High-Dimensional Tensor Regression With Oracle Properties

Anonymous Author
Anonymous Institution

Abstract

Tensor regression is a powerful tool for modeling linear relationships between multi-dimensional variables in complex data analysis. In this paper, we study a high-dimensional tensor-response tensor regression model under low-dimensional structural assumptions. Specifically, we focus on five key common structural assumptions for the tensor coefficients: element-wise sparsity, fiber-wise sparsity, slice-wise sparsity, mode-wise low-rankness, and slice-wise low-rankness. To effectively handle these structures, we employ non-convex penalties tailored to each case, yielding refined statistical convergence rates. We derive general error bounds for the resulting estimators under mild assumptions, showing that they can achieve oracle rates. Furthermore, we propose a non-convex accelerated proximal gradient algorithm to compute the estimators. Extensive numerical experiments validate the theoretical findings and show good performance on both synthetic and real-world datasets, demonstrating the efficiency and robustness of the proposed methods.

1 INTRODUCTION

Tensors, or multi-dimension arrays, generalize matrices to higher dimensions and have become fundamental in various domains of data analysis (McCullagh, 2018). They have been proven useful in numerous fields, including biostatistics (Hore et al., 2016), chemometrics (Acar et al., 2011), image signal processing (Westin et al., 1999), etc. Tensor-based methods are particularly instrumental because they preserve the multi-linear structures inherent in the data. While applying classical methods to tensors typically requires unfolding tensors into matrices, which distorts the structure and results in information loss. Among tensor-based models, tensor regression is beneficial to uncover linear relationships between sets of high-dimensional variables and has significant applications across various fields (Liu et al., 2022). For example, in image processing, tensor regression has been successfully applied to tasks such as denoising (Zhang et al., 2023), image inpainting (Fan et al., 2017), and medical image analysis (Hua Zhou and Zhu, 2013). In multitask learning, tensor regression leverages shared information across multiple learning tasks to develop more accurate models for each task, achieving superior performance compared to learning them individually (Yang and Hospedales, 2017).

Tensor regression models frequently encounter significant challenges due to the complexity and scale of the high-order data involved, leading to pronounced issues of ill-posedness (Auddy et al., 2024). This problem is further exacerbated in high-dimensional settings where the number of model parameters substantially exceeds the number of observations (Raskutti et al., 2019). To address this issue, it is critical to impose structural assumptions on tensor models. One effective approach is to apply regularization penalties to tensor regression models to capture underlying structural data. Common penalties include sparsity (Tibshirani, 1996) or low-rankness penalties (Recht et al., 2010; Candes and Recht, 2012). Sparsity in tensors can manifest at the element-wise (Zhang et al., 2019; Raskutti et al., 2019), fiber-wise (Raskutti et al., 2019) or slice-wise level (Raskutti et al., 2019), each with different analytical implications. Raskutti et al. (2019) investigated various sparsity structures, establishing both general risk bounds and specific upper bounds in scenarios using convex optimization techniques. Similarly, low-rankness can be enforced on different matricized forms of tensors, depending on the specific application (Farias and Li, 2017; Raskutti et al., 2019; Luo and Zhang, 2022). In their work, Raskutti et al. (2019) explored penalties for mode-wise and slice-wise low-rankness, deriving comprehensive risk and bound analysis across different scenarios through convex op-

Preliminary work. Under review by AISTATS 2025. Do not distribute.

timization. Another approach to capturing structures is via tensor decomposition, such as CANDECOMP/PARAFAC (CP) decomposition (Kiers, 2000) and Tucker decomposition (Tucker, 1966), etc. However, methods based on tensor decomposition are generally non-unique and involve solving non-convex optimization problems. To further reduce the number of estimated parameters and improve model performance, some studies consider combining both constraints and penalties (Ahmed et al., 2020; Xu, 2020).

In this paper, we investigate a high-dimensional tensor-on-tensor regression model incorporating penalty regularizers. Existing works, such as (Zhang et al., 2019; Raskutti et al., 2019), have adapted a convex approach inspired by classical regularization methods—such as Lasso (Tibshirani, 1996) and nuclear norm-based approaches (Recht et al., 2010; Candes and Recht, 2012). Despite their popularity, these convex methods introduce non-negligible biases that adversely affect estimator accuracy. Departing from these conventional approaches, we focus on non-convex regularization techniques. Specifically, we employ the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001) and the minimax concave penalty (MCP) (Zhang, 2010), which demonstrate superior performance in inducing sparsity and low-rank structures without the drawbacks of convex methods. We propose five distinct non-convex regularizers, and our results show that these estimators not only achieve faster statistical convergence rates but also enjoy oracle properties under mild assumptions. To optimize these estimators, we develop a novel accelerated proximal gradient algorithm to address the challenges of non-convex optimization. This work establishes a unifying and general framework for tensor regression, bridging gaps between empirical results and theoretical understanding in current research. We substantiate our theoretical contributions through extensive numerical experiments on both synthetic data and real-world datasets.

Notations. Scalars and matrices are denoted by standard lowercase letters and boldface capital letters, respectively. For a matrix $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$, its singular values are denoted as $\sigma_1(\mathbf{X}) \geq \ldots \geq \sigma_{\min\{d_1,d_2\}}(\mathbf{X})$. The nuclear norm is defined as $\|\mathbf{X}\|_* = \sum_{i=1}^{\min\{d_1,d_2\}} \sigma_i$, and the spectral norm is defined as $\|\mathbf{X}\|_{\text{sp}} = \sigma_1(\mathbf{X})$. We denote the i-th column of $\mathbf{X}$ by $\mathbf{X}_{\cdot,i}$, and the j-th row of $\mathbf{X}$ by $\mathbf{X}_{j,\cdot}$. The Frobenius norm of $\mathbf{X}$ is defined as $\mathbf{X} = \sqrt{\sum_{i=1}^{d_2} \|\mathbf{X}_{\cdot,i}\|_2^2}$, where $\|\mathbf{X}_{\cdot,i}\|_2$ denotes the Euclidean norm of vector $\mathbf{X}_{\cdot,i}$. Define $\text{vec}(\mathbf{X})$ as the column-wise vectorization of $\mathbf{X}$.

Tensors are denoted by boldface calligraphy letters. The order of a tensor is determined by the number of dimensions (or modes) it has; hence, scalars, vectors, and matrices can be seen as 0th-order, 1st-order, and 2nd-order tensors, respectively. For a N-th order tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1 \times \cdots \times d_N}$, where the i-th dimension is with size d_i, $i \in \{1, \ldots, N\}$, the entry of $\boldsymbol{\mathcal{X}}$ at position $(i_1, \ldots, i_N)$ is denoted by $\boldsymbol{\mathcal{X}}_{i_1,\ldots,i_N}$. By fixing all indices of tensor $\boldsymbol{\mathcal{X}}$ except for the k-th mode, we obtain mode-(k) fibers, which are vectors. Similarly, mode-(j,k) slices are derived by fixing all indices except for j-th and k-th modes, which are matrices. For any tensor $\boldsymbol{\mathcal{X}}$, it can be unfolded into a matrix via mode-(k) unfolding, denoted as $\boldsymbol{\mathcal{X}}_{(k)} \in \mathbb{R}^{d_k \times \Pi_{i \neq k} d_i}$. This unfolding arranges the mode-(k) fibers of $\boldsymbol{\mathcal{X}}$ into the columns of $\boldsymbol{\mathcal{X}}_{(k)}$. Specifically, the elements of $\boldsymbol{\mathcal{X}}_{(k)}$ satisfy $\left[\boldsymbol{\mathcal{X}}_{(k)}\right]_{j_k,l} = \boldsymbol{\mathcal{X}}_{j_1,\ldots,j_k,\ldots,j_N}$, where the column index $l = 1 + \sum_{\substack{s=1 \\ s \neq k}}^{N} (j_s - 1) \prod_{\substack{m=1 \\ m \neq k}}^{s-1} d_m$. Similarly, the tensor $\boldsymbol{\mathcal{X}}$ can be unfolded into a 3rd-order tensor via mode-(j,k) unfolding, denoted as $\boldsymbol{\mathcal{X}}_{(j,k)} \in \mathbb{R}^{d_j \times d_k \times \prod_{s \neq j,k} d_s}$. This unfolding arranges the mode-(j,k) slices of $\boldsymbol{\mathcal{X}}$ into the frontal slices of $\boldsymbol{\mathcal{X}}_{(j,k)}$. Specifically, the element satisfy $\left[\boldsymbol{\mathcal{X}}_{(j,k)}\right]_{i_j,i_k,l} = \boldsymbol{\mathcal{X}}_{i_1,\ldots,i_N}$, where the slice index $l = 1 + \sum_{\substack{s=1 \\ s \neq j,k}}^{N} (i_s - 1) \prod_{\substack{m=1 \\ m \neq j,k}}^{s-1} d_m$. Define $\text{vec}(\boldsymbol{\mathcal{X}})$ as the vectorization of tensor $\boldsymbol{\mathcal{X}}$, where $\text{vec}(\boldsymbol{\mathcal{X}}) = \text{vec}\left(\boldsymbol{\mathcal{X}}_{(1)}\right)$. The the Frobenius norm of $\boldsymbol{\mathcal{X}}$ is defined as $\|\boldsymbol{\mathcal{X}}\|_{\text{F}} = \sqrt{\sum_{i_1=1}^{d_1} \cdots \sum_{i_N=1}^{d_N} \boldsymbol{\mathcal{X}}_{i_1,\ldots,i_N}^2}$.

For notational simplicity, we adopt a slight abuse of notation by using calligraphic letters to represent subspaces and index sets. The support of a set $\mathcal{S}$ is denoted by $|\mathcal{S}|$. For a function f, $f'(\cdot)$ denotes its derivative, $\nabla f(\cdot)$ represents its gradient, and $\nabla^2 f(\cdot)$ denotes its Hessian. For functionals $f(x)$ and $g(x)$, we denote $f(x) \gtrsim g(x)$ if $f(x) \geq c g(x)$, $f(x) \lesssim g(x)$ if $f(x) \leq C g(x)$, and $f(x) \asymp g(x)$ if $c g(x) \leq f(x) \leq C g(x)$ for some positive constants c and C.

2 PROBLEM FORMULATION

In this section, we present a unified framework for high-dimensional tensor regression with non-convex regularization.

2.1 Tensor Regression Model

We consider the following generic observation model with tensor coefficient $\boldsymbol{\mathcal{A}}^* \in \mathbb{R}^{d_1 \times \cdots \times d_M}$:

$$\boldsymbol{\mathcal{Y}} = \langle \boldsymbol{\mathcal{A}}^*, \boldsymbol{\mathcal{X}} \rangle + \boldsymbol{\mathcal{E}}, \tag{1}$$

where $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1 \times \cdots \times d_N}$ is the predictor variable, $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{d_{N+1} \times \cdots \times d_M}$ is the response variable with $M \geq N$, and $\boldsymbol{\mathcal{E}} \in \mathbb{R}^{d_{N+1} \times \cdots \times d_M}$ is the noise. $\langle \cdot, \cdot \rangle$ is defined as a generalized inner product between two tensors, where its $(i_{N+1}, \ldots, i_M)$-

th element is defined as $[\langle \boldsymbol{\mathcal{A}}^*, \boldsymbol{\mathcal{X}} \rangle]_{j_{N+1},\ldots,j_M} = \sum_{j_1=1}^{d_1} \cdots \sum_{j_N=1}^{d_N} \boldsymbol{\mathcal{A}}^*_{j_1,\ldots,j_N} \boldsymbol{\mathcal{X}}_{j_1,\ldots,j_N,j_{N+1},\ldots,j_M}$. Specifically, when $M = N$, the output is a scalar, in which case (1) becomes a scalar-on-tensor regression model (Hua Zhou and Zhu, 2013; Gui et al., 2016); when $M > N$, the output is an $(M-N)$-th order tensor, in which case (1) is the tensor-on-tensor regression model (Lock, 2018; Raskutti et al., 2019; Huihui Miao and Shi, 2022).

2.2 Proposed Estimators

Given a collection of n samples $\left\{\left(\boldsymbol{\mathcal{Y}}^{(i)}, \boldsymbol{\mathcal{X}}^{(i)}\right)\right\}_{i=1}^{n}$, which is assumed to be generated from the observation model (1), our goal is to estimate the unknown coefficient tensor $\boldsymbol{\mathcal{A}}^*$ by solving the following regularized optimization problem:

$$\min_{\boldsymbol{\mathcal{A}} \in \mathbb{R}^{d_1 \times \cdots \times d_N}} L(\boldsymbol{\mathcal{A}}) + R_\lambda(\boldsymbol{\mathcal{A}}), \tag{2}$$

where $L(\boldsymbol{\mathcal{A}})$ is an empirical loss function defined as $L(\boldsymbol{\mathcal{A}}) = \frac{1}{2n} \sum_{i=1}^{n} \left\| \boldsymbol{\mathcal{Y}}^{(i)} - \left\langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{X}}^{(i)} \right\rangle \right\|_{\mathrm{F}}^2$ and $R_\lambda(\boldsymbol{\mathcal{A}})$ is a regularizer with parameter $\lambda \geq 0$. In this paper, we consider a family of decomposable non-convex penalty functions (Fan and Li, 2001; Zhang, 2010; Negahban et al., 2012), which enforces structural constraints on the regularized M-estimator $\hat{\boldsymbol{\mathcal{A}}}$. We first introduce a class of typical univariate non-convex functions $p_\lambda(\cdot)$ and subsequently introduce how the regularizers $R_\lambda(\boldsymbol{\mathcal{A}})$ are defined based on $p_\lambda(\cdot)$. A substantial body of literature has proposed various non-convex functions$p_\lambda(\cdot)$, with representative examples including SCAD (Fan and Li, 2001) and MCP (Zhang, 2010). These non-convex function $p_\lambda(x)$ can be decomposed as $p_\lambda(x) = \lambda|x| + q_\lambda(x)$, where $|x|$ represents the ℓ_1 penalty and $q_\lambda(x)$ is a concave function. Based on $p_\lambda(\cdot)$, we consider the following regularizers $R_\lambda(\boldsymbol{\mathcal{A}})$ to enforce various structural constraints on $\boldsymbol{\mathcal{A}}$.

2.2.1 Sparsity Penalties

A straightforward method to induce sparsity within a tensor $\boldsymbol{\mathcal{A}}$ is to encourage sparsity at element-wise level (Zhang et al., 2019; Raskutti et al., 2019), similar to the Lasso. This can be achieved by applying a non-convex penalty to each entry of the tensor. Specifically, the *element-wise sparsity penalty* is defined as:

$$R_\lambda(\boldsymbol{\mathcal{A}}) = \sum_{i_1=1}^{d_1} \cdots \sum_{i_M=1}^{d_M} p_\lambda(\boldsymbol{\mathcal{A}}_{i_1,\ldots,i_M}).$$

In many applications, sparsity manifests in structured patterns, such as entire fibers being zero (Li et al., 2015; Raskutti et al., 2019), analogous to the group Lasso (Yuan and Lin, 2006). To capture such structured sparsity, we introduce the fiber-wise sparsity regularizer. Consider unfolding the tensor $\boldsymbol{\mathcal{A}}$ along mode-(k), the *fiber-wise sparsity penalty* is defined as:

$$R_\lambda(\boldsymbol{\mathcal{A}}) = \sum_{l=1}^{\prod_{j \neq k} d_j} p_\lambda\left(\left\| \left[\boldsymbol{\mathcal{A}}_{(k)}\right]_{\cdot,l} \right\|_2 \right).$$

Beyond fiber-wise sparsity, a tensor may exhibit sparsity at the slice level, where the entire slice across two modes is zero (Raskutti et al., 2019). To promote such slice-wise sparsity, we introduce the *slice-wise sparsity penalty*:

$$R_\lambda(\boldsymbol{\mathcal{A}}) = \sum_{l=1}^{\prod_{s \neq j,k} d_s} p_\lambda\left(\left\| \left[\boldsymbol{\mathcal{A}}_{(j,k)}\right]_{\cdot,\cdot,l} \right\|_{\mathrm{F}} \right).$$

2.2.2 Low-rankness Penalties

In addition to promoting sparsity, encouraging low-rankness in tensors can be highly beneficial in various applications. Given the multiple notions of "tensor rank" (Kolda and Bader, 2009), we focus on two specific low-rank penalties, which we detail below.

One common approach to defining tensor rank is through the Tucker decomposition (Tucker, 1966). To encourage low tensor rank, we introduce the mode-wise low-rank regularizer, defined as the sum of the nuclear norms of its unfolded matrices along mode-(k). Building upon this, the *mode-wise low-rank penalty* is defined as:

$$R_\lambda(\boldsymbol{\mathcal{A}}) = \sum_{i=1}^{\min\left\{I_k, \prod_{j \neq k} d_j\right\}} p_\lambda\left(\sigma_i\left(\boldsymbol{\mathcal{A}}_{(k)}\right)\right),$$

where $\sigma_i\left(\boldsymbol{\mathcal{A}}_{(k)}\right)$ denotes the i-th singular value of the mode-(k) unfolding $\boldsymbol{\mathcal{A}}_{(k)}$. This regularizer, also known as the tensor nuclear norm regularizer, encourages low rankness across all modes of the tensor by penalizing the singular values of each unfolding (Raskutti et al., 2019).

Another intuitive approach is to encourage slice-wise low-rankness (Lock, 2018; Raskutti et al., 2019). In many applications, low-rank structures are present within specific slices of a tensor. To capture and exploit this property, we introduce the *slice-wise low-rank regularizer* defined as:

$$R_\lambda(\boldsymbol{\mathcal{A}}) = \sum_{l=1}^{\prod_{m \neq j,k} d_m} \sum_{s=1}^{s^{\mathrm{all}}} p_\lambda\left(\sigma_s\left(\left[\boldsymbol{\mathcal{A}}_{(j,k)}\right]_{\cdot,\cdot,l}\right)\right),$$

where $s^{\mathrm{all}} = \min\left\{d_j d_k, \prod_{l \neq j,k} d_l\right\}$.

3 MAIN THEORY

In this section, we present the primary theoretical results for the general estimator in (2). We first introduce some necessary assumptions regarding the empirical loss function and the non-convex penalty.

3.1 Technical Assumptions

Assumption 1 (Restricted strong convexity (RSC)). *For any tensor $\boldsymbol{\mathcal{B}}$, there exists a constant $\rho^- > 0$ such that*

$$L(\boldsymbol{\mathcal{A}}+\boldsymbol{\mathcal{B}}) - L(\boldsymbol{\mathcal{A}}) \geq \langle \nabla L(\boldsymbol{\mathcal{A}}), \boldsymbol{\mathcal{B}} \rangle + \frac{\rho^-}{2}\|\boldsymbol{\mathcal{B}}\|_{\mathrm{F}}^2.$$

Assumption 2 (Restricted strong smoothness (RSS)). *For any tensor $\boldsymbol{\mathcal{B}}$, there exists a constant ρ^+ satisfying $\rho^+ > \rho^- > 0$ such that*

$$L(\boldsymbol{\mathcal{A}}+\boldsymbol{\mathcal{B}}) - L(\boldsymbol{\mathcal{A}}) \leq \langle \nabla L(\boldsymbol{\mathcal{A}}), \boldsymbol{\mathcal{B}} \rangle + \frac{\rho^+}{2}\|\boldsymbol{\mathcal{B}}\|_{\mathrm{F}}^2.$$

Assumptions 1 and 2, which characterize the curvature behavior of the empirical loss function $L(\cdot)$, are analogous to the classical RSC and RSS conditions commonly used in the literature on simple linear regression problems (Gui et al., 2016; Elenberg et al., 2018). Following the methodology of (Candes and Tao, 2007), it can be proven that the function $L(\cdot)$ satisfies both RSC and RSS conditions with high probability.

Recall that the non-convex penalty function $R_\lambda(\cdot)$ is based on $p_\lambda(\cdot)$. We impose several conditions on the non-convex penalty function $R_\lambda(\cdot)$ in terms of functions $p_\lambda(\cdot)$ and $q_\lambda(\cdot)$.

Assumption 3. *The functions $p_\lambda(t)$ and $q_\lambda(t)$ satisfy the following conditions:*

- There exits a constant $\nu > 0$ such that the penalty function satisfies $p'_\lambda(t) = 0$ for all $t \geq \nu$;
- $q_\lambda(t)$ is symmetric, i.e., $q_\lambda(-t) = q_\lambda(t)$ for all t;
- $q'_\lambda(t)$ is monotone and Lipschitz continuous, i.e., for $t_2 \geq t_1$, there exist two non-negative constant ζ^+, ζ^- such that $-\zeta^- \leq \frac{q'_\lambda(t_2)-q'_\lambda(t_1)}{(t_2-t_1)} \leq -\zeta^+ \leq 0$;
- Both $q_\lambda(t)$ and $q'_\lambda(t)$ pass through the origin, i.e. $q_\lambda(0) = q'_\lambda(0) = 0$;
- There exists a positive constant λ such that $|q'_\lambda(t)| \leq \lambda$ for all t.

The third condition pertains to a curvature property that governs the degree of concavity level of $q'_\lambda(\cdot)$ and, consequently, the non-convexity level of $p_\lambda(\cdot)$. Such a condition is commonly employed in the analysis of non-convex regression problems (Wang et al., 2014; Gui et al., 2016; Fan et al., 2018).

Assumption 4. *Denote $\tilde{\boldsymbol{\mathcal{X}}}$ as $\tilde{\boldsymbol{\mathcal{X}}} = \left(\mathrm{vec}\left(\boldsymbol{\mathcal{X}}^{(1)}\right)^\top, \ldots, \mathrm{vec}\left(\boldsymbol{\mathcal{X}}^{(n)}\right)^\top\right)$ the concatenation of vectorized covariates from n samples. Assume that $\tilde{\boldsymbol{\mathcal{X}}}$ follows a multivariate normal distribution, i.e., $\tilde{\boldsymbol{\mathcal{X}}} \sim \mathbb{N}(0, \boldsymbol{\Sigma})$, with covariance matrix $\boldsymbol{\Sigma} = \mathrm{cov}\left(\tilde{\boldsymbol{\mathcal{X}}}, \tilde{\boldsymbol{\mathcal{X}}}\right)$. We assume that there exists constants $\kappa \geq 1$ such that the eigenvalues of covariance matrix $\boldsymbol{\Sigma}$ satisfy:*

$$\frac{1}{\kappa} \leq \sigma_{\min}(\boldsymbol{\Sigma}) \leq \sigma_{\max}(\boldsymbol{\Sigma}) \leq \kappa.$$

Assumption 4 ensures the positiveness and well-conditioning of the covariance matrix $\boldsymbol{\Sigma}$, which is essential for the non-degeneracy of the parameter space. Such conditions can be verified in various statistical examples (Raskutti et al., 2019). Notably, if the covariate $\left\{\boldsymbol{\mathcal{X}}^{(i)}\right\}_{i=1}^n$ are independent and identically distributed (i.i.d.), the covariance matrix $\boldsymbol{\Sigma}$ exhibits a block-diagonal structure. In this case, Assumption 4 simplifies to similar conditions on the covariance matrix of each sample.

3.2 Statistical Error Analysis

We now establish statistical error bounds for general decomposable regularizer $R_\lambda(\cdot)$.

Theorem 5. *Let the noise tensors $\boldsymbol{\mathcal{E}}^{(i)}$ be i.i.d. with each entry following a Gaussian distribution $\mathbb{N}(0, \eta^2)$. Suppose Assumptions 1∼4 hold. If $\|\boldsymbol{\mathcal{A}}\|_{\mathrm{F}} \leq R_\lambda(\boldsymbol{\mathcal{A}})$, and $\lambda \geq 8\eta\sqrt{\frac{\kappa}{n}}w(\Omega)$, where $\Omega = \{\boldsymbol{\mathcal{A}} \mid R_\lambda(\boldsymbol{\mathcal{A}}) \leq 1\}$ represents the unit norm ball and $w(\mathcal{S}) = \mathbb{E}(\sup_{\boldsymbol{\mathcal{B}}\in\mathcal{S}}\langle \boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{T}}\rangle)$ is the Gaussian width of the set $\mathcal{S}$, then there exists a constant $\gamma > 0$ such that with probability at least $1 - \exp\left(-\gamma w^2(\Omega)\right)$, the estimate $\hat{\boldsymbol{\mathcal{A}}}$ satisfies:*

$$\left\|\hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^*\right\|_{\mathrm{F}} \lesssim \lambda\kappa\Psi(\boldsymbol{\mathcal{A}}),$$

where $\Psi(\boldsymbol{\mathcal{A}}) = \sup_{\mathcal{U}\in\boldsymbol{\mathcal{A}}\setminus\{0\}}\frac{R_\lambda(\mathcal{U})}{\|\mathcal{U}\|_{\mathrm{F}}}$ is the subspace compatibility constant for subspace $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{I_1\times I_2\times\cdots\times I_N}$.

Theorem 5 provides an upper bound on the estimation error between the optimal solution $\hat{\boldsymbol{\mathcal{A}}}$ and the true value $\boldsymbol{\mathcal{A}}^*$. This bound is driven by two terms: the subspace compatibility constant $\Psi(\boldsymbol{\mathcal{A}})$, which captures the intrinsic complexity of the subspace $\boldsymbol{\mathcal{A}}$, and the Gaussian width $w(\Omega)$, while quantifies the size of the unit norm ball Ω under the regularizer $R(\cdot)$. Intuitively, $\Psi(\boldsymbol{\mathcal{A}})$ controls the sensitivity of the regularization to the subspace, while $w(\Omega)$ measures the richness of the subspace in terms of its capacity to fit the data.

Next, we present explicit statistical rates of convergence for different regularizers, providing deeper insights into their effectiveness under differing conditions.

Before conducting a detailed analysis of the convergence rates associated with element-wise sparsity regularizer, we first introduce a valuable theoretical benchmark—the oracle rate. The oracle rate refers to the statistical convergence rate of the oracle estimator, which serves as an idealized benchmark by assuming that the true parameter support set $\mathcal{S}^*$ is known a prior. This assumption allows the oracle estimator to achieve the best theoretical performance. Specifically, the element-wise sparse oracle estimator is defined as $\hat{\boldsymbol{\mathcal{A}}}^O = \arg\min_{\boldsymbol{\mathcal{A}}:\boldsymbol{\mathcal{A}}_{\overline{\mathcal{S}_1}}=\mathbf{0}} L(\boldsymbol{\mathcal{A}})$, where $\overline{\mathcal{S}_1}$ denotes the complement of the support set $\mathcal{S}_1$, i.e., $\mathcal{S}_1 = \left\{(i_1,\ldots,i_M) \mid \boldsymbol{\mathcal{A}}^*_{i_1,\ldots,i_M} \neq 0\right\}$. Now, we have the following result.

Corollary 1 (Element-wise sparsity)**.** *Suppose that Assumptions 1∼4 hold. There exists a constant* $0 < \gamma_1 < \infty$ *such that the Gaussian width satisfies* $w(\Omega) \leq \gamma_1$*. If*

$$\rho^- > \zeta^-, \quad \lambda \asymp \sqrt{\frac{\log(d_1 d_2 \cdots d_N)}{n}},$$

and $\boldsymbol{\mathcal{A}}^*_{i_1,\ldots,i_M}$ *satisfies the condition*

$$\min_{(i_1,\ldots,i_M)\in\mathcal{S}_1} \left|\boldsymbol{\mathcal{A}}^*_{i_1,\ldots,i_M}\right| \geq \nu,$$

the estimator $\hat{\boldsymbol{\mathcal{A}}}$ *satisfies*

$$\left\|\hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^*\right\|_{\mathrm{F}} \lesssim \sqrt{\frac{|\mathcal{S}_1|}{n}}.$$

A similar argument applies to fiber-wise sparsity. The oracle rate refers to the statistical convergence rate of the fiber-wise sparse oracle estimator, which is assumed to know the true support set $\mathcal{S}_2$ in advance, where $\mathcal{S}_2 = \left\{i \mid \left\|\left[\boldsymbol{\mathcal{A}}_{(k)}\right]_{\cdot,i}\right\|_2 \neq 0\right\}$. Specifically, the fiber-wise sparse oracle estimator is defined as $\hat{\boldsymbol{\mathcal{A}}}^O = \arg\min_{\boldsymbol{\mathcal{A}}:\boldsymbol{\mathcal{A}}_{\overline{\mathcal{S}_2}}=\mathbf{0}} L(\boldsymbol{\mathcal{A}})$.

Corollary 2 (Fiber-wise sparsity)**.** *Suppose that Assumptions 1∼4 hold. There exists a constant* $0 < \gamma_2 < \infty$ *such that the Gaussian width satisfies* $w(\Omega) \leq \gamma_2\sqrt{d_k}$*. If*

$$\rho^- > \zeta^-, \quad \lambda \asymp \sqrt{\frac{d_k}{n}},$$

and $\left[\boldsymbol{\mathcal{A}}_{(k)}\right]_{\cdot,l}$ *satisfies the condition*

$$\min_{l\in\mathcal{S}_2} \left|\left\|\left[\boldsymbol{\mathcal{A}}_{(k)}\right]_{\cdot,l}\right\|_2\right| \geq \nu,$$

the estimator $\hat{\boldsymbol{\mathcal{A}}}$ *satisfies*

$$\left\|\hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^*\right\|_{\mathrm{F}} \lesssim \sqrt{\frac{|\mathcal{S}_2|\, d_k}{n}}.$$

Similarly, for slice-wise sparsity, the oracle rate refers to the statistical convergence rate of the slice-wise sparse oracle estimator, which is assumed to know the true support set $\mathcal{S}_3$ in advance, where $\mathcal{S}_3 = \left\{l \mid \left\|\left[\boldsymbol{\mathcal{A}}_{(j,k)}\right]_{\cdot,\cdot,l}\right\|_{\mathrm{F}} \neq 0\right\}$. Specifically, the slice-wise sparse oracle estimator is defined as $\hat{\boldsymbol{\mathcal{A}}}^O = \arg\min_{\boldsymbol{\mathcal{A}}:\boldsymbol{\mathcal{A}}_{\overline{\mathcal{S}_3}}=\mathbf{0}} L(\boldsymbol{\mathcal{A}})$.

Corollary 3 (Slice-wise sparsity)**.** *Suppose that Assumptions 1∼4 hold. There exists a constant* $0 < \gamma_3 < \infty$ *such that the Gaussian width satisfies* $w(\Omega) \leq \gamma_3\sqrt{d_j d_k}$*. If*

$$\rho^- > \zeta^-, \quad \lambda \asymp \sqrt{\frac{d_j d_k}{n}},$$

and $\left[\boldsymbol{\mathcal{A}}_{(j,k)}\right]_{\cdot,\cdot,l}$ *satisfies the condition*

$$\min_{l\in\mathcal{S}_3} \left|\left\|\left[\boldsymbol{\mathcal{A}}_{(j,k)}\right]_{\cdot,\cdot,l}\right\|_{\mathrm{F}}\right| \geq \nu,$$

the estimator $\hat{\boldsymbol{\mathcal{A}}}$ *satisfies*

$$\left\|\hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^*\right\|_{\mathrm{F}} \lesssim \sqrt{\frac{|\mathcal{S}_3|\, d_j d_k}{n}}.$$

In the following, we first introduce the singular value decomposition (SVD). Consider the matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$, the SVD is given by $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{m\times r}$ contains the left singular vectors, $\mathbf{V} \in \mathbb{R}^{n\times r}$ contains the right singular vectors, and $\boldsymbol{\Sigma} = \mathrm{Diag}(\sigma_1(\mathbf{X}),\ldots,\sigma_r(\mathbf{X})) \in \mathbb{R}^{r\times r}$ is a diagonal matrix. We further define a subspace $\mathcal{F}$ along with its orthogonal complement $\mathcal{F}^\perp$ as follows:

$$\mathcal{F}(\mathbf{X}) = \{\mathbf{W} \mid \mathrm{row}(\mathbf{W}) \subseteq \mathbf{V}, \mathrm{col}(\mathbf{W}) \subseteq \mathbf{U}\},$$
$$\mathcal{F}^\perp(\mathbf{X}) = \{\mathbf{W} \mid \mathrm{row}(\mathbf{W}) \perp \mathbf{V}, \mathrm{col}(\mathbf{W}) \perp \mathbf{U}\}.$$

For brevity, we adopt shorthand notations $\mathcal{F}$ and $\mathcal{F}^\perp$ when the dependence on $\mathbf{X}$ is clear from context. We denote by $\Pi_{\mathcal{F}}(\cdot)$ and $\Pi_{\mathcal{F}^\perp}(\cdot)$ the projection operators onto the subspace $\mathcal{F}$ and $\mathcal{F}^\perp$, respectively. Additionally, we introduce the linear operator $\mathfrak{X}(\boldsymbol{\mathcal{A}}) : \mathbb{R}^{d_1\times\cdots\times d_N} \to \mathbb{R}^{n\times d_{N+1}\times\cdots\times d_M}$ defined as $\mathfrak{X}(\boldsymbol{\mathcal{A}}) = \left(\left\langle\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{X}}^{(1)}\right\rangle,\ldots,\left\langle\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{X}}^{(n)}\right\rangle\right)^\top$ along with its adjoint operator $\mathfrak{X}^*(\boldsymbol{\mathcal{E}})$. With these fundamental definitions, we present conclusions regarding tensor regression with low-rankness regularizers.

The oracle rate refers to the statistical convergence rate of the mode-wise low-rank oracle estimator, which

is assumed to know the true rank subspace $\mathcal{S}_4 = \left\{ i \mid \sigma_i \left(\mathcal{F} \left(\boldsymbol{\mathcal{A}}^*_{(k)} \right) \right) \neq 0 \right\}$ in advance. Specifically, the tensor nuclear norm oracle estimator is defined as $\hat{\boldsymbol{\mathcal{A}}}^O = \arg\min_{\boldsymbol{\mathcal{A}}: \boldsymbol{\mathcal{A}}_{\overline{\mathcal{S}_4}} = \mathbf{0}} L(\boldsymbol{\mathcal{A}})$.

Corollary 4 (Mode-wise low-rankness)**.** *Suppose that Assumptions 1∼4 hold. There exists a constant* $0 < \gamma_4 < \infty$ *such that the Gaussian width satisfies* $w(\Omega) \leq \frac{\gamma_3}{\rho^- \sqrt{n}} \tau_k$, *where* $\tau_k = \left\| \Pi_{\mathcal{F}} \left([\mathfrak{X}^*(\boldsymbol{\mathcal{E}})]_{(k)} \right) \right\|_{\text{sp}}$. *If*

$$\rho^- > \zeta^-, \quad \lambda \gtrsim \frac{1}{n} \sqrt{|\mathcal{S}_4|} \left\| [\mathfrak{X}^*(\boldsymbol{\mathcal{E}})]_{(k)} \right\|_{\text{sp}},$$

and $\sigma_i \left(\boldsymbol{\mathcal{A}}^*_{(k)} \right)$ *satisfies the condition*

$$\left| \sigma_i \left(\boldsymbol{\mathcal{A}}^*_{(k)} \right) \right| \geq \nu + \frac{2\sqrt{|\mathcal{S}_4|}}{n\rho^-} \left\| [\mathfrak{X}^*(\boldsymbol{\mathcal{E}})]_{(k)} \right\|_{\text{sp}},$$

the estimator $\hat{\boldsymbol{\mathcal{A}}}$ *satisfinumer*

$$\left\| \hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^* \right\|_{\text{F}} \lesssim \frac{\tau_k \sqrt{|\mathcal{S}_4|}}{n}.$$

The oracle rate refers to the statistical convergence rate of the slice-wise low-rank oracle estimator, which is assumed to know the true rank subspace $\mathcal{S}_5 = \left\{ s \mid \sigma_s \left(\mathcal{F} \left(\left[\boldsymbol{\mathcal{A}}^*_{(j,k)} \right]_{\cdot,\cdot,l} \right) \right) \neq 0 \right\}$ in advance. Specifically, the slice-wise low-rank oracle estimator is defined as $\hat{\boldsymbol{\mathcal{A}}}^O = \arg\min_{\boldsymbol{\mathcal{A}}: \boldsymbol{\mathcal{A}}_{\overline{\mathcal{S}_5}} = \mathbf{0}} L(\boldsymbol{\mathcal{A}})$.

Corollary 5 (Slice-wise low-rankness)**.** *Suppose that Assumptions 1∼4 hold. There exists a constant* $0 < \gamma_5 < \infty$ *such that the Gaussian width satisfies* $w(\Omega) \leq \frac{\gamma_5}{\rho^- \sqrt{n}} \tau_{(j,k)}$, *where* $\tau_{(j,k)} = \max_l \left\| \Pi_{\mathcal{F}} \left([\mathfrak{X}^*(\boldsymbol{\mathcal{E}})]_{(j,k)} \right)_{\cdot,\cdot,l} \right\|_{\text{sp}}$. *If*

$$\rho^- > \zeta^-, \quad \lambda \gtrsim \frac{1}{n} \sqrt{|\mathcal{S}_5|} \left\| \left[[\mathfrak{X}^*(\boldsymbol{\mathcal{E}})]_{(j,k)} \right]_{\cdot,\cdot,l} \right\|_{\text{sp}},$$

and $\sigma_s \left(\left[\boldsymbol{\mathcal{A}}^*_{(j,k)} \right]_{\cdot,\cdot,l} \right)$ *satisfies the condition*

$$\left| \sigma_s \left(\left[\boldsymbol{\mathcal{A}}^*_{(j,k)} \right]_{\cdot,\cdot,l} \right) \right| \geq \nu + \frac{2\sqrt{|\mathcal{S}_5|}}{n\rho^-} \left\| \left[[\mathfrak{X}^*(\boldsymbol{\mathcal{E}})]_{(j,k)} \right]_{\cdot,\cdot,l} \right\|_{\text{sp}},$$

the estimator $\hat{\boldsymbol{\mathcal{A}}}$ *satisfies*

$$\left\| \hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^* \right\|_{\text{F}} \lesssim \frac{\tau_{(j,k)} \sqrt{|\mathcal{S}_5|}}{n}.$$

Corollaries 1, 2, 3, 4, and 5 are direct consequences of Theorem 5. Furthermore, it is easy to see that the proposed estimators achieve a faster statistical rate of convergence—matching the oracle rate—compared to existing estimators that employ an ℓ_1 penalty (Raskutti et al., 2019) when evaluated under the Frobenius norm.

Algorithm 1: Proximal Gradient Algorithm

Require: $\eta \in (0, \frac{1}{l_p}), \delta \in (0, \frac{1}{\eta} - l_p)$;

1 $\boldsymbol{\mathcal{A}}_0 = \boldsymbol{\mathcal{A}}_0 = \mathbf{0}$;
2 **for** $t = 1, \ldots, T$ **do**
3 $\quad \boldsymbol{\mathcal{Y}}_t = \boldsymbol{\mathcal{A}}_t + \frac{t-1}{t+2}(\boldsymbol{\mathcal{A}}_t - \boldsymbol{\mathcal{A}}_{t-1})$;
4 $\quad \Delta_t = \max_{s=\max(1,t-q),\ldots,t} H(\boldsymbol{\mathcal{A}}_t)$;
5 $\quad$ **if** $H(\boldsymbol{\mathcal{Y}}_t) \leq \Delta_t$ **then**
6 $\quad\quad \mathcal{V}_t = \boldsymbol{\mathcal{Y}}_t$;
7 $\quad$ **else**
8 $\quad\quad \mathcal{V}_t = \boldsymbol{\mathcal{A}}_t$;
9 $\quad$ **end**
10 $\quad \mathcal{Z}_t = \mathcal{V}_t - \eta \nabla L(\mathcal{V}_t)$;
11 $\quad \boldsymbol{\mathcal{A}}_t = \text{prox}_{\eta \mathcal{R}_\lambda}(\mathcal{Z}_t)$
12 **end**

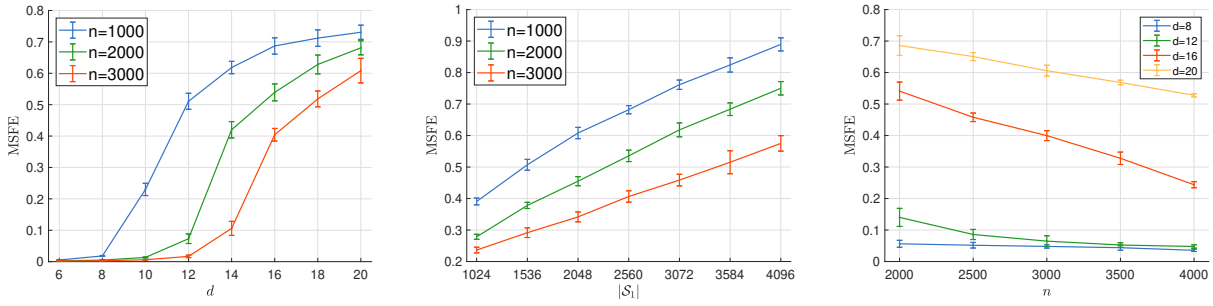
Output: $\boldsymbol{\mathcal{A}}_{T+1}$

4 OPTIMIZATION ALGORITHM

In this section, we present a proximal gradient homotopy algorithm, adapted from (Yao et al., 2017), to solve the proposed estimators (2). The detailed steps are provided in Algorithm 1. We define $H(\boldsymbol{\mathcal{A}}) = L(\boldsymbol{\mathcal{A}}) + R_\lambda(\boldsymbol{\mathcal{A}})$ for clarity and the proximal operator as $\text{prox}_{\eta R_\lambda}(v) = \arg\min_x \left\{ \frac{1}{2} \|x - v\|_2^2 + \eta R_\lambda(x) \right\}$. At each iteration t,the accelerated proximal gradient algorithm updates x_{t+1} through the following steps:

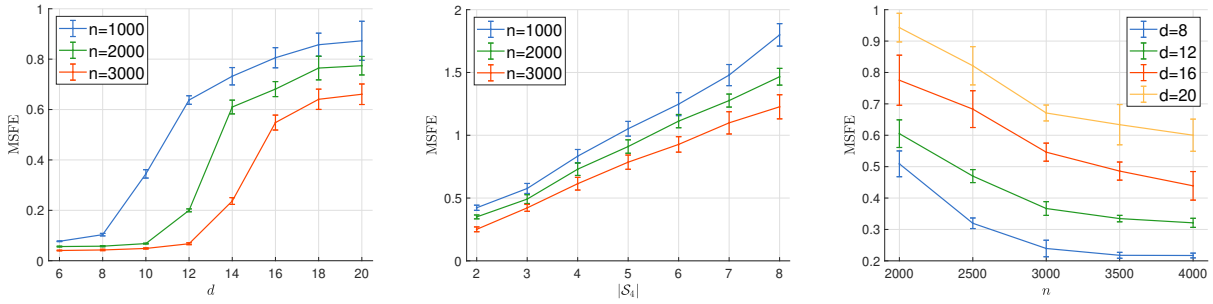
$$y_t = x_t + \theta_t (x_t - x_{t-1}),$$
$$x_{t+1} = \text{prox}_{\eta R_\lambda} (y_t - \eta \nabla L(y_t)),$$

where $\theta_t = \frac{t-1}{t+2}$. When $\theta_t = 0$, the accelerated proximal gradient reduces to the standard proximal gradient algorithm. However, when extending to nonconvex problems, a significant challenge arises due to the extrapolation term $\boldsymbol{\mathcal{Y}}_t$. The presence of $\boldsymbol{\mathcal{Y}}_t$ prevents guaranteeing a sufficient decrease of the objective function in each iteration. To address this issue, Yao et al. (2017) proposed to evaluate the objective value before the proximal step, instead of checking after it as was previously done, demonstrating that this modification guarantees a sufficient decrease under certain conditions. Moreover, to effectively implement the Algorithm 1, we need to choose an appropriate step size η. The parameter η is related to the Lipschitz constant l_p of $L(\cdot)$ (i.e., $\|\nabla L(\boldsymbol{\mathcal{X}}) - \nabla L(\boldsymbol{\mathcal{Y}})\|_{\text{F}} \leq l_p \|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{Y}}\|_{\text{F}}$). Choosing $\eta \leq \frac{1}{l_p}$ ensures that the gradient step does not overshoot, which is critical for the convergence of the algorithm. In line 4, q is an adjustable parameter. Setting $q = 0$ represents the simplest approach, while a larger q permits the objective function $H(\boldsymbol{\mathcal{Y}}_t)$ to occasionally increase. This flexibility is inspired by the Barzilai-Borwein scheme for unconstrained smooth minimization Grippo and Sciandrone (2002). We adopt $q = 5$ in our experiments.



(a) (b) (c)

Figure 1: Element-wise sparsity regularizer with the error bars of MSFE ± standard deviation.



(a) (b) (c)

Figure 2: Mode-wise lowrankness regularizer with the error bars of MSFE ± standard deviation.

5 NUMERICAL EXPERIMENTS

In this section, we perform extensive numerical experiments to evaluate the performance of the proposed tensor regression models with various regularization techniques. For each regularizer, we take the SCAD penalty as the non-convex penalty. The tuning parameter λ and the hyperparameter associated with the SCAD penalty are selected via ten-fold cross-validation, aiming to minimize the estimation error quantified by the Mean Squared Frobenius norm Error (MSFE). Specifically, the MSFE is defined as $\frac{\|\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}\|_{\mathrm{F}}}{\prod_{i=1}^{N} d_i}$, where $\boldsymbol{\mathcal{A}}^*$ is the true parameter tensor and $\hat{\boldsymbol{\mathcal{A}}}$ is the estimated one. All experiments are implemented in MATLAB, and the reported results are averaged on 100 Monte Carlo realizations to ensure statistical robustness.

5.1 Synthetic Data

In synthetic data experiments, the covariate tensors $\boldsymbol{\mathcal{X}}^{(i)}$ are independently drawn from standard Gaussian ensembles. Additionally, the i.i.d. Gaussian noise terms are generated from the distribution $\mathbb{N}\left(0, \eta^2\right)$. Beyond MSFE, we also employ the Root Mean Square Error (RMSE) between the predicted response and the true response. Specifically, the RMSE is calculated as $\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{\mathcal{Y}}^{(i)} - \left\langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{X}}^{(i)} \right\rangle\right\|_{\mathrm{F}}^2}$.

In the subsequent experiments, we employ 3rd-order tensors $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{d\times d\times d}$ to evaluate the effectiveness of entry-wise sparsity, fiber-wise sparsity and mode-wise low-tensor-rankness regularizers. To better show the performance of slice-wise sparsity and low-rankness regularizers, we consider tensors $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{d\times d\times s}$, where the number of slices s is fixed.

Fig. 1 illustrates the performance of the estimation with the element-wise sparsity regularizer with the variation of dimensions d, element-wise sparsity s^* and the number of samples n, respectively. In Fig. 1a and Fig. 1b, three lines correspond to sample size $n = \{1000, 2000, 3000\}$. The proportion of non-zero elements $\frac{s^*}{d^3}$ is set to 0.5 in Fig. 1a and Fig. 1c. Fig. 1a demonstrates the estimation error decreases consistently with an increase in sample size n. Conversely, Fig. 1b shows that an increase in the cardinality of the support set leads to an increase in estimation error. The observed trends in Fig. 1 align with Corollary 1, thereby validating our theoretical conclusions.

Table 1: Comparisons between non-convex regularizers and convex regularizers.

Structures	Methods		Synthetic Data					Real-world Data	
			size	$\lvert\mathcal{S}\rvert$	η	MSFE $\pm$ std	RMSE $\pm$ std	MSFE $\pm$ std	MPRE $\pm$ std
Sparsity	Element-wise	Non-convex	$16 \times 16 \times 16$	2048	0.1	0.4042 ± 0.0201	0.0992 ± 0.0021	134.5864 ± 11.2950	8.8072 ± 0.0313
		Convex				0.6938 ± 0.0297	0.1004 ± 0.0023	144.7160 ± 14.9947	7.7498 ± 0.0457
	Fiber-wise	Non-convex	$16 \times 16 \times 16$	8	0.1	0.4406 ± 0.0157	0.0993 ± 0.0012	90.3068 ± 7.3006	4.7161 ± 0.0103
		Convex				0.7512 ± 0.0439	0.0995 ± 0.0019	102.7019 ± 9.2188	5.0330 ± 0.0118
	Slice-wise	Non-convex	$16 \times 16 \times 20$	8	0.1	0.5761 ± 0.0289	0.0997 ± 0.0027	43.8705 ± 3.0257	1.8909 ± 0.0043
		Convex				0.7201 ± 0.0314	0.1005 ± 0.0039	48.4585 ± 3.8834	1.9250 ± 0.0045
Low- rankness	Mode-wise	Non-convex	$16 \times 16 \times 16$	5	1	0.5482 ± 0.0395	0.1002 ± 0.0012	35.5536 ± 1.4889	1.0330 ± 0.0022
		Convex				1.7411 ± 0.0953	0.1096 ± 0.0020	41.2719 ± 3.5079	1.1027 ± 0.0024
	Slice-wise	Non-convex	$16 \times 16 \times 20$	5	1	0.9214 ± 0.0736	0.1004 ± 0.0010	8.9348 ± 0.7493	0.0436 ± 0.0002
		Convex				1.8261 ± 0.1066	0.1113 ± 0.0031	10.1655 ± 0.9050	0.6348 ± 0.0009

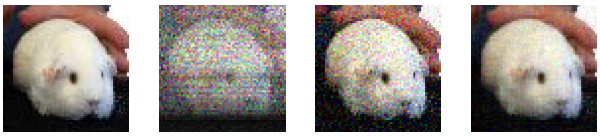
Additionally, we assess the performance of other sparsity regularizers, with detailed results presented in Appendix A.

Fig. 2 illustrates the results of the mode-wise low-rankness regularizer. In Fig. 2b, x-axis r represent the rank of the mode-(k) unfolded matrix. Fig. 2a and Fig. 2c sets the rank of the mode-(k) unfolded matrix to 5, while Fig. 2b and Fig. 2c fix the dimension of each mode to 16. Three distinct lines correspond to the estimation errors for sample size $n = \{1000, 2000, 3000\}$. These experimental results validate our theoretical findings as in Corollary 4. We also report results for the slice-wise low-rank regularizer in Appendix A.

In Table 1, we compare the performance of our proposed non-convex regularizers against traditional convex regularizers. For sparsity regularization, we set the $\eta = 0.1$, and for low-rankness regularizers, we set $\eta = 1$. We configure the tensor dimension such that tensors with slices-wise structures $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{d\times d\times s}$ and the others $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{d\times d\times d}$, where $d = 16$, $s = 20$. Depending on the tensor structure, the sparsity or the rank of the tensors varies accordingly. The results in Table 1 demonstrate that non-convex regularizers achieve lower MSFE for parameter estimation and lower RMSE for predictions compared to their convex counterparts. These empirical findings are in strong agreement with our theoretical analysis.

5.2 Real-world Datasets

To illustrate the effectiveness of the proposed methods, we conduct experiments using the ImageNet 2012 dataset (Russakovsky et al., 2015). In our regression framework, the tensor parameter is a 3rd-order color image, denoted as $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{64\times 64\times 3}$. We utilize $n = 4000$ samples in our experiments. In addition to MSFE, we evaluate performance using the Mean Prediction Relative Error (MPRE), defined as $\frac{1}{n}\sum_{i=1}^{n} \frac{\left\|\boldsymbol{\mathcal{Y}}^* - \hat{\boldsymbol{\mathcal{Y}}}^{(i)}\right\|_{\mathrm{F}}}{\|\boldsymbol{\mathcal{Y}}^*\|_{\mathrm{F}}}$, where $\boldsymbol{\mathcal{Y}}^* = \langle \boldsymbol{\mathcal{A}}^*, \boldsymbol{\mathcal{X}} \rangle$ represents the original tensor response with the RGB image and $\hat{\boldsymbol{\mathcal{Y}}}$ is the estimated response. Fig. 3 illustrates the estimated image, and Table 1 presents the comparative results.



(a) original (b) noisy (c) convex (d) nonconvex (prop.)

Figure 3: The original image, noisy image, and denoised images using convex and non-convex methods.

6 CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we propose a comprehensive framework for tensor regression estimation using non-convex penalties. Our findings demonstrate that estimators employing non-convex penalties exhibit faster convergence rates compared to those with convex penalties. Furthermore, we show that under several mild conditions, our proposed estimator possesses the oracle property. Extensive experimental results validate our theoretical claims, showcasing a close alignment between the theoretical predictions and the observed numerical performance of our estimators. Currently, we are limited to applying regularization penalties to tensor regression models. It would be desirable to derive some theoretical guarantees for alternative methods that capture structure in the tensor regression models, such as tensor decomposition; this is the aim of our future work. To conclude, our work effectively bridge the gap between practical applications and theoretical analysis of tensor-on-tensor regression with non-convex penalties. To the best of the our knowledge, this is the first work to obtain the oracle statistical rate of convergence for the tensor regression problem.

References

Evrim Acar, Daniel M Dunlavy, Tamara G Kolda, and Morten Mørup. Scalable tensor factorizations for incomplete data. *Chemometrics and Intelligent Laboratory Systems*, 106(1):41–56, 2011.

Talal Ahmed, Haroon Raja, and Waheed U Bajwa. Tensor regression using low-rank and sparse tucker decompositions. *SIAM Journal on Mathematics of Data Science*, 2(4):944–966, 2020.

Arnab Auddy, Dong Xia, and Ming Yuan. Tensor methods in high dimensional data analysis: Opportunities and challenges. *arXiv preprint arXiv:2405.18412*, 2024.

Emmanuel Candes and Benjamin Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119, 2012.

Emmanuel Candes and Terence Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

Ethan R Elenberg, Rajiv Khanna, Alexandros G Dimakis, and Sahand Negahban. Restricted strong convexity implies weak submodularity. *The Annals of Statistics*, 46(6B):3539–3568, 2018.

Haiyan Fan, Yunjin Chen, Yulan Guo, Hongyan Zhang, and Gangyao Kuang. Hyperspectral image restoration using low-rank tensor recovery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(10):4589–4604, 2017.

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

Jianqing Fan, Han Liu, Qiang Sun, and Tong Zhang. I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *The Annals of Statistics*, 46(2):814, 2018.

Vivek Farias and Andrew Li. Optimal recovery of tensor slices. In *Artificial Intelligence and Statistics*, pages 1394–1402. PMLR, 2017.

Yehoram Gordon. On milman's inequality and random subspaces which escape through a mesh in $\mathbf{r}^n$. In *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 1986–87*, pages 84–106. Springer, 1988.

Luigi Grippo and Marco Sciandrone. Nonmonotone globalization techniques for the Barzilai-Borwein gradient method. *Computational Optimization and Applications*, 23:143–169, 2002.

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

Victoria Hore, Ana Vinuela, Alfonso Buil, Julian Knight, Mark I McCarthy, Kerrin Small, and Jonathan Marchini. Tensor decomposition for multiple-tissue gene expression experiments. *Nature genetics*, 48(9):1094–1100, 2016.

Lexin Li Hua Zhou and Hongtu Zhu. Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108 (502):540–552, 2013.

Bing Li Huihui Miao, Andi Wang and Jianjun Shi. Structural tensor-on-tensor regression with interaction effects and its application to a hot rolling process. *Journal of Quality Technology*, 54(5):547–560, 2022.

Henk Kiers. Towards a standardized notation and terminology in multiway analysis. *Journal of Chemometrics*, 14:105–122, 05 2000.

Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM Review*, 51(3): 455–500, 2009.

Yanming Li, Bin Nan, and Ji Zhu. Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics*, 71(2):354–363, 2015.

Yipeng Liu, Jiani Liu, Zhen Long, and Ce Zhu. *Tensor regression*. Springer, 2022.

Eric F. Lock. Tensor-on-tensor regression. *Journal of Computational and Graphical Statistics*, 27(3):638–647, June 2018. ISSN 1537-2715.

Yuetian Luo and Anru R Zhang. Tensor-on-tensor regression: Riemannian optimization, over-parameterization, statistical-computational gap, and their interplay. *arXiv preprint arXiv:2206.08756*, 2022.

Pascal Massart. *Concentration inequalities and model selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*. Springer, 2007.

Peter McCullagh. *Tensor methods in statistics: Monographs on statistics and applied probability*. Chapman and Hall/CRC, 2018.

Sahand Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers. *Statistical Science*, 27 (4):538 – 557, 2012.

Garvesh Raskutti, Ming Yuan, and Han Chen. Convex regularization for high-dimensional multiresponse tensor regression. *The Annals of Statistics*, 47(3): 1554–1584, 2019.

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

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

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

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

C F Westin, Stephan E Maier, Besam Khidhir, Peter Everett, Ferenc A Jolesz, and Ron Kikinis. Image processing for diffusion tensor magnetic resonance imaging. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI'99: Second International Conference, Cambridge, UK, September 19-22, 1999. Proceedings 2*, pages 441–452. Springer, 1999.

Da Xu. Sparse symmetric tensor regression for functional connectivity analysis. *arXiv preprint arXiv:2010.14700*, 2020.

Yongxin Yang and Timothy Hospedales. Deep multitask representation learning: A tensor factorisation approach. In *5th International Conference on Learning Representations*, 2017.

Quanming Yao, James T Kwok, Fei Gao, Wei Chen, and Tie-Yan Liu. Efficient inexact proximal gradient algorithm for nonconvex problems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 3308–3314, 2017.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67, 2006.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, pages 894–942, 2010.

Hao Zhang, Ting-Zhu Huang, Xi-Le Zhao, Wei He, Jae Kyu Choi, and Yu-Bang Zheng. Hyperspectral image denoising: Reconciling sparse and low-tensor-ring-rank priors in the transformed domain. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–13, 2023.

Jiaqi Zhang, Yinghao Cai, Zhaoyang Wang, and Beilun Wang. Sparse and low-rank high-order tensor regression via parallel proximal method. *arXiv preprint arXiv:1911.12965*, 2019.

Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Yes]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

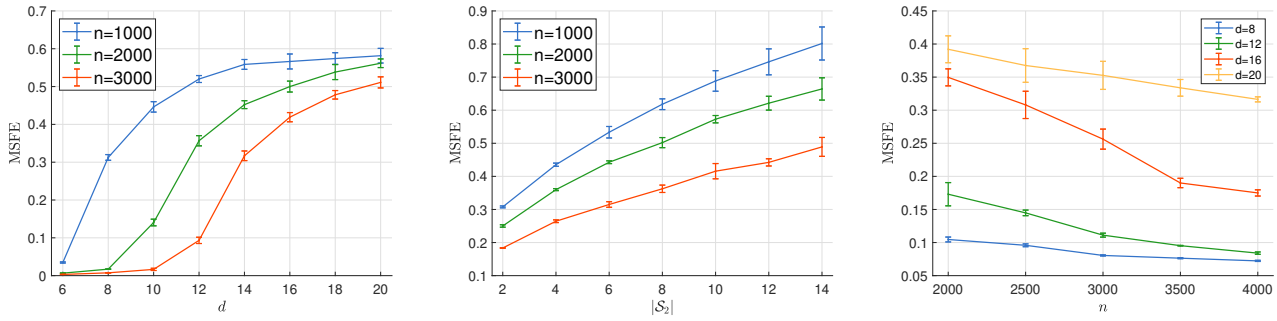
High-Dimensional Tensor Regression With Oracle Properties: Supplementary Materials

Contents

A ADDITIONAL NUMERICAL RESULTS

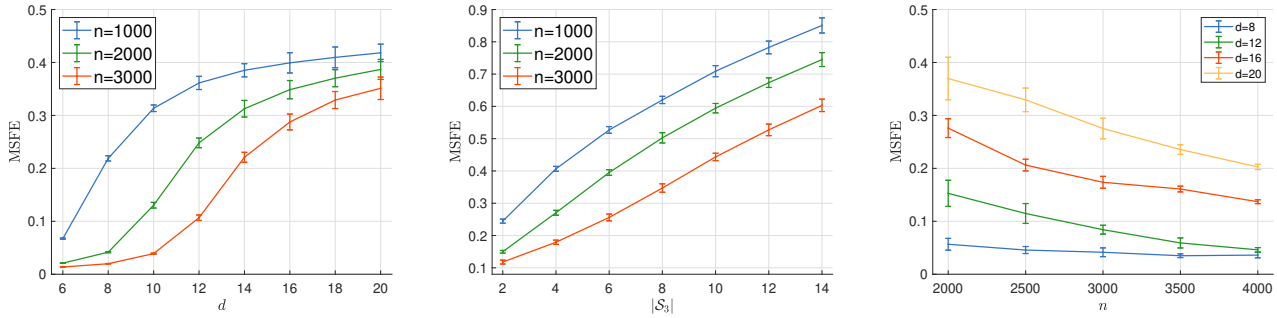
In this section, we provide additional results for the proposed regularizers besides the results in Section 5. In Section A.1, the performance of different regularizers based on synthetic data will be discussed, followed by the analysis of the experimental results. Moreover, the results of the regularizers for the real-world data are illustrated in Section A.2.

A.1 Synthetic Data



(a) (b) (c)

Figure 4: Fiber-wise sparsity regularizer with the error bars of MSFE ± standard deviation .

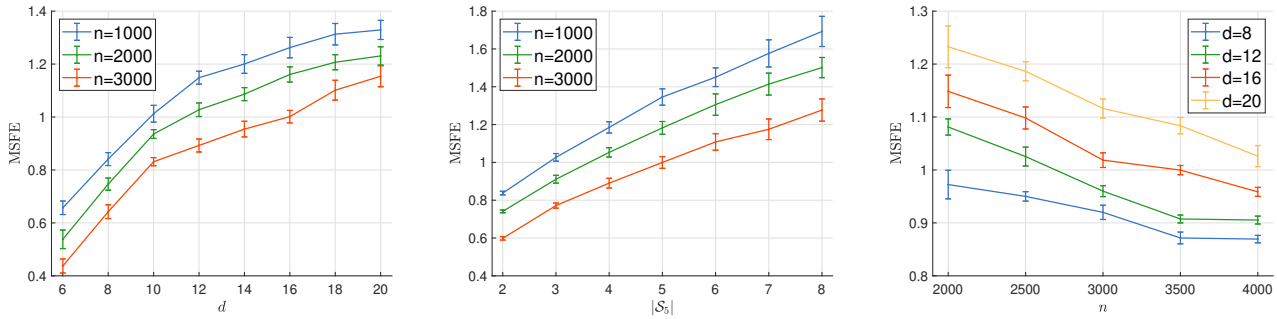


(a) (b) (c)

Figure 5: Slice-wise sparsity regularizer with the error bars of MSFE ± standard deviation .

Figure 4 presents the results related to the fiber-wise sparsity regularizer. We display the results of the Mean Squared Frobenius norm Error (MSFE) versus the dimension d, the fiber-wise sparsity $|\mathcal{S}_2|$ and the sample size n, respectively. In these experiments, all tensors are 3rd-order tensors, with the dimension $d = 16$. In Figure 4a and Figure 4b, we fix the fiber-wise sparsity $|\mathcal{S}_2| = 4$. The three lines in different colors represent varying sample sizes $n = \{1000, 2000, 3000\}$. From Figure 4a, we observe that MSFE increases as the tensor dimension d increases. From Figure 4b, we find that increasing the sample size n decreases the MSFE. This demonstrates that larger sample sizes improve the accuracy of the tensor estimation, as expected. In Figure 4c, we see that increasing the fiber-wise-sparsity $|\mathcal{S}_2|$ leads to an increase in the estimate error. This suggests that tensors with more nonzero fibers are harder to estimate accurately. Furthermore, the standard deviation of the estimation error follows the same trend, increasing with fiber-wise sparsity.

Figure 5 presents the results of the slice-wise sparsity regularizer. In Figure 5a and Figure 5b, the number of slices is uniformly set to $s = 20$. And we set the slice-wise sparsity $|\mathcal{S}_3| = 4$ in Figure 5a and Figure 5c. We select three sample sizes while varying the dimension d or the number of non-zero slices $|\mathcal{S}_3|$. The results indicate that



(a) (b) (c)

Figure 6: Slice-wise lowrankness regularizer with the error bars of MSFE ± standard deviation .

the estimation error increases with increments in d or $|\mathcal{S}_3|$. The standard deviation of the MSFE also rises as the MSFE increases. Furthermore, as observed in Figure 5c, increasing the sample size reduces estimation errors when the dimension is fixed.

Figure 6 demonstrates the results of the slice-wise low-rankness regularizer. In Figure 6b, the x-axis $|\mathcal{S}_5|$ represents the rank of each slice of the tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{d\times d\times d}$. Figure 6a and Figure 6c fix the rank of each slice to 5. The three distinct lines correspond to the estimation errors for sample size $n = \{1000, 2000, 3000\}$. From Figure 6a, we observe that with a fixed rank and sample size, the estimation error increases as the dimension d enlarges. Furthermore, Figure 6b shows that the estimation errors increase with the rank. Figure 6c demonstrates that with more samples, the estimation errors decrease.

Table 2: The Frobenius norm $\left\|\hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^*\right\|_{\mathrm{F}}$ with standard variance changing the noise parameter

Structures	Methods	$d = 10\times 10\times 10$			$d = 20\times 20\times 20$		
		$\eta = 0.1$	$\eta = 1$	$\eta = 5$	$\eta = 0.1$	$\eta = 1$	$\eta = 5$
Sparsity	Element-wise	1.8909 ± 0.2004	1.9021 ± 0.2214	1.9015 ± 0.2084	19.4215 ± 2.2710	18.6899 ± 2.7556	19.2904 ± 2.3523
	Fiber-wise	1.8622 ± 0.2813	1.8461 ± 0.2783	1.8500 ± 0.2431	19.8920 ± 2.6721	19.3542 ± 2.8909	19.7628 ± 2.4745
	Slice-wise	2.0509 ± 0.3510	2.0421 ± 0.3242	1.9927 ± 0.3666	20.0062 ± 2.7153	20.4267 ± 2.7248	19.4231 ± 2.6420
Low-rankness	Mode-wise	2.6311 ± 0.3008	2.6947 ± 0.3254	2.6265 ± 0.3410	24.6129 ± 2.7010	23.8932 ± 2.7108	24.6571 ± 2.5114
	Slice-wise	3.0150 ± 0.3227	3.0045 ± 0.3754	3.0184 ± 0.3365	25.8502 ± 3.7601	25.6691 ± 3.5732	25.9134 ± 3.8282

For Corollaries 4 and 5, the error bounds contain terms involving η. However, the presence of conjugate operators, projection operators, and nuclear norm regularization complicates a direct analysis of these terms. Specifically, from equation (140), we observe that as the noise parameter η increases, the associated error term also increases, resulting in a larger overall error bound. In the table 2, we include synthetic data experiments with varying noise intensity. The results, which demonstrate the effect of changing η on the error bound, are provided in the table.

Table 3: The Frobenius norm $\left\|\hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^*\right\|_{\mathrm{F}}$ with standard variance for higher dimension

Structures	Methods	3-order		4-order		5-order	
		$d = 8$	$d = 16$	$d = 8$	$d = 16$	$d = 8$	$d = 16$
Sparsity	Element-wise	1.0509 ± 0.1004	1.9021 ± 0.2214	7.9015 ± 1.2084	19.4215 ± 2.2710	58.6899 ± 7.7556	192.2904 ± 17.3523
	Fiber-wise	1.0622 ± 0.0813	1.8461 ± 0.2783	7.8500 ± 1.2431	19.8920 ± 2.6721	59.3542 ± 8.2909	190.7628 ± 16.4745
	Slice-wise	1.0909 ± 0.1510	2.0421 ± 0.3242	8.0927 ± 1.3666	20.0062 ± 2.7153	60.4267 ± 8.1248	193.4231 ± 17.6420
Low-rankness	Mode-wise	1.6311 ± 0.3008	2.6947 ± 0.3254	8.6265 ± 1.3410	34.6129 ± 3.4010	63.8932 ± 9.7108	224.6571 ± 21.5114
	Slice-wise	1.8150 ± 0.3227	2.7045 ± 0.3754	9.0184 ± 1.3365	35.8502 ± 3.7601	65.6691 ± 9.5732	245.9134 ± 22.8282

As outlined in the five corollaries of our paper, our theoretical framework is inherently generalizable to tensors of any order. Although the scope of this paper did not include experimental results for higher-order tensors, in table 3 we have conducted supplementary experiments that demonstrate promising outcomes for these cases.

Also, to explore whether the proposed methods performs robustness under a unknown structure, in table 4, we

Table 4: The Frobenius norm $\left\|\hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^*\right\|_{\mathrm{F}}$ with standard variance changing ground data structure of our proposed methods

Structures	Methods	Tensor Data Structures				
		element-sp	fiber-sp	slice-sp	lr-mode	lr-slice
Sparsity	Element-wise	$\mathbf{1.0509 \pm 0.1004}$	1.0680 ± 0.1027	1.1263 ± 0.2046	1.8991 ± 0.4002	1.9367 ± 0.3979
	Fiber-wise	1.0931 ± 0.1421	$\mathbf{1.0622 \pm 0.0813}$	1.1305 ± 0.2488	1.9054 ± 0.3865	1.9274 ± 0.4410
	Slice-wise	1.1014 ± 0.1852	1.1226 ± 0.2200	$\mathbf{1.0909 \pm 0.1510}$	2.0221 ± 0.4518	2.3185 ± 0.4477
Low-rankness	Mode-wise	6.8502 ± 1.2101	6.9333 ± 1.2565	6.9068 ± 1.1987	$\mathbf{1.6311 \pm 0.3008}$	14.9490 ± 1.4555
	Slice-wise	7.1481 ± 1.3061	7.1636 ± 1.2989	7.0701 ± 1.3004	15.5770 ± 1.3435	$\mathbf{1.8150 \pm 0.3227}$

implement experiments on the proposed methods for each tensor structures, and the results are shown in the table.

A.2 Real-world Data

We have chosen several real-world images from the ImageNet 2012 dataset (Russakovsky et al., 2015) besides the image used in Section 5.2. We implement experiments based on different regularizers, revealing the following performance. In Figure 7, we pick four different images from the dataset, and Figure 8 shows these images with noise added. Figure 10 illustrates the images denoised with the proposed non-convex regularizers. As a comparison, the images denoised with convex methods are revealed in Figure 9.

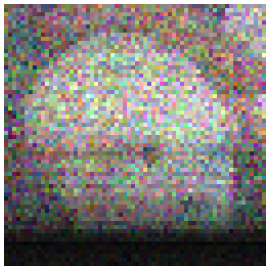
(a) "rabbit"

(b) "human"

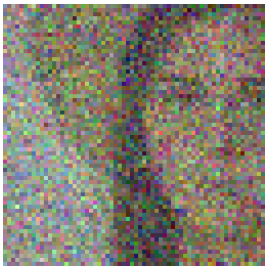
(c) "car"

(d) "scene"

Figure 7: The original images.



(a) noisy "rabbit"



(b) noisy "human"

(c) noisy "car"

(d) noisy "scene"

Figure 8: The noisy images.

We have also implemented additional real-world data experiments with the proposed methods. In table 5, the real data is considered the tensor to be estimate. Regarding the initialization of the covariate tensors $\boldsymbol{\mathcal{A}}$ in the

(a) denoised "rabbit" (cvx. fiber-sp)

(b) denoised "human" (cvx. slice-sp)

(c) denoised "car" (cvx. mode-lr)

(d) denoised "scene" (cvx. slice-lr)

Figure 9: The denoised images using convex methods.

(a) denoised "rabbit" (ncvx. fiber-sp)

(b) denoised "human" (ncvx. slice-sp)

(c) denoised "car" (ncvx. mode-lr)

(d) denoised "scene" (ncvx. slice-lr)

Figure 10: The denoised images using non-convex methods.

Table 5: The MSFE of the climate data(10 years) observation and alcoholic genetic predisposition data with our proposed methods

Dataset	Penalties	Sparsity		Low-rankness		
		Element	Fiber	Slice	Mode	Slice
NA-1990-2002-Monthly	SCAD	7.4556 ± 0.8235	8.0716 ± 0.9456	8.4187 ± 0.9491	11.6809 ± 2.0437	9.9750 ± 1.2203
	MCP	7.6087 ± 1.0003	7.9554 ± 0.9884	8.1305 ± 0.9050	12.6281 ± 2.1882	10.2314 ± 2.0015
	Convex	8.2502 ± 1.4887	8.7425 ± 1.7264	9.4577 ± 1.3004	14.5808 ± 3.3435	13.4508 ± 2.6688
EEG Database	SCAD	12.6865 ± 2.3544	13.8878 ± 2.2412	14.5640 ± 2.4898	18.7983 ± 5.0977	16.3202 ± 4.8331
	MCP	13.0024 ± 1.9973	14.0368 ± 2.0241	16.0431 ± 3.9314	18.4546 ± 4.8020	16.4002 ± 4.7771
	Convex	13.8001 ± 2.6764	14.5716 ± 2.1379	16.8890 ± 4.3051	19.6404 ± 5.3317	18.5783 ± 5.3854

real-data experiments, the number of covariate tensors $\boldsymbol{\mathcal{A}}$ corresponds to the sample size $n = 5000$, and the noise term $\boldsymbol{\mathcal{E}}$ are drawn independently from a Gaussian distribution with mean 0 and variance equal to $\eta = 0.01$.

The experimental data employed in this study were sourced from the University of Southern California's Viterbi School of Engineering repository and the UCI Machine Learning Repository's EEG Database. Specifically, the datasets can be accessed via the following links: https://archive.ics.uci.edu/dataset/121/eeg+database, https://viterbi-web.usc.edu/~liu32/data.html.

B Background

This section provides a concise overview of the foundational concepts essential for understanding the subsequent proof.

Definition 1 (Gaussian Width)**.** Let $\mathcal{S} \subset \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_M}$ be a subset of the tensor space. The Gaussian width of $\mathcal{S}$ is defined as

$$w(\mathcal{S}) = \mathbb{E}\left(\sup_{\boldsymbol{\mathcal{B}} \in \mathcal{S}} \langle \boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{T}} \rangle\right),$$

where $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_M}$ is a tensor whose entry $\mathcal{T}(i_1, \ldots, i_M)$ are independent standard Gaussian random variable. We refer readers to (Gordon, 1988; Negahban et al., 2012) for a comprehensive review.

Definition 2 (Subspace Compatibility Constant)**.** For any subspace $\mathcal{M}$ of $\mathbb{R}^{d_1 \times d_2 \times \cdots \times d_M}$, the subspace compatibility constant $\Psi(\mathcal{M})$ is given by

$$\Psi(\mathcal{M}) = \sup_{\boldsymbol{\mathcal{U}} \in \mathcal{M} \setminus \{0\}} \frac{R(\boldsymbol{\mathcal{U}})}{\|\boldsymbol{\mathcal{U}}\|_{\mathrm{F}}}.$$

This constant quantifies the degree of compatibility between the regularizer $R(\boldsymbol{\mathcal{U}})$ and the Frobenius norm within the subspace subspace $\mathcal{M}$. It can be interpreted as a measure of the intrinsic dimensionality of $\mathcal{M}$.

Singular Value Decomposition. Let $\mathbf{X} \in \mathbb{R}^{m \times n}$ be a matrix with singular value decomposition (SVD) given by $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$, where

- $\mathbf{U} \in \mathbb{R}^{m \times r}$ contains the left singular vectors,
- $\mathbf{V} \in \mathbb{R}^{n \times r}$ contains the right singular vectors,
- $\boldsymbol{\Sigma} = \operatorname{diag}(\boldsymbol{\sigma}) \in \mathbb{R}^{r \times r}$ is a diagonal matrix with singular values $\boldsymbol{\sigma} = [\sigma_1(\mathbf{X}), \ldots, \sigma_r(\mathbf{X})]$,
- $r = \operatorname{rank}(\mathbf{X})$.

We define the subspaces $\mathcal{F}$ and its orthogonal complement $\mathcal{F}^\perp$ as follows:

$$\begin{aligned}
\mathcal{F}(\mathbf{X}) &= \{\mathbf{W} \mid \operatorname{row}(\mathbf{W}) \subseteq \operatorname{span}(\mathbf{V}), \operatorname{col}(\mathbf{W}) \subseteq \operatorname{span}(\mathbf{U})\}, \\
\mathcal{F}^\perp(\mathbf{X}) &= \{\mathbf{W} \mid \operatorname{row}(\mathbf{W}) \perp \operatorname{span}(\mathbf{V}), \operatorname{col}(\mathbf{W}) \perp \operatorname{span}(\mathbf{U})\}.
\end{aligned}$$

Here, $\operatorname{row}(\mathbf{W})$ and $\operatorname{col}(\mathbf{W})$ denote the row space and column space of $\mathbf{W}$, respectively, $\operatorname{span}(\mathbf{V})$ denotes the subspace spanned by $\mathbf{V}$, and $\perp$ denotes orthogonality with respect to the Euclidean inner product. For brevity, we will use the shorthand notations of $\mathcal{F}$ and $\mathcal{F}^\perp$ when the dependence on $\mathbf{X}$ is clear from the context.

The projection operators onto the subspace $\mathcal{F}$ and $\mathcal{F}^\perp$ are denoted by $\Pi_{\mathcal{F}}(\cdot)$ and $\Pi_{\mathcal{F}^\perp}(\cdot)$, respectively, and are defined as:

$$\begin{aligned}
\Pi_{\mathcal{F}}(\mathbf{X}) &= \mathbf{U}\mathbf{U}^\top\mathbf{X}\mathbf{V}\mathbf{V}^\top, \\
\Pi_{\mathcal{F}^\perp}(\mathbf{X}) &= \left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top\right)\mathbf{X}\left(\mathbf{I} - \mathbf{V}\mathbf{V}^\top\right),
\end{aligned}$$

where $\mathbf{I}$ denotes the identity matrix with contextually appropriate dimensions.

C Proof of the Corollary 1

We begin by demonstrating that the element-wise sparsity penalty can be reformulated as the sum of the ℓ_1 penalty and a concave part. Specifically, we have

$$R_\lambda(\boldsymbol{\mathcal{A}}) = \sum_{i_1=1}^{d_1} \cdots \sum_{i_M=1}^{d_M} p_\lambda(\boldsymbol{\mathcal{A}}_{i_1,\ldots,i_M}) = \lambda \|\boldsymbol{\mathcal{A}}_{i_1,\ldots,i_M}\|_1 + Q_\lambda(\boldsymbol{\mathcal{A}}),$$

where $\|\boldsymbol{\mathcal{A}}_{i_1,\ldots,i_M}\|_1 = \sum_{i_1=1}^{d_1} \cdots \sum_{i_M=1}^{d_M} |\boldsymbol{\mathcal{A}}_{i_1,\ldots,i_M}|$ and $Q_\lambda(\boldsymbol{\mathcal{A}}) = \sum_{i_1=1}^{d_1} \cdots \sum_{i_M=1}^{d_M} q_\lambda(\boldsymbol{\mathcal{A}}_{i_1,\ldots,i_M})$. Define $\widetilde{L}(\boldsymbol{\mathcal{A}}) = L(\boldsymbol{\mathcal{A}}) + Q_\lambda(\boldsymbol{\mathcal{A}})$.

Lemma 1. *Under Assumptions 1 and 2, the loss function $\widetilde{L}(\boldsymbol{\mathcal{A}})$ satisfies the restricted strong convexity*

$$\widetilde{L}(\boldsymbol{\mathcal{A}}') \geq \widetilde{L}(\boldsymbol{\mathcal{A}}) + \left\langle \nabla \widetilde{L}(\boldsymbol{\mathcal{A}}), \boldsymbol{\mathcal{A}}' - \boldsymbol{\mathcal{A}} \right\rangle + \frac{\rho^- - \zeta^-}{2} \left\| \boldsymbol{\mathcal{A}}' - \boldsymbol{\mathcal{A}} \right\|_{\mathrm{F}}^2,$$

and the restricted strong smoothness

$$\widetilde{L}(\boldsymbol{\mathcal{A}}') \leq \widetilde{L}(\boldsymbol{\mathcal{A}}) + \left\langle \nabla \widetilde{L}(\boldsymbol{\mathcal{A}}), \boldsymbol{\mathcal{A}}' - \boldsymbol{\mathcal{A}} \right\rangle + \frac{\rho^+ - \zeta^+}{2} \left\| \boldsymbol{\mathcal{A}}' - \boldsymbol{\mathcal{A}} \right\|_{\mathrm{F}}^2.$$

Proof. Recall that $Q_\lambda(\boldsymbol{\mathcal{A}})$ represents the concave component of the non-convex penalty $R_\lambda(\boldsymbol{\mathcal{A}}_{i_1,\ldots,i_N})$, implying that $-Q_\lambda(\boldsymbol{\mathcal{A}})$ is convex. Specifically, $Q_\lambda(\boldsymbol{\mathcal{A}})$ can be expressed as a sum over its elements $Q_\lambda(\boldsymbol{\mathcal{A}}) = \sum_{i_1=1}^{d_1} \cdots \sum_{i_N=1}^{d_N} q_\lambda(\boldsymbol{\mathcal{A}}_{i_1,\ldots,i_N})$, where $q_\lambda(\boldsymbol{\mathcal{A}}_{i_1,\ldots,i_N})$ satisfies the third regularity condition specified in Assumption 3. From this assumption, we have

$$-\zeta^- \left(\boldsymbol{\mathcal{A}}'_{i_1,\ldots,i_N} - \boldsymbol{\mathcal{A}}_{i_1,\ldots,i_N}\right)^2 \leq \left(q'_\lambda\left(\boldsymbol{\mathcal{A}}'_{i_1,\ldots,i_N}\right) - q'_\lambda\left(\boldsymbol{\mathcal{A}}_{i_1,\ldots,i_N}\right)\right)\left(\boldsymbol{\mathcal{A}}'_{i_1,\ldots,i_N} - \boldsymbol{\mathcal{A}}_{i_1,\ldots,i_N}\right) \leq -\zeta^+ \left(\boldsymbol{\mathcal{A}}'_{i_1,\ldots,i_N} - \boldsymbol{\mathcal{A}}_{i_1,\ldots,i_N}\right)^2. \tag{3}$$

By aggregating over all elements, we deduce that the convex function $-Q_\lambda(\boldsymbol{\mathcal{A}})$ satisfies the following inequality

$$\left\langle \left(\nabla\left(-Q_\lambda(\boldsymbol{\mathcal{A}}')\right) - \nabla\left(-Q_\lambda(\boldsymbol{\mathcal{A}})\right)\right)^\top, \boldsymbol{\mathcal{A}}' - \boldsymbol{\mathcal{A}} \right\rangle \leq \zeta^- \left\| \boldsymbol{\mathcal{A}}' - \boldsymbol{\mathcal{A}} \right\|_{\mathrm{F}}^2, \tag{4}$$

$$\left\langle \left(\nabla\left(-Q_\lambda(\boldsymbol{\mathcal{A}}')\right) - \nabla\left(-Q_\lambda(\boldsymbol{\mathcal{A}})\right)\right)^\top, \boldsymbol{\mathcal{A}}' - \boldsymbol{\mathcal{A}} \right\rangle \geq \zeta^+ \left\| \boldsymbol{\mathcal{A}}' - \boldsymbol{\mathcal{A}} \right\|_{\mathrm{F}}^2. \tag{5}$$

Inequalities (4) and (5) correspond to the definitions of RSC and RSS for the function $-Q_\lambda(\boldsymbol{\mathcal{A}})$, respectively. Specifically, they imply that $-Q_\lambda(\boldsymbol{\mathcal{A}})$ is both ζ^--smooth and ζ^+-strongly convex. Consequently, we have the following

$$-Q_\lambda(\boldsymbol{\mathcal{A}}') \leq -Q_\lambda(\boldsymbol{\mathcal{A}}) - \left\langle \nabla Q_\lambda(\boldsymbol{\mathcal{A}}), \boldsymbol{\mathcal{A}}' - \boldsymbol{\mathcal{A}} \right\rangle + \frac{\zeta^-}{2} \left\| \boldsymbol{\mathcal{A}}' - \boldsymbol{\mathcal{A}} \right\|_{\mathrm{F}}^2, \tag{6}$$

$$-Q_\lambda(\boldsymbol{\mathcal{A}}') \geq -Q_\lambda(\boldsymbol{\mathcal{A}}) - \left\langle \nabla Q_\lambda(\boldsymbol{\mathcal{A}}), \boldsymbol{\mathcal{A}}' - \boldsymbol{\mathcal{A}} \right\rangle + \frac{\zeta^+}{2} \left\| \boldsymbol{\mathcal{A}}' - \boldsymbol{\mathcal{A}} \right\|_{\mathrm{F}}^2. \tag{7}$$

For the loss function $L(\boldsymbol{\mathcal{A}})$, applying Taylor's theorem and the mean value theorem yields

$$L(\boldsymbol{\mathcal{A}}') = L(\boldsymbol{\mathcal{A}}) + \left\langle \nabla L(\boldsymbol{\mathcal{A}}), \boldsymbol{\mathcal{A}}' - \boldsymbol{\mathcal{A}} \right\rangle + \frac{1}{2} \left\langle \nabla^2 L\left(\beta \boldsymbol{\mathcal{A}}' + (1-\beta)\boldsymbol{\mathcal{A}}\right), \left(\boldsymbol{\mathcal{A}}' - \boldsymbol{\mathcal{A}}\right) \otimes \left(\boldsymbol{\mathcal{A}}' - \boldsymbol{\mathcal{A}}\right) \right\rangle, \tag{8}$$

for some $\beta \in [0,1]$. Here, $\otimes$ denotes the Kronecker product. Given two tensors $\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{A}}' \in \mathbb{R}^{d_1 \times \cdots \times d_M}$, their Kronecker product results in a tensor $\boldsymbol{\mathcal{A}}''$ of dimension $(d_1 d_1) \times \ldots (d_M d_M)$. Each element $\boldsymbol{\mathcal{A}}''_{i_1 j_1 i_2 j_2 \ldots i_M j_M}$ is defined as $\boldsymbol{\mathcal{A}}_{i_1 i_2 \ldots i_M} \times \boldsymbol{\mathcal{A}}'_{j_1 j_2 \ldots j_M}$.

Under Assumptions 1 and 2, we have

$$L(\boldsymbol{\mathcal{A}} + \boldsymbol{\mathcal{B}}) - L(\boldsymbol{\mathcal{A}}) \geq \langle \nabla L(\boldsymbol{\mathcal{A}}), \boldsymbol{\mathcal{B}} \rangle + \frac{\rho^-}{2} \|\boldsymbol{\mathcal{B}}\|_{\mathrm{F}}^2, \tag{9}$$

$$L(\boldsymbol{\mathcal{A}} + \boldsymbol{\mathcal{B}}) - L(\boldsymbol{\mathcal{A}}) \leq \langle \nabla L(\boldsymbol{\mathcal{A}}), \boldsymbol{\mathcal{B}} \rangle + \frac{\rho^+}{2} \|\boldsymbol{\mathcal{B}}\|_{\mathrm{F}}^2. \tag{10}$$

Recall that $\widetilde{L}(\boldsymbol{\mathcal{A}}) = L(\boldsymbol{\mathcal{A}}) + Q_\lambda(\boldsymbol{\mathcal{A}})$. Thus, we obtain

$$\widetilde{L}(\boldsymbol{\mathcal{A}}') \geq \widetilde{L}(\boldsymbol{\mathcal{A}}) + \left\langle \nabla \widetilde{L}(\boldsymbol{\mathcal{A}}), \boldsymbol{\mathcal{A}}' - \boldsymbol{\mathcal{A}} \right\rangle + \frac{\rho^- - \zeta^-}{2} \left\| \boldsymbol{\mathcal{A}}' - \boldsymbol{\mathcal{A}} \right\|_{\mathrm{F}}^2, \tag{11}$$

$$\widetilde{L}(\boldsymbol{\mathcal{A}}') \leq \widetilde{L}(\boldsymbol{\mathcal{A}}) + \left\langle \nabla \widetilde{L}(\boldsymbol{\mathcal{A}}), \boldsymbol{\mathcal{A}}' - \boldsymbol{\mathcal{A}} \right\rangle + \frac{\rho^+ - \zeta^+}{2} \left\| \boldsymbol{\mathcal{A}}' - \boldsymbol{\mathcal{A}} \right\|_{\mathrm{F}}^2, \tag{12}$$

□

Lemma 2. *Let C be a constant. Suppose there exists an integer $\tilde{s}_1 > C\left|\boldsymbol{S}_1\right|$, and that $\boldsymbol{\mathcal{A}}$ satisfies $\left\|\boldsymbol{\mathcal{A}}_{\overline{\boldsymbol{S}_1}}\right\|_0 \leq \widetilde{s_1}$, $\omega\left(\boldsymbol{\mathcal{A}}\right) \leq \frac{\lambda}{2}$, where $\omega\left(\boldsymbol{\mathcal{A}}\right) = \min_{\boldsymbol{\mathcal{H}} \in \partial\|\boldsymbol{\mathcal{A}}\|_1} \left\{\left\|\nabla \widetilde{L}\left(\boldsymbol{\mathcal{A}}\right) + \lambda \boldsymbol{\mathcal{H}}\right\|_{\max}\right\}$, and $\left\|\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right\|_{\max} \leq \lambda/8$. Under Assumptions 1 and 2, $\boldsymbol{\mathcal{A}}$ satisfies*

$$\left\|\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right\|_{\mathrm{F}} \leq \frac{21/8}{\rho^- - \zeta^-} \lambda \sqrt{\left|\boldsymbol{S}_1\right|}.$$

Proof. Given that $\left\|\boldsymbol{\mathcal{A}}_{\overline{\boldsymbol{S}_1}}\right\|_0 \leq \tilde{s}_1$ and $\left\|\boldsymbol{\mathcal{A}}^*_{\overline{\boldsymbol{S}_1}}\right\|_0 = 0$, it follows that $\left\|\left(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right)_{\overline{\boldsymbol{S}_1}}\right\|_0 \leq \tilde{s}_1$. Based on Lemma 1, we can derive the following inequalities

$$\widetilde{L}\left(\boldsymbol{\mathcal{A}}^*\right) \geq \widetilde{L}\left(\boldsymbol{\mathcal{A}}\right) + \left\langle \nabla \widetilde{L}\left(\boldsymbol{\mathcal{A}}\right), \boldsymbol{\mathcal{A}}^* - \boldsymbol{\mathcal{A}} \right\rangle + \frac{\rho^- - \zeta^-}{2} \left\|\boldsymbol{\mathcal{A}}^* - \boldsymbol{\mathcal{A}}\right\|_{\mathrm{F}}^2, \tag{13}$$

$$\widetilde{L}\left(\boldsymbol{\mathcal{A}}\right) \geq \widetilde{L}\left(\boldsymbol{\mathcal{A}}^*\right) + \left\langle \nabla \widetilde{L}\left(\boldsymbol{\mathcal{A}}^*\right), \boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^* \right\rangle + \frac{\rho^- - \zeta^-}{2} \left\|\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right\|_{\mathrm{F}}^2. \tag{14}$$

Adding (13) and (14), we obtain

$$\left\langle \nabla \widetilde{L}\left(\boldsymbol{\mathcal{A}}\right), \boldsymbol{\mathcal{A}}^* - \boldsymbol{\mathcal{A}} \right\rangle \geq \left\langle \nabla \widetilde{L}\left(\boldsymbol{\mathcal{A}}^*\right), \boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^* \right\rangle + \left(\rho^- - \zeta^-\right) \left\|\boldsymbol{\mathcal{A}}^* - \boldsymbol{\mathcal{A}}\right\|_{\mathrm{F}}^2. \tag{15}$$

Let $\boldsymbol{\mathcal{H}} \in \partial \left\|\boldsymbol{\mathcal{A}}\right\|_1$ denote the sub-gradient and $\boldsymbol{\mathcal{S}}$ be a set. According to the Karush-Kuhn-Tucker (KKT) condition, we have

$$\nabla \widetilde{L}\left(\boldsymbol{\mathcal{A}}\right) + \lambda \boldsymbol{\mathcal{H}} = \mathbf{0}. \tag{16}$$

Determining the optimal solution is challenging, therefore, we introduce a measure of sub-optimality

$$\omega\left(\boldsymbol{\mathcal{A}}\right) = \min_{\boldsymbol{\mathcal{H}}' \in \partial\|\boldsymbol{\mathcal{A}}\|_1} \max_{\boldsymbol{\mathcal{A}}' \in \boldsymbol{\mathcal{S}}} \left\{ \frac{1}{\left\|\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}'\right\|_1} \left\langle \boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}', \nabla \widetilde{L}\left(\boldsymbol{\mathcal{A}}\right) + \lambda \boldsymbol{\mathcal{H}}' \right\rangle \right\}. \tag{17}$$

We define our algorithm's stopping criterion as $\omega\left(\boldsymbol{\mathcal{A}}\right) \leq \varepsilon$. Consequently, the sub-optimality can be expressed as

$$\omega\left(\boldsymbol{\mathcal{A}}\right) = \max_{\boldsymbol{\mathcal{A}}' \in \boldsymbol{\mathcal{S}}} \left\{ \frac{1}{\left\|\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}'\right\|_1} \left\langle \boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}', \nabla \widetilde{L}\left(\boldsymbol{\mathcal{A}}\right) + \lambda \boldsymbol{\mathcal{H}} \right\rangle \right\}. \tag{18}$$

Adding $\lambda \left\langle \boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*, \boldsymbol{\mathcal{H}}' \right\rangle$ to the both sides of (15), we obtain

$$\left\langle \boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*, \nabla \widetilde{L}\left(\boldsymbol{\mathcal{A}}\right) + \lambda \boldsymbol{\mathcal{H}} \right\rangle \geq \left\langle \boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*, \nabla \widetilde{L}\left(\boldsymbol{\mathcal{A}}\right) \right\rangle + \left(\rho^- - \zeta^-\right) \left\|\boldsymbol{\mathcal{A}}^* - \boldsymbol{\mathcal{A}}\right\|_{\mathrm{F}}^2 + \lambda \left\langle \boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*, \boldsymbol{\mathcal{H}} \right\rangle. \tag{19}$$

Since $\boldsymbol{\mathcal{A}}^* \in \boldsymbol{\mathcal{S}}$, we have

$$\frac{1}{\left\|\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right\|_1} \left\langle \boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*, \nabla \widetilde{L}\left(\boldsymbol{\mathcal{A}}\right) + \lambda \boldsymbol{\mathcal{H}} \right\rangle \leq \max_{\boldsymbol{\mathcal{A}}' \in \boldsymbol{\mathcal{S}}} \left\{ \frac{1}{\left\|\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}'\right\|_1} \left\langle \boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}', \nabla \widetilde{L}\left(\boldsymbol{\mathcal{A}}\right) + \lambda \boldsymbol{\mathcal{H}} \right\rangle \right\} = v\left(\boldsymbol{\mathcal{A}}\right). \tag{20}$$

Recall that we assume $v\left(\boldsymbol{\mathcal{A}}\right) \leq \frac{\lambda}{2}$, we obtain

$$\left\langle \boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*, \nabla \widetilde{L}\left(\boldsymbol{\mathcal{A}}\right) + \lambda \boldsymbol{\mathcal{H}} \right\rangle \leq v\left(\boldsymbol{\mathcal{A}}\right) \leq \frac{\lambda}{2} \left\|\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right\|_1. \tag{21}$$

Combining (15) and (21), we obtain

$$\frac{\lambda}{2} \left\|\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right\|_1 \geq \underbrace{\left\langle \boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*, \nabla \widetilde{L}\left(\boldsymbol{\mathcal{A}}^*\right) \right\rangle}_{\mathrm{I}} + \left(\rho^- - \zeta^-\right) \left\|\boldsymbol{\mathcal{A}}^* - \boldsymbol{\mathcal{A}}\right\|_{\mathrm{F}}^2 + \underbrace{\lambda \left\langle \boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*, \boldsymbol{\mathcal{H}} \right\rangle}_{\mathrm{II}}. \tag{22}$$

For term I, separating the support of $\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*$ into $\boldsymbol{\mathcal{S}}_1$ and $\overline{\boldsymbol{\mathcal{S}}_1}$, we have

$$\begin{aligned}
\left\langle \boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*, \nabla \widetilde{L}\left(\boldsymbol{\mathcal{A}}^*\right)\right\rangle &= \left\langle \boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*, \nabla \widetilde{L}\left(\boldsymbol{\mathcal{A}}^*\right)\right\rangle + \left\langle \boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*, \nabla Q_\lambda\left(\boldsymbol{\mathcal{A}}^*\right)\right\rangle && (23)\\
&\geq - \left\|\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right\|_1 \left\|\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right\|_{\max} + \left\langle \boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*, \nabla Q_\lambda\left(\boldsymbol{\mathcal{A}}^*\right)\right\rangle && (24)\\
&= - \left\|\left(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right)_{\overline{\boldsymbol{\mathcal{S}}_1}}\right\|_1 \left\|\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right\|_{\max} - \left\|\left(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right)_{\boldsymbol{\mathcal{S}}_1}\right\|_1 \left\|\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right\|_{\max} && (25)\\
&\quad + \left\langle \left(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right)_{\boldsymbol{\mathcal{S}}_1}, \left(\nabla Q_\lambda\left(\boldsymbol{\mathcal{A}}^*\right)\right)_{\boldsymbol{\mathcal{S}}_1}\right\rangle + \left\langle \left(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right)_{\overline{\boldsymbol{\mathcal{S}}_1}}, \left(\nabla Q_\lambda\left(\boldsymbol{\mathcal{A}}^*\right)\right)_{\overline{\boldsymbol{\mathcal{S}}_1}}\right\rangle \\
&\geq - \left\|\left(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right)_{\overline{\boldsymbol{\mathcal{S}}_1}}\right\|_1 \left\|\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right\|_{\max} - \left\|\left(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right)_{\boldsymbol{\mathcal{S}}_1}\right\|_1 \left\|\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right\|_{\max} && (26)\\
&\quad - \left\|\left(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right)_{\boldsymbol{\mathcal{S}}_1}\right\|_1 \left\|\nabla Q_\lambda\left(\boldsymbol{\mathcal{A}}^*\right)\right\|_{\max}.
\end{aligned}$$

For term II, separating the support of $\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*$ into $\boldsymbol{\mathcal{S}}_1$ and $\overline{\boldsymbol{\mathcal{S}}_1}$, we have

$$\begin{aligned}
\lambda \left\langle \boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*, \boldsymbol{\mathcal{H}}\right\rangle &= \lambda \left\langle \left(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right)_{\boldsymbol{\mathcal{S}}_1}, \left(\boldsymbol{\mathcal{H}}\right)_{\boldsymbol{\mathcal{S}}_1}\right\rangle + \lambda \left\langle \left(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right)_{\overline{\boldsymbol{\mathcal{S}}_1}}, \left(\boldsymbol{\mathcal{H}}\right)_{\overline{\boldsymbol{\mathcal{S}}_1}}\right\rangle && (27)\\
&\geq - \lambda \left\|\left(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right)_{\boldsymbol{\mathcal{S}}_1}\right\|_1 \left\|\left(\boldsymbol{\mathcal{H}}\right)_{\boldsymbol{\mathcal{S}}_1}\right\|_{\max} + \lambda \left\langle \left(\boldsymbol{\mathcal{A}}\right)_{\overline{\boldsymbol{\mathcal{S}}_1}}, \left(\boldsymbol{\mathcal{H}}\right)_{\overline{\boldsymbol{\mathcal{S}}_1}}\right\rangle && (28)\\
&\geq - \lambda \left\|\left(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right)_{\boldsymbol{\mathcal{S}}_1}\right\|_1 + \lambda \sum_{\left(i_1,\ldots,i_M \in \overline{\boldsymbol{\mathcal{S}}_1}\right)} \left|\boldsymbol{\mathcal{A}}_{i_1,\ldots,i_M}\right| && (29)\\
&= - \lambda \left\|\left(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right)_{\boldsymbol{\mathcal{S}}_1}\right\|_1 + \lambda \left\|\left(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right)_{\overline{\boldsymbol{\mathcal{S}}_1}}\right\|_1. && (30)
\end{aligned}$$

Thus, we obtain

$$\begin{aligned}
\frac{\lambda}{2}\left\|\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right\|_1 \geq &- \left\|\left(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right)_{\overline{\boldsymbol{\mathcal{S}}_1}}\right\|_1 \left\|\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right\|_{\max} - \left\|\left(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right)_{\boldsymbol{\mathcal{S}}_1}\right\|_1 \left\|\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right\|_{\max} && (31)\\
&- \left\|\left(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right)_{\boldsymbol{\mathcal{S}}_1}\right\|_1 \left\|\nabla Q_\lambda\left(\boldsymbol{\mathcal{A}}^*\right)\right\|_{\max} + \left(\rho^- - \zeta^-\right)\left\|\boldsymbol{\mathcal{A}}^* - \boldsymbol{\mathcal{A}}\right\|_{\mathrm{F}}^2 \\
&- \lambda \left\|\left(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right)_{\boldsymbol{\mathcal{S}}_1}\right\|_1 + \lambda \left\|\left(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right)_{\overline{\boldsymbol{\mathcal{S}}_1}}\right\|_1.
\end{aligned}$$

We separate the left-hand side of (31) as

$$\frac{\lambda}{2}\left\|\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right\|_1 = \frac{\lambda}{2}\left\|\left(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right)_{\boldsymbol{\mathcal{S}}_1}\right\|_1 + \frac{\lambda}{2}\left\|\left(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right)_{\overline{\boldsymbol{\mathcal{S}}_1}}\right\|_1. \tag{32}$$

Rearranging the terms, we obtain

$$\begin{aligned}
&\left(\rho^- - \zeta^-\right)\left\|\boldsymbol{\mathcal{A}}^* - \boldsymbol{\mathcal{A}}\right\|_{\mathrm{F}}^2 + \left(\frac{\lambda}{2} - \left\|\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right\|_{\max}\right)\left\|\left(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right)_{\overline{\boldsymbol{\mathcal{S}}_1}}\right\|_1 && (33)\\
\leq &\left(\frac{3\lambda}{2} + \left\|\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right\|_{\max} + \left\|\nabla Q_\lambda\left(\boldsymbol{\mathcal{A}}^*\right)\right\|_{\max}\right)\left\|\left(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right)_{\boldsymbol{\mathcal{S}}_1}\right\|_1.
\end{aligned}$$

Recall that $\left\|\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right\|_{\max} \leq \frac{\lambda}{8}$, we have

$$\begin{aligned}
\left(\rho^- - \zeta^-\right)\left\|\boldsymbol{\mathcal{A}}^* - \boldsymbol{\mathcal{A}}\right\|_{\mathrm{F}}^2 &\leq \left(\frac{3\lambda}{2} + \left\|\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right\|_{\max} + \left\|\nabla Q_\lambda\left(\boldsymbol{\mathcal{A}}^*\right)\right\|_{\max}\right)\left\|\left(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right)_{\boldsymbol{\mathcal{S}}_1}\right\|_1 && (34)\\
&\leq \left(\frac{3\lambda}{2} + \frac{\lambda}{8} + \lambda\right)\left\|\left(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right)_{\boldsymbol{\mathcal{S}}_1}\right\|_1 && (35)\\
&\leq \frac{21\lambda}{8}\sqrt{\left|\boldsymbol{\mathcal{S}}_1\right|}\left\|\left(\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right)_{\boldsymbol{\mathcal{S}}_1}\right\|_{\mathrm{F}} && (36)\\
&\leq \frac{21\lambda}{8}\sqrt{\left|\boldsymbol{\mathcal{S}}_1\right|}\left\|\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right\|_{\mathrm{F}}. && (37)
\end{aligned}$$

Given that $\rho^- - \zeta^- > 0$, we have

$$\left\|\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right\|_{\mathrm{F}} \leq \frac{21/8}{\rho^- - \zeta^-}\lambda\sqrt{\left|\boldsymbol{\mathcal{S}}_1\right|}. \tag{38}$$

□

Lemma 3. *Consider the regularization parameter λ and assume that the derivative of the non-convex penalty satisfies $p'_\lambda(\mathcal{A}_{d_1,\ldots,d_M}) = 0$ whenever $|\mathcal{A}_{d_1,\ldots,d_M}| \geq \nu$ for some $\nu > 0$. Let $\mathcal{S}_1^{\mathrm{I}} \cup \mathcal{S}_1^{\mathrm{II}} = \mathcal{S}_1$. For indices $(i_1,\ldots,i_M) \in \mathcal{S}_1^{\mathrm{I}} \subseteq \mathcal{S}_1$, we assume $|\mathcal{A}^*_{d_1,\ldots,d_M}| \geq \nu$, and for indices $(i_1,\ldots,i_M) \in \mathcal{S}_1^{\mathrm{II}} \subseteq \mathcal{S}_1$, we assume $|\mathcal{A}^*_{d_1,\ldots,d_M}| \leq \nu$. Under Assumptions 1∼4, we derive the following bound*

$$\left\|\hat{\mathcal{A}} - \mathcal{A}^*\right\|_{\mathrm{F}} \leq \frac{1}{\rho^- - \zeta^-}\left\|(\nabla L(\mathcal{A}^*))_{\mathcal{S}_1^{\mathrm{I}}}\right\|_{\mathrm{F}} + \frac{3}{\rho^- - \zeta^-}\lambda\sqrt{\left|\mathcal{S}_1^{\mathrm{II}}\right|}.$$

Proof. Define the sub-gradients $\mathcal{H}^* \in \partial\|\mathcal{A}^*\|_1$ and $\hat{\mathcal{H}} \in \partial\left\|\hat{\mathcal{A}}\right\|_1$.

Note that $\hat{\mathcal{A}}$ satisfies the optimality condition that $\omega\left(\hat{\mathcal{A}}\right) \leq 0$, we have

$$\max_{\mathcal{A}' \in \mathcal{S}}\left\{\left\langle \hat{\mathcal{A}} - \mathcal{A}', \nabla\widetilde{L}\left(\hat{\mathcal{A}}\right) + \lambda\hat{\mathcal{H}}\right\rangle\right\} \leq 0. \tag{39}$$

Given that $\left\|\left(\hat{\mathcal{A}}\right)_{\overline{\mathcal{S}_1}}\right\|_0 \leq \tilde{s}_1$, since $\left\|\left(\hat{\mathcal{A}} - \mathcal{A}^*\right)_{\overline{\mathcal{S}_1}}\right\|_0 \leq \tilde{s}_1$, according to Lemma 1, we obtain

$$\widetilde{L}\left(\hat{\mathcal{A}}\right) \geq \widetilde{L}(\mathcal{A}^*) + \left\langle \nabla\widetilde{L}(\mathcal{A}^*), \hat{\mathcal{A}} - \mathcal{A}^*\right\rangle + \frac{\rho^- - \zeta^-}{2}\|\hat{\mathcal{A}} - \mathcal{A}^*\|_{\mathrm{F}}^2, \tag{40}$$

$$\widetilde{L}(\mathcal{A}^*) \geq \widetilde{L}\left(\hat{\mathcal{A}}\right) + \left\langle \nabla\widetilde{L}\left(\hat{\mathcal{A}}\right), \mathcal{A}^* - \hat{\mathcal{A}}\right\rangle + \frac{\rho^- - \zeta^-}{2}\|\mathcal{A}^* - \hat{\mathcal{A}}\|_{\mathrm{F}}^2. \tag{41}$$

By the convexity of ℓ_1 norm, we have

$$\lambda\left\|\hat{\mathcal{A}}\right\|_1 \leq \lambda\|\mathcal{A}^*\|_1 + \lambda\left\langle \hat{\mathcal{A}} - \mathcal{A}^*, \mathcal{H}^*\right\rangle, \tag{42}$$

$$\lambda\|\mathcal{A}^*\|_1 \leq \lambda\left\|\hat{\mathcal{A}}\right\|_1 + \lambda\left\langle \mathcal{A}^* - \hat{\mathcal{A}}, \hat{\mathcal{H}}\right\rangle. \tag{43}$$

Adding (40) $\sim$ (43), we obtain

$$0 \geq \underbrace{\left\langle \nabla L(\mathcal{A}^*) + \nabla\mathcal{Q}_\lambda\left(\mathcal{A}^*_{i_1\ldots i_M}\right) + \lambda\mathcal{H}^*, \hat{\mathcal{A}} - \mathcal{A}^*\right\rangle}_{\text{(i)}} + \underbrace{\left\langle \nabla\widetilde{L}\left(\hat{\mathcal{A}}\right) + \lambda\mathcal{H}^*, \mathcal{A}^* - \hat{\mathcal{A}}\right\rangle}_{\text{(ii)}} + \left(\rho^- - \zeta^-\right)\left\|\hat{\mathcal{A}} - \mathcal{A}^*\right\|_{\mathrm{F}}^2. \tag{44}$$

From the optimality condition (39), we have

$$\left\langle \nabla\widetilde{L}\left(\hat{\mathcal{A}}\right) + \lambda\hat{\mathcal{H}}, \mathcal{A}^* - \hat{\mathcal{A}}\right\rangle \leq \max_{\mathcal{A}' \in \mathcal{S}}\left\{\left\langle \hat{\mathcal{A}} - \mathcal{A}', \nabla\widetilde{L}\left(\hat{\mathcal{A}}\right) + \lambda\hat{\mathcal{H}}\right\rangle\right\} \leq 0, \tag{45}$$

which implies the term (ii) in (44) is non-negative. Consequently, we can arrange (44) to obtain

$$\begin{aligned}
&\left(\rho^- - \zeta^-\right)\left\|\hat{\mathcal{A}} - \mathcal{A}^*\right\|_{\mathrm{F}}^2 \\
\leq &\left\langle \nabla L(\mathcal{A}^*) + \nabla\mathcal{Q}_\lambda(\mathcal{A}^*) + \lambda_t\mathcal{H}^*, \hat{\mathcal{A}} - \mathcal{A}^*\right\rangle \\
\leq &\min_{\mathcal{H}^* \in \partial\|\mathcal{A}^*\|_1}\left\{\sum_{i_1=1}^{I_1}\cdots\sum_{i_M=1}^{I_M}\left|\left(\nabla L(\mathcal{A}^*) + \nabla\mathcal{Q}_\lambda(\mathcal{A}^*) + \lambda_t\mathcal{H}^*\right)_{i_1\ldots i_M}\right| \cdot \left|\left(\mathcal{A}^* - \hat{\mathcal{A}}\right)_{i_1\ldots i_M}\right|\right\}.
\end{aligned} \tag{46}$$

We proceed by decomposing the summation on the right-hand side of (46) into three distinct parts

- $(i_1,\ldots,i_M) \in \overline{\mathcal{S}_1}$,
- $(i_1,\ldots,i_M) \in \mathcal{S}_1^{\mathrm{I}}$,

- $(i_1, \ldots, i_M) \in \mathcal{S}_1^{\mathrm{II}}$.

Here, $\mathcal{S}_1^{\mathrm{I}} = \left\{(i_1, \ldots, i_M) \mid \left|\boldsymbol{\mathcal{A}}_{(i_1,\ldots,i_M)}\right| \geq \nu\right\}$, $\mathcal{S}_1^{\mathrm{I}} = \left\{(i_1, \ldots, i_M) \mid \left|\boldsymbol{\mathcal{A}}_{(i_1,\ldots,i_M)}\right| < \nu\right\}$, and $\nu > 0$ is defined in Assumption 3.

(i) For any index $(i_1, \ldots, i_M) \in \overline{\mathcal{S}_1}$, the regularity condition yields

$$\nabla \mathcal{Q}_\lambda \left(\boldsymbol{\mathcal{A}}^*\right)_{i_1 \ldots i_M} = q_\lambda' \left(\left(\boldsymbol{\mathcal{A}}^*\right)_{i_1 \ldots i_M}\right) = q_\lambda' \left(0\right), \quad \text{for} \quad j \in \overline{\mathcal{S}_1}. \tag{47}$$

Assuming that $\left\|\nabla L \left(\boldsymbol{\mathcal{A}}^*\right)\right\|_{\max} \leq \frac{\lambda}{8}$, it follows that

$$\max_{(i_1,\ldots,i_M)\in\overline{\mathcal{S}_1}} \left|\left(\nabla L \left(\boldsymbol{\mathcal{A}}^*\right)\right)_{i_1,\ldots,i_M}\right| \leq \left\|\nabla L \left(\boldsymbol{\mathcal{A}}^*\right)\right\|_{\max} \leq \frac{\lambda}{8} \leq \lambda. \tag{48}$$

Therefore,

$$\max_{(i_1,\ldots,i_M)\in\overline{\mathcal{S}_1}} \left|\left(\nabla L \left(\boldsymbol{\mathcal{A}}^*\right) + Q_\lambda \left(\boldsymbol{\mathcal{A}}^*\right)\right)_{i_1,\ldots,i_M}\right| \leq \lambda. \tag{49}$$

Moreover, since $\boldsymbol{\mathcal{H}}^* \in \partial \left\|\boldsymbol{\mathcal{A}}^*\right\|_1$, it holds that $\lambda \boldsymbol{\mathcal{H}}^*_{i_1,\ldots,i_M} \in [-\lambda, \lambda]$. Consequently, for each $(i_1, \ldots, i_M) \in \overline{\mathcal{S}_1}$, we can select $\boldsymbol{\mathcal{H}}^*_{i_1,\ldots,i_M}$ such that

$$\left|\left(\nabla L \left(\boldsymbol{\mathcal{A}}^*\right) + \nabla \mathcal{Q}_\lambda \left(\boldsymbol{\mathcal{A}}^*\right)\right)_{i_1 \ldots i_M} + \lambda \boldsymbol{\mathcal{H}}^*_{i_1 \ldots i_M}\right| = 0. \tag{50}$$

This implies

$$\min_{\boldsymbol{\mathcal{H}}^* \in \partial \left\|\boldsymbol{\mathcal{A}}^*\right\|_1} \left\{\left|\left(\nabla L \left(\boldsymbol{\mathcal{A}}^*\right) + \nabla \mathcal{Q}_\lambda \left(\boldsymbol{\mathcal{A}}^*\right) + \lambda \boldsymbol{\mathcal{H}}^*\right)_{i_1 \ldots i_M}\right|\right\} = 0, \quad \text{for} \quad (i_1 \ldots i_M) \in \overline{\mathcal{S}_1}. \tag{51}$$

Therefore, we obtain

$$\min_{\boldsymbol{\mathcal{H}}^* \in \partial \left\|\boldsymbol{\mathcal{A}}^*\right\|_1} \left\{\sum_{(i_1,\ldots,i_M)\in\overline{\mathcal{S}_1}} \left|\left(\nabla L \left(\boldsymbol{\mathcal{A}}^*\right) + \nabla \mathcal{Q}_\lambda \left(\boldsymbol{\mathcal{A}}^*\right) + \lambda \boldsymbol{\mathcal{H}}^*\right)_{i_1 \ldots i_M}\right| \cdot \left|\left(\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}\right)_{i_1 \ldots i_M}\right|\right\} = 0. \tag{52}$$

(ii) For indices $(i_1, \ldots, i_M) \in \mathcal{S}_1^{\mathrm{I}}$, we have $\left|\boldsymbol{\mathcal{A}}^*_{i_1,\ldots,i_M}\right| \geq \nu$. Given that $R\left(\boldsymbol{\mathcal{A}}\right) = \lambda \left\|\boldsymbol{\mathcal{A}}\right\|_1 + \mathcal{Q}_\lambda \left(\boldsymbol{\mathcal{A}}_{i_1 \ldots i_M}\right)$, our assumption on $R\left(\boldsymbol{\mathcal{A}}\right)$ ensures that

$$\left(\nabla \mathcal{Q}_\lambda \left(\boldsymbol{\mathcal{A}}^*\right) + \lambda \boldsymbol{\mathcal{H}}^*\right)_{i_1 \ldots i_M} = p_\lambda' \left(\boldsymbol{\mathcal{A}}^*_{i_1 \ldots i_M}\right) = 0, \quad \text{for} \quad (i_1 \ldots i_M) \in \mathcal{S}_1^{\mathrm{I}}. \tag{53}$$

This leads to

$$\min_{\boldsymbol{\mathcal{H}}^* \in \partial \left\|\boldsymbol{\mathcal{A}}^*\right\|_1} \left\{\sum_{(i_1 \ldots i_M)\in\mathcal{S}_1^{\mathrm{I}}} \left|\left(\nabla L \left(\boldsymbol{\mathcal{A}}^*\right) + \nabla \mathcal{Q}_\lambda \left(\boldsymbol{\mathcal{A}}^*\right) + \lambda \boldsymbol{\mathcal{H}}^*\right)_{i_1 \ldots i_M}\right| \cdot \left|\left(\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}\right)_{i_1 \ldots i_M}\right|\right\} \tag{54}$$

$$= \sum_{(i_1 \ldots i_M)\in\mathcal{S}_1^{\mathrm{I}}} \left|\left(\nabla L \left(\boldsymbol{\mathcal{A}}^*\right)\right)_{i_1 \ldots i_M}\right| \cdot \left|\left(\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}\right)_{i_1 \ldots i_M}\right| \tag{55}$$

$$\leq \left\|\left(\nabla L \left(\boldsymbol{\mathcal{A}}^*\right)\right)_{\mathcal{S}_1^{\mathrm{I}}}\right\|_{\mathrm{F}} \cdot \left\|\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}\right\|_{\mathrm{F}}. \tag{56}$$

(iii) For indices $(i_1, \ldots, i_M) \in \mathcal{S}_1^{\mathrm{II}}$, we have $\left|\boldsymbol{\mathcal{A}}^*_{(i_1,\ldots,i_M)}\right| < \nu$. Given that $\left\|\nabla L \left(\boldsymbol{\mathcal{A}}^*\right)\right\|_{\max} \leq \frac{\lambda}{8}$, we have

$$\max_{(i_1 \ldots i_M)\in\mathcal{S}_1^{\mathrm{II}}} \left|\left(\nabla L \left(\boldsymbol{\mathcal{A}}^*\right)\right)_{i_1 \ldots i_M}\right| \leq \left|\left(\nabla L \left(\boldsymbol{\mathcal{A}}^*\right)\right)_{i_1 \ldots i_M}\right|_{\max} \leq \lambda/8. \tag{57}$$

Meanwhile, we have

$$\max_{(i_1\dots i_M)\in\mathcal{S}_1^{\mathrm{II}}}\left|\left(\nabla\mathcal{Q}_\lambda\left(\boldsymbol{\mathcal{A}}^*\right)\right)_{i_1\dots i_M}\right| = \max_{(i_1\dots i_M)\in\mathcal{S}_1^{\mathrm{II}}}\left|q'_\lambda\left(\left(\boldsymbol{\mathcal{A}}^*\right)_{i_1\dots i_M}\right)\right| \le \max\left|q'_\lambda\left(\left(\boldsymbol{\mathcal{A}}^*\right)_{i_1\dots i_M}\right)\right| \le \lambda, \tag{58}$$

Additionally, since $\boldsymbol{\mathcal{H}}^* \in \partial\left\|\boldsymbol{\mathcal{A}}^*\right\|_1$, it follows that $\left|\boldsymbol{\mathcal{H}}^*_{i_1,\dots,i_M}\right| \le 1$. Therefore, for each $(i_1,\dots,i_M)\in\mathcal{S}_1^{\mathrm{II}}$, we obtain

$$\begin{aligned}\left|\left(\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)+\nabla\mathcal{Q}_\lambda\left(\boldsymbol{\mathcal{A}}^*\right)+\lambda\boldsymbol{\mathcal{H}}^*\right)_{i_1\dots i_M}\right| &\le \max_{(i_1\dots i_N)\in\mathcal{S}_1^{\mathrm{II}}}\left|\left(\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right)_{i_1\dots i_M}\right| + \max_{(i_1\dots i_N)\in\mathcal{S}_1^{\mathrm{II}}}\left|\left(\nabla\mathcal{Q}_\lambda\left(\boldsymbol{\mathcal{A}}^*\right)\right)_{i_1\dots i_M}\right| + \lambda \\ &\le 3\lambda,\end{aligned} \tag{59}$$

which implies

$$\min_{\boldsymbol{\mathcal{H}}^*\in\partial\left\|\boldsymbol{\mathcal{A}}^*\right\|_1}\left\{\sum_{(i_1\dots i_N)\in\mathcal{S}_1^{\mathrm{II}}}\left|\left(\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)+\nabla\mathcal{Q}_\lambda\left(\boldsymbol{\mathcal{A}}^*\right)+\lambda\boldsymbol{\mathcal{H}}^*\right)_{i_1\dots i_M}\right|\cdot\left|\left(\boldsymbol{\mathcal{A}}^*-\hat{\boldsymbol{\mathcal{A}}}\right)_{i_1\dots i_M}\right|\right\} \tag{60}$$

$$\le 3\lambda\left|\left(\boldsymbol{\mathcal{A}}^*-\hat{\boldsymbol{\mathcal{A}}}\right)_{i_1\dots i_M}\right| \tag{61}$$

$$= 3\lambda\left\|\left(\boldsymbol{\mathcal{A}}^*-\hat{\boldsymbol{\mathcal{A}}}\right)_{\overline{\mathcal{S}_1^{\mathrm{II}}}}\right\|_{\mathrm{F}} \tag{62}$$

$$\le 3\lambda\sqrt{\left|\mathcal{S}_1\right|}\left\|\left(\boldsymbol{\mathcal{A}}^*-\hat{\boldsymbol{\mathcal{A}}}\right)_{\overline{\mathcal{S}_1^{\mathrm{II}}}}\right\|_{\mathrm{F}} \tag{63}$$

$$\le 3\lambda\sqrt{\left|\mathcal{S}_1\right|}\left\|\boldsymbol{\mathcal{A}}^*-\hat{\boldsymbol{\mathcal{A}}}\right\|_{\mathrm{F}}. \tag{64}$$

Substituting the bounds from (52) to (64) into the right-hand side of (46), we obtain

$$\left\|\boldsymbol{\mathcal{A}}^*-\hat{\boldsymbol{\mathcal{A}}}\right\|_{\mathrm{F}} \le \frac{1}{\rho^- - \zeta^-}\left(\left\|\left(\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right)_{\mathcal{S}_1^{\mathrm{I}}}\right\|_{\mathrm{F}} + 3\lambda\sqrt{\left|\mathcal{S}_1^{\mathrm{II}}\right|}\right). \tag{65}$$

□

Lemma 4. *For least-squares regression with sub-Gaussian noise, we assume that the columns of $\overline{\boldsymbol{\mathcal{X}}}$ are normalized in such a way that $\max_{j\in\{1,\dots,d_1d_2\dots d_N\}}\left\|\overline{\boldsymbol{\mathcal{X}}}_{\cdot j}\right\|_2 \le \sqrt{n}$, where $\overline{\boldsymbol{\mathcal{X}}} = \left(\mathrm{vec}\left(\boldsymbol{\mathcal{X}}^{(1)}\right),\dots,\mathrm{vec}\left(\boldsymbol{\mathcal{X}}^{(n)}\right)\right)^\top$. If $\lambda \asymp \sqrt{\frac{\log(d_1d_2\cdots d_M)}{n}}$, then we have*

$$\left\|\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right\|_{\mathrm{F}} \lesssim \sqrt{\frac{\left|\mathcal{S}_1\right|}{n}}.$$

Proof. We begin by establishing an upper bound on the probability that the maximum entry of the gradient $\mathbb{P}\left(\left\|\nabla L\left(\boldsymbol{\mathcal{A}}\right)\right\|_{\max} \ge \frac{\lambda}{8}\right)$, where $\nabla L\left(\boldsymbol{\mathcal{A}}\right) = \frac{1}{n}\left\langle\overline{\boldsymbol{\mathcal{X}}},\overline{\boldsymbol{\mathcal{E}}}\right\rangle$ and $\overline{\boldsymbol{\mathcal{E}}} = \left(\mathrm{vec}\left(\boldsymbol{\mathcal{E}}^{(1)}\right),\dots,\mathrm{vec}\left(\boldsymbol{\mathcal{E}}^{(n)}\right)\right)^\top$.

For $\lambda \asymp \sqrt{\frac{\log(d_1d_2\cdots d_M)}{n}}$, using the union bound, we obtain

$$\mathbb{P}\left(\left\|\nabla L\left(\boldsymbol{\mathcal{A}}\right)\right\|_{\max} \ge \frac{\lambda}{8}\right) \le \mathbb{P}\left(\left\|\frac{1}{n}\left\langle\overline{\boldsymbol{\mathcal{X}}},\overline{\boldsymbol{\mathcal{E}}}\right\rangle\right\|_{\max} \ge \frac{c\sqrt{\log d/n}}{8}\right) \tag{66}$$

$$\le \sum_{j=1}^{d_1d_2\times d_N}\mathbb{P}\left(\left|\frac{1}{n}\left\langle\overline{\boldsymbol{\mathcal{X}}},\overline{\boldsymbol{\mathcal{E}}}\right\rangle\right|_{j\ge\frac{c\sqrt{\log d/n}}{8}}\right). \tag{67}$$

Let's define $\theta_k = \left|\left\langle\overline{\boldsymbol{\mathcal{X}}},\overline{\boldsymbol{\mathcal{E}}}\right\rangle\right|_k$, where k is composite coordinate. Since $\overline{\boldsymbol{\mathcal{E}}}_j$ is sub-Gaussian$\left(0,\eta^2\right)$, it follows that for any $t_0 > 0$,

$$\mathbb{E}\left(\exp\left\{t_0\theta_k\right\}+\exp\left\{-t_0\theta_k\right\}\right) \le 2\exp\left\{\frac{1}{n^2}\left\|\overline{\boldsymbol{\mathcal{X}}}_{\cdot k}\right\|^2\eta^2t_0^2/2\right\}. \tag{68}$$

Taking $t_0 = \frac{tn^2}{\left\|\overline{\boldsymbol{\mathcal{X}}}_{\cdot k}\right\|^2 \eta^2 t_0^2}$ yields that

$$\mathbb{P}\left(|\theta_k| \geq t\right) \leq 2\exp\left\{-\frac{n^2t^2}{2\left\|\overline{\boldsymbol{\mathcal{X}}}_{\cdot k}\right\|^2 \eta^2}\right\}. \tag{69}$$

Further taking $t = \frac{\lambda}{8}$ results

$$\mathbb{P}\left(\left\|\nabla L\left(\boldsymbol{\mathcal{A}}\right)\right\|_{\max} \geq \frac{\lambda}{8}\right) \leq 2\left(d_1 \times \cdots \times d_N\right)^{-c^2/\left(128\eta^2\right)}. \tag{70}$$

Applying the Hanson-Wright inequality yields that

$$\begin{aligned}
&\mathbb{P}\left(\left|\langle \boldsymbol{\mathcal{E}}, \langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{E}}\rangle\rangle - \mathbb{E}\langle \boldsymbol{\mathcal{E}}, \langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{E}}\rangle\rangle\right| > \mathbb{E}\langle \boldsymbol{\mathcal{E}}, \langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{E}}\rangle\rangle\right) \\
\leq & 2\exp\left[-C\min\left\{\frac{\mathbb{E}\langle \boldsymbol{\mathcal{E}}, \langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{E}}\rangle\rangle}{\eta^2\left\|\boldsymbol{\mathcal{A}}\right\|_{\mathrm{F}}}, \left(\frac{\mathbb{E}\langle \boldsymbol{\mathcal{E}}, \langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{E}}\rangle\rangle}{\eta^2\left\|\boldsymbol{\mathcal{A}}\right\|_{\mathrm{F}}}\right)^2\right\}\right],
\end{aligned} \tag{71}$$

where C is a universal constant.

Combining the above two inequalities, we have

$$\left\|\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right\|_{\mathrm{F}} = \sqrt{\frac{\langle \overline{\boldsymbol{\mathcal{E}}}, \overline{\boldsymbol{\mathcal{X}}}\rangle}{n}} \leq \sqrt{\frac{2\mathbb{E}\langle \overline{\boldsymbol{\mathcal{E}}}, \overline{\boldsymbol{\mathcal{X}}}\rangle}{n}} \leq \sqrt{2\rho^+}\eta\sqrt{\frac{|\mathcal{S}_1|}{n}}. \tag{72}$$

□

Building upon Lemma 1 through Lemma 4, we derive Corollary 1.

D Proof of the Corollary 2

We begin by demonstrating that the fiber-wise sparsity penalty can be reformulated as the sum of the ℓ_1 penalty and a concave part. Specifically, we have:

$$R_\lambda\left(\boldsymbol{\mathcal{A}}\right) = \sum_{l=1}^{\prod_{j\neq k} d_j} p_\lambda\left(\left\|\left[\boldsymbol{\mathcal{A}}_{(k)}\right]_{\cdot,l}\right\|_2\right) = \sum_{l=1}^{\prod_{j\neq k} d_j} \lambda\left\|\left[\boldsymbol{\mathcal{A}}_{(k)}\right]_{\cdot,l}\right\|_2 + Q_\lambda\left(\left\|\left[\boldsymbol{\mathcal{A}}_{(k)}\right]_{\cdot,l}\right\|_2\right),$$

where $Q_\lambda\left(\left\|\left[\boldsymbol{\mathcal{A}}_{(k)}\right]_{\cdot,l}\right\|_2\right) = \sum_{l=1}^{\prod_{j\neq k} d_j} q_\lambda\left(\left\|\left[\boldsymbol{\mathcal{A}}_{(k)}\right]_{\cdot,l}\right\|_2\right)$.

Lemma 5. *Under Assumptions 1 and 2, the loss function $\widetilde{L}\left(\boldsymbol{\mathcal{A}}\right)$ satisfies the restricted strong convexity*

$$\widetilde{L}\left(\boldsymbol{\mathcal{A}}'\right) \geq \widetilde{L}\left(\boldsymbol{\mathcal{A}}\right) + \left\langle \nabla\widetilde{L}\left(\boldsymbol{\mathcal{A}}\right), \boldsymbol{\mathcal{A}}' - \boldsymbol{\mathcal{A}}\right\rangle + \frac{\rho^- - \zeta^-}{2}\left\|\boldsymbol{\mathcal{A}}' - \boldsymbol{\mathcal{A}}\right\|_{\mathrm{F}}^2,$$

and the restricted strong smoothness

$$\widetilde{L}\left(\boldsymbol{\mathcal{A}}'\right) \leq \widetilde{L}\left(\boldsymbol{\mathcal{A}}\right) + \left\langle \nabla\widetilde{L}\left(\boldsymbol{\mathcal{A}}\right), \boldsymbol{\mathcal{A}}' - \boldsymbol{\mathcal{A}}\right\rangle + \frac{\rho^+ - \zeta^+}{2}\left\|\boldsymbol{\mathcal{A}}' - \boldsymbol{\mathcal{A}}\right\|_{\mathrm{F}}^2.$$

Proof. Since the proof closely mirrors that of Lemma 1, it is omitted here for brevity. □

Lemma 6. *Let C be a constant. Suppose there exists an integer $\tilde{s}_2 > C\left|\boldsymbol{S}_2\right|$, and that $\boldsymbol{\mathcal{A}}$ satisfies $\left\|\boldsymbol{\mathcal{A}}_{\overline{\boldsymbol{S}_2}}\right\|_0 \leq \tilde{s}_2$, $\omega\left(\boldsymbol{\mathcal{A}}\right) \leq \frac{\lambda}{2}$, where $\omega\left(\boldsymbol{\mathcal{A}}\right) = \min_{\boldsymbol{\mathcal{H}}\in\partial\left\|\left[\boldsymbol{\mathcal{A}}_{(k)}\right]_{\cdot,l}\right\|_2}\left\{\left\|\nabla\widetilde{L}\left(\boldsymbol{\mathcal{A}}\right) + \lambda\boldsymbol{\mathcal{H}}\right\|_{\max}\right\}$, and $\left\|\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right\|_{\max} \leq \lambda/8$. Under Assumptions 1 and 2, $\boldsymbol{\mathcal{A}}$ satisfies*

$$\left\|\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{A}}^*\right\|_{\mathrm{F}} \leq \frac{21/8}{\rho^- - \zeta^-}\lambda\sqrt{\left|\boldsymbol{S}_2\right|}.$$

Proof. We omit the proof here for brevity, as it closely mirrors that of Lemma 2. □

Lemma 7. *Consider the regularization parameter λ and assume that the derivative of the non-convex penalty satisfies $p'_\lambda\left(\left\|\left[\boldsymbol{\mathcal{A}}_{(k)}\right]_{\cdot,l}\right\|_2\right)=0$ whenever $\left\|\left[\boldsymbol{\mathcal{A}}_{(k)}\right]_{\cdot,l}\right\|_2 \geq \nu$ for some $\nu>0$. Let $\boldsymbol{\mathcal{S}}_2^{\mathrm{I}} \cup \boldsymbol{\mathcal{S}}_2^{\mathrm{II}}=\boldsymbol{\mathcal{S}}_2$. For indices $(i_1,\ldots,i_M)\in\boldsymbol{\mathcal{S}}_2^{\mathrm{I}}\subseteq\boldsymbol{\mathcal{S}}_2$, we assume $\min_l\left[\boldsymbol{\mathcal{A}}^*_{(k)}\right]_{\cdot,l}\geq\nu$, and for indices $(i_1,\ldots,i_M)\in\boldsymbol{\mathcal{S}}_2^{\mathrm{II}}\subseteq\boldsymbol{\mathcal{S}}_2$, we assume $\min_l\left[\boldsymbol{\mathcal{A}}^*_{(k)}\right]_{\cdot,l}\leq\nu$. Under Assumptions 1∼4, we derive the following bound:*

$$\left\|\hat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}}^*\right\|_{\mathrm{F}}\leq\frac{1}{\rho^--\zeta^-}\left\|\left(\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right)_{\boldsymbol{\mathcal{S}}_2^{\mathrm{I}}}\right\|_{\mathrm{F}}+\frac{3}{\rho^--\zeta^-}\lambda\sqrt{\left|\boldsymbol{\mathcal{S}}_2^{\mathrm{II}}\right|}.$$

Proof. For brevity, we omit the proof here, as it closely resembles that of Lemma 3. □

Lemma 8. *For least-squares regression with sub-Gaussian noise, we assume that the columns of $\overline{\boldsymbol{\mathcal{X}}}$ are normalized in such a way that $\max_{j\in\{1,\ldots,d_1d_2\ldots d_N\}}\left\|\overline{\boldsymbol{\mathcal{X}}}_{\cdot j}\right\|_2\leq\sqrt{n}$, where $\overline{\boldsymbol{\mathcal{X}}}=\left(\mathrm{vec}\left(\boldsymbol{\mathcal{X}}^{(1)}\right),\ldots,\mathrm{vec}\left(\boldsymbol{\mathcal{X}}^{(n)}\right)\right)^\top$. If $\lambda\asymp\sqrt{\frac{\log(d_k)}{n}}$, then we have*

$$\left\|\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right\|_{\mathrm{F}}\lesssim\sqrt{\frac{|\mathcal{S}_2|}{n}}.$$

Proof. For brevity, the proof is omitted here as it closely follows the methodology established in Lemma 4. □

E Proof of the Corollary 3

We begin by demonstrating that the fiber-wise sparsity penalty can be reformulated as the sum of the ℓ_1 penalty and a concave part. Specifically, we have:

$$R_\lambda\left(\boldsymbol{\mathcal{A}}\right)=\sum_{i=1}^{\prod_{s\neq j,k}d_s}p_\lambda\left(\left\|\left[\boldsymbol{\mathcal{A}}_{(j,k)}\right]_{\cdot,\cdot,l}\right\|_{\mathrm{F}}\right)=\sum_{i=1}^{\prod_{s\neq j,k}d_s}\lambda\left\|\left[\boldsymbol{\mathcal{A}}_{(j,k)}\right]_{\cdot,\cdot,l}\right\|_{\mathrm{F}}+Q_\lambda\left(\left\|\left[\boldsymbol{\mathcal{A}}_{(j,k)}\right]_{\cdot,\cdot,l}\right\|_{\mathrm{F}}\right),$$

where $Q_\lambda\left(\left\|\left[\boldsymbol{\mathcal{A}}_{(j,k)}\right]_{\cdot,\cdot,l}\right\|_{\mathrm{F}}\right)=\sum_{i=1}^{\prod_{s\neq j,k}d_s}q_\lambda\left(\left\|\left[\boldsymbol{\mathcal{A}}_{(j,k)}\right]_{\cdot,\cdot,l}\right\|_{\mathrm{F}}\right)$.

Lemma 9. *Under Assumptions 1 and 2, the loss function $\widetilde{L}\left(\boldsymbol{\mathcal{A}}\right)$ satisfies the restricted strong convexity*

$$\widetilde{L}\left(\boldsymbol{\mathcal{A}}'\right)\geq\widetilde{L}\left(\boldsymbol{\mathcal{A}}\right)+\left\langle\nabla\widetilde{L}\left(\boldsymbol{\mathcal{A}}\right),\boldsymbol{\mathcal{A}}'-\boldsymbol{\mathcal{A}}\right\rangle+\frac{\rho^--\zeta^-}{2}\left\|\boldsymbol{\mathcal{A}}'-\boldsymbol{\mathcal{A}}\right\|_{\mathrm{F}}^2,$$

and the restricted strong smoothness

$$\widetilde{L}\left(\boldsymbol{\mathcal{A}}'\right)\leq\widetilde{L}\left(\boldsymbol{\mathcal{A}}\right)+\left\langle\nabla\widetilde{L}\left(\boldsymbol{\mathcal{A}}\right),\boldsymbol{\mathcal{A}}'-\boldsymbol{\mathcal{A}}\right\rangle+\frac{\rho^+-\zeta^+}{2}\left\|\boldsymbol{\mathcal{A}}'-\boldsymbol{\mathcal{A}}\right\|_{\mathrm{F}}^2.$$

Proof. The proof can be demonstrated similarly to the proof in Lemma 1. Hence, we omit it here. □

Lemma 10. *Let C be a constant. Suppose there exists an integer $\tilde{s}_3>C\left|\boldsymbol{S}_3\right|$, and that $\boldsymbol{\mathcal{A}}$ satisfies $\left\|\boldsymbol{\mathcal{A}}_{\overline{\boldsymbol{S}_3}}\right\|_0\leq\widetilde{s_3}$, $\omega\left(\boldsymbol{\mathcal{A}}\right)\leq\frac{\lambda}{2}$, where $\omega\left(\boldsymbol{\mathcal{A}}\right)=\min_{\boldsymbol{\mathcal{H}}\in\partial\left\|\left[\boldsymbol{\mathcal{A}}_{(j,k)}\right]_{\cdot,\cdot,l}\right\|_{\mathrm{F}}}\left\{\left\|\nabla\widetilde{L}\left(\boldsymbol{\mathcal{A}}\right)+\lambda\boldsymbol{\mathcal{H}}\right\|_{\max}\right\}$, and $\left\|\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right\|_{\max}\leq\lambda/8$. Under Assumptions 1 and 2, $\boldsymbol{\mathcal{A}}$ satisfies*

$$\left\|\boldsymbol{\mathcal{A}}-\boldsymbol{\mathcal{A}}^*\right\|_{\mathrm{F}}\leq\frac{21/8}{\rho^--\zeta^-}\lambda\sqrt{\left|\boldsymbol{S}_3\right|}.$$

Proof. For the sake of brevity, we omit the proof here, as it closely follows that of Lemma 2. □

Lemma 11. *Consider the regularization parameter λ and assume that the derivative of the non-convex penalty satisfies $p'_\lambda\left(\left\|\left[\boldsymbol{\mathcal{A}}_{(j,k)}\right]_{\cdot,\cdot,l}\right\|_{\mathrm{F}}\right)=0$ whenever $\left\|\left[\boldsymbol{\mathcal{A}}_{(j,k)}\right]_{\cdot,\cdot,l}\right\|_{\mathrm{F}}\geq\nu$ for some $\nu>0$. Let $\boldsymbol{\mathcal{S}}_3^{\mathrm{I}}\cup\boldsymbol{\mathcal{S}}_3^{\mathrm{II}}=\boldsymbol{\mathcal{S}}_3$. For indices $(i_1,\ldots,i_M)\in\boldsymbol{\mathcal{S}}_3^{\mathrm{I}}\subseteq\boldsymbol{\mathcal{S}}_3$, we assume $\min_l\left\|\left[\boldsymbol{\mathcal{A}}_{(j,k)}\right]_{\cdot,\cdot,l}\right\|_{\mathrm{F}}\geq\nu$, and for indices $(i_1,\ldots,i_M)\in\boldsymbol{\mathcal{S}}_3^{\mathrm{II}}\subseteq\boldsymbol{\mathcal{S}}_3$, we assume $\min_l\left\|\left[\boldsymbol{\mathcal{A}}_{(j,k)}\right]_{\cdot,\cdot,l}\right\|_{\mathrm{F}}\leq\nu$. Under Assumptions 1∼4, we derive the following bound:*

$$\left\|\hat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}}^*\right\|_{\mathrm{F}}\leq\frac{1}{\rho^--\zeta^-}\left\|\left(\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right)_{\boldsymbol{\mathcal{S}}_3^{\mathrm{I}}}\right\|_{\mathrm{F}}+\frac{3}{\rho^--\zeta^-}\lambda\sqrt{\left|\boldsymbol{\mathcal{S}}_3^{\mathrm{II}}\right|}.$$

Proof. For brevity, we omit the proof here, as it closely resembles that of Lemma 3. □

Lemma 12. *For least-squares regression with sub-Gaussian noise, we assume that the columns of $\overline{\boldsymbol{\mathcal{X}}}$ are normalized in such a way that $\max_{j\in\{1,\ldots,d_1d_2\ldots d_N\}}\left\|\overline{\boldsymbol{\mathcal{X}}}_{\cdot j}\right\|_2\leq\sqrt{n}$, where $\overline{\boldsymbol{\mathcal{X}}}=\left(\mathrm{vec}\left(\boldsymbol{\mathcal{X}}^{(1)}\right),\ldots,\mathrm{vec}\left(\boldsymbol{\mathcal{X}}^{(n)}\right)\right)^\top$. If $\lambda\asymp\sqrt{\frac{\log(d_jd_k)}{n}}$, then we have*

$$\left\|\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right\|_{\mathrm{F}}\lesssim\sqrt{\frac{|\mathcal{S}_3|}{n}}.$$

Proof. For brevity, the proof is omitted here as it closely follows the methodology established in Lemma 4. □

F Proof of the Corollary 4

The proposed mode-wise low-rankness penalty can be reformulated as the sum of a scaled norm and a concave function. Specifically, we have

$$R_\lambda\left(\boldsymbol{\mathcal{A}}\right)=\sum_{i=1}^{\min\left\{I_k,\prod_{j\neq k}d_j\right\}}p_\lambda\left(\sigma_i\left(\boldsymbol{\mathcal{A}}_{(k)}\right)\right)=\lambda\left\|\boldsymbol{\mathcal{A}}_{(k)}\right\|_*+Q_\lambda\left(\boldsymbol{\mathcal{A}}_{(k)}\right),$$

where $\sigma_i\left(\boldsymbol{\mathcal{A}}_{(k)}\right)$ denotes the i-th singular value of the mode-(k) unfolding $\boldsymbol{\mathcal{A}}_{(k)}$. For the estimation problem, we define

$$\tilde{L}\left(\boldsymbol{\mathcal{A}}\right)=L\left(\boldsymbol{\mathcal{A}}\right)+Q_\lambda\left(\boldsymbol{\mathcal{A}}_{(k)}\right),$$

where $Q_\lambda\left(\boldsymbol{\mathcal{A}}_{(k)}\right)=\sum_{i=1}^{\min\left\{I_k,\prod_{j\neq k}d_j\right\}}q_\lambda\left(\sigma_i\left(\boldsymbol{\mathcal{A}}_{(k)}\right)\right)$, and $\min\left\{I_k,\prod_{j\neq k}d_j\right\}$ is the number of the singular values.

Based on the restrict strongly convexity of $L\left(\cdot\right)$ in Assumption 1 and the parameter for regularity condition in Assumption 3, if $\rho^->\zeta^-$, we have the restrict strongly convexity of $\tilde{L}\left(\cdot\right)$.

Besides, for the RSC and RSS assumption, we define the following cone of directions

$$\boldsymbol{\mathcal{C}}=\left\{\boldsymbol{\mathcal{B}}\in\mathbb{R}^{d_1\cdots d_N}\,|\,\left\|\Pi_{\mathcal{F}^\perp}\left(\boldsymbol{\mathcal{B}}\right)\right\|_*\leq5\left\|\Pi_{\mathcal{F}}\left(\boldsymbol{\mathcal{B}}\right)\right\|_*\right\}$$

Lemma 13. *Under Assumption 1, if $\boldsymbol{\mathcal{B}}\in\boldsymbol{\mathcal{C}}$,we have*

$$\tilde{L}\left(\boldsymbol{\mathcal{A}}+\boldsymbol{\mathcal{B}}\right)\geq\tilde{L}\left(\boldsymbol{\mathcal{A}}\right)+\left\langle\nabla\tilde{L}\left(\boldsymbol{\mathcal{A}}\right),\boldsymbol{\mathcal{B}}\right\rangle+\frac{\rho^--\zeta^-}{2}\left\|\boldsymbol{\mathcal{B}}\right\|_{\mathrm{F}}^2.$$

Proof. Based on Assumption 1, we have

$$L\left(\boldsymbol{\mathcal{A}}+\boldsymbol{\mathcal{B}}\right)\leq L\left(\boldsymbol{\mathcal{A}}\right)+\left\langle\nabla L\left(\boldsymbol{\mathcal{A}}\right),\boldsymbol{\mathcal{B}}\right\rangle+\frac{\rho^-}{2}\left\|\boldsymbol{\mathcal{B}}\right\|_{\mathrm{F}}.\tag{73}$$

Moreover, considering the singular values of the unfolded matrices $\boldsymbol{\mathcal{A}}_{(k)}$ and $\boldsymbol{\mathcal{B}}_{(k)}$, we obtain

$$-\zeta^-\leq\frac{q'_\lambda\left(\sigma_i\left(\boldsymbol{\mathcal{A}}_{(k)}\right)\right)-q'_\lambda\left(\sigma_I\left(\left[\boldsymbol{\mathcal{A}}+\boldsymbol{\mathcal{B}}\right]_{(k)}\right)\right)}{\left[\sigma_i\left(\boldsymbol{\mathcal{A}}_{(k)}\right)-\sigma_i\left(\left[\boldsymbol{\mathcal{A}}+\boldsymbol{\mathcal{B}}\right]_{(k)}\right)\right]},\tag{74}$$

which is similar to the proof for Lemma 1. This inequality leads to

$$\left\langle \left(-\nabla Q_\lambda\left(\boldsymbol{\mathcal{A}}_{(k)}\right)\right) - \left(-\nabla Q_\lambda\left(\left[\boldsymbol{\mathcal{A}}+\boldsymbol{\mathcal{B}}\right]_{(k)}\right)\right), \boldsymbol{\mathcal{B}}_{(k)} \right\rangle \leq \zeta^- \left\|\boldsymbol{\mathcal{B}}_{(k)}\right\|_{\mathrm{F}}. \tag{75}$$

This inequality characterizes the strong smoothness of $-Q(\cdot)$, which is equivalent to

$$Q_\lambda\left[\left[\boldsymbol{\mathcal{A}}+\boldsymbol{\mathcal{B}}\right]_{(k)}\right] = Q_\lambda\left(\boldsymbol{\mathcal{A}}_{(k)}+\boldsymbol{\mathcal{B}}_{(k)}\right) \geq Q_\lambda\left(\boldsymbol{\mathcal{A}}_{(k)}\right) + \left\langle \nabla Q_\lambda\left(\boldsymbol{\mathcal{A}}_{(k)}\right), \boldsymbol{\mathcal{B}}_{(k)} \right\rangle - \frac{\zeta^-}{2}\left\|\boldsymbol{\mathcal{B}}_{(k)}\right\|_{\mathrm{F}}^2. \tag{76}$$

Noting that the Frobenius norm satisfies $\|\boldsymbol{\mathcal{B}}_{(k)}\|_{\mathrm{F}}^2 = \|\boldsymbol{\mathcal{B}}\|_{\mathrm{F}}^2$. Let $\boldsymbol{\mathcal{A}}' = \boldsymbol{\mathcal{A}}+\boldsymbol{\mathcal{B}}$, adding (73) and (76), we have

$$\begin{aligned}
\tilde{L}\left(\boldsymbol{\mathcal{A}}'\right) &= L\left(\boldsymbol{\mathcal{A}}'\right) + Q_\lambda\left(\boldsymbol{\mathcal{A}}'_{(k)}\right) && (77)\\
&\geq L\left(\boldsymbol{\mathcal{A}}\right) + Q_\lambda\left(\boldsymbol{\mathcal{A}}_{(k)}\right) + \left\langle \nabla L\left(\boldsymbol{\mathcal{A}}\right), \boldsymbol{\mathcal{B}} \right\rangle + \left\langle \nabla Q_\lambda\left(\boldsymbol{\mathcal{A}}_{(k)}\right), \boldsymbol{\mathcal{B}}_{(k)} \right\rangle + \frac{\rho^- - \zeta^-}{2}\|\boldsymbol{\mathcal{B}}\|_{\mathrm{F}}^2 && (78)\\
&= \tilde{L}\left(\boldsymbol{\mathcal{A}}\right) + \left\langle \nabla L\left(\boldsymbol{\mathcal{A}}\right), \boldsymbol{\mathcal{B}} \right\rangle + \left\langle \nabla Q_\lambda\left(\boldsymbol{\mathcal{A}}_{(k)}\right), \boldsymbol{\mathcal{B}}_{(k)} \right\rangle + \frac{\rho^- - \zeta^-}{2}\|\boldsymbol{\mathcal{B}}\|_{\mathrm{F}}^2. && (79)
\end{aligned}$$

□

Lemma 14. *Under Assumption 1, if $\rho^- > \zeta^-$ and the regularization parameter $\lambda \geq \frac{\left\|\mathfrak{X}^*\left(\boldsymbol{\mathcal{E}}\right)_{(k)}\right\|_{\mathrm{sp}}}{2n}$, we have*

$$\left\|\Pi_{\mathcal{F}^\perp}\left(\hat{\boldsymbol{\mathcal{A}}}_{(k)} - \boldsymbol{\mathcal{A}}^*_{(k)}\right)\right\|_* \leq 5\left\|\Pi_{\mathcal{F}}\left(\hat{\boldsymbol{\mathcal{A}}}_{(k)} - \boldsymbol{\mathcal{A}}^*_{(k)}\right)\right\|_*.$$

Proof. By Lemma 13 , we have

$$\tilde{L}\left(\hat{\boldsymbol{\mathcal{A}}}\right) - \tilde{L}\left(\boldsymbol{\mathcal{A}}^*\right) \geq \left\langle \nabla L\left(\boldsymbol{\mathcal{A}}^*\right), \hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^* \right\rangle + \left\langle \nabla Q\left(\boldsymbol{\mathcal{A}}^*_{(k)}\right), \left[\hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^*\right]_{(k)} \right\rangle. \tag{80}$$

We proceed to bound the right-hand side of inequality (80). By decomposing the inner products using the projections onto two orthogonal subspaces, we have

$$\begin{aligned}
&\left\langle \nabla L\left(\boldsymbol{\mathcal{A}}^*\right), \hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^* \right\rangle + \left\langle \nabla Q\left(\boldsymbol{\mathcal{A}}^*_{(k)}\right), \left[\hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^*\right]_{(k)} \right\rangle\\
=&\left\langle \left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)} + \nabla Q\left(\boldsymbol{\mathcal{A}}^*_{(k)}\right), \Pi_{\mathcal{F}}\left(\left[\hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^*\right]_{(k)}\right) \right\rangle + \left\langle \left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)} + \nabla Q\left(\boldsymbol{\mathcal{A}}^*_{(k)}\right), \Pi_{\mathcal{F}^\perp}\left(\left[\hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^*\right]_{(k)}\right) \right\rangle\\
\geq& -\left\|\Pi_{\mathcal{F}}\left(\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)} + \nabla Q\left(\boldsymbol{\mathcal{A}}^*_{(k)}\right)\right)\right\|_{\mathrm{sp}} \left\|\Pi_{\mathcal{F}}\left(\left[\hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^*\right]_{(k)}\right)\right\|_* && (81)\\
&-\left\|\Pi_{\mathcal{F}^\perp}\left(\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)} + \nabla Q\left(\boldsymbol{\mathcal{A}}^*_{(k)}\right)\right)\right\|_{\mathrm{sp}} \left\|\Pi_{\mathcal{F}^\perp}\left(\left[\hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^*\right]_{(k)}\right)\right\|_*. && (82)
\end{aligned}$$

For (81), due to $\lambda \geq \frac{1}{2n}\left\|\left[\mathfrak{X}^*\left(\boldsymbol{\mathcal{E}}\right)\right]_{(k)}\right\|_{\mathrm{sp}}$, we see that $\left\|\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)}\right\|_{\mathrm{sp}} \leq \lambda/2$. According to Assumption 3, we have

$$\left\|\Pi_{\mathcal{F}}\left(\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)} + \nabla Q\left(\boldsymbol{\mathcal{A}}^*_{(k)}\right)\right)\right\|_{\mathrm{sp}} \leq \frac{3}{2}\lambda. \tag{83}$$

For (82), since $\Pi_{\mathcal{F}^\perp}\left(\boldsymbol{\mathcal{A}}^*_{(k)}\right) = 0$, we obtain

$$\left\|\Pi_{\mathcal{F}^\perp}\left(\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)} + \nabla Q\left(\boldsymbol{\mathcal{A}}^*_{(k)}\right)\right)\right\|_{\mathrm{sp}} \leq \frac{1}{2}\lambda. \tag{84}$$

Combine (83) and (84), we have

$$\begin{aligned}
&\left\langle \nabla L\left(\boldsymbol{\mathcal{A}}^*\right), \hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^* \right\rangle + \left\langle \nabla Q\left(\boldsymbol{\mathcal{A}}^*_{(k)}\right), \left[\hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^*\right]_{(k)} \right\rangle && (85)\\
\geq& -\frac{3}{2}\lambda\left\|\Pi_{\mathcal{F}}\left(\left[\hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^*\right]_{(k)}\right)\right\|_* - \frac{1}{2}\lambda\left\|\Pi_{\mathcal{F}^\perp}\left(\left[\hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^*\right]_{(k)}\right)\right\|_* && (86)
\end{aligned}$$

Moreover, noting that $\lambda\left\|\hat{\boldsymbol{\mathcal{A}}}\right\|_* - \lambda\left\|\boldsymbol{\mathcal{A}}^*\right\|_* \geq -\lambda\left\|\Pi_{\mathcal{F}}\left(\left[\hat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}}^*\right]_{(k)}\right)\right\| + \lambda\left\|\Pi_{\mathcal{F}^\perp}\left(\left[\hat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}}^*\right]_{(k)}\right)\right\|_*$, and combining with (80) , we obtain

$$\left\langle \nabla L\left(\boldsymbol{\mathcal{A}}^*\right) + \nabla Q\left(\boldsymbol{\mathcal{A}}^*_{(k)}\right), \hat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}}^*\right\rangle + \lambda\left\|\hat{\boldsymbol{\mathcal{A}}}\right\|_* - \lambda\left\|\boldsymbol{\mathcal{A}}^*\right\|_* \tag{87}$$

$$\geq -\frac{5}{2}\lambda\left\|\Pi_{\mathcal{F}}\left(\left[\hat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}}^*\right]_{(k)}\right)\right\|_* + \frac{1}{2}\lambda\left\|\Pi_{\mathcal{F}^\perp}\left(\left[\hat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}}^*\right]_{(k)}\right)\right\|_*. \tag{88}$$

Since $\hat{\boldsymbol{\mathcal{A}}}$ is the global minimizer of the general estimator (2) and given that $\rho^- > \zeta^-$, it follows that

$$\widetilde{L}\left(\hat{\boldsymbol{\mathcal{A}}}\right) + \lambda\left\|\hat{\boldsymbol{\mathcal{A}}}\right\|_* - \widetilde{L}\left(\boldsymbol{\mathcal{A}}^*\right) - \lambda\left\|\boldsymbol{\mathcal{A}}^*\right\|_* \leq 0. \tag{89}$$

Substituting (80) and (89) into (88), we obtain

$$\frac{1}{2}\lambda\left\|\Pi_{\mathcal{F}^\perp}\left(\left[\hat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}}^*\right]_{(k)}\right)\right\|_* \leq \frac{5}{2}\lambda\left\|\Pi_{\mathcal{F}}\left(\left[\hat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}}^*\right]_{(k)}\right)\right\|_*. \tag{90}$$

Since $\lambda > 0$, we obtain

$$\left\|\Pi_{\mathcal{F}^\perp}\left(\left[\hat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}}^*\right]_{(k)}\right)\right\|_* \leq 5\left\|\Pi_{\mathcal{F}}\left(\left[\hat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}}^*\right]_{(k)}\right)\right\|_*. \tag{91}$$

□

Lemma 15. *Considering the mode-wise low-rankness regularizer, under Assumptions 1∼4, for the estimated parameter tensor $\hat{\boldsymbol{\mathcal{A}}}$ and the true parameter tensor $\boldsymbol{\mathcal{A}}^*$, we have*

$$\left\|\hat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}}^*\right\|_{\mathrm{F}} \leq \frac{1}{(\rho^- - \zeta^-)}\left[\sqrt{\left|\mathcal{S}_4^{\mathrm{I}}\right|}\left\|\Pi_{\mathcal{F}}\left(\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)}\right)\right\|_{\mathrm{sp}} + 3\lambda\sqrt{\left|\mathcal{S}_4^{\mathrm{II}}\right|}\right].$$

where $\mathcal{S}_4^{\mathrm{I}}$ and $\mathcal{S}_4^{\mathrm{II}}$ are subsets of the support set of $\mathcal{S}_4$. The set $\mathcal{S}_4^{\mathrm{I}}$ include all indices $i \in \mathcal{S}_4^{\mathrm{I}}$ which satisfy $\sigma_i\left(\boldsymbol{\mathcal{A}}^*_{(k)}\right) \geq \nu$, and $\mathcal{S}_4^{\mathrm{II}}$ includes all indices with $\sigma_i\left(\boldsymbol{\mathcal{A}}^*_{(k)}\right) < \nu$.

Proof. Since $\|\cdot\|_*$ is convex, we have

$$\lambda\left\|\hat{\boldsymbol{\mathcal{A}}}_{(k)}\right\|_* \geq \lambda\left\|\boldsymbol{\mathcal{A}}^*_{(k)}\right\|_* + \lambda\left\langle\left[\hat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}}^*\right]_{(k)}, \boldsymbol{\mathcal{H}}^*\right\rangle. \tag{92}$$

$$\lambda\left\|\boldsymbol{\mathcal{A}}^*_{(k)}\right\|_* \geq \lambda\left\|\hat{\boldsymbol{\mathcal{A}}}_{(k)}\right\|_* + \lambda\left\langle\left[\boldsymbol{\mathcal{A}}^*-\hat{\boldsymbol{\mathcal{A}}}\right]_{(k)}, \hat{\boldsymbol{\mathcal{H}}}\right\rangle. \tag{93}$$

where $\boldsymbol{\mathcal{H}}^* \in \partial\|\boldsymbol{\mathcal{A}}^*_{(k)}\|_*$ and $\hat{\boldsymbol{\mathcal{H}}} \in \partial\|\hat{\boldsymbol{\mathcal{A}}}_{(k)}\|_*$. From (92) and (93), we have

$$\lambda\left\|\hat{\boldsymbol{\mathcal{A}}}_{(k)}\right\|_* + \lambda\left\|\boldsymbol{\mathcal{A}}^*_{(k)}\right\|_* \geq \lambda\left\|\boldsymbol{\mathcal{A}}^*_{(k)}\right\|_* + \lambda\left\|\hat{\boldsymbol{\mathcal{A}}}_{(k)}\right\|_* + \lambda\left\langle\left[\hat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}}^*\right]_{(k)}, \boldsymbol{\mathcal{H}}^*\right\rangle + \lambda\left\langle\left[\boldsymbol{\mathcal{A}}^*-\hat{\boldsymbol{\mathcal{A}}}\right]_{(k)}, \hat{\boldsymbol{\mathcal{H}}}\right\rangle. \tag{94}$$

This equals to

$$0 \geq \left(\lambda\left\langle\left[\hat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}}^*\right]_{(k)}, \boldsymbol{\mathcal{H}}^*\right\rangle + \lambda\left\langle\left[\boldsymbol{\mathcal{A}}^*-\hat{\boldsymbol{\mathcal{A}}}\right]_{(k)}, \hat{\boldsymbol{\mathcal{H}}}\right\rangle\right). \tag{95}$$

Moreover, according to Lemma 13, we have

$$\widetilde{L}\left(\hat{\boldsymbol{\mathcal{A}}}\right) \geq \widetilde{L}\left(\boldsymbol{\mathcal{A}}^*\right) + \left\langle\nabla L\left(\boldsymbol{\mathcal{A}}^*\right), \hat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}}^*\right\rangle + \left\langle\nabla Q_\lambda\left(\boldsymbol{\mathcal{A}}^*_{(k)}\right), \left[\hat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}}^*\right]_{(k)}\right\rangle + \frac{\rho^- - \zeta^-}{2}\left\|\hat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}}^*\right\|_{\mathrm{F}}^2. \tag{96}$$

$$\widetilde{L}\left(\boldsymbol{\mathcal{A}}^*\right) \geq \widetilde{L}\left(\hat{\boldsymbol{\mathcal{A}}}\right) + \left\langle \nabla L\left(\hat{\boldsymbol{\mathcal{A}}}\right), \boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}} \right\rangle + \left\langle \nabla Q_\lambda\left(\hat{\boldsymbol{\mathcal{A}}}_{(k)}\right), \left[\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}\right]_{(k)} \right\rangle + \frac{\rho^- - \zeta^-}{2} \left\|\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}\right\|_{\mathrm{F}}^2. \tag{97}$$

Summing (95), (96), and (97), we have

$$\begin{aligned} 0 \geq & \left\langle \nabla L\left(\boldsymbol{\mathcal{A}}^*\right), \hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^* \right\rangle + \left\langle \nabla L\left(\hat{\boldsymbol{\mathcal{A}}}\right), \boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}} \right\rangle + \left(\rho^- - \zeta^-\right) \left\|\hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^*\right\|_{\mathrm{F}}^2 \\ & + \left(\left\langle \nabla Q_\lambda\left(\boldsymbol{\mathcal{A}}^*_{(k)}\right) + \lambda \boldsymbol{\mathcal{H}}^*, \left[\hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^*\right]_{(k)} \right\rangle + \left\langle \nabla Q_\lambda\left(\hat{\boldsymbol{\mathcal{A}}}_{(k)}\right) + \lambda \hat{\boldsymbol{\mathcal{H}}}, \left[\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}\right]_{(k)} \right\rangle \right). \end{aligned} \tag{98}$$

Since $\hat{\boldsymbol{\mathcal{A}}}$ is the solution to the estimation problem and $\hat{\boldsymbol{\mathcal{A}}}$ satisfies the optimality condition, for any $\boldsymbol{\mathcal{A}}' \in \mathbb{R}^{d_1 \times \cdots \times d_N}$, it holds that

$$\max_{\boldsymbol{\mathcal{A}}'} \left\{ \left\langle \nabla L\left(\hat{\boldsymbol{\mathcal{A}}}\right), \hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}' \right\rangle + \left\langle \nabla Q_\lambda\left(\hat{\boldsymbol{\mathcal{A}}}_{(k)}\right) + \lambda \hat{\boldsymbol{\mathcal{H}}}, \left[\hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}'\right]_{(k)} \right\rangle \right\} \leq 0. \tag{99}$$

which implies

$$\left\langle \nabla L\left(\hat{\boldsymbol{\mathcal{A}}}\right), \boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}} \right\rangle + \left\langle \nabla Q_\lambda\left(\hat{\boldsymbol{\mathcal{A}}}_{(k)}\right) + \lambda \hat{\boldsymbol{\mathcal{H}}}, \left[\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}\right]_{(k)} \right\rangle \geq 0. \tag{100}$$

Since $\left\langle \nabla L\left(\boldsymbol{\mathcal{A}}^*\right), \boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}} \right\rangle = \left\langle \left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)}, \left[\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}\right]_{(k)} \right\rangle$, we have

$$\begin{aligned} \left(\rho^- - \zeta^-\right) \left\|\hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^*\right\|_{\mathrm{F}}^2 &\leq \left[\left\langle \left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)}, \left[\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}\right]_{(k)} \right\rangle + \left\langle \nabla Q_\lambda\left(\boldsymbol{\mathcal{A}}^*_{(k)}\right) + \lambda \boldsymbol{\mathcal{H}}^*, \left[\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}\right]_{(k)} \right\rangle \right] \\ &\leq \left\langle \Pi_{\mathcal{F}^\perp} \left[\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)} + \nabla Q_\lambda\left(\boldsymbol{\mathcal{A}}^*_{(k)}\right) + \lambda \boldsymbol{\mathcal{H}}^* \right], \left[\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}\right]_{(k)} \right\rangle \\ &\quad + \left\langle \Pi_{\mathcal{F}} \left(\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)} + \nabla Q_\lambda\left(\boldsymbol{\mathcal{A}}^*_{(k)}\right) + \lambda \boldsymbol{\mathcal{H}}^* \right), \left[\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}\right]_{(k)} \right\rangle. \end{aligned} \tag{101}$$

We have defined $\sigma_i\left(\boldsymbol{\mathcal{A}}^*_{(k)}\right)$ as the i-th singular value of tensor $\boldsymbol{\mathcal{A}}^*$. With regard to the magnitudes of the singular values of $\boldsymbol{\mathcal{A}}^*$, we can decompose (101) into three parts:

- $i \in \mathcal{S}_4^{\mathrm{I}}$ that $\sigma_i\left(\boldsymbol{\mathcal{A}}^*_{(k)}\right) \geq \nu$,
- $i \in \mathcal{S}_4^{\mathrm{II}}$ that $\nu \geq \sigma_i\left(\boldsymbol{\mathcal{A}}^*_{(k)}\right) > 0$,
- $i \in \mathcal{S}_4^{\mathrm{c}}$ that $\sigma_i\left(\boldsymbol{\mathcal{A}}^*_{(k)}\right) = 0$.

(i) For $i \in \mathcal{S}_4^{\mathrm{I}}$ that $\sigma_i\left(\boldsymbol{\mathcal{A}}^*_{(k)}\right) \geq \nu$, define a subspace of $\mathcal{F}$ associated with $\mathcal{S}_4^{\mathrm{I}}$ as follows

$$\mathcal{F}_{\mathcal{S}_4^{\mathrm{I}}}\left(\mathbf{U}^*, \mathbf{V}^*\right) := \left\{ \mathbf{W} \middle| \operatorname{row}\left(\mathbf{W}\right) \subset \mathbf{V}_{\mathrm{I}}^*, \operatorname{col}\left(\mathbf{W}\right) \subset \mathbf{U}_{\mathrm{I}}^* \right\}, \tag{102}$$

where $\mathbf{V}_{\mathrm{I}}^*$ and $\mathbf{U}_{\mathrm{I}}^*$ is the matrix with the i-th row of $\mathbf{V}_{\mathrm{I}}^*$ and $\mathbf{U}_{\mathrm{I}}^*$ with $i \in \mathcal{S}_4^{\mathrm{I}}$.

Recall that $R_\lambda\left(\boldsymbol{\mathcal{A}}^*_{(k)}\right) = \lambda \left\|\boldsymbol{\mathcal{A}}^*_{(k)}\right\|_* + Q_\lambda\left(\boldsymbol{\mathcal{A}}^*_{(k)}\right)$, we have

$$\nabla R_\lambda\left(\boldsymbol{\mathcal{A}}^*_{(k)}\right) = \nabla Q_\lambda\left(\boldsymbol{\mathcal{A}}^*_{(k)}\right) + \lambda_k \left(\mathbf{U}_{\mathrm{I}}^* \mathbf{V}_{\mathrm{I}}^{*\top} + \mathbf{Z}_{\mathrm{I}}^*\right), \tag{103}$$

where $\mathbf{Z}_{\mathrm{I}}^{*}=-\lambda^{-1}\Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{I}}}}\left(\left[\nabla L\left(\boldsymbol{\mathcal{A}}^{*}\right)\right]_{(k)}\right)$. Since $\|\mathbf{Z}_{\mathrm{I}}^{*}\|\leq 1$ and $\mathbf{Z}_{\mathrm{I}}^{*}\in\mathcal{F}_{\mathcal{S}_4^{\mathrm{I}}}$, which satisfies the condition of $\mathbf{W}^{*}$ to be sub-gradient of $\left\|\boldsymbol{\mathcal{A}}_{(k)}^{*}\right\|$. Projecting $R_\lambda\left(\boldsymbol{\mathcal{A}}_{(k)}^{*}\right)$ into the subspace $\mathcal{F}_{\mathcal{S}_4^{\mathrm{I}}}$, we have

$$\Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{I}}}}\left(\nabla R_\lambda\left(\boldsymbol{\mathcal{A}}_{(k)}^{*}\right)\right)=\Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{I}}}}\left(\nabla Q_\lambda\left(\boldsymbol{\mathcal{A}}_{(k)}^{*}\right)+\lambda\mathbf{U}_{\mathrm{I}}^{*}\mathbf{V}_{\mathrm{I}}^{*\top}+\lambda\mathbf{Z}_{\mathrm{I}}^{*}\right) \tag{104}$$

$$=\mathbf{U}_{\mathrm{I}}^{*}q_\lambda'\left(\mathbf{\Sigma}_{\mathrm{I}}^{*}\right)\mathbf{V}_{\mathrm{I}}^{*\top}+\lambda\mathbf{U}_{\mathrm{I}}^{*}\mathbf{V}_{\mathrm{I}}^{*\top} \tag{105}$$

$$=\mathbf{U}_{\mathrm{I}}^{*}\left[q_\lambda'\left(\mathbf{\Sigma}_{\mathrm{I}}^{*}\right)+\lambda\mathbf{I}_{\mathrm{I}}\right]\mathbf{V}_{\mathrm{I}}^{*\top}, \tag{106}$$

where $\mathbf{I}_{\mathrm{I}}$ is an identity matrix with the size $\min\{d_k,\Pi_{j\neq k}d_j\}$ and $\left(q_\lambda'\left(\mathbf{\Sigma}_{\mathrm{I}}^{*}\right)+\lambda\mathbf{I}_{\mathrm{I}}\right)$ is a diagonal matrix that for $i\notin\mathcal{S}_4^{\mathrm{I}}$, the i-th element on the diagonal equals 0, i.e. $\left[q_\lambda'\left(\mathbf{\Sigma}_{\mathrm{I}}^{*}\right)+\lambda\mathbf{I}_{\mathrm{I}}\right]_{ii}=0$, and for all $i\in\mathcal{S}_4^{\mathrm{I}}$, we have

$$\left[q_\lambda'\left(\mathbf{\Sigma}_{\mathrm{I}}^{*}\right)+\lambda\mathbf{I}_{\mathrm{I}}\right]_{ii}=q_\lambda'\left(\sigma_i\left(\boldsymbol{\mathcal{A}}_{(k)}^{*}\right)\right)+\lambda=p_\lambda'\left(\sigma_i\left(\boldsymbol{\mathcal{A}}_{(k)}^{*}\right)\right)=0. \tag{107}$$

The last equality is derived from fact that $i\in\mathcal{S}_4^{\mathrm{I}}$ satisfies Assumption 3, $p_\lambda'(t)=0$. Therefore, we have $q_\lambda'\left(\mathbf{\Sigma}_{\mathrm{I}}^{*}\right)+\lambda\mathbf{I}_{\mathrm{I}}=0$, which indicates that $\Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{I}}}}\left(\nabla R_\lambda\left(\boldsymbol{\mathcal{A}}_{(k)}^{*}\right)\right)=0$. For $\boldsymbol{\mathcal{H}}^{*}=\mathbf{U}_{\mathrm{I}}^{*}\mathbf{V}_{\mathrm{I}}^{*\top}+\mathbf{Z}_{\mathrm{I}}^{*}\in\partial\left\|\boldsymbol{\mathcal{A}}_{(k)}^{*}\right\|_{*}$, we have

$$\left\langle\Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{I}}}}\left[\left[\nabla L\left(\boldsymbol{\mathcal{A}}^{*}\right)\right]_{(k)}+\lambda\boldsymbol{\mathcal{H}}^{*}+\nabla Q_\lambda\left(\boldsymbol{\mathcal{A}}_{(k)}^{*}\right)\right],\left[\boldsymbol{\mathcal{A}}^{*}-\hat{\boldsymbol{\mathcal{A}}}\right]_{(k)}\right\rangle \tag{108}$$

$$=\left\langle\Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{I}}}}\left[\left[\nabla L\left(\boldsymbol{\mathcal{A}}^{*}\right)\right]_{(k)}+\nabla R_\lambda\left(\boldsymbol{\mathcal{A}}_{(k)}^{*}\right)\right],\left[\boldsymbol{\mathcal{A}}^{*}-\hat{\boldsymbol{\mathcal{A}}}\right]_{(k)}\right\rangle \tag{109}$$

$$=\left\langle\Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{I}}}}\left(\left[\nabla L\left(\boldsymbol{\mathcal{A}}^{*}\right)\right]_{(k)}\right),\Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{I}}}}\left(\left[\boldsymbol{\mathcal{A}}^{*}-\hat{\boldsymbol{\mathcal{A}}}\right]_{(k)}\right)\right\rangle \tag{110}$$

$$\leq\left\|\Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{I}}}}\left(\left[\nabla L\left(\boldsymbol{\mathcal{A}}^{*}\right)\right]_{(k)}\right)\right\|_{\mathrm{sp}}\cdot\left\|\Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{I}}}}\left(\left[\boldsymbol{\mathcal{A}}^{*}-\hat{\boldsymbol{\mathcal{A}}}\right]_{(k)}\right)\right\|_{*}, \tag{111}$$

where the last inequality is derived from the Hölder inequality. For $\left\|\Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{I}}}}\left(\left[\boldsymbol{\mathcal{A}}^{*}-\hat{\boldsymbol{\mathcal{A}}}\right]_{(k)}\right)\right\|_{*}$, from the properties of projection on to the subspace $\mathcal{F}_{\mathcal{S}_4^{\mathrm{I}}}$, we have

$$\left\|\Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{I}}}}\left(\left[\boldsymbol{\mathcal{A}}^{*}-\hat{\boldsymbol{\mathcal{A}}}\right]_{(k)}\right)\right\|_{*}\leq\sqrt{\left|\mathcal{S}_4^{\mathrm{I}}\right|}\left\|\Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{I}}}}\left(\left[\boldsymbol{\mathcal{A}}^{*}-\hat{\boldsymbol{\mathcal{A}}}\right]_{(k)}\right)\right\|_{\mathrm{F}} \tag{112}$$

$$\leq\sqrt{\left|\mathcal{S}_4^{\mathrm{I}}\right|}\left\|\left[\boldsymbol{\mathcal{A}}^{*}-\hat{\boldsymbol{\mathcal{A}}}\right]_{(k)}\right\|_{\mathrm{F}}=\sqrt{\left|\mathcal{S}_4^{\mathrm{I}}\right|}\left\|\boldsymbol{\mathcal{A}}^{*}-\hat{\boldsymbol{\mathcal{A}}}\right\|_{\mathrm{F}}. \tag{113}$$

We obtain the second inequality from that the rank of the matrix $\Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{I}}}}\left(\left[\boldsymbol{\mathcal{A}}^{*}-\hat{\boldsymbol{\mathcal{A}}}\right]_{(k)}\right)\leq\left|\mathcal{S}_4^{\mathrm{I}}\right|$.

Thus, we have

$$\left\langle\Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{I}}}}\left[\left[\nabla L\left(\boldsymbol{\mathcal{A}}^{*}\right)\right]_{(k)}+\nabla Q_\lambda\left(\boldsymbol{\mathcal{A}}_{(k)}^{*}\right)+\lambda\boldsymbol{\mathcal{H}}^{*}\right],\left[\boldsymbol{\mathcal{A}}^{*}-\hat{\boldsymbol{\mathcal{A}}}\right]_{(k)}\right\rangle$$

$$\leq\sqrt{\left|\mathcal{S}_4^{\mathrm{I}}\right|}\left\|\Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{I}}}}\left(\left[\boldsymbol{\mathcal{A}}^{*}-\hat{\boldsymbol{\mathcal{A}}}\right]_{(k)}\right)\right\|_{\mathrm{sp}}\cdot\left\|\boldsymbol{\mathcal{A}}^{*}-\hat{\boldsymbol{\mathcal{A}}}\right\|_{\mathrm{F}}. \tag{114}$$

(ii) For $i\in\mathcal{S}_4^{\mathrm{II}}$, $\nu\geq\sigma_i\left(\boldsymbol{\mathcal{A}}_{(k)}^{*}\right)>0$, define a subspace of $\mathcal{F}$ associated with $\mathcal{S}_4^{\mathrm{II}}$ as follows

$$\mathcal{F}_{\mathcal{S}_4^{\mathrm{II}}}\left(\mathbf{U}^{*},\mathbf{V}^{*}\right):=\left\{\mathbf{W}\middle|\,\mathrm{row}\left(\mathbf{W}\right)\subset\mathbf{V}_{\mathrm{II}}^{*},\mathrm{col}\left(\mathbf{W}\right)\subset\mathbf{U}_{\mathrm{II}}^{*}\right\}.$$

where $\mathbf{V}_{\mathrm{II}}^{*}$ and $\mathbf{U}_{\mathrm{II}}^{*}$ is the matrix with the i-th row of $\mathbf{U}^{*}$ and $\mathbf{V}^{*}$ with $i\in\mathcal{S}_4^{\mathrm{II}}$. Obviously, for all $\mathbf{W}$, the following decomposition holds

$$\Pi_{\mathcal{F}}\left(\mathbf{W}\right)=\Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{I}}}}\left(\mathbf{W}\right)+\Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{II}}}}\left(\mathbf{W}\right).$$

In addition, since $\mathbf{U}^*$, $\mathbf{V}^*$ are unitary matrices, for subspace $\mathcal{F}_{\mathcal{S}_4^{\mathrm{I}}}$ and $\mathcal{F}_{\mathcal{S}_4^{\mathrm{II}}}$, we have the complementary subspace $\mathcal{F}_{\mathcal{S}_4^{\mathrm{I}}}^{\perp}$, $\mathcal{F}_{\mathcal{S}_4^{\mathrm{II}}}^{\perp}$, thus we have

$$\mathcal{F}_{\mathcal{S}_4^{\mathrm{I}}} \subset \mathcal{F}_{\mathcal{S}_4^{\mathrm{I}}}^{\perp}, \text{ and } \mathcal{F}_{\mathcal{S}_4^{\mathrm{II}}} \subset \mathcal{F}_{\mathcal{S}_4^{\mathrm{II}}}^{\perp}.$$

Similar to analysis in (i) on $\mathcal{S}_4^{\mathrm{I}}$, we have

$$\Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{II}}}}\left(\nabla Q_\lambda\left(\boldsymbol{\mathcal{A}}_{(k)}^*\right)\right) = \mathbf{U}_{\mathrm{II}}^* q_\lambda'\left(\boldsymbol{\Sigma}_{\mathrm{II}}^*\right)\mathbf{V}_{\mathrm{II}}^{*\top}. \tag{115}$$

where $q_\lambda'\left(\boldsymbol{\Sigma}_{\mathrm{II}}^*\right)$ is a diagonal matrix that $\left[q_\lambda'\left(\boldsymbol{\Sigma}_{\mathrm{II}}^*\right)\right]_{ii} = 0$ for $i \notin \mathcal{S}_4^{\mathrm{II}}$, and for all $i \in \mathcal{S}_4^{\mathrm{II}}$,

$$\left[q_\lambda'\left(\boldsymbol{\Sigma}_{\mathrm{II}}^*\right)\right]_{ii} = \left[q_\lambda'\left(\sigma_i\left(\boldsymbol{\mathcal{A}}_{(k)}^*\right)\right)\right]_{ii} \leq \lambda. \tag{116}$$

Since $\sigma_i\left(\boldsymbol{\mathcal{A}}_{(k)}^*\right) \leq \nu$, and $q_\lambda(\cdot)$ satisfies the regularity Assumption 3, $\left|q_\lambda'(t)\right| \leq \lambda$. Therefore

$$\left\|\Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{II}}}}\left(\nabla Q_\lambda\left(\boldsymbol{\mathcal{A}}_{(k)}^*\right)\right)\right\|_{\mathrm{sp}} = \max_{i\in\mathcal{S}_4^{\mathrm{II}}}\left[q_\lambda'\left(\boldsymbol{\Sigma}_{\mathrm{II}}^*\right)\right]_{ii} \leq \lambda. \tag{117}$$

Meanwhile, because of the fact that $\mathcal{F}_{\mathcal{S}_4^{\mathrm{II}}} \subset \mathcal{F}_{\mathcal{S}_4}$, we have

$$\left\|\Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{II}}}}\left(\lambda\boldsymbol{\mathcal{H}}^*\right)\right\|_{\mathrm{sp}} \leq \left\|\Pi_{\mathcal{F}}\left(\lambda\mathbf{U}_{\mathrm{II}}^*\mathbf{V}_{\mathrm{II}}^{*\top}\right)\right\|_{\mathrm{sp}}. \tag{118}$$

Since $\left\|\mathbf{U}^*\mathbf{V}^{*\top}\right\|_{\mathrm{sp}} = 1$, we have

$$\left\|\Pi_{\mathcal{F}}\left(\lambda\mathbf{U}_{\mathrm{II}}^*\mathbf{V}_{\mathrm{II}}^{*\top}\right)\right\|_{\mathrm{sp}} = \lambda. \tag{119}$$

Thus, from (118) and (119), we have

$$\left\|\Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{II}}}}\left(\lambda\boldsymbol{\mathcal{H}}^*\right)\right\|_{\mathrm{sp}} \leq \lambda. \tag{120}$$

Additionally, due to the fact that $\left\|\Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{II}}}}\left(\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)}\right)\right\|_{\mathrm{sp}} \leq \left\|\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)}\right\|_{\mathrm{sp}} \leq \lambda$, which indicates that

$$\begin{aligned}
&\left\langle \Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{II}}}}\left(\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)} + \nabla Q_\lambda\left(\boldsymbol{\mathcal{A}}_{(k)}^*\right) + \lambda\boldsymbol{\mathcal{H}}^*\right), \left[\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}\right]_{(k)} \right\rangle \\
&= \left\langle \Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{II}}}}\left(\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)}\right), \left[\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}\right]_{(k)} \right\rangle + \left\langle \Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{II}}}}\left(\nabla Q_\lambda\left(\boldsymbol{\mathcal{A}}_{(k)}^*\right)\right), \left[\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}\right]_{(k)} \right\rangle + \left\langle \Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{II}}}}\left(\lambda\boldsymbol{\mathcal{H}}^*\right), \left[\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}\right]_{(k)} \right\rangle \\
&\leq \left(\left\|\Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{II}}}}\left(\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)}\right)\right\|_{\mathrm{sp}} + \left\|\Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{II}}}}\left(\nabla Q_\lambda\left(\boldsymbol{\mathcal{A}}_{(k)}^*\right)\right)\right\|_{\mathrm{sp}} + \left\|\Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{II}}}}\left(\lambda\boldsymbol{\mathcal{H}}^*\right)\right\|_{\mathrm{sp}} \right) \cdot \left\|\Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{II}}}}\left(\left[\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}\right]_{(k)}\right)\right\|_*,
\end{aligned}$$

where the last inequality is derived from the Hölder inequality. Since we have obtained the bound for each term, as in (119) and (120), we have

$$\begin{aligned}
\left\langle \Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{II}}}}\left(\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)} + \nabla Q_\lambda\left(\boldsymbol{\mathcal{A}}_{(k)}^*\right) + \lambda\boldsymbol{\mathcal{H}}^*\right), \left[\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}\right]_{(k)} \right\rangle &\leq 3\lambda \left\|\Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{II}}}}\left(\left[\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}\right]_{(k)}\right)\right\|_* \\
&\leq 3\lambda\sqrt{\left|\mathcal{S}_4^{\mathrm{II}}\right|}\left\|\left[\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}\right]_{(k)}\right\|_{\mathrm{F}} \\
&= 3\lambda\sqrt{\left|\mathcal{S}_4^{\mathrm{II}}\right|}\left\|\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}\right\|_{\mathrm{F}}.
\end{aligned} \tag{121}$$

where the second inequality utilizes the fact that $\operatorname{rank}\left(\Pi_{\mathcal{F}_{\mathcal{S}_4^{\mathrm{II}}}}\left(\left[\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}\right]_{(k)}\right)\right) \leq \left|\mathcal{S}_4^{\mathrm{II}}\right|$.

(iii) For $i \in \mathcal{S}_4^{\mathrm{c}}$, which correspond to the projector $\Pi_{\mathcal{F}^\perp}$ since $\sigma_i\left(\Pi_{\mathcal{F}^\perp}\left(\boldsymbol{\mathcal{A}}^*_{(k)}\right)\right) = 0$.

Based on Assumption 3, $q_\lambda(0) = q'_\lambda(0) = 0$. We have that $\nabla Q_\lambda\left(\boldsymbol{\mathcal{A}}^*_{(k)}\right) = \mathbf{U}^*_{\mathrm{c}} q'_\lambda\left(\boldsymbol{\Sigma}^*_{\mathrm{c}}\right)\mathbf{V}^{*\top}_{\mathrm{c}}$, where $\boldsymbol{\Sigma}^*_{\mathrm{c}} \in \mathbb{R}^{r\times r}$ is a diagonal matrix and $r = \min\{d_k, \Pi_{j\neq k} d_j\}$. Now we have

$$\Pi_{\mathcal{F}^\perp}\left(\nabla Q_\lambda\left(\boldsymbol{\mathcal{A}}^*_{(k)}\right)\right) = \left(\mathbf{I}_{\mathrm{c}} - \mathbf{U}^*_{\mathrm{c}}\mathbf{U}^{*\top}_{\mathrm{c}}\right)\mathbf{U}^*_k q'_\lambda\left(\boldsymbol{\Sigma}^*_{\mathrm{c}}\right)\mathbf{V}^{*\top}_{\mathrm{c}}\left(\mathbf{I}_{\mathrm{c}} - \mathbf{V}^*_{\mathrm{c}}\mathbf{V}^{*\top}_{\mathrm{c}}\right) \tag{122}$$

$$= \left(\mathbf{U}^*_{\mathrm{c}} - \mathbf{U}^*_{\mathrm{c}}\right) q'_\lambda\left(\boldsymbol{\Sigma}^*_{\mathrm{c}}\right)\left(\mathbf{V}^{*\top}_{\mathrm{c}} - \mathbf{V}^{*\top}_{\mathrm{c}}\right) \tag{123}$$

$$= 0, \tag{124}$$

where $\mathbf{I}_{\mathrm{c}}$ is the identity matrix. Meanwhile, since

$$\left\|\Pi_{\mathcal{F}^\perp}\left(\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)}\right)\right\|_{\mathrm{sp}} \leq \left\|\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)}\right\|_{\mathrm{sp}} = \frac{\left\|\left[\mathfrak{X}^*\left(\boldsymbol{\mathcal{E}}\right)\right]_{(k)}\right\|_{\mathrm{sp}}}{n} \leq \lambda. \tag{125}$$

For $\mathbf{Z}^*_{\mathrm{c}} = -\lambda^{-1}\Pi_{\mathcal{F}^\perp}\left(\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)}\right)$ and $\boldsymbol{\mathcal{H}}^* = \mathbf{U}^*_{\mathrm{c}}\mathbf{V}^{*\top}_{\mathrm{c}} + \mathbf{Z}^*_{\mathrm{c}} \in \partial\left\|\boldsymbol{\mathcal{A}}^*_{(k)}\right\|_*$, we have

$$\Pi_{\mathcal{F}^\perp}\left[\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)} + \lambda\boldsymbol{\mathcal{H}}^*\right] = \Pi_{\mathcal{F}^\perp}\left(\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)}\right) + \lambda\mathbf{Z}^*_{\mathrm{c}} = 0, \tag{126}$$

which implies that

$$\left\langle \Pi_{\mathcal{F}^\perp}\left(\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)} + \lambda\boldsymbol{\mathcal{H}}^* + \nabla Q_\lambda\left(\boldsymbol{\mathcal{A}}^*_{(k)}\right)\right), \left[\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}\right]_{(k)}\right\rangle = \left\langle 0, \left[\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}\right]_{(k)}\right\rangle = 0. \tag{127}$$

Adding (114), (121) and (127), which indicate that

$$\left(\rho^- - \zeta^-\right)\left\|\hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^*\right\|_{\mathrm{F}} \tag{128}$$

$$\leq \left\langle \Pi_{\mathcal{F}}\left(\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)} + \nabla Q_\lambda\left(\boldsymbol{\mathcal{A}}^*_{(k)}\right) + \lambda\boldsymbol{\mathcal{H}}^*\right), \left[\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}\right]_{(k)}\right\rangle \tag{129}$$

$$\leq \sqrt{\left|\mathcal{S}_4^{\mathrm{I}}\right|}\left\|\Pi_{\mathcal{S}_4^{\mathrm{I}}}\left(\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)}\right)\right\|_{\mathrm{sp}} \cdot \left\|\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}\right\|_{\mathrm{F}} + 3\lambda\sqrt{\left|\mathcal{S}_4^{\mathrm{II}}\right|}\left\|\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}\right\|_{\mathrm{F}} \tag{130}$$

$$= \left\|\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}\right\|_{\mathrm{F}}\sqrt{\left|\mathcal{S}_4^{\mathrm{I}}\right|}\left\|\Pi_{\mathcal{S}_4^{\mathrm{I}}}\left(\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)}\right)\right\|_{\mathrm{sp}} + 3\lambda\sqrt{\left|\mathcal{S}_4^{\mathrm{II}}\right|}. \tag{131}$$

Thus, we have

$$\left\|\hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^*\right\|_{\mathrm{F}} \leq \frac{1}{\rho^- - \zeta^-}\left[\sqrt{\left|\mathcal{S}_4^{\mathrm{I}}\right|}\left\|\Pi_{\mathcal{S}_4^{\mathrm{I}}}\left(\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)}\right)\right\|_{\mathrm{sp}} + 3\lambda\sqrt{\left|\mathcal{S}_4^{\mathrm{II}}\right|}\right].$$

□

Lemma 16. *Suppose $\boldsymbol{\mathcal{A}}^* \in \mathbb{R}^{d_1\times\cdots\times d_N}$ with rank of each mode-(k) unfolding $|\mathcal{S}_4|$. Then the error bound between the oracle estimator $\hat{\boldsymbol{\mathcal{A}}}^O$ and the true $\boldsymbol{\mathcal{A}}^*$ satisfies*

$$\left\|\hat{\boldsymbol{\mathcal{A}}}^O - \boldsymbol{\mathcal{A}}^*\right\|_{\mathrm{F}} = \left\|\left[\hat{\boldsymbol{\mathcal{A}}}^O - \boldsymbol{\mathcal{A}}^*\right]_{(k)}\right\|_{\mathrm{F}} \leq \frac{2\sqrt{|\mathcal{S}_4|}\left\|\Pi_{\mathcal{F}}\left(\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)}\right)\right\|_{\mathrm{sp}}}{\rho^-}. \tag{132}$$

Proof. Let $\boldsymbol{\mathcal{B}}' = \hat{\boldsymbol{\mathcal{A}}}^O - \boldsymbol{\mathcal{A}}^*$. According to the general estimator (2) and the definition of the adjoint operator $\mathfrak{X}(\cdot)$, we can express the difference in loss as follows

$$L\left(\hat{\boldsymbol{\mathcal{A}}}^O\right) - L\left(\boldsymbol{\mathcal{A}}^*\right) = \frac{1}{2n}\sum_{i=1}^n\left[\boldsymbol{\mathcal{Y}}^{(i)} - \mathfrak{X}^{(i)}\left(\boldsymbol{\mathcal{A}}^* + \boldsymbol{\mathcal{B}}'\right)\right]^2 - \frac{1}{2n}\sum_{i=1}^n\left[\boldsymbol{\mathcal{Y}}^{(i)} - \mathfrak{X}^{(i)}\left(\boldsymbol{\mathcal{A}}^*\right)\right]^2 \tag{133}$$

$$= \frac{1}{2n}\sum_{i=1}^n\left[\boldsymbol{\mathcal{E}}^{(i)} - \mathfrak{X}^{(i)}\left(\boldsymbol{\mathcal{B}}'\right)\right]^2 - \frac{1}{2n}\sum_{i=1}^n\boldsymbol{\mathcal{E}}^{(i)} \tag{134}$$

$$= \frac{1}{2n}\left\|\left[\mathfrak{X}\left(\boldsymbol{\mathcal{B}}'\right)\right]_{(k)}\right\|_{\mathrm{sp}}^2 - \frac{1}{n}\left\langle\mathfrak{X}^*\left(\boldsymbol{\mathcal{E}}\right), \boldsymbol{\mathcal{B}}'\right\rangle. \tag{135}$$

Given that $\hat{\boldsymbol{\mathcal{A}}}^O$ minimizes $L(\cdot)$ over the subspace $\mathcal{F}$ and $\boldsymbol{\mathcal{A}}^*_{(k)} \in \mathcal{F}$, we have

$$L\left(\hat{\boldsymbol{\mathcal{A}}}^O\right) - L\left(\boldsymbol{\mathcal{A}}^*\right) \leq 0. \tag{136}$$

Thus, it follows that

$$\frac{1}{2n}\left\|\left[\mathfrak{X}\left(\boldsymbol{\mathcal{B}}'\right)\right]_{(k)}\right\|_{\text{sp}}^2 \leq \frac{1}{n}\left\langle \mathfrak{X}^*\left(\boldsymbol{\mathcal{E}}\right), \boldsymbol{\mathcal{B}}'\right\rangle. \tag{137}$$

By the RSC condition 1, we know that

$$L\left(\boldsymbol{\mathcal{A}}+\boldsymbol{\mathcal{B}}\right) - L\left(\boldsymbol{\mathcal{A}}\right) \geq \left\langle \nabla L\left(\boldsymbol{\mathcal{A}}\right), \boldsymbol{\mathcal{B}}\right\rangle + \frac{\rho^-}{2}\left\|\boldsymbol{\mathcal{B}}\right\|_{\text{F}}^2. \tag{138}$$

Applying this to $\boldsymbol{\mathcal{B}}'$,

$$\begin{aligned}\frac{\rho^-}{2}\left\|\boldsymbol{\mathcal{B}}'\right\|_{\text{F}}^2 &\leq L\left(\boldsymbol{\mathcal{B}}'\right) - L\left(\boldsymbol{\mathcal{A}}^*\right) - \left\langle \nabla L\left(\boldsymbol{\mathcal{A}}^*\right), \boldsymbol{\mathcal{B}}'\right\rangle \\ &= \frac{1}{2n}\left\|\left[\mathfrak{X}\left(\boldsymbol{\mathcal{B}}'\right)\right]_{(k)}\right\|_{\text{sp}}^2 - \frac{1}{n}\left\langle \mathfrak{X}^*\left(\boldsymbol{\mathcal{E}}\right), \boldsymbol{\mathcal{B}}'\right\rangle - \left\langle \nabla L\left(\boldsymbol{\mathcal{A}}^*\right), \boldsymbol{\mathcal{B}}'\right\rangle.\end{aligned} \tag{139}$$

Substituting (137) into (139) gives

$$\frac{\rho^-}{2}\left\|\boldsymbol{\mathcal{B}}'\right\|_{\text{F}}^2 \leq \frac{1}{2n}\left\|\left[\mathfrak{X}\left(\boldsymbol{\mathcal{B}}'\right)\right]_{(k)}\right\|_{\text{sp}}^2 \leq \frac{1}{n}\left\langle \mathfrak{X}^*\left(\boldsymbol{\mathcal{E}}\right), \boldsymbol{\mathcal{B}}'\right\rangle. \tag{140}$$

Therefore, we have

$$\left\|\boldsymbol{\mathcal{B}}'\right\|_{\text{F}}^2 \leq \frac{2\left\langle \Pi_{\mathcal{F}}\left(\left[\mathfrak{X}^*\left(\boldsymbol{\mathcal{E}}\right)\right]_{(k)}\right), \boldsymbol{\mathcal{B}}'\right\rangle}{n\rho^-} \leq \frac{2\left\|\Pi_{\mathcal{F}}\left(\left[\mathfrak{X}^*\left(\boldsymbol{\mathcal{E}}\right)\right]_{(k)}\right)\right\|_{\text{sp}} \cdot \left\|\boldsymbol{\mathcal{B}}'\right\|_*}{n\rho^-}. \tag{141}$$

Using the fact that $\text{rank}\left(\boldsymbol{\mathcal{B}}'\right) = |\mathcal{S}_4|$, we have

$$\left\|\boldsymbol{\mathcal{B}}'_{(k)}\right\|_* \leq \sqrt{|\mathcal{S}_4|}\left\|\boldsymbol{\mathcal{B}}'_{(k)}\right\|_{\text{F}}^2. \tag{142}$$

Thus, it follows that

$$\left\|\boldsymbol{\mathcal{B}}'_{(k)}\right\|_{\text{F}}^2 \leq \frac{2\sqrt{|\mathcal{S}_4|}\left\|\Pi_{\mathcal{F}}\left(\left[\mathfrak{X}^*\left(\boldsymbol{\mathcal{E}}\right)\right]_{(k)}\right)\right\|_{\text{sp}} \cdot \left\|\boldsymbol{\mathcal{B}}'_{(k)}\right\|_{\text{F}}^2}{n\rho^-}. \tag{143}$$

Recalling that $\nabla L\left(\boldsymbol{\mathcal{A}}^*\right) = -\frac{\mathfrak{X}^*(\boldsymbol{\mathcal{E}})}{n}$, we conclude

$$\left\|\boldsymbol{\mathcal{B}}'_{(k)}\right\|_{\text{F}} \leq \frac{2\sqrt{|\mathcal{S}_4|}\left\|\Pi_{\mathcal{F}}\left(\left[\mathfrak{X}^*\left(\boldsymbol{\mathcal{E}}\right)\right]_{(k)}\right)\right\|_{\text{sp}}}{n\rho^-} = \frac{2\sqrt{|\mathcal{S}_4|}\left\|\Pi_{\mathcal{F}}\left(\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)}\right)\right\|_{\text{sp}}}{\rho^-}. \tag{144}$$

Thus, since $\left\|\boldsymbol{\mathcal{B}}'_{(k)}\right\|_{\text{F}} = \left\|\boldsymbol{\mathcal{B}}'\right\|_{\text{F}}$, we have the desired error bound

$$\left\|\hat{\boldsymbol{\mathcal{A}}}^O - \boldsymbol{\mathcal{A}}^*\right\|_{\text{F}} = \left\|\boldsymbol{\mathcal{B}}'\right\|_{\text{F}} \leq \frac{2\sqrt{|\mathcal{S}_4|}\left\|\Pi_{\mathcal{F}}\left(\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)}\right)\right\|_{\text{sp}}}{\rho^-}. \tag{145}$$

□

Proof of Corollary 4

Proof. Suppose $\hat{\boldsymbol{\mathcal{H}}} \in \partial \left\| \left(\hat{\boldsymbol{\mathcal{A}}}_{(k)} \right) \right\|_*$, since $\hat{\boldsymbol{\mathcal{A}}}$ satisfies the optimality condition, for any $\boldsymbol{A}' \in \mathbb{R}^{d_1 \times \cdots \times d_N}$, it holds that

$$\max_{\boldsymbol{A}'} \left\{ \left\langle \nabla L \left(\hat{\boldsymbol{\mathcal{A}}} \right), \hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{A}' \right\rangle + \left\langle \nabla Q_\lambda \left(\hat{\boldsymbol{\mathcal{A}}}_{(k)} \right) + \lambda \hat{\boldsymbol{\mathcal{H}}}, \left[\hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{A}' \right]_{(k)} \right\rangle \right\} \leq 0. \tag{146}$$

In the following, we will show some $\hat{\boldsymbol{\mathcal{H}}}^O \in \partial \left\| \hat{\boldsymbol{\mathcal{A}}}^O_{(k)} \right\|_*$ satisfy that

$$\max_{\boldsymbol{A}'} \left\{ \left\langle \left[\nabla L \left(\hat{\boldsymbol{\mathcal{A}}}^O \right) \right]_{(k)} + \nabla Q_\lambda \left(\hat{\boldsymbol{\mathcal{A}}}^O_{(k)} \right) + \lambda \hat{\boldsymbol{\mathcal{H}}}^O, \left[\hat{\boldsymbol{\mathcal{A}}}^O - \boldsymbol{A}' \right]_{(k)} \right\rangle \right\} \leq 0. \tag{147}$$

Recall that $\tilde{L}(\boldsymbol{\mathcal{A}}) = L(\boldsymbol{\mathcal{A}}) + Q_\lambda \left(\boldsymbol{\mathcal{A}}_{(k)} \right)$. Projecting the components of the inner product of the LHS in (147) into two complementary spaces $\mathcal{F}$ and $\mathcal{F}^\perp$, we have the following decomposition

$$\begin{aligned}
&\left\langle \left[\nabla L \left(\hat{\boldsymbol{\mathcal{A}}}^O \right) \right]_{(k)} + \nabla Q_\lambda \left(\hat{\boldsymbol{\mathcal{A}}}^O_{(k)} \right) + \lambda \hat{\boldsymbol{\mathcal{H}}}^O, \left[\hat{\boldsymbol{\mathcal{A}}}^O - \boldsymbol{A}' \right]_{(k)} \right\rangle \\
=&\underbrace{\left\langle \left[\nabla L \left(\hat{\boldsymbol{\mathcal{A}}}^O \right) \right]_{(k)} + \nabla Q_\lambda \left(\hat{\boldsymbol{\mathcal{A}}}^O_{(k)} \right) + \lambda \hat{\boldsymbol{\mathcal{H}}}^O, \Pi_{\mathcal{F}} \left(\left[\hat{\boldsymbol{\mathcal{A}}}^O - \boldsymbol{A}' \right]_{(k)} \right) \right\rangle}_{P_1} \\
+&\underbrace{\left\langle \left[\nabla L \left(\hat{\boldsymbol{\mathcal{A}}}^O \right) \right]_{(k)} + \nabla Q_\lambda \left(\hat{\boldsymbol{\mathcal{A}}}^O_{(k)} \right) + \lambda \hat{\boldsymbol{\mathcal{H}}}^O, \Pi_{\mathcal{F}^\perp} \left(\left[\hat{\boldsymbol{\mathcal{A}}}^O - \boldsymbol{A}' \right]_{(k)} \right) \right\rangle}_{P_2}.
\end{aligned} \tag{148}$$

For Term P_1.

By applying Weyl's inequality for singular values, we obtain

$$\max_l \left| \sigma_i \left(\boldsymbol{\mathcal{A}}^*_{(k)} \right) - \sigma_i \left(\hat{\boldsymbol{\mathcal{A}}}^O_{(k)} \right) \right| \leq \left\| \boldsymbol{\mathcal{A}}^*_{(k)} - \hat{\boldsymbol{\mathcal{A}}}^O_{(k)} \right\|_{\text{sp}}. \tag{149}$$

Further, from the properties of the Frobenius norm, we have

$$\left\| \boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}^O \right\|_{\text{F}} = \left\| \left[\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}^O \right]_{(k)} \right\|_{\text{F}}. \tag{150}$$

From Lemma 16, the estimation error $\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}^O$ yields

$$\max_l \left| \sigma_i \left(\boldsymbol{\mathcal{A}}^*_{(k)} \right) - \sigma_i \left(\hat{\boldsymbol{\mathcal{A}}}^O_{(k)} \right) \right| \leq \frac{2\sqrt{|\mathcal{S}_4|} \left\| \left[\mathfrak{X}^* (\boldsymbol{\mathcal{E}}) \right]_{(k)} \right\|_{\text{sp}}}{n\rho^-}, \tag{151}$$

where $|\mathcal{S}_4|$ denotes the rank of the unfolded matrix $\boldsymbol{\mathcal{A}}^*_{(k)}$. Utilizing the weak condition of the singular values, we find

$$\min_{i \in \mathcal{S}_4} \left| \sigma_i \left(\boldsymbol{\mathcal{A}}^*_{(k)} \right) \right| \geq \nu + \frac{2\sqrt{|\mathcal{S}_4|}}{n\rho^-} \left\| \left[\mathfrak{X}^* (\boldsymbol{\mathcal{E}}) \right]_{(k)} \right\|_{\text{sp}}. \tag{152}$$

Applying the triangle inequality, we derive

$$\begin{aligned}
\min_{i \in \mathcal{S}_4} \left| \sigma_i \left(\hat{\boldsymbol{\mathcal{A}}}^O_{(k)} \right) \right| &= \min_{i \in \mathcal{S}_4} \left| \sigma_i \left(\hat{\boldsymbol{\mathcal{A}}}^O_{(k)} \right) - \sigma_i \left(\boldsymbol{\mathcal{A}}^*_{(k)} \right) + \sigma_i \left(\boldsymbol{\mathcal{A}}^*_{(k)} \right) \right| \\
&\geq - \max_{i \in \mathcal{S}_4} \left| \sigma_i \left(\hat{\boldsymbol{\mathcal{A}}}^O_{(k)} \right) - \sigma_i \left(\boldsymbol{\mathcal{A}}^*_{(k)} \right) \right| + \min_{i \in \mathcal{S}_4} \left| \sigma_i \left(\boldsymbol{\mathcal{A}}^*_{(k)} \right) \right| \\
&\geq - \frac{2\sqrt{|\mathcal{S}_4|}}{n\rho^-} \left\| \left[\mathfrak{X}^* (\boldsymbol{\mathcal{E}}) \right]_{(k)} \right\|_{\text{sp}} + \nu + \frac{2\sqrt{|\mathcal{S}_4|}}{n\rho^-} \left\| \left[\mathfrak{X}^* (\boldsymbol{\mathcal{E}}) \right]_{(k)} \right\|_{\text{sp}} \\
&= \nu.
\end{aligned} \tag{153}$$

Considering the definition of oracle estimator , $\hat{\boldsymbol{\mathcal{A}}}^O \in \mathcal{F}$, which implies the tensor rank of each mode-(k) unfolding rank $\left(\hat{\boldsymbol{\mathcal{A}}}^O_{(k)}\right) = |\mathcal{S}_4|$. And we have the singular value decomposition $\hat{\boldsymbol{\mathcal{A}}}^O_{(k)} = \mathbf{U}^* \hat{\boldsymbol{\Sigma}}^O \mathbf{V}^{*\top}$. Since $R_\lambda\left(\boldsymbol{\mathcal{A}}_{(k)}\right) = \lambda \left\|\boldsymbol{\mathcal{A}}_{(k)}\right\|_* + Q_\lambda\left(\boldsymbol{\mathcal{A}}_{(k)}\right)$, for $\hat{\mathbf{Z}}^O \in \mathcal{F}^\perp$, $\left\|\hat{\mathbf{Z}}^O\right\|_{\text{sp}} \leq 1$, and $\left(\hat{\boldsymbol{\Sigma}}^O\right)_{\mathcal{S}_4} \in \mathbb{R}^{|\mathcal{S}_4|\times|\mathcal{S}_4|}$ is a diagonal matrix where $\Pi_{\mathcal{F}}\left(q'_\lambda\left(\hat{\boldsymbol{\Sigma}}^O\right)\right) = q'_\lambda\left(\left(\hat{\boldsymbol{\Sigma}}^O\right)_{\mathcal{S}_4}\right)$. Based on the definition of $\nabla Q_\lambda(\cdot)$ and $\partial\|\cdot\|_*$, we have

$$\begin{aligned}
\Pi_{\mathcal{F}}\left(\nabla R_\lambda\left(\hat{\boldsymbol{\mathcal{A}}}^O_{(k)}\right)\right) &= \Pi_{\mathcal{F}}\left(Q_\lambda\left(\hat{\boldsymbol{\mathcal{A}}}^O_{(k)}\right)\right) + \lambda\partial\left\|\hat{\boldsymbol{\mathcal{A}}}^O_{(k)}\right\|_* \\
&= \Pi_{\mathcal{F}}\left(\mathbf{U}^* q'_\lambda\left(\hat{\boldsymbol{\Sigma}}^O\right)\mathbf{V}^{*\top} + \lambda\mathbf{U}^*\mathbf{V}^{*\top} + \lambda\hat{\mathbf{Z}}^O\right) \\
&= \mathbf{U}^*\left(q'_\lambda\left(\left(\hat{\boldsymbol{\Sigma}}^O\right)_{\mathcal{S}_4}\right) + \lambda\mathbf{I}_{\mathcal{S}_4}\right)\mathbf{V}^{*\top}, \qquad (154)
\end{aligned}$$

where the second equality in (154) is to simply project each component into the subspace $\mathcal{F}$. $\mathbf{I}_{\mathcal{S}_4}$ is the identity matrix and $\mathbf{I}_{\mathcal{S}_4} \in \mathbb{R}^{|\mathcal{S}_4|\times|\mathcal{S}_4|}$. Since $p_\lambda(t) = q_\lambda(t) + \lambda|t|$, we have $p'_\lambda(t) = q'_\lambda(t) + \lambda t$ for all $t > 0$. Consider the diagonal matrix $q'_\lambda\left(\left(\hat{\boldsymbol{\Sigma}}^O\right)_{\mathcal{S}_4}\right) + \lambda\mathbf{I}_{\mathcal{S}_4}$, we have the i-th ($i \in \mathcal{S}_4$) element on the diagonal that

$$\left[q'_\lambda\left(\left(\hat{\boldsymbol{\Sigma}}^O\right)_{\mathcal{S}_4}\right) + \lambda\mathbf{I}_{\mathcal{S}_4}\right]_{ii} = q'_\lambda\left(\sigma_i\left(\hat{\boldsymbol{\mathcal{A}}}^O_{(k)}\right)\right) + \lambda = p'_\lambda\left(\sigma_i\left(\hat{\boldsymbol{\mathcal{A}}}^O_{(k)}\right)\right). \qquad (155)$$

Since $p_\lambda(\cdot)$ satisfies the regularity condition (iii) in Assumption 3 that $p'_\lambda(t) = 0$ for all $t \geq \nu$, we have $p'_\lambda\left(\sigma_i\left(\hat{\boldsymbol{\mathcal{A}}}^O_{(k)}\right)\right) = 0$ for $i \in \mathcal{S}_4$, due to the fact that $\sigma_i\left(\hat{\boldsymbol{\mathcal{A}}}^O_{(k)}\right) \geq \nu > 0$.

Therefore, the diagonal matrix $q'_\lambda\left(\left(\hat{\boldsymbol{\Sigma}}^O\right)_{\mathcal{S}_4}\right) + \lambda\mathbf{I}_{\mathcal{S}_4} = 0$, substituting which in to (154) yields

$$\Pi_{\mathcal{F}}\left(\nabla R_\lambda\left(\hat{\boldsymbol{\mathcal{A}}}^O_{(k)}\right)\right) = 0. \qquad (156)$$

Since $\hat{\boldsymbol{\mathcal{A}}}^O$ is the estimator over $\mathcal{F}$, we have the optimality condition that for any $\boldsymbol{\mathcal{A}}' \in \mathbb{R}^{d_1\times\cdots\times d_N}$, it holds that

$$\max_{\boldsymbol{\mathcal{A}}'}\left\langle\left[\nabla L\left(\hat{\boldsymbol{\mathcal{A}}}^O\right)\right]_{(k)}, \Pi_{\mathcal{F}}\left(\left[\hat{\boldsymbol{\mathcal{A}}}^O - \boldsymbol{\mathcal{A}}'\right]_{(k)}\right)\right\rangle \leq 0. \qquad (157)$$

Substitute (156) and (157) into P_1, for all $\hat{\boldsymbol{\mathcal{H}}}^O \in \partial\left\|\hat{\boldsymbol{\mathcal{A}}}^O_{(k)}\right\|_*$ we have

$$\begin{aligned}
&\max_{\boldsymbol{\mathcal{A}}'}\left\langle\left[\nabla L\left(\hat{\boldsymbol{\mathcal{A}}}^O\right)\right]_{(k)} + \nabla Q_\lambda\left(\hat{\boldsymbol{\mathcal{A}}}^O_{(k)}\right) + \lambda\hat{\boldsymbol{\mathcal{H}}}^O, \Pi_{\mathcal{F}}\left(\left[\hat{\boldsymbol{\mathcal{A}}}^O - \boldsymbol{\mathcal{A}}'\right]_{(k)}\right)\right\rangle \\
=&\max_{\boldsymbol{\mathcal{A}}'}\left\langle\left[\nabla L\left(\hat{\boldsymbol{\mathcal{A}}}^O\right)\right]_{(k)}, \Pi_{\mathcal{F}}\left(\left[\hat{\boldsymbol{\mathcal{A}}}^O - \boldsymbol{\mathcal{A}}'\right]_{(k)}\right)\right\rangle + \max_{\boldsymbol{\mathcal{A}}'}\left\langle\Pi_{\mathcal{F}}\left(\nabla R_\lambda\left(\hat{\boldsymbol{\mathcal{A}}}^O_{(k)}\right)\right), \Pi_{\mathcal{F}}\left(\left[\hat{\boldsymbol{\mathcal{A}}}^O - \boldsymbol{\mathcal{A}}'\right]_{(k)}\right)\right\rangle \\
\leq& 0. \qquad (158)
\end{aligned}$$

For Term P_2.

By definition of $\nabla Q_\lambda(\cdot)$, and the regularity condition (v) in Assumption 3 , we do the decomposition that $\nabla Q_\lambda\left(\hat{\boldsymbol{\mathcal{A}}}^O_{(k)}\right) = \mathbf{U}^* q'_\lambda\left(\hat{\boldsymbol{\Sigma}}^O\right)\mathbf{V}^{*\top}$, where $\hat{\boldsymbol{\Sigma}}^O$ is diagonal matrix. Projecting $\nabla Q_\lambda\left(\hat{\boldsymbol{\mathcal{A}}}^O_{(k)}\right)$ into $\mathcal{F}^\perp$ yields that

$$\begin{aligned}
\Pi_{\mathcal{F}^\perp}\left(\nabla Q_\lambda\left(\hat{\boldsymbol{\mathcal{A}}}^O_{(k)}\right)\right) &= \left(\mathbf{I}_{\mathcal{S}_4} - \mathbf{U}^*\mathbf{U}^{*\top}\right)\mathbf{U}^* q'_\lambda\left(\hat{\boldsymbol{\Sigma}}^O\right)\mathbf{V}^{*\top}\left(\mathbf{I}_{\mathcal{S}_4} - \mathbf{V}^*\mathbf{V}^{*\top}\right) \\
&= \left(\mathbf{U}^* - \mathbf{U}^*\right) q'_\lambda\left(\left(\hat{\boldsymbol{\Sigma}}^O\right)_{\mathcal{S}_4}\right)\left(\mathbf{V}^{*\top} - \mathbf{V}^{*\top}\right) \\
&= 0. \qquad (159)
\end{aligned}$$

Therefore,

$$P_2 = \left\langle \Pi_{\mathcal{F}^\perp}\left(\left[\nabla L\left(\hat{\boldsymbol{\mathcal{A}}}^O\right)\right]_{(k)} + \lambda\hat{\boldsymbol{\mathcal{H}}}^O\right), \Pi_{\mathcal{F}^\perp}\left(\left[\hat{\boldsymbol{\mathcal{A}}}^O - \boldsymbol{\mathcal{A}}'\right]_{(k)}\right)\right\rangle. \tag{160}$$

Moreover, with the triangle inequality, we have

$$\begin{aligned}\left\|\left[\nabla L\left(\hat{\boldsymbol{\mathcal{A}}}^O\right)\right]_{(k)}\right\|_{\text{sp}} &\leq \left\|\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)}\right\|_{\text{sp}} + \left\|\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)} - \left[\nabla L\left(\hat{\boldsymbol{\mathcal{A}}}^O\right)\right]_{(k)}\right\|_{\text{sp}} \\ &\leq \left\|\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)}\right\|_{\text{sp}} + \left\|\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)} - \left[\nabla L\left(\hat{\boldsymbol{\mathcal{A}}}^O\right)\right]_{(k)}\right\|_{\text{F}},\end{aligned} \tag{161}$$

where the second inequality comes from the fact that

$$\left\|\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)} - \left[\nabla L\left(\hat{\boldsymbol{\mathcal{A}}}^O\right)\right]_{(k)}\right\|_{\text{sp}} \leq \left\|\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)} - \left[\nabla L\left(\hat{\boldsymbol{\mathcal{A}}}^O\right)\right]_{(k)}\right\|_{\text{F}}. \tag{162}$$

From Restricted Strong Smoothness in Assumption 2 where $\left\|\nabla L\left(\boldsymbol{\mathcal{A}}\right) - \nabla L\left(\boldsymbol{\mathcal{A}} + \boldsymbol{\mathcal{B}}'\right)\right\|_{\text{F}} \leq \rho^+ \left\|\boldsymbol{\mathcal{B}}'\right\|_{\text{F}}$, we can substitute it into (161), and we have

$$\left\|\left[\nabla L\left(\hat{\boldsymbol{\mathcal{A}}}^O\right)\right]_{(k)}\right\|_{\text{sp}} \leq \left\|\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)}\right\|_{\text{sp}} + \rho^+ \left\|\boldsymbol{\mathcal{A}}^* - \hat{\boldsymbol{\mathcal{A}}}^O\right\|_{\text{F}}. \tag{163}$$

Since $\Pi_{\mathcal{F}^\perp}\left(\boldsymbol{\mathcal{B}}'\right) = 0$, it is evident that $\boldsymbol{\mathcal{B}}' \in \boldsymbol{\mathcal{C}}$. Substitute (132) from Lemma 16 into (163), from the choice of λ, we have

$$\begin{aligned}\left\|\Pi_{\mathcal{F}^\perp}\left(\left[\nabla L\left(\hat{\boldsymbol{\mathcal{A}}}^O\right)\right]_{(k)}\right)\right\|_{\text{sp}} &\leq \left\|\left[\nabla L\left(\hat{\boldsymbol{\mathcal{A}}}^O\right)\right]_{(k)}\right\|_{\text{sp}} &(164)\\ &\leq \left\|\left[\nabla L\left(\boldsymbol{\mathcal{A}}^*\right)\right]_{(k)}\right\|_{\text{sp}} + \frac{2\sqrt{|\mathcal{S}_4|}\rho^+}{n\rho^-}\left\|\left[\mathfrak{X}^*\left(\boldsymbol{\mathcal{E}}\right)\right]_{(k)}\right\|_{\text{sp}} &(165)\\ &\leq \lambda. &(166)\end{aligned}$$

By setting $\hat{\mathbf{Z}}^O = -\lambda^{-1}\Pi_{\mathcal{F}^\perp}\left(\left[\nabla L\left(\hat{\boldsymbol{\mathcal{A}}}^O\right)\right]_{(k)}\right)$, such that $\hat{\boldsymbol{\mathcal{H}}}^O = \mathbf{U}^*\mathbf{V}^{*\top} + \hat{\mathbf{Z}}^O \in \partial\left\|\hat{\boldsymbol{\mathcal{A}}}^O_{(k)}\right\|_*$, since $\hat{\mathbf{Z}}^O$ satisfies the condition $\hat{\mathbf{Z}}^O \in \mathcal{F}^\perp$. $\left\|\hat{\mathbf{Z}}^O\right\|_{\text{sp}} \leq 1$, we have

$$\Pi_{\mathcal{F}^\perp}\left(\left[\nabla L\left(\hat{\boldsymbol{\mathcal{A}}}^O\right)\right]_{(k)} + \lambda\hat{\boldsymbol{\mathcal{H}}}^O\right) = 0, \tag{167}$$

which implies that

$$P_2 = \left\langle 0, \Pi_{\mathcal{F}^\perp}\left(\left[0 - \boldsymbol{\mathcal{A}}'\right]_{(k)}\right)\right\rangle = 0. \tag{168}$$

Substituting (158) and (168) into (148), we obtain (147) that

$$\max_{\boldsymbol{\mathcal{A}}'}\left\langle\left[\nabla L\left(\hat{\boldsymbol{\mathcal{A}}}^O\right)\right]_{(k)} + Q_\lambda\left(\hat{\boldsymbol{\mathcal{A}}}^O_{(k)}\right) + \lambda\hat{\boldsymbol{\mathcal{H}}}^O, \left[\hat{\boldsymbol{\mathcal{A}}}^O - \boldsymbol{\mathcal{A}}'\right]_{(k)}\right\rangle \leq 0. \tag{169}$$

Now we are going to prove that $\hat{\boldsymbol{\mathcal{A}}}^O = \hat{\boldsymbol{\mathcal{A}}}$ and the error bound between $\hat{\boldsymbol{\mathcal{A}}}^O$ and $\boldsymbol{\mathcal{A}}^*$.

Similar to the proof of Lemma 15, since $\|\cdot\|_*$ is convex, and applying Lemma 13, we have

$$\begin{aligned}0 \geq &\left\langle\nabla L\left(\hat{\boldsymbol{\mathcal{A}}}\right), \hat{\boldsymbol{\mathcal{A}}}^O - \hat{\boldsymbol{\mathcal{A}}}\right\rangle + \left\langle\nabla Q_\lambda\left(\hat{\boldsymbol{\mathcal{A}}}_{(k)}\right) + \lambda\hat{\boldsymbol{\mathcal{H}}}, \left[\hat{\boldsymbol{\mathcal{A}}}^O - \hat{\boldsymbol{\mathcal{A}}}\right]_{(k)}\right\rangle \\ &+ \left\langle\nabla L\left(\hat{\boldsymbol{\mathcal{A}}}^O\right), \hat{\boldsymbol{\mathcal{A}}} - \hat{\boldsymbol{\mathcal{A}}}^O\right\rangle + \left\langle\nabla Q_\lambda\left(\hat{\boldsymbol{\mathcal{A}}}^O_{(k)}\right) + \lambda\hat{\boldsymbol{\mathcal{H}}}^O, \left[\hat{\boldsymbol{\mathcal{A}}} - \hat{\boldsymbol{\mathcal{A}}}^O\right]_{(k)}\right\rangle + \left(\rho^- - \zeta^-\right)\left\|\hat{\boldsymbol{\mathcal{A}}}^O - \hat{\boldsymbol{\mathcal{A}}}\right\|_{\text{F}}^2.\end{aligned} \tag{170}$$

From (146), we have

$$\begin{aligned}&\left\langle \nabla L\left(\hat{\boldsymbol{\mathcal{A}}}\right), \hat{\boldsymbol{\mathcal{A}}} - \hat{\boldsymbol{\mathcal{A}}}^{O}\right\rangle + \left\langle \nabla Q_{\lambda}\left(\hat{\boldsymbol{\mathcal{A}}}_{(k)}\right) + \lambda\hat{\boldsymbol{\mathcal{H}}}, \left[\hat{\boldsymbol{\mathcal{A}}} - \hat{\boldsymbol{\mathcal{A}}}^{O}\right]_{(k)}\right\rangle \\ &\leq \max_{\boldsymbol{\mathcal{A}}'}\left\{\left\langle \nabla L\left(\hat{\boldsymbol{\mathcal{A}}}\right), \hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}'\right\rangle + \left\langle \nabla Q_{\lambda}\left(\hat{\boldsymbol{\mathcal{A}}}_{(k)}\right) + \lambda\hat{\boldsymbol{\mathcal{H}}}, \left[\hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}'\right]_{(k)}\right\rangle\right\} \leq 0. \end{aligned} \tag{171}$$

Therefore, in (170),

$$\left\langle \nabla L\left(\hat{\boldsymbol{\mathcal{A}}}\right), \hat{\boldsymbol{\mathcal{A}}}^{O} - \hat{\boldsymbol{\mathcal{A}}}\right\rangle + \left\langle \nabla Q_{\lambda}\left(\hat{\boldsymbol{\mathcal{A}}}_{(k)}\right) + \lambda\hat{\boldsymbol{\mathcal{H}}}, \left[\hat{\boldsymbol{\mathcal{A}}}^{O} - \hat{\boldsymbol{\mathcal{A}}}\right]_{(k)}\right\rangle \geq 0. \tag{172}$$

From (147), we have

$$\begin{aligned}&\left\langle \nabla L\left(\hat{\boldsymbol{\mathcal{A}}}^{O}\right), \hat{\boldsymbol{\mathcal{A}}}^{O} - \hat{\boldsymbol{\mathcal{A}}}\right\rangle + \left\langle \nabla Q_{\lambda}\left(\hat{\boldsymbol{\mathcal{A}}}^{O}_{(k)}\right) + \lambda\hat{\boldsymbol{\mathcal{H}}}, \left[\hat{\boldsymbol{\mathcal{A}}}^{O} - \hat{\boldsymbol{\mathcal{A}}}\right]_{(k)}\right\rangle \\ &\leq \max_{\boldsymbol{\mathcal{A}}'}\left\{\left\langle \nabla L\left(\hat{\boldsymbol{\mathcal{A}}}^{O}\right), \hat{\boldsymbol{\mathcal{A}}}^{O} - \boldsymbol{\mathcal{A}}'\right\rangle + \left\langle \nabla Q_{\lambda}\left(\hat{\boldsymbol{\mathcal{A}}}^{O}_{(k)}\right) + \lambda\hat{\boldsymbol{\mathcal{H}}}, \left[\hat{\boldsymbol{\mathcal{A}}}^{O} - \boldsymbol{\mathcal{A}}'\right]_{(k)}\right\rangle\right\} \leq 0. \end{aligned} \tag{173}$$

Therefore, in (170),

$$\left\langle \nabla L\left(\hat{\boldsymbol{\mathcal{A}}}^{O}\right), \hat{\boldsymbol{\mathcal{A}}} - \hat{\boldsymbol{\mathcal{A}}}^{O}\right\rangle + \left\langle \nabla Q_{\lambda}\left(\hat{\boldsymbol{\mathcal{A}}}^{O}_{(k)}\right) + \lambda\hat{\boldsymbol{\mathcal{H}}}^{O}, \left[\hat{\boldsymbol{\mathcal{A}}} - \hat{\boldsymbol{\mathcal{A}}}^{O}\right]_{(k)}\right\rangle \geq 0. \tag{174}$$

Substituting (170) and (171) into (173) such that

$$\left(\rho^{-} - \zeta^{-}\right)\left\|\hat{\boldsymbol{\mathcal{A}}}^{O} - \hat{\boldsymbol{\mathcal{A}}}\right\|_{\mathrm{F}}^{2} \geq 0. \tag{175}$$

Since $\rho^{-} > \zeta^{-}$, the inequation holds only if

$$\hat{\boldsymbol{\mathcal{A}}}^{O} = \hat{\boldsymbol{\mathcal{A}}}.$$

And by Lemma 16, we obtain the statistical oracle bound for the penalty

$$\left\|\hat{\boldsymbol{\mathcal{A}}}^{O} - \boldsymbol{\mathcal{A}}^{*}\right\|_{\mathrm{F}} = \left\|\left[\hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^{*}\right]_{(k)}\right\|_{\mathrm{F}} \leq \frac{2\sqrt{|\mathcal{S}_4|}\left\|\Pi_{\mathcal{F}}\left(\left[\mathfrak{X}^{*}\left(\boldsymbol{\mathcal{E}}\right)\right]_{(k)}\right)\right\|_{\mathrm{sp}}}{n\rho^{-}}. \tag{176}$$

Furthermore, we can have

$$\left\|\hat{\boldsymbol{\mathcal{A}}}^{O} - \boldsymbol{\mathcal{A}}^{*}\right\|_{\mathrm{F}} = \frac{2\sqrt{|\mathcal{S}_4|}\tau_k}{n\rho^{-}}, \tag{177}$$

where $\tau_k = \left\|\Pi_{\mathcal{F}}\left(\left[\mathfrak{X}^{*}\left(\boldsymbol{\mathcal{E}}\right)\right]_{(k)}\right)\right\|_{\mathrm{sp}}$, which completes the proof. □

G Proof of the Corollary 5

Recall that the proposed slice-wise lowrankness penalty can be reformulated as the sum of the ℓ_1 penalty and a concave part. Specifically, we have:

$$R_{\lambda}\left(\boldsymbol{\mathcal{A}}\right) = \sum_{l=1}^{\prod_{m\neq j,k} d_m}\sum_{s=1}^{s^{\mathrm{all}}} p_{\lambda}\left(\sigma_s\left(\left[\boldsymbol{\mathcal{A}}_{(j,k)}\right]_{\cdot,\cdot,l}\right)\right) = \sum_{l=1}^{\prod_{m\neq j,k} d_m}\left[\lambda\left\|\left[\boldsymbol{\mathcal{A}}_{(j,k)}\right]_{\cdot,\cdot,l}\right\|_{*} + Q_{\lambda}\left(\left[\boldsymbol{\mathcal{A}}_{(j,k)}\right]_{\cdot,\cdot,l}\right)\right],$$

where $s^{\text{all}} = \min\left\{d_j d_k, \prod_{l\neq j,k} d_l\right\}$, $\left[\boldsymbol{\mathcal{A}}_{(j,k)}\right]_{\cdot,\cdot,l}$ denotes the l-th slice of the mode-(j,k) unfolding $\boldsymbol{\mathcal{A}}_{(j,k)}$ and $\sigma_s\left(\left[\boldsymbol{\mathcal{A}}_{(j,k)}\right]_{\cdot,\cdot,l}\right)$ denotes the s-th singular value of the slice. For the estimation problem, we define

$$\tilde{L}\left(\boldsymbol{\mathcal{A}}\right) = L\left(\boldsymbol{\mathcal{A}}\right) + \sum_{l=1}^{\prod_{m\neq j,k} d_m} Q_\lambda\left(\left[\boldsymbol{\mathcal{A}}_{(j,k)}\right]_{\cdot,\cdot,l}\right),$$

where $Q_\lambda\left(\left[\boldsymbol{\mathcal{A}}_{(j,k)}\right]_{\cdot,\cdot,l}\right) = \sum_{s=1}^{s^{\text{all}}} q_\lambda\left(\sigma_s\left(\left[\boldsymbol{\mathcal{A}}_{(j,k)}\right]_{\cdot,\cdot,l}\right)\right)$.

Based on Lemma 13, for slice-wise lowrankness regularizer, we can similarly prove the following lemmas

Lemma 17. *Under Assumption 1 , $\rho^- > \zeta^-$, and the regularization parameter $\lambda \geq \frac{\left\|\left[\left[\mathfrak{X}^*(\boldsymbol{\mathcal{E}})\right]_{(j,k)}\right]_{\cdot,\cdot,l}\right\|_{\text{sp}}}{2n}$, we have*

$$\left\|\Pi_{\mathcal{F}^\perp}\left(\left[\hat{\boldsymbol{\mathcal{A}}}_{(j,k)}\right]_{\cdot,\cdot,l} - \left[\boldsymbol{\mathcal{A}}^*_{(j,k)}\right]_{\cdot,\cdot,l}\right)\right\|_* \leq 5\left\|\Pi_{\mathcal{F}}\left[\left[\hat{\boldsymbol{\mathcal{A}}}_{(j,k)}\right]_{\cdot,\cdot,l} - \left[\boldsymbol{\mathcal{A}}^*_{(j,k)}\right]_{\cdot,\cdot,l}\right]\right\|_*.$$

Proof. Similar to the proof of Lemma 14, we can prove the Lemma 17 □

From Lemma 13 and Lemma 17, we can prove the following general deterministic bound.

Lemma 18. *For the estimated parameter tensor $\hat{\boldsymbol{\mathcal{A}}}$ and the true parameter tensor $\boldsymbol{\mathcal{A}}^*$, we have*

$$\left\|\hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^*\right\|_{\text{F}} \leq \frac{1}{(\rho^- - \zeta^-)}\sqrt{\sum_{l=1}^{\prod_{m\neq j,k} d_m}\left[\sqrt{|\mathcal{S}_5^{\text{I}}|}\left\|\Pi_{\mathcal{F}}\left(\left[\left[\mathfrak{X}^*(\boldsymbol{\mathcal{E}})\right]_{(j,k)}\right]_{\cdot,\cdot,l}\right)\right\|_{\text{sp}} + 3\lambda\sqrt{|\mathcal{S}_5^{\text{II}}|}\right]^2}.$$

Proof. □

Similar to the proof for Lemma 16, we can derive the error bound for the slice-wise low-rankness regularizer.

Lemma 19. *Suppose $\boldsymbol{\mathcal{A}}^* \in \mathbb{R}^{d_1\times\cdots\times d_N}$ with rank of each slices $|\mathcal{S}_5|$. Then the error bound between the oracle estimator $\hat{\boldsymbol{\mathcal{A}}}^O$ and the true $\boldsymbol{\mathcal{A}}^*$ satisfies*

$$\left\|\hat{\boldsymbol{\mathcal{A}}}^O - \boldsymbol{\mathcal{A}}^*\right\|_{\text{F}} = \sqrt{\sum_{l=1}^{\prod_{m\neq j,k} d_m}\left\|\left[\left[\hat{\boldsymbol{\mathcal{A}}}^O - \boldsymbol{\mathcal{A}}^*\right]_{(j,k)}\right]_{\cdot,\cdot,l}\right\|_{\text{F}}^2} \lesssim \frac{2\sqrt{|\mathcal{S}_5|}\left\|\Pi_{\mathcal{F}}\left(\left[\left[\mathfrak{X}^*(\boldsymbol{\mathcal{E}})\right]_{(j,k)}\right]_{\cdot,\cdot,l}\right)\right\|_{\text{sp}}}{n\rho^-}.$$

Proof. With Lemma 19, we can also obtain that $\hat{\boldsymbol{\mathcal{A}}}^O = \hat{\boldsymbol{\mathcal{A}}}$. Similarly, we can prove the Corollary 5. □

H Proof of the Theorem 5

Lemma 20 (Slepian's Lemma)**.** *Let $\{G_s, s\in\mathcal{S}\}$ and $\{H_s, s\in\mathcal{S}\}$ be two centered Gaussian processes defined over the same index set $\mathcal{S}$. Suppose that both processes are almost surely bounded. For any $s,t\in\mathcal{S}$, if*

$$\mathbb{E}(G_s - G_t)^2 \leq \mathbb{E}(H_s - H_t)^2,$$

then it follows that

$$\mathbb{E}\big[\sup_{s\in\mathcal{S}} G_s\big] \leq \mathbb{E}\big[\sup_{s\in\mathcal{S}} H_s\big].$$

Furthermore, if the second moments of both processes are equal, i.e., $\mathbb{E}(G_s^2) = \mathbb{E}(H_s^2)$ for all $s\in\mathcal{S}$, then for all $x>0$, the following inequality holds:

$$\mathbb{P}\left\{\sup_{s\in\mathcal{S}} G_s > x\right\} \leq \mathbb{P}\left\{\sup_{s\in\mathcal{S}} H_s > x\right\}.$$

Lemma 21. *Let Z denote a centralized χ^2-distributed random variable with m degrees of freedom. For all $x \geq 0$, the following tail bounds hold:*

$$\mathbb{P}[Z - m \geq 2\sqrt{mx} + 2x] \leq \exp(-x), \quad \text{and}$$
$$\mathbb{P}[Z - m \leq -2\sqrt{mx}] \leq \exp(-x).$$

Lemma 22 (Theorem 3.8 in Massart (2007))**.** *Let $\alpha \sim \mathbb{N}(0, I_{d\times d})$ denote a d-dimensional Gaussian random variable. Consider a function $F : \mathbb{R}^d \to \mathbb{R}$ that satisfies a Lipschitz condition, i.e.,*

$$|F(x) - F(y)| \leq L\|x - y\|_2 \quad \text{for all } x, y \in \mathbb{R}^d,$$

where L is the Lipschitz constant. Then for all $t > 0$, we have the following concentration inequality for the deviation of $F(\alpha)$ from its expectation:

$$\mathbb{P}\big[|F(\alpha) - \mathbb{E}[F(\alpha)]| \geq t\big] \leq 2\exp\left(-\frac{t^2}{2L^2}\right).$$

Lemma 23. *Let $\boldsymbol{\mathcal{B}}$ be an tensor with each entry following a Gaussian distribution. Suppose that $\lambda \geq 4\eta\sqrt{\frac{\kappa}{n}}\mathbb{E}[R^*_\lambda(\boldsymbol{\mathcal{B}})]$, where $R^*_\lambda(\boldsymbol{\mathcal{B}})$ is the dual norm of $R_\lambda(\boldsymbol{\mathcal{B}})$. Then, with probability at least $1 - \exp\left(-\mathbb{E}[R^*_\lambda(\boldsymbol{\mathcal{B}})]^2\right)$, we have*

$$\lambda \geq 4R^*_\lambda\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{\mathcal{E}}^{(i)} \otimes \boldsymbol{\mathcal{X}}^{(i)}\right).$$

Proof. Recall that we have set

$$\lambda \geq 4\eta\sqrt{\frac{\kappa}{n}}\mathbb{E}[R^*_\lambda(\boldsymbol{\mathcal{B}})].$$

First we show that $\lambda \geq 2\eta\sqrt{\frac{\kappa}{n}}R^*_\lambda(\boldsymbol{\mathcal{B}})$ with high probability using concentration of Lipschitz functions for Gaussian random variables, due to Lemma 22. First we prove that $f(\boldsymbol{\mathcal{B}}) = R^*_\lambda(\boldsymbol{\mathcal{B}}) = \sup_{\boldsymbol{\mathcal{A}}:R_\lambda(\boldsymbol{\mathcal{A}})\leq 1}\langle \boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{A}}\rangle$ is a 1-Lipschitz function in terms of $\boldsymbol{\mathcal{B}}$. In particular note that:

$$f(\boldsymbol{\mathcal{B}}) - f(\boldsymbol{\mathcal{B}}') = \sup_{\boldsymbol{\mathcal{A}}:R_\lambda(\boldsymbol{\mathcal{A}})\leq 1}\langle \boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{A}}\rangle - \sup_{\boldsymbol{\mathcal{A}}:R_\lambda(\boldsymbol{\mathcal{A}})\leq 1}\langle \boldsymbol{\mathcal{B}}', \boldsymbol{\mathcal{A}}\rangle.$$

Let $\tilde{\boldsymbol{\mathcal{A}}} := \arg\max_{\boldsymbol{\mathcal{A}}:R_\lambda(\boldsymbol{\mathcal{A}})\leq 1}\langle \boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{A}}\rangle$. Then

$$\sup_{\boldsymbol{\mathcal{A}}:R_\lambda(\boldsymbol{\mathcal{A}})\leq 1}\langle \boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{A}}\rangle - \sup_{\boldsymbol{\mathcal{A}}:R_\lambda(\boldsymbol{\mathcal{A}})\leq 1}\langle \boldsymbol{\mathcal{B}}', \boldsymbol{\mathcal{A}}\rangle \leq \langle \boldsymbol{\mathcal{B}}, \tilde{\boldsymbol{\mathcal{A}}}\rangle - \langle \boldsymbol{\mathcal{B}}', \tilde{\boldsymbol{\mathcal{A}}}\rangle \leq \sup_{\boldsymbol{\mathcal{A}}:R_\lambda(\boldsymbol{\mathcal{A}})\leq 1}\langle \boldsymbol{\mathcal{B}} - \boldsymbol{\mathcal{B}}', \boldsymbol{\mathcal{A}}\rangle \leq \|\boldsymbol{\mathcal{B}} - \boldsymbol{\mathcal{B}}'\|_{\mathrm{F}},$$

Therefore, applying Lemma 22, we obtain

$$\mathbb{P}\left\{\left|\sup_{\boldsymbol{\mathcal{A}}:R_\lambda(\boldsymbol{\mathcal{A}})\leq 1}\langle \boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{A}}\rangle - \mathbb{E}\left[\sup_{\boldsymbol{\mathcal{A}}:R_\lambda(\boldsymbol{\mathcal{A}})\leq 1}\langle \boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{A}}\rangle\right]\right| > w(\Omega)\right\} \leq 2\exp\left(-\frac{1}{2}w^2[\Omega]\right).$$

Therefore,

$$\lambda \geq 2\eta\sqrt{\frac{\kappa}{n}}R^*_\lambda(\boldsymbol{\mathcal{B}})$$

with probability at least $1 - 2\exp\left(-\frac{w^2(\Omega)}{2}\right)$.

To complete the proof, we use a Gaussian comparison inequality between the supremum of the process $\sqrt{\frac{\kappa}{n}}\langle \boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{A}}\rangle$ and $\sum_{i=1}^n\langle \boldsymbol{\mathcal{E}}^{(i)} \otimes \boldsymbol{\mathcal{X}}^{(i)}, \boldsymbol{\mathcal{A}}\rangle/n$ over the set Ω. Recall that:

$$R^*_\lambda\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{\mathcal{E}}^{(i)} \otimes \boldsymbol{\mathcal{X}}^{(i)}\right) = \sup_{\boldsymbol{\mathcal{E}}\in\Omega}\left\langle \boldsymbol{\mathcal{A}}, \frac{1}{n}\sum_{i=1}^n \boldsymbol{\mathcal{E}}^{(i)} \otimes \boldsymbol{\mathcal{X}}^{(i)}\right\rangle.$$

Recall that each $\boldsymbol{\mathcal{E}}^{(i)} \in \mathbb{R}^{d_{M+1}+d_{M+2}\cdots+d_N}$ with each entry having variance η^2 and $\mathrm{vec}(\boldsymbol{\mathcal{X}}) \in \mathbb{R}^{nd_1d_2\cdots d_M}$ is a Gaussian vector covariance $\boldsymbol{\Sigma}$. First we condition on all the $\boldsymbol{\mathcal{E}}^{(i)}$'s, which are independent of the $\boldsymbol{\mathcal{X}}^{(i)}$'s. Further

let $\{u^{(i)} : i = 1, \ldots, n\}$ be i.i.d. standard normal Gaussian tensors where $u^{(i)} \in \mathbb{R}^{d_d \times d_2 \times \cdots \times d_N}$. First we condition on all the $\boldsymbol{\mathcal{E}}^{(i)}$'s, which are independent of the $\boldsymbol{\mathcal{X}}^{(i)}$'s. Using a standard Gaussian comparison inequality, if we condition on the $\boldsymbol{\mathcal{E}}^{(i)}$'s we have

$$\mathbb{P}\left\{\sup_{\boldsymbol{\mathcal{A}}:R_\lambda(\boldsymbol{\mathcal{A}})\leq 1} \frac{1}{n}\sum_{i=1}^{n}\langle \boldsymbol{\mathcal{E}}^{(i)} \otimes \boldsymbol{\mathcal{X}}^{(i)}, \boldsymbol{\mathcal{A}}\rangle > x\right\} \leq \mathbb{P}\left\{\sup_{\boldsymbol{\mathcal{A}}:R_\lambda(\boldsymbol{\mathcal{A}})\leq 1} \frac{1}{n}\sum_{i=1}^{n}\langle \boldsymbol{\mathcal{E}}^{(i)} \otimes u^{(i)}, \boldsymbol{\mathcal{A}}\rangle > \frac{x}{\sqrt{\kappa}}\right\},$$

since $\mathrm{cov}(\tilde{\boldsymbol{\mathcal{X}}}) = \boldsymbol{\Sigma} \preceq \kappa \boldsymbol{I}$.

Now we define the $u_j \in \mathbb{R}^n$ as the standard random vector where $1 \leq j \leq d_M$ and $u_j = (u_j^{(1)}, \ldots, u_j^{(n)})$. Conditioning on the $u^{(i)}$'s and dealing with the randomness in the $\boldsymbol{\mathcal{E}}^{(i)}$'s,

$$\frac{1}{n}\sum_{i=1}^{n}\langle \boldsymbol{\mathcal{E}}^{(i)} \otimes u^{(i)}, \boldsymbol{\mathcal{A}}\rangle \leq \max_{1\leq j\leq d_M} \frac{||u_j||_{l^2}}{\sqrt{n}} \frac{\eta}{\sqrt{n}}\langle \boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{A}}\rangle,$$

where $\boldsymbol{\mathcal{B}} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}$ is an i.i.d. standard normal tensor, since the $\boldsymbol{\mathcal{E}}^{(i)}$'s are i.i.d. standard normal. Now we upper bound

$$\max_{1\leq j\leq d_M} \frac{||u_j||_{l^2}}{\sqrt{n}},$$

using standard χ^2 tail bounds. Since $||u_j||_{l^2}^2/n$ is a χ^2 random variable with n degrees of freedom, for each j,

$$\mathbb{P}\left\{||u_j||_{l^2}^2/n \geq 4\right\} \leq \exp(-n),$$

using χ^2 tail bounds.

Now taking the union bound over d_M,

$$\mathbb{P}\left\{\max_{1\leq j\leq d_M} \frac{||u_j||_{l^2}}{\sqrt{n}} \geq 2\right\} \leq \exp(\log(d_M) - n),$$

and provided $n \geq 2\log(d_M)$, it follows that with probability greater than $1 - \exp(-n/2)$, $\max_{1\leq j\leq d_M} ||u_j||_{l^2} \leq 2$. Therefore, with probability at least $1 - \exp(-n/2)$,

$$\frac{1}{n}\sum_{i=1}^{n}\langle \boldsymbol{\mathcal{E}}^{(i)} \otimes u^{(i)}, \boldsymbol{\mathcal{A}}\rangle \leq \frac{2\eta}{\sqrt{n}}\langle \boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{A}}\rangle.$$

Now we apply Lemma 20 to complete the proof,

$$\mathbb{P}\left\{\sup_{R_\lambda(A)\leq 1} \frac{1}{n}\sum_{i=1}^{n}\langle \boldsymbol{\mathcal{E}}^{(i)} \otimes u^{(i)}, \boldsymbol{\mathcal{A}}\rangle > x\right\} \leq \mathbb{P}\left\{\sup_{R_\lambda(A)\leq 1} \frac{2\eta}{\sqrt{n}}\langle \boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{A}}\rangle > x\right\},$$

for all $x > 0$. Substituting x by $x/\sqrt{\kappa}$ means that

$$\mathbb{P}\left\{R_\lambda^*\left(\frac{1}{n}\sum_{i=1}^{n}\langle \boldsymbol{\mathcal{E}}^{(i)} \otimes u^{(i)}, \boldsymbol{\mathcal{A}}\rangle\right) > x\right\} \leq \mathbb{P}\left\{2\eta\sqrt{\frac{\kappa}{n}}R_\lambda^*(\boldsymbol{\mathcal{B}}) > x\right\},$$

for any $x > 0$. This completes the proof.

In light of Lemma 11, for the remainder of the proof, we can condition on the event that

$$\lambda \geq 2R_\lambda^*\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\mathcal{E}}^{(i)} \otimes \boldsymbol{\mathcal{X}}^{(i)}\right).$$

Under this event,

$$\frac{1}{2n}\sum_{i=1}^{n}||\boldsymbol{\mathcal{X}}^{(i)}, \boldsymbol{\Delta}||_{\mathrm{F}}^2 \leq \frac{1}{2}R_\lambda(\boldsymbol{\Delta}) + \lambda(R_\lambda(\boldsymbol{\Delta_0}) - R_\lambda(\boldsymbol{\Delta}^\perp)) \leq \left(\frac{3}{2}\right)\mathcal{R}(\Delta_0) - \left(c_R - \frac{1}{\eta_R}\right)\mathcal{R}(\Delta^\perp),$$

where $\mathbf{\Delta}^{\perp}$ represents a projection into the complementary space of $\mathbf{\Delta}$ and $\mathbf{\Delta}_0$ represents a projection into the zero space of $\mathbf{\Delta}$.

Since

$$\frac{1}{2n}\sum_{i=1}^{n}||\langle \mathbf{\Delta}, \boldsymbol{\mathcal{X}}^{(i)}\rangle||_{\mathrm{F}}^2 > 0,$$

we get

$$R_\lambda(\mathbf{\Delta}^{\perp}) \leq 3R_\lambda(\mathbf{\Delta_0}).$$

Here we define the cone

$$\mathcal{C} = \left\{\mathbf{\Delta} \mid R_\lambda(\mathbf{\Delta}^{\perp}) \leq 3R_\lambda(\mathbf{\Delta}_0)\right\}$$

and know that $\mathbf{\Delta} \in \mathcal{C}$. Hence

$$\frac{1}{2n}\sum_{i=1}^{n}\|\langle \boldsymbol{\mathcal{X}}^{(i)}, \mathbf{\Delta}\rangle\|_{\mathrm{F}}^2 \leq \frac{3}{2}\lambda R_\lambda(\mathbf{\Delta}_0) \leq \frac{3}{2}(s(\boldsymbol{\mathcal{A}})\lambda)^{-1/4}\|\mathbf{\Delta}\|_{\mathrm{F}}.$$

Let us define the following set:

$$C(\delta_{\mathrm{F}}) := \{\mathbf{\Delta} \in \mathbb{R}^{d_d \times d_2 \times \cdots \times d_N} | R_\lambda(\mathbf{\Delta}^{\perp}) \leq 3R_\lambda(\mathbf{\Delta_0})\}.$$

For convenience, in the remainder of this proof let

$$\delta_n := 3(s(\mathcal{A})\lambda)^{-1/4}.$$

Further, let us define the event:

$$E(\delta_{\mathrm{F}}) := \left\{\mathbf{\Delta} \in C(\delta_{\mathrm{F}}) \mid ||\mathbf{\Delta}||_{\mathrm{F}} \geq \sqrt{\kappa}\delta_{\mathrm{F}}\right\}.$$

We claim that it suffices to show that $E(\boldsymbol{\delta}_{\mathrm{F}})$ holds with probability at least $1 - \exp(-2c)$ for some constant $c > 0$. In particular, given an arbitrary non-zero $\mathbf{\Delta} \in C(\boldsymbol{\delta}_{\mathrm{F}})$, consider the re-scaled tensor

$$\tilde{\mathbf{\Delta}} = \kappa\eta\frac{\mathbf{\Delta}}{||\mathbf{\Delta}||_{\mathrm{F}}}.$$

Since $\mathbf{\Delta} \in C(\delta_{\mathrm{F}})$ and $C(\delta_{\mathrm{F}})$ is star-shaped, we have $\tilde{\mathbf{\Delta}} \in C(\mathbf{\Delta}_{\mathrm{F}})$ and $||\tilde{\mathbf{\Delta}}||_{\mathrm{F}} = \frac{\kappa\mathbf{\Delta}_{\mathrm{F}}}{||\mathbf{\Delta}||_{\mathrm{F}}}$ by construction. Consequently, it is sufficient to prove that $E(\delta_{\mathrm{F}})$ holds with high probability.

□

Lemma 24. *Suppose that Assumptions 1 and 2 hold. Assume that for any $c > 0$, there exists an n such that $\sqrt{s}\lambda < c$. Then there exists a $\tilde{c} > 0$ such that*

$$\mathbb{P}(E(\delta_{\mathrm{F}})) \geq 1 - \exp(-2\tilde{c}).$$

Proof. We define the random variable

$$Z_n(C(\delta_{\mathrm{F}})) = \sup_{\Delta \in C(\delta_{\mathrm{F}})}\left\{\frac{\kappa^4\delta_{\mathrm{F}}^2}{4} - \frac{1}{n}\sum_{i=1}^{n}||\langle \mathbf{\Delta}, \boldsymbol{\mathcal{X}}^{(i)}\rangle||_{\mathrm{F}}^2\right\},$$

then it suffices to show that

$$Z_n(C(\delta_{\mathrm{F}})) \leq \frac{\kappa^4\delta_{\mathrm{F}}^2}{2}.$$

With a slight abuse of notation, it follows that

$$||\mathbf{\Delta}||_{\mathrm{F}}^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{d_N}{d_M}\sum_{m=1}^{d_N/d_M}\langle \tilde{\mathbf{\Delta}}_m, \mathrm{vec}(\boldsymbol{\mathcal{X}}^{(i)})\rangle^2,$$

where $\tilde{\boldsymbol{\Delta}}_m \in \mathbb{R}^{d_M}$ and clearly $\text{vec}(\boldsymbol{\mathcal{X}}^{(i)}) \in \mathbb{R}^{d_M}$. In order to complete the proof we make use of a truncation argument. For a constant $\tau > 0$ to be chosen later, consider the truncated function

$$\phi_\tau \left(\tilde{\boldsymbol{\Delta}}_m, \text{vec}(\boldsymbol{\mathcal{X}}) \right)$$

and define

$$\boldsymbol{\Delta}_{m,\tau}(\boldsymbol{\mathcal{X}}) = \text{sign} \left(\langle \tilde{\boldsymbol{\Delta}}_m, \text{vec}(X) \rangle \right) \sqrt{\phi_\tau \left(\langle \tilde{\boldsymbol{\Delta}}_m, \text{vec}(\boldsymbol{\mathcal{X}}) \right)}.$$

Further let

$$\boldsymbol{\Delta}_r(\boldsymbol{\mathcal{X}}) = \left(\boldsymbol{\Delta}_{1,\tau}(\boldsymbol{\mathcal{X}}), \boldsymbol{\Delta}_{2,\tau}(\boldsymbol{\mathcal{X}}), \ldots, \boldsymbol{\Delta}_{d_N/d_M,\tau}(\boldsymbol{\mathcal{X}}) \right).$$

The remainder of the proof consists of showing that for a suitable τ,

$$||\boldsymbol{\Delta}_r||_{\text{F}}^2 \geq \frac{3}{4} ||\boldsymbol{\Delta}||_{\text{F}}^2,$$

for all $\boldsymbol{\Delta} \in C(\boldsymbol{\Delta}_{\text{F}})$ and

$$P \left\{ Z_n \geq \frac{\kappa^2 \delta_{\text{F}}^2}{4} \right\} \leq c_1 \exp(-c_2 n \delta_{\text{F}}^2),$$

where $Z_n := \sup_{\boldsymbol{\Delta} \in C(\delta_{\text{F}})} ||\boldsymbol{\Delta}||_{\text{F}}^2 - ||\boldsymbol{\Delta}_r||_{\text{F}}^2$. By definition

$$||\boldsymbol{\Delta}_m||_{\text{F}}^2 - ||\boldsymbol{\Delta}_r||_{\text{F}}^2 \leq \mathbb{E}[\left(\langle \tilde{\boldsymbol{\Delta}}_m, \text{vec}(\boldsymbol{\mathcal{X}}) \rangle \right)^2 1 \left[\langle \tilde{\boldsymbol{\Delta}}_m, \text{vec}(\boldsymbol{\mathcal{X}}) \rangle \geq \tau \right]] \leq \frac{\mathbb{E}[\langle \tilde{\boldsymbol{\Delta}}_m, \text{vec}(\boldsymbol{\mathcal{X}}) \rangle]^4}{\tau^2},$$

where the last inequality follows from the Cauchy-Schwarz inequality and the final inequality follows from Markov's inequality. Since $\langle \tilde{\boldsymbol{\Delta}}_m, \text{vec}(\boldsymbol{\mathcal{X}}) \rangle$ is a Gaussian random variable,

$$E \left[\langle \tilde{\boldsymbol{\Delta}}_m, \text{vec}(\boldsymbol{\mathcal{X}}) \rangle \right] \leq 3E[||\tilde{\boldsymbol{\Delta}}_m||_{\text{F}}].$$

Therefore $||\boldsymbol{\Delta}||_{\text{F}}^2 \geq \frac{3}{4} ||\boldsymbol{\Delta}||_{\text{F}}^2$.

From Raskutti et al. (2019), we see the high probability bound on Z_n by first upper bounding $E[Z_n]$. Thus, we obtain

$$\left\| \hat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^* \right\|_{\text{F}} \leq 3\lambda\kappa\Psi \left(\boldsymbol{\mathcal{A}} \right),$$

This completes the proof.

□