{t_0}) = H(\tilde{\mathcal{X}})$. Note that H decreases monotonically, then $\mathcal{X}^t = \mathcal{X}^{t_0} = \tilde{\mathcal{X}}$ holds true for all $t > t_0$.

(ii) Let $\Delta_t := \sum_{j=t}^\infty \|\mathcal{X}^{j+1} - \mathcal{X}^j\|_F$. According to (130), it can be directly seen that $\Delta_t \ge \|\mathcal{X}^t - \tilde{\mathcal{X}}\|_F$. Then, (129) implies that

$$\begin{aligned}\Delta_t = \sum_{j=t}^\infty \|\mathcal{X}^{j+1} - \mathcal{X}^j\|_F &\le \|\mathcal{X}^t - \mathcal{X}^{t-1}\|_F + \frac{\mu_1}{\rho_1(1-\alpha)}(H(\mathcal{X}^t) - H(\tilde{\mathcal{X}}))^{1-\alpha} \\ &\le \|\mathcal{X}^t - \mathcal{X}^{t-1}\|_F + \frac{\mu_1}{\rho_1(1-\alpha)}\left(\mu_1\nu\|\mathcal{X}^t - \mathcal{X}^{t-1}\|_F\right)^{\frac{1-\alpha}{\alpha}} \\ &= (\Delta_{t-1} - \Delta_t) + a_1(\Delta_{t-1} - \Delta_t)^{\frac{1-\alpha}{\alpha}},\end{aligned} \tag{133}$$

where the second inequality follows from (132) and the last equality holds according to letting $a_1 := \frac{\mu_1}{\rho_1(1-\alpha)}(\mu_1\nu)^{\frac{1-\alpha}{\alpha}}$.

If $0 < \alpha \le \frac{1}{2}$, then $\frac{1-\alpha}{\alpha} \ge 1$. Similar to the proof in [70, Theorem 2(ii)], by (133), we can easily show that $\Delta_t \le \frac{a_2}{a_2+1}\Delta_{t-1}$ holds, where $a_2 := a_1 + 1$. Denote $w := \Delta_0$ and $\vartheta := \frac{a_2}{a_2+1} \in (0, 1)$. It follows from Lemma 17(ii) that

$$\|\mathcal{X}^t - \tilde{\mathcal{X}}\|_F \le \Delta_t \le \vartheta\Delta_{t-1} \le \cdots \le \vartheta^t\Delta_0 = w\vartheta^t.$$

(iii) If $\frac{1}{2} < \alpha < 1$, then $0 < \frac{1-\alpha}{\alpha} < 1$. Let $b_t := \sum_{j=0}^t \|\mathcal{X}^{j+1} - \mathcal{X}^j\|_F$. By setting $t = 1$ in (129), we can obtain that

$$\begin{aligned}\sum_{j=0}^\infty \|\mathcal{X}^{j+1} - \mathcal{X}^j\|_F &= \sum_{j=1}^\infty \|\mathcal{X}^{j+1} - \mathcal{X}^j\|_F + \|\mathcal{X}^1 - \mathcal{X}^0\|_F \\ &\le 2\|\mathcal{X}^1 - \mathcal{X}^0\|_F + \frac{\mu_1}{\rho_1(1-\alpha)}(H(\mathcal{X}^1) - H(\tilde{\mathcal{X}}))^{1-\alpha} < +\infty,\end{aligned} \tag{134}$$

which implies that $\lim_{t\to\infty} b_t = \sum_{j=0}^\infty \|\mathcal{X}^{j+1} - \mathcal{X}^j\|_F$ exists. Observe that

$$\Delta_t = \sum_{j=t}^\infty \|\mathcal{X}^{j+1} - \mathcal{X}^j\|_F = \sum_{j=0}^\infty \|\mathcal{X}^{j+1} - \mathcal{X}^j\|_F - b_{t-1}.$$

Hence, we have

$$\lim_{t\to\infty} \Delta_t = \sum_{j=0}^\infty \|\mathcal{X}^{j+1} - \mathcal{X}^j\|_F - \lim_{t\to\infty} b_{t-1} = 0. \tag{135}$$

Consequently, there exists some positive integer T_0 such that for any $t > T_0$,

$$(\Delta_{t-1} - \Delta_t) \le (\Delta_{t-1} - \Delta_t)^{\frac{1-\alpha}{\alpha}}.$$

This together with (133) leads to

$$\Delta_t^{\frac{\alpha}{1-\alpha}} \leq a_3(\Delta_{t-1} - \Delta_t), \tag{136}$$

where $a_3 = (1 + a_1)^{\frac{\alpha}{1-\alpha}}$.

Given a constant c with $c \in (1, \infty)$. Define $\gamma(s) := s^{-\frac{\alpha}{1-\alpha}}, s > 0$.

Case I. Assume that $\gamma(\Delta_t) \leq c\gamma(\Delta_{t-1})$, by (136), we obtain

$$\begin{aligned} a_3^{-1} &\leq \Delta_t^{-\frac{\alpha}{1-\alpha}}(\Delta_{t-1} - \Delta_t) = \gamma(\Delta_t)(\Delta_{t-1} - \Delta_t) \\ &\leq \int_{\Delta_t}^{\Delta_{t-1}} c\gamma(\Delta_{t-1})ds \leq \int_{\Delta_t}^{\Delta_{t-1}} c\gamma(s)ds \\ &= \frac{c(1-\alpha)}{1-2\alpha}\left(\Delta_{t-1}^{\frac{1-2\alpha}{1-\alpha}} - \Delta_t^{\frac{1-2\alpha}{1-\alpha}}\right). \end{aligned} \tag{137}$$

Let $\tau := \frac{1-2\alpha}{1-\alpha}$ and $e_1 := a_3^{-1}\frac{2\alpha-1}{c(1-\alpha)}$. Note that $\tau < 0$. Observe from (137) that

$$\Delta_t^\tau - \Delta_{t-1}^\tau \geq e_1 > 0. \tag{138}$$

Case II. Assume that $\gamma(\Delta_t) > c\gamma(\Delta_{t-1})$, i.e., $\Delta_t^{-\frac{\alpha}{1-\alpha}} > c\Delta_{t-1}^{-\frac{\alpha}{1-\alpha}}$, which is equivalent to $\Delta_t < c^{-\frac{1-\alpha}{\alpha}}\Delta_{t-1}$. By the definition of $\tau = \frac{1-2\alpha}{1-\alpha} < 0$, we get that

$$\Delta_t^\tau > (c^{-\frac{1-\alpha}{\alpha}})^\tau \Delta_{t-1}^\tau,$$

which implies that

$$\Delta_t^\tau - \Delta_{t-1}^\tau > ((c^{-\frac{1-\alpha}{\alpha}})^\tau - 1)\Delta_{t-1}^\tau. \tag{139}$$

Notice that $c^{-\frac{1-\alpha}{\alpha}} \in (0, 1)$. Therefore, we obtain $(c^{-\frac{1-\alpha}{\alpha}})^\tau - 1 > 0$.

It follows from (135) that there exist a constant $a_4 > 0$ and a positive integer T_1 such that $\Delta_{t-1} \leq a_4$ for any $t > T_1$. Consequently, we can deduce that $((c^{-\frac{1-\alpha}{\alpha}})^\tau - 1)\Delta_{t-1}^\tau > ((c^{-\frac{1-\alpha}{\alpha}})^\tau - 1)a_4^\tau$. This taken together with (139) leads to

$$\Delta_t^\tau - \Delta_{t-1}^\tau > e_2 > 0, \tag{140}$$

where $e_2 := ((c^{-\frac{1-\alpha}{\alpha}})^\tau - 1)a_4^\tau$.

Let $\tilde{e} := \min\{e_1, e_2\}$. Combining (138) with (140), we obtain $\Delta_t^\tau - \Delta_{t-1}^\tau \geq \tilde{e}$. By a similar argument as [70, Theorem 2(iii)], we get that $\Delta_t^\tau \geq \frac{\tilde{e}t}{2}$, where $t > 2\max\{T_0, T_1\}$. Then we can conclude that

$$\|\mathcal{X}^t - \tilde{\mathcal{X}}\|_F \leq \Delta_t \leq wt^{\frac{1}{\tau}} = wt^{-\frac{1-\alpha}{2\alpha-1}},$$

where $w := (\frac{\tilde{e}}{2})^{-\frac{1-\alpha}{2\alpha-1}}$.

References

[1] M. Ahn, J.-S. Pang, and J. Xin. Difference-of-convex learning: Directional stationarity, optimality, and sparsity. *SIAM J. Optim.*, 27(3):1637–1665, 2017.

[2] H. Attouch and J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Math. Program.*, 116(1-2):5–16, 2009.

[3] H. Attouch, J. Bolte, and B. F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss-Seidel methods. *Math. Program.*, 137(1-2):91–129, 2013.

[4] A. Banerjee, S. Chen, F. Fazayeli, and V. Sivakumar. Estimation with norm regularization. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, pages 1556–1564, Cambridge, MA, USA, 2014.

[5] A. Beck. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, 2017.

[6] J. A. Bengua, H. N. Phien, H. D. Tuan, and M. N. Do. Efficient tensor completion for color image and video recovery: Low-rank tensor train. *IEEE Trans. Image Process.*, 26(5):2466–2479, 2017.

[7] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Program.*, 146(1-2):459–494, 2014.

[8] C. Broadbent, T. Song, and R. Kuang. Deciphering high-order structures in spatial transcriptomes with graph-guided Tucker decomposition. *Bioinformatics*, 40:i529–i538, 2024.

[9] E. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

[10] E. J. Candès and Y. Plan. Matrix completion with noise. *Proc. IEEE*, 98(6):925–936, 2010.

[11] E. J. Candès, M. B. Wakin, and S. P. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *J. Fourier Anal. Appl.*, 14:877–905, 2008.

[12] J. D. Carroll and J.-J. Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika*, 35(3):283–319, 1970.

[13] C. Chen, K. Batselier, C.-Y. Ko, and N. Wong. A support tensor train machine. In *2019 International Joint Conference on Neural Networks*, pages 1–8, 2019.

[14] Y. Chen, Y. Chi, J. Fan, and C. Ma. Spectral methods for data science: A statistical perspective. *Found. Trends Mach. Learn.*, 14(5):566–806, 2021.

[15] C. Ding, D. Sun, J. Sun, and K.-C. Toh. Spectral operators of matrices: Semismoothness and characterizations of the generalized Jacobian. *SIAM J. Optim.*, 30(1):630–659, 2020.

[16] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.*, 96(456):1348–1360, 2001.

[17] M. Fazel, T. K. Pong, D. Sun, and P. Tseng. Hankel matrix rank minimization with applications to system identification and realization. *SIAM J. Matrix Anal. Appl.*, 34(3):946–977, 2013.

[18] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Springer, New York, 2013.

[19] S. Gandy, B. Recht, and I. Yamada. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Probl.*, 27(2):025010, 2011.

[20] K. Gao and Z.-H. Huang. Tensor robust principal component analysis via tensor fibered rank and ℓ_p minimization. *SIAM J. Imaging Sci.*, 16(1):423–460, 2023.

[21] R. Glowinski and A. Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. *Revue Française D'automatique, Informatique, Recherche Opérationnelle. Analyse Numérique*, 9(R2):41–76, 1975.

[22] P. Gong, C. Zhang, Z. Lu, J. Z. Huang, and J. Ye. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *Proceedings of the 30th International Conference on International Conference on Machine Learning*, volume 28, pages 37–45. JMLR, 2013.

[23] H. Gui, J. Han, and Q. Gu. Towards faster rates and oracle property for low-rank matrix estimation. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 2300–2309. PMLR, 2016.

[24] P. Hartman. On functions representable as a difference of convex functions. *Pac. J. Math.*, 9(3):707–713, 1959.

[25] L. He, X. Kong, P. S. Yu, X. Yang, A. B. Ragin, and Z. Hao. Dusk: A dual structure-preserving kernel for supervised tensor learning with applications to neuroimages. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 127–135, 2014.

[26] C. J. Hillar and L.-H. Lim. Most tensor problems are NP-hard. *J. ACM*, 60(6):45, 2013.

[27] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, Cambridge, 1991.

[28] T.-X. Jiang, M. K. Ng, X.-L. Zhao, and T.-Z. Huang. Framelet representation of tensor nuclear norm for third-order tensor completion. *IEEE Trans. Image Process.*, 29:7233–7244, 2020.

[29] E. Kernfeld, M. Kilmer, and S. Aeron. Tensor–tensor products with invertible linear transforms. *Linear Algebra Appl.*, 485:545–570, 2015.

[30] M. E. Kilmer, L. Horesh, H. Avron, and E. Newman. Tensor-tensor algebra for optimal representation and compression of multiway data. *Proc. Natl. Acad. Sci.*, 118(28):e2015851118, 2021.

[31] M. E. Kilmer and C. D. Martin. Factorization strategies for third-order tensors. *Linear Algebra Appl.*, 435(3):641–658, 2011.

[32] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Rev.*, 51(3):455–500, 2009.

[33] H. A. Le Thi, T. Pham Dinh, H. M. Le, and X. T. Vo. DC approximation approaches for sparse optimization. *Eur. J. Oper. Res.*, 244(1):26–46, 2015.

[34] A. S. Lewis and H. S. Sendov. Nonsmooth analysis of singular values. part I: Theory. *Set-Valued Anal.*, 13(3):213–241, 2005.

[35] B.-Z. Li, X.-L. Zhao, T.-Y. Ji, X.-J. Zhang, and T.-Z. Huang. Nonlinear transform induced tensor nuclear norm for tensor completion. *J. Sci. Comput.*, 92(3):83, 2022.

[36] G. Li and T. Pong. Calculus of the exponent of kurdyka-łojasiewicz inequality and its applications to linear convergence of first-order methods. *Found. Comput. Math.*, 18(5):1199–1232, 2018.

[37] H. Lian. Learning rate for convex support tensor machines. *IEEE Trans. Neural Netw. Learn. Syst.*, 32(8):3755–3760, 2021.

[38] J. Liu, P. Musialski, P. Wonka, and J. Ye. Tensor completion for estimating missing values in visual data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):208–220, 2013.

[39] P.-L. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Stat.*, 40(3):1637–1664, 2012.

[40] P.-L. Loh and M. J. Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *J. Mach. Learn. Res.*, 16(19):559–616, 2015.

[41] C. Lu, J. Feng, Y. Chen, W. Liu, Z. Lin, and S. Yan. Tensor robust principal component analysis with a new tensor nuclear norm. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(4):925–938, 2020.

[42] C. D. Martin, R. Shafer, and B. LaRue. An order-p tensor factorization with applications in imaging. *SIAM J. Sci. Comput.*, 35(1):A474–A490, 2013.

[43] W. Miao. *Matrix Completion Models with Fixed Basis Coefficients and Rank Regularized Problems with Hard Constraints*. PhD thesis, National University of Singapore, 2013.

[44] C. Mu, B. Huang, J. Wright, and D. Goldfarb. Square deal: Lower bounds and improved relaxations for tensor recovery. In *Proceedings of the 31st International Conference on Machine Learning*, pages 73–81. PMLR, 2014.

[45] G. S. Mudholkar, M. Freimer, and P. Subbaiah. An extension of Holder's inequality. *J. Math. Anal. Appl.*, 102(2):435–441, 1984.

[46] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Stat. Sci.*, 27(4):538–557, 2012.

[47] I. V. Oseledets. Tensor-train decomposition. *SIAM J. Sci. Comput.*, 33(5):2295–2317, 2011.

[48] Y. Panagakis, J. Kossaifi, G. G. Chrysos, J. Oldfield, M. A. Nicolaou, A. Anandkumar, and S. Zafeiriou. Tensor methods in computer vision and deep learning. *Proc. IEEE*, 109(5):863–890, 2021.

[49] W. Qin, H. Wang, F. Zhang, J. Wang, X. Luo, and T. Huang. Low-rank high-order tensor completion with applications in visual data. *IEEE Trans. Image Process.*, 31:2433–2448, 2022.

[50] D. Qiu, M. Bai, M. K. Ng, and X. Zhang. Nonlocal robust tensor recovery with nonconvex regularization. *Inverse Probl.*, 37(3):035001, 2021.

[51] D. Qiu, M. Bai, M. K. Ng, and X. Zhang. Robust low transformed multi-rank tensor methods for image alignment. *J. Sci. Comput.*, 87(1):24, 2021.

[52] D. Qiu, B. Yang, and X. Zhang. Robust tensor completion via dictionary learning and generalized nonconvex regularization for visual data recovery. *IEEE Trans. Circuits Syst. Video Technol.*, 2024.

[53] Y. Qiu, G. Zhou, Q. Zhao, and S. Xie. Noisy tensor completion via low-rank tensor ring. *IEEE Trans. Neural Netw. Learning Syst.*, 35(1):1127–1141, 2024.

[54] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52(3):471–501, 2010.

[55] O. Rivasplata. Subgaussian random variables: An expository note. Technical report, https://www.homepages.ucl.ac.uk/~ucabriv/pubs/note/subgaussians.pdf, 2012.

[56] R. T. Rockafellar and R. J. B. Wets. *Variational Analysis*. Springer, Berlin, 2009.

[57] B. Romera-Paredes and M. Pontil. A new convex relaxation for tensor completion. In *Procedings of Neural Information Processing Systems*, volume 26, pages 2967–2975, 2013.

[58] O. Semerci, N. Hao, M. E. Kilmer, and E. L. Miller. Tensor-based formulation and nuclear norm regularization for multienergy computed tomography. *IEEE Trans. Image Process.*, 23(4):1678–1693, 2014.

[59] G. Song, M. K. Ng, and X. Zhang. Robust tensor completion using transformed tensor singular value decomposition. *Numer. Linear Algebra Appl.*, 27(3):e2299, 2020.

[60] G.-J. Song, M. K. Ng, and X. Zhang. Tensor completion by multi-rank via unitary transformation. *Appl. Comput. Harmon. Anal.*, 65:348–373, 2023.

[61] X. Tan, Y. Zhang, S. Tang, J. Shao, F. Wu, and Y. Zhuang. Logistic tensor regression for classification. In *International Conference on Intelligent Science and Intelligent Data Engineering*, pages 573–581. Springer, 2013.

[62] P. Tang, C. Wang, D. Sun, and K.-C. Toh. A sparse semismooth newton based proximal majorization-minimization algorithm for nonconvex square-root-loss regression problems. *J. Mach. Learn. Res.*, 21(226):1–38, 2020.

[63] L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

[64] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, Cambridge, 2019.

[65] H. Wang and N. Ahuja. A tensor approximation approach to dimensionality reduction. *Inter. J. Comput. Vis.*, 76:217–229, 2008.

[66] H. Wang, F. Zhang, J. Wang, T. Huang, J. Huang, and X. Liu. Generalized nonconvex approach for low-tubal-rank tensor recovery. *IEEE Trans. Neural Netw. Learn. Syst.*, 33(8):3305–3319, 2022.

[67] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, Apr. 2004.

[68] B. Wen, X. Chen, and T. K. Pong. A proximal difference-of-convex algorithm with extrapolation. *Comput. Optim. Appl.*, 69(2):297–324, 2017.

[69] K. Wimalawarne, R. Tomioka, and M. Sugiyama. Theoretical and experimental analyses of tensor-based regression and classification. *Neural Comput.*, 28(4):686–715, 2016.

[70] S. Xia, D. Qiu, and X. Zhang. Tensor factorization via transformed tensor-tensor product for image alignment. *Numer. Algorithms*, 95(3):1251–1289, 2024.

[71] Y. Xu, R. Hao, W. Yin, and Z. Su. Parallel matrix factorization for low-rank tensor completion. *Inverse Probl. Imaging*, 9(2):601–624, 2015.

[72] Q. Yao, Y. Wang, B. Han, and J. T. Kwok. Low-rank tensor learning with nonconvex overlapped nuclear norm regularization. *J. Mach. Learn. Res.*, 23(136):1–60, 2022.

[73] M. Yin, D. Zeng, J. Gao, Z. Wu, and S. Xie. Robust multinomial logistic regression based on RPCA. *IEEE J. Sel. Topics Signal Process.*, 12(6):1144–1154, 2018.

[74] L. Yuan, C. Li, D. Mandic, J. Cao, and Q. Zhao. Tensor ring decomposition with rank minimization on latent space: An efficient approach for tensor completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9151–9158, 2019.

[75] C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.*, 38(2):894–942, 2010.

[76] X. Zhang. A nonconvex relaxation approach to low-rank tensor completion. *IEEE Trans. Neural Netw. Learn. Syst.*, 30(6):1659–1671, 2019.

[77] X. Zhang and M. K. Ng. A corrected tensor nuclear norm minimization method for noisy low-rank tensor completion. *SIAM J. Imaging Sci.*, 12(2):1231–1273, 2019.

[78] X. Zhang and M. K. Ng. Low rank tensor completion with Poisson observations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(8):4239–4251, 2022.

[79] X. Zhang and M. K. Ng. Sparse nonnegative tensor factorization and completion with noisy observations. *IEEE Trans. Inf. Theory*, 68(4):2551–2572, 2022.

[80] Q. Zhao, G. Zhou, S. Xie, L. Zhang, and A. Cichocki. Tensor ring decomposition. *arXiv:1606.05535*, 2016.

[81] X. Zhao, M. Bai, D. Sun, and L. Zheng. Robust tensor completion: Equivalent surrogates, error bounds, and algorithms. *SIAM J. Imaging Sci.*, 15(2):625–669, 2022.

[82] Y.-B. Zheng, T.-Z. Huang, X.-L. Zhao, Q. Zhao, and T.-X. Jiang. Fully-connected tensor network decomposition and its application to higher-order tensor completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11071–11078, 2021.