COMMUNICATION-EFFICIENT DISTRIBUTED MULTILINEAR MULTITASK LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

In multi-task learning (MTL), the parameter of each task is often multidimensional and exhibits high-order intra-task relatedness. Meanwhile, multiple multidimensional parameters in MTL naturally induce the multi-way inter-task relatedness. This paper proposes a distributed multilinear MTL method that models these two types of relatedness through a shared multilinear representation, which models the parameter of each task through mode-wise factors shared across tasks and a task-specific core tensor. We instantiate our method for the multi-task high-dimensional tensor linear regression (MTTLR) problem. A two-stage communication-efficient distributed method, termed D-AltMin-PGD, is developed to minimize the induced non-convex loss to jointly learn the models. For the Gaussian input and additive noise, we prove that from initialization via the proposed distributed power-iteration, D-AltMin-PGD converges linearly to a statistical neighborhood of truth parameters despite nonconvexity. The resulting statistical error separates into a shared-factor term that benefits from the samples across all tasks and a task-specific core term determined by the per-task samples, thereby establishing a provable multi-task gain. We further propose a preconditioned variant, D-AltMin-P2GD, to accelerate convergence in ill-conditioned regimes with negligible extra communication cost. Finally, we establish minimax lower bound for proposed MTTLR under Gaussian input and noise. Extensive experiments validate our theoretical results and demonstrate the efficiency of the proposed method.

1 INTRODUCTION

Multi-task learning (MTL) (Caruana, 1997) is a learning paradigm that jointly learns multiple related tasks to improve the performance over learning each task individually. In recent years, MTL has become a versatile framework across a wide range of domains, including computer vision (Vandenhende et al., 2021; Zhang et al., 2019), natural language processing (Chen et al., 2024; Zhou et al., 2022), speech and audio (Geng et al., 2025; Kim et al., 2017), reinforcement learning (Huang et al., 2022; Sodhani et al., 2021), and large language models (Audibert et al., 2023; Wang et al., 2025; 2024). A central premise behind these successes is that MTL can exploit relatedness among tasks as an inductive bias (Caruana, 1997). From the perspective of how task relatedness is characterized (Zhang & Yang, 2018; 2021), existing MTL methods can be broadly categorized into feature-based (Argyriou et al., 2006; Zhang et al., 2025a), instance-based (Bickel et al., 2008; Grégoire et al., 2024), and parameter-based algorithms (Argyriou et al., 2007; Evgeniou & Pontil, 2004; Zhou & Zhao, 2015). Among these categories, parameter-based approaches are widely used in practice (Ruder, 2017; Zhang & Yang, 2018) and the main aim of this paper is to study parameter-based MTL.

Parameter-based MTL approaches encode task-relatedness by imposing constraints, priors, or regularizers on model parameters (Yu et al., 2025; Zhang & Yang, 2018). Consequently, the performance gain of MTL largely depends on how well algorithms can capture and leverage correlations among task parameters (Sodhani et al., 2021; Yu et al., 2025). For parameter-based MTL, each task is typically parameterized by a weight vector (Yang & Hospedales, 2017b; Yu et al., 2025), and task parameters are stacked into a matrix whose structure is regularized to capture relatedness across tasks (Zhang & Yang, 2021). Representative regularizations include low-rank constraints, manifold and graph-based regularization (Agarwal et al., 2010; Ando et al., 2005; Chen et al., 2009; Han & Zhang, 2016; Lin et al., 2026; Pong et al., 2010). This matrix-based relatedness has provided a principled foundation for both applications and theoretical analysis of MTL (Zhang & Yang, 2018; 2021).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

However, the learnable parameters of each task are naturally represented as a high-order tensor rather than single vector. For instance, convolutional kernels in CNNs (Krizhevsky et al., 2012) and projection weights in attention modules are multi-way parameters (Zhang et al., 2025b). Tensor parameters in each task exhibit informative multi-way structure, which we refer to as intra-task correlations, such as inter-layer correlations (KALLE et al., 2025; Novikov et al., 2015) and structured redundancy across attention heads in Transformers (Ma et al., 2019; Zhang et al., 2025b). Thus, naively flattening tensor parameters as long vectors and then applying matrix-based MTL can discard these intra-task correlations and fail to capture the multi-way structure of each task (Yang & Hospedales, 2017b), which leads to inferior characterization of inter-task relatedness and suboptimal learning performance (Liu et al., 2021; Long et al., 2017).

In addition, in real-world deployments, the multiple tasks are often distributed in different locations due to various restrictions (Liu et al., 2017; Smith et al., 2017), such as memory constraints or data privacy concerns (Liao et al., 2024; Lu et al., 2024; Wu et al.). Thus, designing algorithms for MTL in the distributed paradigm is essential. Therefore, this naturally raises the question: *How can we design an efficient distributed MTL method that can preserve and exploit the intrinsic intra-task correlations within each task to capture inter-task relatedness compactly?*

Motivated by the above question, we propose a new distributed multilinear multitask learning model that leverages a shared multilinear representation to exploit both intra-task and inter-task correlations. Specifically, we consider the multitask tensor linear regression (MTTLR) where K tasks are distributed on K locations, and each task has m training samples $\left\{\left(\boldsymbol{\mathcal{A}}_i^k, y_i^k\right)\right\}_{i\in[m]}$ given by

$$y_i^k = \left\langle \boldsymbol{\mathcal{A}}_i^k, \boldsymbol{\mathcal{X}}_k^\star \right\rangle + \xi_{k,i},\ \forall (i,k) \in [m]\times[K], \quad \text{s.t. } \operatorname{rank}\left(\boldsymbol{X}_{(n)}^\star\right) = r_n,\ \forall n \in [N]. \tag{1}$$

$\boldsymbol{\mathcal{X}}_k^\star$ is unknown parameter to be learned for k-th task and $\boldsymbol{X}_{(n)}^\star := \left[\left(\boldsymbol{X}_1^\star\right)_{(n)}, \cdots, \left(\boldsymbol{X}_K^\star\right)_{(n)}\right] \in \mathbb{R}^{d_n \times K d_n^-}$ denotes the stacking of unfolding matrices of all tasks parameters along mode-n. $\boldsymbol{\mathcal{A}}_i^k \in \mathbb{R}^{d_1\times d_2\times\cdots\times d_N}$ and y_i^k are respectively an order N tensor as an input and the corresponding scalar output. $\xi_{k,i}$ is independent additive noise. The aim of MTTLR is collaboratively to learn $\{\boldsymbol{\mathcal{X}}_k^\star\}_{k\in K}$ using training samples in distributed setting. We model the task-relatedness by the low rank constraint $\operatorname{rank}\left(\boldsymbol{X}_{(n)}^\star\right) = r_n \ll d_n$, which captures both intra-task and inter-task correlations simultaneously by requiring all task parameters to lie in a low-rank subspace along each mode. Given this linear MTL model (1), the question is:

Can we design an efficient distributed MTL method to learn $\{\boldsymbol{\mathcal{X}}_k^\star\}_{k\in[K]}$ with provable guarantees to demonstrate the benefit of MTL for the proposed MTTLR model (1)?

To answer the above question, our main contributions are as follows.

- We introduce a distributed multilinear MTL framework for high-dimensional MTTLR, in which tasks share mode-wise factor matrices and retain task-specific core tensors. The resulting shared multilinear representation preserves intra-task structure and captures multi-way task-relatedness.
- We develop a two-stage communication-efficient distributed algorithm: distributed tensor power iteration for spectral initialization, followed by D-AltMin-PGD, which alternates between exact local core updates and server-side projected gradient updates of shared factors. Under Gaussian designs and additive Gaussian noise, we establish nonasymptotic guarantees of D-AltMin-PGD that characterize both its linear optimization convergence rate and statistical error rate.
- We propose a preconditioned variant, D-AltMin-P2GD, that achieves condition number independent linear convergence rate while preserving the same statistical error rate as D-AltMin-PGD, with negligible extra communication and computation.
- We show that per-task sample complexity decreases with the number of tasks and the statistical error consists of a shared-factor term with rate $1/\sqrt{mK}$ and a task-specific core term with rate $1/\sqrt{m}$, revealing a provable multi-task gain. We further establish a minimax lower bound for noisy MTTLR, confirming the optimal $1/\sqrt{mK}$ multi-task gain and the unavoidable $1/\sqrt{m}$ task-specific error. Theoretical comparison with existing MTL models in Table 1 and further experiments validate the advantages of the proposed model and methods.

Methods	Computational Complexity	Communication Complexity	Sample Complexity
MTLR (Thekumparampil et al., 2021; Vaswani, 2024)	$\frac{d^4r^4}{K} + d^2r^4$	d^2r^2	$\frac{(d^2+K)r^4}{K} + \frac{(d^2+K)r^2}{K}\log\frac{1}{\epsilon}$
MTTPR (Wu et al., 2025)	$\frac{d^3(d+K)r^2}{K} + \left(\frac{d^4r^3}{K} + d^3r^3\right)\log\frac{1}{\epsilon}$	$d^2r\log\frac{1}{\epsilon}$	$\frac{d(d+K)r^2}{K} + \left(\frac{d^2r}{K} + r\right)\log\frac{1}{\epsilon}$
Proposed MTTLR	$\frac{d^4r^5}{K} + \left(\frac{d^3r^3}{K} + r^2\right)\log\frac{1}{\epsilon}$	$dr\log\frac{1}{\epsilon}$	$\frac{dr^{\frac{3}{2}}}{\sqrt{K}} + \left(\frac{dr^2}{K} + r^2\right)\log\frac{1}{\epsilon}$

Table 1: Comparison of existing multi-task linear regression methods for achieving $\|\hat{\boldsymbol{\mathcal{X}}}_k - \boldsymbol{\mathcal{X}}_k^\star\|_F \leq \epsilon$ when solving (1) under noiseless setting. The comparison is specialized to the $N = 2$ setting and $d_1 = d_2 = d$, $r_1 = r_2 = r$ with $r \ll d$.

2 Notations and Preliminaries

We use x, $\boldsymbol{x}$, $\boldsymbol{X}$, and $\boldsymbol{\mathcal{X}}$ to denote a scalar, vector, matrix, and tensor, respectively. For an N-order tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1 \times \cdots \times d_N}$, let $\mathrm{Vec}(\boldsymbol{\mathcal{X}}) \in \mathbb{R}^{\prod_{n=1}^N d_n \times 1}$ denote its vectorization and let $\boldsymbol{X}_{(n)} \in \mathbb{R}^{d_n \times d_n^-}$ denote its mode-n unfolding, where d_n^- is defined as $d_n^- := \prod_{i \neq n} d_i$. The multilinear rank of $\boldsymbol{\mathcal{X}}$ is $\mathrm{rank}(\boldsymbol{\mathcal{X}}) = (\mathrm{rank}(\boldsymbol{X}_{(1)}), \ldots, \mathrm{rank}(\boldsymbol{X}_{(N)}))$. For $\boldsymbol{U} \in \mathbb{R}^{d_n' \times d_n}$, the mode-$n$ product $\boldsymbol{\mathcal{X}} \times_n \boldsymbol{U}$ satisfies $(\boldsymbol{\mathcal{X}} \times_n \boldsymbol{U})_{(n)} = \boldsymbol{U}\boldsymbol{X}_{(n)}$. We write $\boldsymbol{\mathcal{X}} \times \{\boldsymbol{U}_n\}_{n\in[N]} = \boldsymbol{\mathcal{X}} \times_1 \boldsymbol{U}_1 \times \cdots \times_N \boldsymbol{U}_N$ and $\boldsymbol{\mathcal{X}} \times_{-n} \{\boldsymbol{U}_n\}_{n\in[N]} = \boldsymbol{\mathcal{X}} \times_1 \boldsymbol{U}_1 \times \cdots \times_{n-1} \boldsymbol{U}_{n-1} \times_{n+1} \boldsymbol{U}_{n+1} \times \cdots \times_N \boldsymbol{U}_N$. Define $d_{\max} := \max_{n\in[N]} d_n$, $d_{\min} := \min_{n\in[N]} d_n$, $r_{\max} := \max_{n\in[N]} r_n$, $r_{\min} := \min_{n\in[N]} r_n$. We use $\langle \cdot, \cdot \rangle$ for the tensor inner product and $\|\boldsymbol{\mathcal{X}}\|_F := \sqrt{\langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}} \rangle}$ for the Frobenius norm. For a matrix $\boldsymbol{X}$, let $\sigma_{\max}(\boldsymbol{X})$ and $\sigma_{\min}(\boldsymbol{X})$ denote its largest and smallest nonzero singular values. For a tensor $\boldsymbol{\mathcal{X}}$, define $\sigma_{\max}(\boldsymbol{\mathcal{X}}) = \max_{n\in[N]} \sigma_{\max}(\boldsymbol{X}_{(n)})$ and $\sigma_{\min}(\boldsymbol{\mathcal{X}}) = \min_{n\in[N]} \sigma_{\min}(\boldsymbol{X}_{(n)})$. Finally, define $\bigotimes_{n=1}^N \boldsymbol{U}_n = \boldsymbol{U}_N \bigotimes \cdots \bigotimes \boldsymbol{U}_1$ and $\bigotimes_{j\neq n}^N \boldsymbol{U}_j = \boldsymbol{U}_N \bigotimes \cdots \boldsymbol{U}_{n+1} \bigotimes \boldsymbol{U}_{n-1} \bigotimes \cdots \boldsymbol{U}_1$.

3 Related Works

Low-rank parameter-based MTL. Parameter-based MTL commonly exploits low-rank task relatedness (Zhang & Yang, 2018; 2021). Early works focused on vector-valued task parameters and modeled inter-task structure through low-rank matrix representations or trace norm regularization (Agarwal et al., 2010; Ando et al., 2005; Caruana, 1997; Chen et al., 2009; Pong et al., 2010; Vaswani, 2024). Multilinear MTL extended this idea to multi-index tasks by imposing tensor norms or low-rank tensor decompositions (Liu et al., 2023a; Romera-Paredes et al., 2013; Wimalawarne et al., 2014; Wu & Sun; Zhang et al., 2023; Zhao et al., 2019). More recent works model relatedness directly over tensor-valued task parameters using trace norms, layer-wise tensor representations, tensor-variate priors, or coupled tensor decompositions (Ahmad et al., 2022; Borsoi et al., 2024; 2023; Cichocki et al., 2015; Fu et al., 2014; Kuang et al., 2023; Lin et al., 2016a; Long et al., 2017; Wu et al., 2025; Yang & Hospedales, 2017a;b; Zhang et al., 2026; 2021; 2022). In contrast, we study distributed shared multilinear representation with non-convex optimization, communication and statistical guarantees.

Theory for MTL. Main MTL theory establishes improved generalization or excess risk over single-task learning (Ando et al., 2005; Bak & Koo, 2023; Boursier et al., 2022; Chen et al., 2022; Du et al., 2021; Duchi et al., 2022; Knight & Duan, 2023; Liu, 2023; Maurer, 2006; Pontil & Maurer, 2013; Tian et al., 2025; Tripuraneni et al., 2021; Vaswani, 2024; Watkins et al., 2023; Wimalawarne et al., 2014; Wu & Sun; Wu et al., 2025; Yousefi et al., 2018; Zhang et al., 2023; 2021; 2022; Zhang, 2015). Most existing guarantees, however, are either optimization-independent or confined to convex models (Bak & Koo, 2023; Boursier et al., 2022; Liu, 2023; Pontil & Maurer, 2013; Wimalawarne et al., 2014; Yousefi et al., 2018; Zhang et al., 2023; 2021; 2022; Zhang, 2015). Recent optimization-dependent analyses of nonconvex shared-representation models are more closely related (Collins et al., 2021; Thekumparampil et al., 2021; Vaswani, 2024; Wu et al., 2025; Zhang et al., 2024), but they focus on vector- or matrix-valued parameters with linear shared representations. Our work instead studies arbitrary order tensor-valued parameters with a multilinear shared representation in a distributed setting, together with communication-efficient spectral initialization and a preconditioned variant.

4 PROPOSED METHOD

We first formulate the non-convex problem for solving the MTTLR model in (1). Then we introduce a two-stage communication-efficient distributed algorithm to solve the induced non-convex problem.

4.1 PROPOSED MODEL

Based on HOSVD Tucker decomposition (De Lathauwer et al., 2000) and constraint in (1), we can parameterize true parameters $\{\boldsymbol{\mathcal{X}}_k^\star\}_{k\in[K]}$ via the following shared multilinear representation

$$\boldsymbol{\mathcal{X}}_k^\star = \boldsymbol{\mathcal{S}}_k^\star \times_1 \boldsymbol{U}_1^\star \times \cdots \times_N \boldsymbol{U}_N^\star,\ {\boldsymbol{U}_n^\star}^\top \boldsymbol{U}_n^\star = \boldsymbol{I}_{r_n}, \boldsymbol{U}_n^\star \in \mathbb{R}^{d_n \times r_n}, \forall k \in [K], \forall n \in [N], \quad (2)$$

which the mode-wise tasks correlations are captured by shared factors $\{\boldsymbol{U}_n^\star\}_{n\in[N]}$ and each core tensor $\boldsymbol{\mathcal{S}}_k^\star \in \mathbb{R}^{r_1\times\cdots\times r_N}$ encodes the task-specific information. There is key quantity in MTTLR (1) that characterizes the difficulty of the associated optimization problem, and we define it as follows.

Definition 1. *The condition number of MTTLR is defined as* $\kappa := \max_{n\in[N]} \sigma_{\max}\left(\boldsymbol{X}_{(n)}^\star\right) / \min_{n\in[N]} \sigma_{\min}\left(\boldsymbol{X}_{(n)}^\star\right)$.

Recall that we denote $\sigma_{\max} := \max_{n\in[N]} \sigma_{\max}\left(\boldsymbol{X}_{(n)}^\star\right), \sigma_{\min} := \min_{n\in[N]} \sigma_{\min}\left(\boldsymbol{X}_{(n)}^\star\right)$. Based on shared multilinear representation in (2), we can induce the following non-convex loss

$$\min_{\substack{\{\boldsymbol{\mathcal{S}}_k\}_{k\in K},\\ \{\boldsymbol{U}_n\}_{n\in[N]}}} F\left(\{\boldsymbol{\mathcal{S}}_k\}_{k\in[K]}, \{\boldsymbol{U}_n\}_{n\in[N]}\right) := \sum_{k=1}^K f_k\left(\boldsymbol{\mathcal{S}}_k, \{\boldsymbol{U}_n\}_{n\in[N]}\right), \quad \text{s.t. } \boldsymbol{U}_n^\top \boldsymbol{U}_n = \boldsymbol{I}_{r_n}, \forall n \in [N] \quad (3)$$

to learn true parameters $\{\boldsymbol{\mathcal{X}}_k^\star\}_{k\in[K]}$ in (1), where $\boldsymbol{\mathcal{S}}_k \in \mathbb{R}^{r_1\times\cdots\times r_N}, \forall k \in [K], \boldsymbol{U}_n \in \mathbb{R}^{d_n\times r_n}, \forall n \in [N]$, m is per-task sample size. Each $f_k\left(\boldsymbol{\mathcal{S}}_k, \{\boldsymbol{U}_n\}_{n\in N}\right)$ is the loss for k-th task, which is defined as

$$f_k\left(\boldsymbol{\mathcal{S}}_k, \{\boldsymbol{U}_n\}_{n\in[N]}\right) := \sum_{i=1}^m \left(\left\langle \boldsymbol{\mathcal{A}}_i^k, \boldsymbol{\mathcal{S}}_k \times \{\boldsymbol{U}_n\}_{n\in[N]} \right\rangle - y_i^k\right)^2. \quad (4)$$

We can observe that the above loss in (3) involves more variables than linear shared representation and is highly non-convex due to shared multilinear representation.

4.2 DISTRIBUTED SPECTRAL INITIALIZATION

The aim in the first stage is to estimate shared factors $\left\{\boldsymbol{U}_n^0\right\}_{n\in[N]}$ that accordingly align with the true subspace $\{\boldsymbol{U}_n^\star\}_{n\in N}$. We denote the samples used in this stage as $\left\{\left(\boldsymbol{\mathcal{A}}_{k,i}^0, y_{k,i}^0\right)\right\}_{k\in[K], i\in[m_0]}$. We first construct the following aggregated second-order moment matrix along the n-th mode as

$$\hat{\boldsymbol{W}}_n := \frac{\sum_{k=1}^K \hat{\boldsymbol{W}}_{n,k}}{2m_0(m_0-1)}, \quad \hat{\boldsymbol{W}}_{n,k} := \sum_{i<i'} y_{k,i}^0 y_{k,i'}^0 \left(\left(\boldsymbol{A}_{k,i}^0\right)_{(n)} \left(\boldsymbol{A}_{k,i'}^0\right)_{(n)}^\top + \left(\boldsymbol{A}_{k,i'}^0\right)_{(n)} \left(\boldsymbol{A}_{k,i}^0\right)_{(n)}^\top\right). \quad (5)$$

The leading r_n eigenspace of $\hat{\boldsymbol{W}}_n$ estimates the subspace spanned by $\boldsymbol{U}_n^\star$. To avoid communicating $\sum_{n=1}^N d_n^2$ entries for server-side eigendecomposition, Algorithm 1 uses distributed power iteration, requiring only $\sum_{n=1}^N d_n r_n$ entries per round.

4.3 DISTRIBUTED ALTERNATIVE MINIMIZATION AND PROJECTED GD: D-ALTMIN-PGD

From the initialization $\{\boldsymbol{U}_n^0\}_{n\in[N]}$ obtained by Algorithm 1, we further propose a distributed local search method to alternatively update the shared factors and task-specific cores iteratively.

Exact Minimization for Task-Specific Cores. Given $\{\boldsymbol{U}_n^t\}_{n\in[N]}$, each task independently updates $\boldsymbol{\mathcal{S}}_k^t$ by solving the following local least-squares problem.

$$\boldsymbol{s}_k^t = \operatorname*{arg\,min}_{\boldsymbol{s}\in\mathbb{R}^{r_1\cdots r_N}} \left\| \boldsymbol{A}_k^t \left(\bigotimes_{n\neq 1}^N \boldsymbol{U}_n^t\right) \boldsymbol{s} - \boldsymbol{y}_k^t \right\|_F^2 \quad (6)$$

Algorithm 1 Distributed Spectral Initialization

Require: Initialization samples $\{(\boldsymbol{\mathcal{A}}_{k,i}^0, y_{k,i}^0)\}_{i\in[m_0],k\in[K]}$, ranks $\{r_n\}_{n\in[N]}$, iterations T.

Initialize: $\forall n \in [N]$, draw $\boldsymbol{Y}_n^0 \in \mathbb{R}^{d_n \times r_n}$ from Gaussian distribution and set $\boldsymbol{Z}_n^0 = \mathrm{Orth}(\boldsymbol{Y}_n^0)$.
for $t = 0, \dots, T-1$ **do**
  Broadcast: Server sends $\{\boldsymbol{Z}_n^t\}_{n\in[N]}$ to all tasks.
  for each task $k \in [K]$ **in parallel do**
    for $n = 1, \dots, N$ **do**
      $\boldsymbol{Y}_{k,n}^{t+1} = \frac{1}{2m_0(m_0-1)} \hat{\boldsymbol{W}}_{n,k} \boldsymbol{Z}_n^t$, where $\hat{\boldsymbol{W}}_{n,k}$ is defined in (5).
    end for
  end for
  Aggregation: Each task sends $\{\boldsymbol{Y}_{k,n}^{t+1}\}_{n\in[N]}$ to the server.
  for $n = 1, \dots, N$ **do**
    $\boldsymbol{Y}_n^{t+1} = \sum_{k=1}^K \boldsymbol{Y}_{k,n}^{t+1}$ and $\boldsymbol{Z}_n^{t+1} = \mathrm{Orth}(\boldsymbol{Y}_n^{t+1})$.
  end for
end for

Ensure: $\boldsymbol{U}_n^0 = \boldsymbol{Z}_n^T$, $\forall n \in [N]$.

The $\boldsymbol{A}_k^t := \left[\mathrm{Vec}\left(\boldsymbol{\mathcal{A}}_{k,i}^t\right)^\top; \cdots; \mathrm{Vec}\left(\boldsymbol{\mathcal{A}}_{k,m}^t\right)^\top\right] \in \mathbb{R}^{m \times \prod_{n=1}^N d_n}$ is the stacked vectorization of $\boldsymbol{\mathcal{A}}_{k,i}^t$.

Projected GD for Shared Factors. After locally updating $\{\boldsymbol{\mathcal{S}}_k^t\}_{k\in[K]}$, the server updates the shared factors by projected GD. Each task resamples $\{\tilde{\boldsymbol{\mathcal{A}}}_{k,i}^t, \tilde{y}_{k,i}^t\}_{i\in[m]}$ and computes the local gradient

$$\nabla_{\boldsymbol{U}_n} f_k\left(\boldsymbol{\mathcal{S}}_k^t, \{\boldsymbol{U}_n^t\}_{n\in[N]}\right) = \sum_{i=1}^m \left(\left\langle \tilde{\boldsymbol{\mathcal{A}}}_{k,i}^t, \boldsymbol{\mathcal{S}}_k^t \times \{\boldsymbol{U}_n^t\}_{n\in[N]} \right\rangle - \tilde{y}_{k,i}^t\right) \left(\tilde{\boldsymbol{A}}_{k,i}^t\right)_{(n)} \left(\bigotimes_{n\neq 1}^N \boldsymbol{U}_n^t\right) \left(\boldsymbol{S}_k^t\right)_{(n)}^\top. \tag{7}$$

The server then takes a PGD step using aggregated gradients and projects via QR decomposition:

$$\hat{\boldsymbol{U}}_n^{t+1} = \boldsymbol{U}_n^t - \eta \sum_{k=1}^K \nabla_{\boldsymbol{U}_n} f_k\left(\boldsymbol{\mathcal{S}}_k^t, \{\boldsymbol{U}_n^t\}_{n\in N}\right), \qquad \boldsymbol{U}_n^{t+1} = \mathrm{QR}\left(\hat{\boldsymbol{U}}_n^{t+1}\right). \tag{8}$$

Finally, the server broadcasts the updated factors. Thus, each round communicates only the gradients and the updated shared factors to maintain communication efficiency.

Remark 1. *Fresh samples in initialization and at each iteration of Algorithm 2 are used only to decouple iterate dependence in the analysis. Similar sample-splitting or resampling schemes are common in prior MTL theory (Collins et al., 2021; Thekumparampil et al., 2021; Vaswani, 2024; Wu et al., 2025; Zhang et al., 2024).*

4.4 Preconditioned D-AltMin-PGD

To further improve efficiency of optimization and mitigate the dependency of the convergence rate on the condition number $\kappa > 1$, we further propose a preconditioned variant, D-AltMin-P2GD, that replaces the aggregated gradient in the update of shared factors in (7) by its preconditioned version:

$$\hat{\boldsymbol{U}}_n^{t+1} = \boldsymbol{U}_n^t - \eta \left(\sum_{k=1}^K \nabla_{\boldsymbol{U}_n^t} f_k\left(\boldsymbol{\mathcal{S}}_k^t, \{\boldsymbol{U}_n^t\}_{n\in[N]}\right)\right) \left(\boldsymbol{S}_{(n)}^t \boldsymbol{S}_{(n)}^{t\top}\right)^{-1}, \quad \boldsymbol{U}_n^{t+1} = \mathrm{QR}\left(\hat{\boldsymbol{U}}_n^{t+1}\right), \tag{9}$$

where $\left(\boldsymbol{S}_{(n)}^t \boldsymbol{S}_{(n)}^{t\top}\right)^{-1}$ is the preconditioner with $\boldsymbol{S}_{(n)}^t := \left[(\boldsymbol{S}_1^t)_{(n)}, \dots, (\boldsymbol{S}_K^t)_{(n)}\right]$. Algorithm 2 summarizes D-AltMin-P2GD. Since $\boldsymbol{S}_{(n)}^t \boldsymbol{S}_{(n)}^{t\top} = \sum_{k=1}^K (\boldsymbol{S}_k^t)_{(n)} (\boldsymbol{S}_k^t)_{(n)}^\top$, each task communicates only an additional $\sum_{n=1}^N r_n^2$ entries per round, which is negligible since $r_n \ll d_n$.

5 Theoretical Guarantees

To characterize convergence of optimization error, we define the following metric.

Algorithm 2 D-AltMin-PGD / D-AltMin-P2GD for solving (3)

Require: Initialization iterations T_0, optimization iterations T_1, samples $\{(\boldsymbol{\mathcal{A}}_{k,i}^0, y_{k,i}^0)\}_{i\in[m_0],k\in[K]}$ and $\{(\boldsymbol{\mathcal{A}}_{k,i}^t, y_{k,i}^t), (\tilde{\boldsymbol{\mathcal{A}}}_{k,i}^t, \tilde{y}_{k,i}^t)\}_{i\in[m],k\in[K],t\in[T_1]}$, ranks $\{r_n\}_{n\in[N]}$, step size η.

Initialize: Obtain $\{\boldsymbol{U}_n^0\}_{n\in[N]}$ by Algorithm 1 using T_0 and $\{(\boldsymbol{\mathcal{A}}_{k,i}^0, y_{k,i}^0)\}_{i\in[m_0],k\in[K]}$.
for $t = 0, \ldots, T_1 - 1$ **do**
  Broadcast: Server sends $\{\boldsymbol{U}_n^t\}_{n\in[N]}$ to all tasks.
  for each task $k \in [K]$ **in parallel do**
    Update $\boldsymbol{\mathcal{S}}_k^t$ via (6) using $\{(\tilde{\boldsymbol{\mathcal{A}}}_{k,i}^t, \tilde{y}_{k,i}^t)\}_{i\in[m]}$.
    Compute $\{\nabla_{\boldsymbol{U}_n} f_k\}_{n\in[N]}$ via (7) using $\{(\boldsymbol{\mathcal{A}}_{k,i}^{t+1}, y_{k,i}^{t+1})\}_{i\in[m]}$.
    if D-AltMin-P2GD **then**
      Send $\{\nabla_{\boldsymbol{U}_n} f_k,\ (\boldsymbol{S}_k^t)_{(n)}(\boldsymbol{S}_k^t)_{(n)}^\top\}_{n\in[N]}$ to the server.
    else
      Send $\{\nabla_{\boldsymbol{U}_n} f_k\}_{n\in[N]}$ to the server.
    end if
  end for
  if D-AltMin-P2GD **then**
    Aggregation: Server updates $\{\boldsymbol{U}_n^{t+1}\}_{n\in[N]}$ via (9).
  else
    Aggregation: Server updates $\{\boldsymbol{U}_n^{t+1}\}_{n\in[N]}$ via (8).
  end if
end for
Ensure: $\{\boldsymbol{\mathcal{X}}_k^{T_1}\}_{k\in[K]}$, where $\boldsymbol{\mathcal{X}}_k^{T_1} = \boldsymbol{\mathcal{S}}_k^{T_1} \times \{\boldsymbol{U}_n^{T_1}\}_{n\in[N]}$.

Definition 2 (**Multiple Subspaces Distance**). *For column-orthogonal matrices $\{\boldsymbol{U}_n, \hat{\boldsymbol{U}}_n\}_{n\in N}$ with $\boldsymbol{U}_n, \hat{\boldsymbol{U}}_n \in \mathbb{R}^{d_n\times r_n}$, we define* $\delta\Big(\{\boldsymbol{U}_n, \hat{\boldsymbol{U}}_n\}_{n\in N}\Big) := \max_{n\in[N]} \delta(\boldsymbol{U}_n, \hat{\boldsymbol{U}}_n) := \max_{n\in[N]} \Big\|\big(\boldsymbol{I}_{d_n} - \boldsymbol{U}_n\boldsymbol{U}_n^\top\big)\hat{\boldsymbol{U}}_n\Big\|$.

This metric measures the alignment of multilinear subspaces along the worst case mode. We also impose the following assumption.

Assumption 1. *There exists a universal constant μ such that* $\max_{k\in[K]} \|\boldsymbol{\mathcal{X}}_k^\star\|_F \le \mu\sigma_{\max}^\star\sqrt{\frac{r_{\max}}{K}}$.

This assumption ensures that the tensor parameters of all tasks are balanced and uniformly bounded, and is standard in related MTL analyses (Boursier et al., 2022; Thekumparampil et al., 2021; Tripuraneni et al., 2021; Vaswani, 2024; Wu & Sun; Wu et al., 2025).

5.1 Guarantee of Distributed Spectral Initialization

The following theorem establishes the guarantee of the distributed initialization stage in Algorithm 1.

Theorem 1. *For the MTTLR in (1) under Assumption 1 and high-dimensional setting that* $\min_n d_n \gg \max(\log d_{\max}, \log N)$, *suppose $\{\boldsymbol{\mathcal{A}}_{k,i}^0\}$ have i.i.d. standard Gaussian entries and noise $\left\{\xi_{k,i}^0\right\}$ are i.i.d with $\xi_{k,i}^0 \sim \mathcal{N}\left(0, \sigma_v^2\right)$. If the signal-to-noise ratio SNR:= $\frac{\sigma_{\min}^2}{K\sigma_v^2}$ satisfies SNR $\gtrsim \max\left\{\frac{N\kappa^6 d_{\max} r_{\max}}{m_0 K}, \frac{N\kappa^2 r_{\max}^{\frac{1}{2}}\left(\sqrt{\prod_{n=1}^N d_n} + d_{\max}\right)}{m_0\sqrt{K}}\right\}$, and the per-task sample size m_0 satisfies*

$$m_0 \gtrsim \frac{N\mu^2\kappa^4 r_{\max}^{\frac{3}{2}}\left(\sqrt{\prod_{n=1}^N d_n} + d_{\max}\right)\log\left(m_0 N K d_{\max}\right)}{\sqrt{K}} + \frac{N\mu^2\kappa^8 d_{\max} r_{\max}^2}{K}, \tag{10}$$

Algorithm 1 needs at most $\mathcal{O}\left(\log\left(d_{\max}\sqrt{N r_{\max}}\kappa^2\right)\right)$ iterations to achieve

$$\delta\left(\left\{\boldsymbol{U}_n^0, \boldsymbol{U}_n^\star\right\}_{n\in[N]}\right) \le \frac{0.01}{\sqrt{N r_{\max}}\kappa^2} \tag{11}$$

with probability at least $1 - d_{\max}^{-10}$.

Communication Complexity. Algorithm 1 achieves (11) only in logarithmic communication rounds, which means that per task only transits $\mathcal{O}(\sum_n d_n r_n \log(d_{\max}\sqrt{N r_{\max}}\kappa^2))$ entries in this stage.

Sample Complexity. For $K = 1$, our result in (10) is the order of $\sqrt{\prod_n d_n} + d_{\max}$, which matches the best known result for single-task tensor linear regression (Han et al., 2022; Tong et al., 2022).

Benefit of MTL. Increasing K reduces the sample complexity in (10) and relaxes the SNR conditions. Since initialization is a major bottleneck in high-dimensional tensor estimation (Tong et al., 2022), this result provides a clear benefit of MTL in per-task sample efficiency, noise tolerance while maintaining communication efficiency.

5.2 Guarantee of D-AltMin-PGD

The following theorem establishes linear convergence of the optimization error from initialization obtained by Algorithm 1 and characterizes the resulting statistical error of learned parameter.

Theorem 2. *For the MTTLR problem in (1) under Assumption 1 and high-dimensional setting that* $\min_n d_n \gg \max(\log d_{\max}, \log N)$*, suppose* $\{\boldsymbol{\mathcal{A}}_{k,i}^t\}$ *have* $i.i.d$ *standard Gaussian entries and noise* $\left\{\xi_{k,i}^t\right\}$ *are* $i.i.d$ *with* $\xi_{k,i}^0 \sim \mathcal{N}\left(0, \sigma_v^2\right)$ *and SNR* $\gtrsim \max\left\{\frac{N d_{\max} r_{\max} \kappa^6}{mK}, \frac{\kappa^2 \prod_{n=1}^N r_n \log(K d_{\max})}{m}\right\}$*. If the initialization* $\{\boldsymbol{U}_n^0\}$ *satisfies (11),* $\eta \leq \frac{0.1}{m\sigma_{\max}^{\star 2}}$ *and the per-task sample size* m *satisfies*

$$m \gtrsim \left(\frac{\mu^2 \kappa^4 N d_{\max} r_{\max}^2}{K} + \prod_{n=1}^N r_n\right) \log\left(K d_{\max}\right), \quad \rho_1 := 1 - \frac{m\eta\sigma_{\min}^2}{2} \in (0,1), \tag{12}$$

then D-AltMin-PGD with probability at least $1 - \frac{t}{d_{\max}^{10}}$ *has*

$$\delta\left(\left\{\boldsymbol{U}_n^t, \boldsymbol{U}_n^\star\right\}_{n\in[N]}\right) \leq \rho_1^t \cdot \delta\left(\left\{\boldsymbol{U}_n^0, \boldsymbol{U}_n^\star\right\}_{n\in[N]}\right) + C_1\kappa\sqrt{\frac{d_{\max}}{mK\,\mathrm{SNR}}} \tag{13}$$

$$\max_{k\in[K]} \left\|\boldsymbol{\mathcal{X}}_k^t - \boldsymbol{\mathcal{X}}_k^\star\right\|_F \leq \frac{\mu\sigma_{\max}}{\sqrt{K}} \cdot \rho_1^t + C_2\sigma_v \left(\mu\kappa^2\sqrt{\frac{N d_{\max} r_{\max}}{mK}} + \sqrt{\frac{\prod_{n=1}^N r_n + \log\left(K d_{\max}\right)}{m}}\right), \tag{14}$$

where C_1, C_2 *are universal constants.*

Communication Complexity. The first term in RHS of (13) and (14) is optimization error, which vanishes at a linear rate $1 - 0.5\eta m\sigma_{\min}^2$. Hence, D-AltMin-PGD only requires $O(\kappa^2 \log(1/\epsilon))$ rounds to reduce the optimization error to ϵ and finally reaching the statistical error floor.

Statistical Error. The statistical error consists of a shared-factor term $O\left(\sigma_v \mu\kappa^2 \sqrt{N d_{\max} r_{\max}/(mK)}\right)$ and a task-specific term $O\left(\sigma_v\sqrt{(\prod_n r_n + \log(K d_{\max}))/m}\right)$. The former enjoys a $1/\sqrt{K}$ gain from jointly learning the shared factors, whereas the latter does not, as each core is estimated from task-specific data. For $K = 1$, our rate matches the minimax-optimal rate in single-task tensor regression (Han et al., 2022) up to logarithmic factors. When there is noiseless $\sigma_v = 0$, our result shows proposed method can recover true parameters with linear rate.

Sample Complexity. The two terms in (12) correspond to estimating the shared factors and task-specific cores, respectively. The former decreases as $1/K$, while the latter is independent of K. When $K = 1$, (12) becomes of order $N d_{\max} r_{\max}^2 + \prod_{n=1}^N r_n$, which is nearly optimal up to an extra factor of $r_{\max}$ compared with the best-known bound $N d_{\max} r_{\max} + \prod_{n=1}^N r_n$ in single-task tensor linear regression (Han et al., 2022; Tong et al., 2022).

Benefits of MTL. The theorem quantifies the benefit of joint learning across related tasks: as the number of tasks increases, both the per-task sample requirement and statistical error associated with shared-factor estimation decrease, while the corresponding SNR requirement is weaker. Thus, when shared-factor estimation is the main bottleneck, MTL enables accurate recovery with fewer samples per task and weaker signals. This gain eventually saturates at the task-specific core estimation error, which does not substantially improve with the number of tasks.

5.3 Benefit of Preconditioned D-AltMin-PGD

D-AltMin-P2GD achieves a κ-independent linear convergence rate, as given in the following theorem.

Theorem 3. *Under the same setting as Theorem 2, if $\eta \le \frac{0.1}{m}$ and $\rho_2 := 1 - 0.5m\eta \in (0,1)$, then D-AltMin-P2GD can achieve*

$$\delta\left(\left\{\boldsymbol{U}_n^t, \boldsymbol{U}_n^\star\right\}_{n\in[N]}\right) \le \rho_2^t \cdot \delta\left(\left\{\boldsymbol{U}_n^0, \boldsymbol{U}_n^\star\right\}_{n\in[N]}\right) + C_3\kappa\sqrt{\frac{d_{\max}}{mK\,\mathrm{SNR}}} \tag{15}$$

$$\max_{k\in[K]} \left\|\boldsymbol{\mathcal{X}}_k^t - \boldsymbol{\mathcal{X}}_k^\star\right\|_F \le \frac{\mu\sigma_{\max}}{\sqrt{K}} \cdot \rho_2^t + C_4\sigma_v\left(\mu\kappa^2\sqrt{\frac{Nd_{\max}r_{\max}}{mK}} + \sqrt{\frac{\prod_{n=1}^N r_n + \log\left(Kd_{\max}\right)}{m}}\right) \tag{16}$$

with probability at least $1 - \frac{t}{d_{\max}^{10}}$, where C_3, C_4 are universal constants.

Benefit of Precondition. Comparing the linear rate $1 - 0.5m\eta\sigma_{\min}^2$ of D-AltMin-PGD in Theorem 2, the linear rate of D-AltMin-P2GD is $1 - 0.5m\eta$, which can use κ-independent logarithmic iterations or communication rounds to achieve same order statistical error as D-AltMin-PGD. This can make that D-AltMin-P2GD can achieve a faster convergence rate for solving ill-conditioned MTTLR problems that have large condition numbers $\kappa \gg 1$.

5.4 Minimax Lower Bound for MTTLR

The following theorem gives the minimax lower bound of statistical error for proposed MTTLR (1).

Theorem 4. *Suppose $\sigma_v > 0$, per-task sample size $m \ge 1$, all input entries are $i.i.d.$ standard Gaussian, and the noises are $i.i.d.$ $\mathcal{N}(0,\sigma_v^2)$. For any $0 < \alpha \le 1/32$, if $m\mathrm{SNR} \ge \max\left\{\frac{1}{K}\sum_{n=1}^N r_n(d_n - r_n),\ \prod_{n=1}^N r_n + \log\frac{K}{\alpha}\right\}$, then*

$$\inf_{\hat{\boldsymbol{\mathcal{X}}}} \sup_{\boldsymbol{\mathcal{X}}\in\mathcal{F}} \mathbb{P}_{\mathcal{X}}\left(\max_{k\in[K]} \|\hat{\boldsymbol{\mathcal{X}}}_k - \boldsymbol{\mathcal{X}}_k\|_F \ge \frac{\sigma_v}{128\sqrt{m}}\sqrt{\prod_{n=1}^N r_n + \frac{1}{K}\sum_{n=1}^N r_n(d_n - r_n) + \log\frac{K}{\alpha}}\right) \ge 2\alpha, \tag{17}$$

where $\mathcal{F} = \Big\{\{\boldsymbol{\mathcal{X}}_k \in \mathbb{R}^{d_1\times\cdots\times d_N}\}_{k\in[K]} : \mathrm{rank}(\mathbf{X}_{(n)}) = r_n, \sigma_{\min}(\boldsymbol{\mathcal{X}}) \ge \frac{\sigma_{\min}}{2}, \sigma_{\max}(\boldsymbol{\mathcal{X}}) \le 2\sigma_{\max}, \max_{k\in[K]}\|\boldsymbol{\mathcal{X}}_k\|_F \le 2\mu\,\sigma_{\max}(\boldsymbol{\mathcal{X}})\sqrt{r_{\max}/K}\Big\}$. The infimum ranges over all estimators using the mK samples, and the probability includes both designs and noises.

For given sufficient total per-task sample size m in Theorem 4 such that $m \ge C_5 m_2 \max\{m_0, m_1\}$, where C_5 is sufficiently large constant, m_0, m_1 satisfy (10), (12). $m_2 := \log\left(\mu\kappa\mathrm{SNR}\Big/\left(\mu\kappa^2\sqrt{\frac{Nd_{\max}r_{\max}}{mK}} + \sqrt{\frac{\prod_{n=1}^N r_n + \log(Kd_{\max})}{m}}\right)\right)$ is the iteration numbers for D-AltMin-P2GD to attain the statistical error in (16) based on linear convergence rate of optimization error. We can use sample splitting to divide the m samples into m_2 disjoint sub-samples for each task that has m/m_2 fresh samples per iteration. Substituting m/m_2 into (14) and (16) and comparing the lower bound in (4), if $\kappa$and μ are assumed constant level and $d_1 = \cdots = d_N = d, r_1 = \cdots = r_n = r$, we can conclude that our proposed methods can achieve the nearly minimax optimal rate of estimation error in the class $\mathcal{F}$ defined in Theorem 4 up to some logarithmic factors.

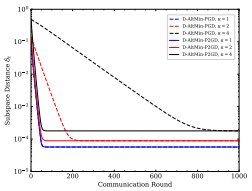
6 Experimental Results

We evaluate the proposed methods on synthetic data and the same samples are reused throughout the iterations in the following experiments. More experiments and comparisons with existing MTL methods on both synthetic and real-world data are deferred to Appendix B.

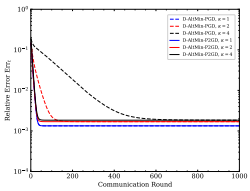
Convergence and preconditioning. We compare D-AltMin-PGD and D-AltMin-P2GD for $\kappa \in \{1,2,4\}$, fixing $N = 3, d = 20, r = 2, K = 50, m = 100, m_0 = 400, \sigma_{\max} = 1, \eta = 0.002$ and Gaussian noise $\sigma_v = 10^{-4}$. Fig. 1 reports the subspace error δ_t and maximum relative estimation error

$\mathrm{Err}_t := \max_{k\in[K]} \|\boldsymbol{\mathcal{X}}_k^t - \boldsymbol{\mathcal{X}}_k^\star\|_F / \|\boldsymbol{\mathcal{X}}_k^\star\|_F$. We can observe that both methods linearly converges to same statistical error floor, while increasing κ does not influence covergence of D-AltMin-P2GD and slow down convergence of D-AltMin-PGD.

Influence of noise. We further investigate the performance of D-AltMin-PGD to noise magnitudes. We fix $N = 3, d = 20, r = 2, K = 50, m_0 = 400, m = 100, \kappa = 2, \eta = 0.002$. and vary Gaussian noise level $\sigma_v \in \{10^{-2}, 10^{-3}, 10^{-4}\}$. The results are shown in Fig. 2, where we can observe that D-AltMin-PGD can linearly converge to different error floors. Decreasing noise magnitude σ_v can reduce both the subspace and relative recovery errors floors. This is consistent with our Theorem 2.

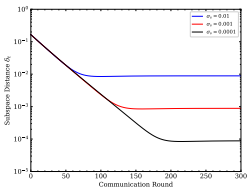


(a) Subspace distance.

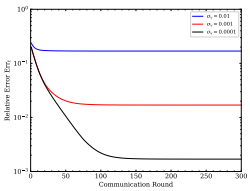


(b) Relative error.

Figure 1: Convergence under different κ.



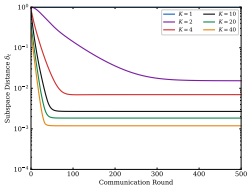
(a) Subspace distance.



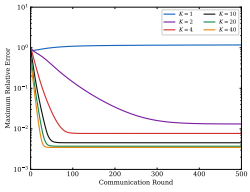
(b) Relative error.

Figure 2: Convergence under different σ_v.

Benefit of MTL. We generate each ground truth tensor with shared factors and Gaussian cores with unit norm. We test D-AltMin-P2GD with $\eta = 0.001$ and vary $K \in \{1, 2, 4, 10, 20, 40\}$, fixing $N = 3, d = 40, r = 2, m = 200, m_0 = 400$, and Gaussian noise $\sigma_v = 0.01$. The results in Fig. 3 shows that when $K = 1$ that single task learning fail. More tasks generally accelerate convergence and reduce statistically floor that illustrates multi-task gains. However, as K grows, the subspace error continues to decrease apparently, whereas the relative recovery errors exhibit diminishing improvement and gradually saturate. This phenomenon is also predicted by our Theorem 3.

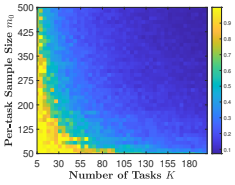


(a) Subspace distance.

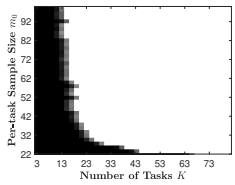


(b) Relative error.

Figure 3: Estimation error under different K.



(a) Initialization.



(b) D-AltMin-PGD

Figure 4: Phase transitions of exact recovery.

Exact recovery in noiseless setting. We consider noiseless third-order tensor recovery with $d_n = 20$, $r_n = 3$, and $\kappa = 1$. For spectral initialization, we vary $K \in \{5, 10, \ldots, 200\}$ and $m_0 \in \{50, 60, \ldots, 500\}$, reporting the mean δ_0 over 10 trials. For recovery, we vary $K \in \{3, 5, \ldots, 81\}$ and $m \in \{22, 24, \ldots, 100\}$, declaring success when $\mathrm{Err}_t < 10^{-3}$. Fig. 4 shows that increasing K reduces both the initialization error and the per-task sample requirement for exact recovery, demonstrating the predicted MTL gain.

7 Conclusion

We proposed a distributed multilinear MTL framework that exploits shared mode-wise factors and task-specific cores to capture both intra- and inter-task structure. For MTTLR, we developed communication-efficient D-AltMin-PGD and its preconditioned variant D-AltMin-P2GD, and established linear convergence to a statistical neighborhood, a provable multi-task gain, and near-minimax statistical rates. Despite these guarantees, our current analysis relies on Gaussian designs and noise, as well as fresh-sample splitting across iterations, and focuses on linear shared-Tucker models. Future work includes removing the sample-splitting requirement, extending the theory to non-Gaussian designs, and studying more general nonlinear and overparameterized multilinear models.

AI USE STATEMENT

We used generative AI tools, including ChatGPT, to assist with literature discovery, language editing, and LaTeX formatting, and as a secondary checking tool for mathematical derivations and proofs developed by the authors. In particular, these tools were used to identify potential inconsistencies, unclear steps, and presentation issues, rather than to formulate mathematical claims, provide critical ingredients for proofs, or determine proof strategies. We also used these tools to provide feedback on experimental design and on the presentation and interpretation of experimental results. Generative AI was not used to generate datasets or to produce, fabricate, or alter experimental results. All AI-assisted feedback was independently evaluated by the authors, and all references, mathematical arguments, and reported results were manually verified. The authors take full responsibility for the final content of this paper.

REFERENCES

Arvind Agarwal, Samuel Gerber, and Hal Daume. Learning multiple tasks using manifold regularization. *Advances in neural information processing systems*, 23, 2010.

Bilal Ahmad, Liana Khamidullina, Alexey A Korobkov, Alla Manina, Jens Haueisen, and Martin Haardt. Joint model order estimation for multiple tensors with a coupled mode and applications to the joint decomposition of eeg, meg magnetometer, and gradiometer tensors. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1186–1190. IEEE, 2022.

Rie Kubota Ando, Tong Zhang, and Peter Bartlett. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of machine learning research*, 6(11), 2005.

Peter Arbenz, Daniel Kressner, and DME Zürich. Lecture notes on solving large scale eigenvalue problems. *D-MATH, EHT Zurich*, 2(3):377, 2012.

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-task feature learning. *Advances in neural information processing systems*, 19, 2006.

Andreas Argyriou, Massimiliano Pontil, Yiming Ying, and Charles Micchelli. A spectral regularization framework for multi-task structure learning. *Advances in neural information processing systems*, 20, 2007.

Alexandre Audibert, Massih R Amini, Konstantin Usevich, and Marianne Clausel. Low-rank updates of pre-trained weights for multi-task learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 7544–7554, 2023.

Kwan-Young Bak and Ja-Yong Koo. Minimax estimation in multi-task regression under low-rank structures. *Journal of Nonparametric Statistics*, 35(1):122–144, 2023.

Steffen Bickel, Jasmina Bogojeska, Thomas Lengauer, and Tobias Scheffer. Multi-task learning for hiv therapy screening. In *Proceedings of the 25th international conference on Machine learning*, pp. 56–63, 2008.

Ricardo A Borsoi, Konstantin Usevich, David Brie, and Tülay Adali. Personalized coupled tensor decomposition for multimodal data fusion: Uniqueness and algorithms. *IEEE Transactions on Signal Processing*, 2024.

Ricardo Augusto Borsoi, Isabell Lehmann, Mohammad ABS Akhonda, Vince D Calhoun, Konstantin Usevich, David Brie, and Tülay Adali. Coupled cp tensor decomposition with shared and distinct components for multi-task fmri data fusion. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Etienne Boursier, Mikhail Konobeev, and Nicolas Flammarion. Trace norm regularization for multi-task learning with scarce data. In *Conference on Learning Theory*, pp. 1303–1327. PMLR, 2022.

Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

Jianhui Chen, Lei Tang, Jun Liu, and Jieping Ye. A convex formulation for learning shared structures from multiple tasks. In *Proceedings of the 26th annual international conference on machine learning*, pp. 137–144, 2009.

Shijie Chen, Yu Zhang, and Qiang Yang. Multi-task learning in natural language processing: An overview. *ACM Computing Surveys*, 56(12):1–32, 2024.

Yifang Chen, Kevin Jamieson, and Simon Du. Active multi-task representation learning. In *International Conference on Machine Learning*, pp. 3271–3298. PMLR, 2022.

Andrzej Cichocki, Danilo Mandic, Lieven De Lathauwer, Guoxu Zhou, Qibin Zhao, Cesar Caiafa, and Huy Anh Phan. Tensor decompositions for signal processing applications: From two-way to multiway component analysis. *IEEE signal processing magazine*, 32(2):145–163, 2015.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International conference on machine learning*, pp. 2089–2099. PMLR, 2021.

Victor De la Pena and Evarist Giné. *Decoupling: from dependence to independence*. Springer Science & Business Media, 2012.

Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.

Simon Shaolei Du, Wei Hu, Sham M. Kakade, Jason D. Lee, and Qi Lei. Few-shot learning via learning the representation, provably. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=pW2Q2xLwIMD.

John C Duchi, Vitaly Feldman, Lunjia Hu, and Kunal Talwar. Subspace recovery from heterogeneous data with non-isotropic noise. *Advances in Neural Information Processing Systems*, 35:5854–5866, 2022.

Theodoros Evgeniou and Massimiliano Pontil. Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 109–117, 2004.

Yifan Fu, Junbin Gao, Xia Hong, and David Tien. Tensor regression based on linked multiway parameter analysis. In *2014 IEEE International Conference on Data Mining*, pp. 821–826. IEEE, 2014.

Yash Garg, Nebiyou Yismaw, Rakib Hyder, Ashley Prater-Bennette, Amit Roy-Chowdhury, and M Salman Asif. Parameter-efficient multi-task and multi-domain learning using factorized tensor networks. *IEEE Open Journal of Signal Processing*, 2025.

Tiantian Geng, Teng Wang, Jinming Duan, Yanfu Zhang, Weili Guan, Feng Zheng, and Ling Shao. Uniav: Unified audio-visual perception for multi-task video event localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Emilie Grégoire, Muhammad Hafeez Chaudhary, and Sam Verboven. Sample-level weighting for multi-task learning with auxiliary tasks. *Applied Intelligence*, 54(4):3482–3501, 2024.

Lei Han and Yu Zhang. Multi-stage multi-task learning with reduced rank. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Rungang Han, Rebecca Willett, and Anru R Zhang. An optimal statistical and computational framework for generalized tensor estimation. *The Annals of Statistics*, 50(1):1–29, 2022.

Hanchi Huang, Deheng Ye, Li Shen, and Wei Liu. Curriculum-based asymmetric multi-task reinforcement learning. *IEEE transactions on pattern analysis and machine intelligence*, 45(6):7258–7269, 2022.

Alper KALLE, Théo Rudkiewicz, Mohamed Ouerfelli, and mohamed Tamaazousti. Distribution-aware tensor decomposition for compression of convolutional neural networks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=ODgWBaErst.

Suyoun Kim, Takaaki Hori, and Shinji Watanabe. Joint ctc-attention based end-to-end speech recognition using multi-task learning. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4835–4839. IEEE, 2017.

Parker Knight and Rui Duan. Multi-task learning with summary statistics. *Advances in neural information processing systems*, 36:54020–54031, 2023.

Vladimir Koltchinskii, Karim Lounici, and Alexandre B Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. 2011.

Jean Kossaifi, Zachary C Lipton, Arinbjorn Kolbeinsson, Aran Khanna, Tommaso Furlanello, and Anima Anandkumar. Tensor regression networks. *Journal of Machine Learning Research*, 21 (123):1–21, 2020.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Li-Dan Kuang, Zhi-Ming He, Jianming Zhang, and Feng Li. Coupled canonical polyadic decomposition of multi-group fmri data with spatial reference and orthonormality constraints. *Biomedical Signal Processing and Control*, 80:104232, 2023.

Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000. doi: 10.1214/aos/1015957395.

Jingyang Li, Jian-Feng Cai, Yang Chen, and Dong Xia. Online tensor learning: Computational and statistical trade-offs, adaptivity and optimal regret. *arXiv preprint arXiv:2306.03372*, 2023.

Xiang Li, Shusen Wang, Kun Chen, and Zhihua Zhang. Communication-efficient distributed svd via local power iterations. In *International Conference on Machine Learning*, pp. 6504–6514. PMLR, 2021.

Tianchi Liao, Lele Fu, Jialong Chen, Zhen Wang, Zibin Zheng, and Chuan Chen. A swiss army knife for heterogeneous federated learning: Flexible coupling via trace norm. *Advances in Neural Information Processing Systems*, 37:139886–139911, 2024.

Kaixiang Lin, Jianpeng Xu, Inci M Baytas, Shuiwang Ji, and Jiayu Zhou. Multi-task feature interaction learning. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1735–1744, 2016a.

Shaohui Lin, Rongrong Ji, Xiaowei Guo, Xuelong Li, et al. Towards convolutional neural networks compression via global error reconstruction. In *IJCAI*, pp. 1753–1759, 2016b.

Xixun Lin, Qing Yu, Yanan Cao, Lixin Zou, Chuan Zhou, Jia Wu, Chenliang Li, Peng Zhang, and Shirui Pan. Generative causality-driven network for graph multi-task learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(1):1029–1044, 2026. doi: 10.1109/TPAMI.2025.3610096.

Jiani Liu, Ce Zhu, Zhen Long, and Yipeng Liu. Tensor regression. *Foundations and Trends in Machine Learning*, 14(4):379–565, 2021.

Jiani Liu, Qinghua Tao, Ce Zhu, Yipeng Liu, and Johan AK Suykens. Tensorized lssvms for multitask regression. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023a.

Renpu Liu, Cong Shen, and Jing Yang. Federated representation learning in the under-parameterized regime. In *International Conference on Machine Learning*, pp. 31808–31843. PMLR, 2024.

Sulin Liu, Sinno Jialin Pan, and Qirong Ho. Distributed multi-task relationship learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 937–946, 2017.

Weiwei Liu. Improved bounds for multi-task learning with trace norm regularization. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 700–714. PMLR, 2023.

Xinling Liu, Jingyao Hou, Jiangjun Peng, Hailin Wang, Deyu Meng, and Jianjun Wang. Tensor compressive sensing fused low-rankness and local-smoothness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8879–8887, 2023b.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Philip S Yu. Learning multiple tasks with multilinear relationship networks. *Advances in neural information processing systems*, 30, 2017.

Yuxiang Lu, Suizhi Huang, Yuwen Yang, Shalayiding Sirejiding, Yue Ding, and Hongtao Lu. Fedhca2: Towards hetero-client federated multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5599–5609, 2024.

Xindian Ma, Peng Zhang, Shuai Zhang, Nan Duan, Yuexian Hou, Ming Zhou, and Dawei Song. A tensorized transformer for language modeling. *Advances in neural information processing systems*, 32, 2019.

Andreas Maurer. Bounds for linear multi-task learning. *The Journal of Machine Learning Research*, 7:117–139, 2006.

Alexander Novikov, Dmitrii Podoprikhin, Anton Osokin, and Dmitry P Vetrov. Tensorizing neural networks. *Advances in neural information processing systems*, 28, 2015.

Ting Kei Pong, Paul Tseng, Shuiwang Ji, and Jieping Ye. Trace norm regularization: Reformulations, algorithms, and multi-task learning. *SIAM Journal on Optimization*, 20(6):3465–3489, 2010.

Massimiliano Pontil and Andreas Maurer. Excess risk bounds for multitask learning with trace norm regularization. In *Conference on Learning Theory*, pp. 55–76. PMLR, 2013.

Bernardino Romera-Paredes, Hane Aung, Nadia Bianchi-Berthouze, and Massimiliano Pontil. Multilinear multitask learning. In *International Conference on Machine Learning*, pp. 1444–1452. PMLR, 2013.

Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

Yousef Saad. *Numerical methods for large eigenvalue problems: revised edition*. SIAM, 2011.

Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. *Advances in neural information processing systems*, 30, 2017.

Shagun Sodhani, Amy Zhang, and Joelle Pineau. Multi-task reinforcement learning with context-based representations. In *International conference on machine learning*, pp. 9767–9779. PMLR, 2021.

Kiran K Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Statistically and computationally efficient linear meta-representation learning. *Advances in Neural Information Processing Systems*, 34:18487–18500, 2021.

Ye Tian, Yuqi Gu, and Yang Feng. Learning from similar linear representations: Adaptivity, minimaxity, and robustness. *Journal of Machine Learning Research*, 26(187):1–125, 2025.

Tian Tong, Cong Ma, Ashley Prater-Bennette, Erin Tripp, and Yuejie Chi. Scaling and scalability: Provable nonconvex low-rank tensor estimation from incomplete measurements. *Journal of Machine Learning Research*, 23(163):1–77, 2022.

Nilesh Tripuraneni, Chi Jin, and Michael Jordan. Provable meta-learning of linear representations. In *International conference on machine learning*, pp. 10434–10443. PMLR, 2021.

Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3614–3633, 2021.

Namrata Vaswani. Efficient federated low rank matrix recovery via alternating gd and minimization: A simple proof. *IEEE Transactions on Information Theory*, 70(7):5162–5167, 2024.

Roman Vershynin. High-dimensional probability: An introduction with applications in data science, 2nd ed. (draft). Available at `https://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-2.pdf`, 2025.

Xuehao Wang, Zhan Zhuang, Feiyang Ye, and Yu Zhang. Mtsam: Multi-task fine-tuning for segment anything model. In *The Thirteenth International Conference on Learning Representations*, 2025.

Zihao Wang, Shaofei Cai, Anji Liu, Yonggang Jin, Jinbing Hou, Bowei Zhang, Haowei Lin, Zhaofeng He, Zilong Zheng, Yaodong Yang, et al. Jarvis-1: Open-world multi-task agents with memory-augmented multimodal language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Austin Watkins, Enayat Ullah, Thanh Nguyen-Tang, and Raman Arora. Optimistic rates for multi-task representation learning. *Advances in Neural Information Processing Systems*, 36:2207–2251, 2023.

Kishan Wimalawarne, Masashi Sugiyama, and Ryota Tomioka. Multitask learning meets tensor factorization: task imputation via convex optimization. *Advances in neural information processing systems*, 27, 2014.

Fan Wu, Xinyu Yan, Jiabei Liu, and Wei Yang Bryan Lim. Fedram: Federated reweighting and aggregation for multi-task learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Tongle Wu and Ying Sun. Non-convex tensor recovery from tube-wise sensing. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Tongle Wu, Ying Sun, and Jicong Fan. Non-convex tensor recovery from local measurements. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21590–21598, 2025.

Dong Xia and Ming Yuan. On polynomial time methods for exact low-rank tensor completion. *Foundations of Computational Mathematics*, 19(6):1265–1313, 2019.

Dong Xia, Ming Yuan, and Cun-Hui Zhang. Statistically optimal and computationally efficient low rank tensor completion from noisy entries. 2021.

Qi Xie, Qian Zhao, Deyu Meng, and Zongben Xu. Kronecker-basis-representation based tensor sparsity and its applications to tensor recovery. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(8): 1888–1902, 2018. doi: 10.1109/TPAMI.2017.2734888.

Jianpeng Xu, Jiayu Zhou, Pang-Ning Tan, Xi Liu, and Lifeng Luo. Spatio-temporal multi-task learning via tensor decomposition. *IEEE Transactions on Knowledge and Data Engineering*, 33 (6):2764–2775, 2019.

Xingyu Xu, Yandi Shen, Yuejie Chi, and Cong Ma. The power of preconditioning in overparameterized low-rank matrix sensing. In *International Conference on Machine Learning*, pp. 38611–38654. PMLR, 2023.

Yongxin Yang and Timothy Hospedales. Trace norm regularised deep multi-task learning. In *5th International Conference on Learning Representations*, 2017a.

Yongxin Yang and Timothy M. Hospedales. Deep multi-task representation learning: A tensor factorisation approach. In *International Conference on Learning Representations*, 2017b. URL `https://openreview.net/forum?id=SkhU2fcll`.

Niloofar Yousefi, Yunwen Lei, Marius Kloft, Mansooreh Mollaghasemi, and Georgios C Anagnostopoulos. Local rademacher complexity-based learning guarantees for multi-task learning. *Journal of Machine Learning Research*, 19(38):1–47, 2018.

Jun Yu, Xiaokang Liu, Chongliang Luo, Jin Huang, Rong Zhou, Yixin Liu, Jie Hu, Jianmin Chen, Ke Zhang, Dazheng Zhang, et al. Multitask learning 1997–2024: Part i fundamentals. *Harvard Data Science Review*, 7(3), 2025.

Anru R. Zhang, Yuetian Luo, Garvesh Raskutti, and Ming Yuan. ISLET: Fast and optimal low-rank tensor regression via importance sketching. *SIAM Journal on Mathematics of Data Science*, 2(2): 444–479, 2020. doi: 10.1137/19M126476X.

Jingdong Zhang, Jiayuan Fan, Peng Ye, Bo Zhang, Hancheng Ye, Baopu Li, Yancheng Cai, and Tao Chen. Bridgenet: Comprehensive and effective feature interactions via bridge feature for multi-task dense predictions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025a.

Thomas TCK Zhang, Leonardo Felipe Toso, James Anderson, and Nikolai Matni. Sample-efficient linear representation learning from non-IID non-isotropic data. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Tr3fZocrI6.

Xiangtao Zhang, Eleftherios Kofidis, Ruituo Wu, Ce Zhu, Le Zhang, and Yipeng Liu. Coupled tensor train decomposition in federated learning. *Pattern Recognition*, 170:112067, 2026.

Xiongjun Zhang, Jin Wu, and Michael K Ng. Multilinear multitask learning by transformed tensor singular value decomposition. *Machine Learning with Applications*, 13:100479, 2023.

Yi Zhang, Yu Zhang, and Wei Wang. Multi-task learning via generalized tensor trace norm. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 2254–2262, 2021.

Yi Zhang, Yu Zhang, and Wei Wang. Learning linear and nonlinear low-rank structure in multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):8157–8170, 2022.

Yifan Zhang, Yifeng Liu, Huizhuo Yuan, Zhen Qin, Yang Yuan, Quanquan Gu, and Andrew C Yao. Tensor product attention is all you need. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=ECTxVRFhUa.

Yu Zhang. Multi-task learning and algorithmic stability. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Yu Zhang and Qiang Yang. An overview of multi-task learning. *National Science Review*, 5(1): 30–43, 2018.

Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE transactions on knowledge and data engineering*, 34(12):5586–5609, 2021.

Zhizhong Zhang, Yuan Xie, Wensheng Zhang, Yongqiang Tang, and Qi Tian. Tensor multi-task learning for person re-identification. *IEEE Transactions on Image Processing*, 29:2463–2477, 2019.

Qian Zhao, Xiangyu Rui, Zhi Han, and Deyu Meng. Multilinear multitask learning by rank-product regularization. *IEEE transactions on neural networks and learning systems*, 31(4):1336–1350, 2019.

Wenjie Zheng, Aurélien Bellet, and Patrick Gallinari. A distributed frank–wolfe framework for learning low-rank matrices with the trace norm. *Machine Learning*, 107(8):1457–1475, 2018.

Chulun Zhou, Yunlong Liang, Fandong Meng, Jie Zhou, Jinan Xu, Hongji Wang, Min Zhang, and Jinsong Su. A multi-task multi-stage transitional training framework for neural chat translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):7970–7985, 2022.

Qiang Zhou and Qi Zhao. Flexible clustered multi-task learning by learning representative tasks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):266–278, 2015.

Appendices

This appendix has following sections. Appendix A contains the detailed introduction of related works. Appendix B shows more experiments of proposed method and compare our method with existing MTL methods in both synthetic and real-world data. We provide the global convergence of D-AltMin-PGD for solving the population loss function of (3) in Appendix C. The proof of Theorem 1 for distributed spectral initialization is given in Appendix D. Appendix E and Appendix F provide proofs for Theorem 2 and Theorem 3, respectively. Appendix G provides the detailed proof of minimax lower bound and Appendix H contains the technical lemmas.

A Detailed Related Works

A.1 Low-rank Parameter-based MTL

The relatedness among multiple tasks can imply the low-rankness of the union task parameters (Zhang & Yang, 2018; 2021), which promotes to the low-rank MTL approaches. In the early study of low-rank-based MTL, most of the work studied the vector formula of task parameters (Caruana, 1997), which are stacked into a matrix along a single task index. They imposed low rank share representation (Ando et al., 2005; Chen et al., 2009; Agarwal et al., 2010; Vaswani, 2024) or trace norm regularization (Pong et al., 2010) on the matrix to capture the inter-task relatedness. However, this matrix representation of task correlations can discard the multi-modal inter-task relatedness for multi-index tasks. Thus, studies in (Romera-Paredes et al., 2013; Wimalawarne et al., 2014) proposed multilinear MTL that stacked the vectors into a tensor and utilized tensor norm or low-rank tensor decomposition (Zhang et al., 2023; Liu et al., 2023a; Zhao et al., 2019; Wu & Sun) to regularize.

In the modern MTL, the model parameters in tasks are always inherently multiway. Vectorization-based matrix regularization would discard the intrinsic multimodel structure (Kossaifi et al., 2020). Recent works stack the layer weights into a high-order tensor and utilize trace-norm regularizations (Yang & Hospedales, 2017a; Zhang et al., 2021; Lin et al., 2016a; Zhang et al., 2022), layer-wise tensor representation (Yang & Hospedales, 2017b; Wu et al., 2025), or tensor-variate priors (Long et al., 2017) to identify the relatedness of parameter tensors. The benefit of tensor-based MTL has been validated in various applications (Garg et al., 2025; Zhang et al., 2019; Xu et al., 2019). Nevertheless, by constructing tensors separately for each block or layer (Yang & Hospedales, 2017a; Zhang et al., 2021), these methods may fail to preserve full intra-task correlation, because dependencies across blocks or across layers are not taken into account (Lin et al., 2016b). Thus, the above matrix-based, lay-wise, or block-wise existing methods result in inferior characterization of inter-task relatedness due to ignoring the intra-task relatedness.

Although there are works that capture the correlation of task parameter tensors directly (Cichocki et al., 2015; Ahmad et al., 2022; Kuang et al., 2023) via the shared representations idea, where they impose all or partial shared factors for coupled tensor decompositions across tasks (Fu et al., 2014; Borsoi et al., 2023; 2024; Zhang et al., 2026), the forms of the shared representations differ from ours. Most importantly, almost all works are empirical and cannot theoretically prove the benefit of low-rank shared representation for multitask learning. In addition, to jointly learn multiple tasks located at different devices (Zhang & Yang, 2021), there are also a lot of works studied the distributed MTL such as (Smith et al., 2017; Zhang et al., 2026; Liao et al., 2024; Zheng et al., 2018; Liu et al., 2024). But the models considered in these works are different with ours.

A.2 Theoretical Analysis of MTL

For the theoretical studies of low-rank parameter-based MTL, most of the mentioned works above are empirical and lack rigorous guarantees to establish the benefits of MTL over single-task learning. There are also a lot of existing MTL works with theoretical guarantees that (Ando et al., 2005; Wimalawarne et al., 2014; Zhang et al., 2021; 2023; 2022; Knight & Duan, 2023; Liu, 2023; Bak & Koo, 2023; Boursier et al., 2022; Chen et al., 2022; Yousefi et al., 2018; Zhang, 2015; Maurer, 2006; Pontil & Maurer, 2013; Watkins et al., 2023; Du et al., 2021; Duchi et al., 2022; Tripuraneni et al., 2021; Tian et al., 2025; Vaswani, 2024; Wu & Sun; Wu et al., 2025) proved the generalization or excess risk bounds in MTL can be improved over single task learning via analysis tools such

as covering number (Ando et al., 2005; Knight & Duan, 2023; Bak & Koo, 2023; Boursier et al., 2022), Rademacher complexity (Wimalawarne et al., 2014; Zhang et al., 2021; 2023; 2022; Maurer, 2006; Pontil & Maurer, 2013; Du et al., 2021), local Rademacher complexity (Watkins et al., 2023; Yousefi et al., 2018), and stability analysis (Zhang, 2015),etc. But either these learning guarantees are optimization algorithms indepedent (Ando et al., 2005; Chen et al., 2022; Maurer, 2006; Pontil & Maurer, 2013; Liu, 2023; Zhang et al., 2022; Yousefi et al., 2018; Bak & Koo, 2023; Zhang et al., 2023; Wimalawarne et al., 2014; Zhang et al., 2021; Knight & Duan, 2023; Watkins et al., 2023; Boursier et al., 2022) or only analysed the convex models (Pontil & Maurer, 2013; Liu, 2023; Zhang et al., 2022; Yousefi et al., 2018; Bak & Koo, 2023; Zhang et al., 2023; Zhang, 2015; Wimalawarne et al., 2014; Zhang et al., 2021; Knight & Duan, 2023; Boursier et al., 2022).

There are also several optimization-dependent theoretical guarantees for non-convex linear share representation models (Collins et al., 2021; Thekumparampil et al., 2021; Vaswani, 2024; Zhang et al., 2024; Wu et al., 2025). There are main differences between these works and ours. For the model perspective, they only considered the linear representation for task parameters in vector or matrix (Wu et al., 2025) spaces, while we study the multilinear share representation for multi-mode tensor parameters, which is motivated by encoding both intra- and intertask correlation simultaneously. For the optimization perspective, they also used the alternative minimization and gradient descent (Thekumparampil et al., 2021) to update the task-specific vector coefficients and task-shared representation matrix, respectively. But our update for task-shared representations are multiple factors due to multilinear representation, which complicates the analysis. In addition, they either have not considered the acceleration optimization methods or have ignored proposing the communication-efficient initialization method, which we both studied in this paper to make the MTL algorithm more efficient.

In addition, existing information-theoretic lower bounds for representation MTL concern vector-valued regression coefficients with a single linear representation for each task (Thekumparampil et al., 2021; Tian et al., 2025), whereas minimax results for low-rank Tucker regression focus on estimating a single tensor (Zhang et al., 2020). In contrast, our MTTLR model couples K tensor-valued task parameters through N shared mode-wise factors and K task-specific cores; to the best of our knowledge, no previous lower bound jointly characterizes these two sources of estimation difficulty. Our lower bound separates the shared-factor term, which enjoys a $1/\sqrt{mK}$ multi-task gain, from the unavoidable $1/\sqrt{m}$ task-specific core term, and, together with our upper bound, establishes the near-minimax optimality of the proposed methods.

B Additional Experiments

B.1 More Synthetic Data Experiments

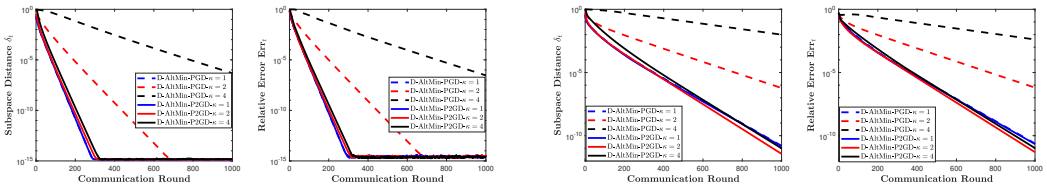
In this section, we provide more synthetic data experiments. We firstly test the linear convergence of D-AltMin-PGD and D-AltMin-P2GD in exact recovery of true parameters in noiseless setting. Then we compare the recovery performance between with and without sample reuse. We finally test the performance of proposed method in misspecified setting.

B.1.1 Linear Convergence to Exact Recovery

We set $N = 3$, $d_n = 20, \sigma_v = 0$ for all $n \in [N]$, $r = 3$, and $\sigma_{\max} = 1$, and evaluate D-AltMin-P2GD and D-AltMin-PGD under $\kappa = 1, 2, 4$. We consider two task settings: $K = 50$ with $m_0 = 400$, $m = 40$, $\eta = 0.04$, and $K = 3$ with $m_0 = 800$, $m = 100$, $\eta = 0.008$. Fig. 5 plots the subspace error and the relative estimation error. In all cases, both methods converge linearly to machine precision from spectral initialization. There is no statistical floor and demonstrates proposed methods can achieve exact recovery under noiseless setting. As κ increases, D-AltMin-PGD becomes slower, whereas D-AltMin-P2GD is nearly unaffected, with almost overlapping curves. This verifies the benefit of preconditioning and supports our theory. Moreover, larger K reduces the per-task sample size needed for exact estimation.

B.1.2 Performance Comparison between sample reuse and Fresh Sampling

We further sample reuse and fresh sampling on noiseless synthetic data using D-AltMin-P2GD with $N = 3$, $d_n = 20$, $r = 3$, $K = 50$, $T = 2500$, and $\eta = 0.04$. Table 2 reports the first communication



(a) Setting 1: $K = 50$, $m_0 = 400$, $m = 40$.

(b) Setting 2: $K = 3$, $m_0 = 800$, $m = 100$.

Figure 5: Convergence of Algorithm 2 under two settings. In each setting, the left panel shows subspace convergence and the right panel shows relative error convergence.

round at which the recovery error falls below 10^{-6}, where m_0 and m denote the initialization and per-iteration sample sizes, respectively. The symbol ✗ indicates that the target accuracy is not reached within 2500 communication rounds.

Table 2: Performance comparison between sample reuse and fresh sampling.

Sample size (m_0, m)	Sample reuse	Fresh sampling
$(400, 25)$	✗	✗
$(400, 30)$	✗	179
$(400, 32)$	✗	93
$(400, 35)$	535	51
$(400, 38)$	212	36

The results show a clear advantage of fresh sampling in both convergence speed and the smallest per-iteration sample size required for successful recovery. In particular, fresh sampling succeeds for $m = 30$ and $m = 32$, whereas sample reuse does not reach the target accuracy within 2500 rounds. When both methods succeed, fresh sampling also requires substantially fewer communication rounds: 51 versus 535 rounds for $m = 35$, and 36 versus 212 rounds for $m = 38$. Meanwhile, sample reuse avoids drawing new independent observations at every iteration, highlighting a trade-off between sample efficiency and convergence speed.

B.1.3 Robustness to Model Misspecification

Table 3: Robustness to approximately shared mode-wise subspaces.

τ	$\bar{\delta}_{\mathrm{mis}}$	δ_T	Err_T
0	1.11×10^{-15}	1.40×10^{-15}	1.04×10^{-15}
10^{-4}	5.0×10^{-4}	7.0×10^{-4}	1.3×10^{-3}
10^{-3}	5.0×10^{-3}	7.0×10^{-3}	1.3×10^{-2}
10^{-2}	5.0×10^{-2}	7.4×10^{-2}	1.31×10^{-1}
2×10^{-2}	1.0×10^{-1}	1.67×10^{-1}	2.58×10^{-1}

Our theoretical analysis assumes that all task parameters share identical mode-wise subspaces. To test the robustness of the proposed method when this assumption is not strictly satisfied, we consider a misspecified setting in which each task has perturbed mode-wise factors

$$\boldsymbol{U}_{n,k}^{\star} = \mathrm{QR}\Big(\boldsymbol{U}_n^{\star} + \tau \boldsymbol{U}_{n,k}^{\star,\perp}\Big),$$

where $\boldsymbol{U}_{n,k}^{\star,\perp}$ is a task-specific component for k-th task and $\tau \geq 0$ controls the deviation from the exactly shared subspace $U_n^{\star}$. Thus, $\tau = 0$ corresponds to the exact shared-subspace model. We use D-AltMin-P2GD with $N = 3$, $d_n = 20$, $r_n = 3$, $\kappa = 4$, $K = 50$, $m_0 = 400$, $m = 35$, $T = 2500$, and $\eta = 0.04$ under noiseless observations. We vary $\tau \in \{0, 10^{-4}, 10^{-3}, 10^{-2}, 2\times 10^{-2}\}$ and measure the average model misspecification $\bar{\delta}_{\mathrm{mis}} := \frac{1}{NK} \sum_{n=1}^{N} \sum_{k=1}^{K} \left\| (\boldsymbol{I} - \boldsymbol{U}_n^{\star} \boldsymbol{U}_n^{\star\top}) \boldsymbol{U}_{n,k}^{\star} \right\|_2$, final shared-subspace error, and final maximum tensor relative error.

Table 3 shows that the proposed method degrades gradually as the deviation from exact shared subspace increases. For mild misspecification, e.g., $\tau \leq 10^{-3}$, both the shared subspace and tensor recovery errors remain small, whereas larger perturbations gradually lead to larger estimation errors. This provides empirical evidence that the method is robust to moderate violations of the exact shared subspace assumption. How to theoretically provide the guarantees for misspecified setting is left for future work.

B.2 Compared with Other MTL methods

In this section, we first compare the proposed method MTTLR with two relevant baseline methods multi-task linear representation learning (MTLR) (Vaswani, 2024) and multi-task tensor product representation learning (MTTPR) (Wu et al., 2025) on synthetic data. Then we evaluate these methods on real-world data recovery problems. Finally, we evaluate our proposed model in multitask classification using convolution neural networks.

B.2.1 Comparison on Synthetic Data

Since compared method MTTPR (Wu et al., 2025) only handles matrix inputs, we focus this experiment on $N = 2$ and set $d_1 = d_2 = d$ and $r_1 = r_2 = r$. We consider three settings, $(d, r) = (30, 3)$, $(50, 4)$, and $(80, 5)$, and vary the number of tasks K. For each task, the ground-truth matrix is generated as $\boldsymbol{X}_k^\star = \boldsymbol{U}^\star \boldsymbol{S}_k^\star \boldsymbol{V}^{\star\top}$, where $\boldsymbol{U}^\star, \boldsymbol{V}^\star \in \mathrm{St}(d, r)$ and $\{\boldsymbol{S}_k^\star\}_{k\in[K]}$ is generated such that $\kappa = 1$.

To jointly evaluate sample and communication efficiency, we report the budget required to achieve $\mathrm{Err}_t \leq 10^{-3}$. For each method, a triple $(m_{\mathrm{req}};\, T_{\mathrm{comm}};\, B_{\mathrm{round}})$ denotes the required per-task sample size, the number of communication rounds, and the transmitted bytes per communication round, respectively. The communication cost assumes 64-bit floating-point transmission.

Setting	K	**MTLR** (Vaswani, 2024)	**MTTPR** (Wu et al., 2025)	**Proposed**
$d = 30,\ r = 3$	10	$(130;\ 1715;\ 2700 \times 8)$	$(120;\ 1643;\ 2700 \times 8)$	$(\mathbf{31};\ \mathbf{318};\ \mathbf{180} \times \mathbf{8})$
	30	$(45;\ 636;\ 2700 \times 8)$	$(115;\ 559;\ 2700 \times 8)$	$(\mathbf{21};\ \mathbf{74};\ \mathbf{180} \times \mathbf{8})$
	50	$(30;\ 414;\ 2700 \times 8)$	$(101;\ 384;\ 2700 \times 8)$	$(\mathbf{17};\ \mathbf{82};\ \mathbf{180} \times \mathbf{8})$
$d = 50,\ r = 4$	10	$(1200;\ 587;\ 10^4 \times 8)$	$(1253;\ 1081;\ 10^4 \times 8)$	$(\mathbf{81};\ \mathbf{72};\ \mathbf{400} \times \mathbf{8})$
	30	$(450;\ 439;\ 10^4 \times 8)$	$(590;\ 639;\ 10^4 \times 8)$	$(\mathbf{51};\ \mathbf{34};\ \mathbf{400} \times \mathbf{8})$
	50	$(261;\ 598;\ 10^4 \times 8)$	$(480;\ 282;\ 10^4 \times 8)$	$(\mathbf{45};\ \mathbf{26};\ \mathbf{400} \times \mathbf{8})$
$d = 80,\ r = 5$	10	$(3750;\ 1576;\ 3.2 \times 10^4 \times 8)$	$(3900;\ 1200;\ 3.2 \times 10^4 \times 8)$	$(\mathbf{169};\ \mathbf{89};\ \mathbf{800} \times \mathbf{8})$
	30	$(1260;\ 856;\ 3.2 \times 10^4 \times 8)$	$(1600;\ 618;\ 3.2 \times 10^4 \times 8)$	$(\mathbf{96};\ \mathbf{35};\ \mathbf{800} \times \mathbf{8})$
	50	$(720;\ 783;\ 3.2 \times 10^4 \times 8)$	$(1300;\ 593;\ 3.2 \times 10^4 \times 8)$	$(\mathbf{70};\ \mathbf{38};\ \mathbf{800} \times \mathbf{8})$

Table 4: Budget required to achieve $\mathrm{Err}_t \leq 10^{-3}$. Each triple denotes (per-task samples; communication rounds; transmitted bytes per round).

Table 4 shows that the advantage of the proposed method extends beyond per-task sample complexity to communication efficiency. Across all settings, our method reaches the target accuracy using fewer samples and substantially fewer communication rounds. It also transmits markedly fewer bytes per round. For example, when $d = 80$ and $r = 5$, the proposed method transmits 800×8 bytes per round, compared with $3.2 \times 10^4 \times 8$ bytes for MTLR and MTTPR. For $K = 50$, it requires only 70 samples and 38 communication rounds, compared with $(720, 783)$ for MTLR and $(1300, 593)$ for MTTPR. These results demonstrate the joint sample and communication efficiency of the proposed multilinear representation and distributed optimization method.

B.2.2 Comparison on Real-World Data

We first consider a collection of compressed sensing tasks constructed from hyperspectral images (HSIs), and then study the same multi-task recovery problem on MRI data. Finally, we perform end-to-end multitask attribute classification on CelebA (Liu et al., 2015) to evaluate whether the proposed shared multilinear representation remains effective beyond linear recovery problems and in a nonlinear deep-learning setting.

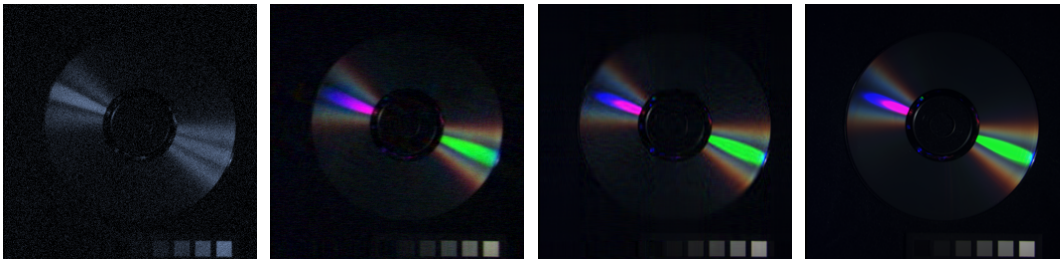
These real-data experiments are model-misspecified relative to our theoretical setting: HSI and MRI signals are better viewed as approximately rather than exactly low-rank, while CelebA classification further departs from the linear MTTLR model. They therefore provide a more challenging test of robustness to approximate multilinear structure and deviations from the exact model assumptions.

B.2.3 Application to HSI Data Recovery

For the HSI data, we use the publicly available CAVE dataset provided by Columbia University.[1] From this dataset, we select four objects *Balloons*, *Lemon*, *CD*, and *Face*. Each object is a third-order tensor of size $512 \times 512 \times 31$, where the first two modes correspond to spatial dimensions and the third mode corresponds to the spectral bands. For computational efficiency, each HSI is resized to $256 \times 256 \times 31$, and all pixel values are normalized to $[0, 1]$.

We treat the recovery of each spectral band image as an individual compressed sensing task. Therefore, for each HSI object, we obtain a collection of 31 related compressed sensing tasks, where all band images are recovered jointly from compressed measurements. To improve memory efficiency, we adopt the Walsh–Hadamard sampling strategy (Liu et al., 2023b), which is attractive due to its hardware-friendly implementation, fast transform, and strong empirical recovery performance. Under this sampling scheme, the compressed observation for task i is given by $\boldsymbol{\mathcal{A}}(\boldsymbol{X}_i^\star) = \boldsymbol{A}_i \operatorname{Vec}(\boldsymbol{X}_i^\star)$, where $\boldsymbol{A}_i = \boldsymbol{D}_i \boldsymbol{F} \boldsymbol{R}_i$. Here, $\boldsymbol{R}_i$ is a random permutation matrix, $\boldsymbol{F}$ is the Walsh–Hadamard transform matrix, and $\boldsymbol{D}_i$ is a random downsampling operator.

Since each task amounts to recovering a matrix-valued signal, we compare the proposed MTTLR method with MTLR (Vaswani, 2024) and MTTPR (Wu et al., 2025). Because MTLR is designed for vector-valued compressed sensing, we reshape each band image into a vector before applying MTLR. For all methods, we fix the compression ratio at 10%. For MTLR and MTTPR, the low-rank parameter is tuned over $r = 2, 3$, respectively, to achieve the best performance. For the proposed method, we use D-AltMin-P2GD and set the core tensor dimensions to $r_1 = r_2 = 50$.



(a) MTLR, PSNR = 22.51 (b) MTTPR, PSNR = 31.72 (c) Ours MTTLR, PSNR = 39.12 (d) Original, PSNR = Inf

Figure 6: Visual comparison of reconstruction results on the *CD* dataset.

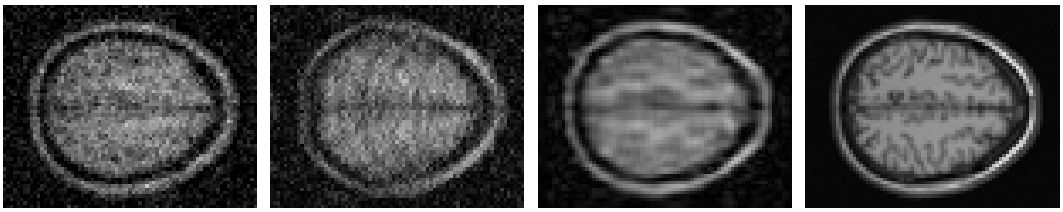
To quantitatively evaluate recovery quality, we adopt four standard image-quality metrics used in (Xie et al., 2018): peak signal-to-noise ratio (PSNR), structural similarity (SSIM), feature similarity (FSIM), and erreur relative globale adimensionnelle de synthèse (ERGAS). Larger values of PSNR, SSIM, and FSIM indicate better recovery quality, while a smaller ERGAS indicates better performance. The quantitative results are reported in Table 5, and visual comparisons are shown in Fig. 6.

From Table 5 and Fig. 6, we observe that the proposed method achieves consistently competitive recovery performance across the tested HSI objects, attaining the best results on most evaluation metrics. In contrast, the vector-based baseline MTLR performs the worst among the three methods. This is likely because vectorizing each band image destroys the intrinsic spatial structure within each task, making it more difficult to effectively exploit correlations across both the spectral and spatial modes.

[1] http://www1.cs.columbia.edu/CAVE/databases/multispectral/

Table 5: Comparison of recovery quality across different hyperspectral objects and methods.

Dataset	Method	PSNR↑	FSIM↑	SSIM↑	ERGAS↓
Balloons	Proposed MTTLR	**41.92**	**0.9583**	**0.9667**	**33.95**
	MTLR (Vaswani, 2024)	26.11	0.6647	0.4661	215.44
	MTTPR (Wu et al., 2025)	41.34	0.9574	0.9548	37.16
Lemon	Proposed MTTLR	**47.49**	0.9798	**0.9866**	**24.69**
	MTLR (Vaswani, 2024)	28.76	0.6991	0.5334	272.37
	MTTPR (Wu et al., 2025)	46.64	**0.9830**	0.9713	28.15
CD	Proposed MTTLR	**39.12**	**0.9331**	**0.9344**	**70.43**
	MTLR (Vaswani, 2024)	22.51	0.6556	0.4065	467.84
	MTTPR (Wu et al., 2025)	31.72	0.7418	0.8364	162.09
Face	Proposed MTTLR	**44.31**	**0.9792**	**0.9747**	**34.17**
	MTLR (Vaswani, 2024)	29.05	0.7728	0.5709	234.20
	MTTPR (Wu et al., 2025)	42.15	0.9588	0.9700	44.03



(a) MTLR PSNR =20.81, FSIM=0.6912 SSIM = 0.6171, ERGAS = 312.73

(b) MTTPR PSNR =19.42, FSIM=0.6567 SSIM = 0.4571, ERGAS = 366.93

(c) Ours MTTLR PSNR =21.81, FSIM=0.7178 SSIM = 0.6120, ERGAS = 270.86

(d) Original

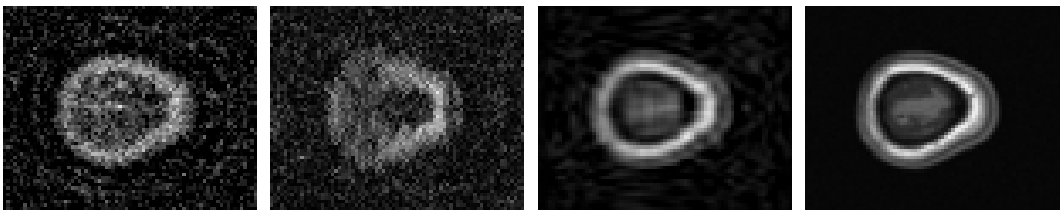
Figure 7: Visual comparison of reconstruction results on the 120-th frame of *MRI* data .

B.2.4 Application to MRI Data

We next evaluate the proposed method on MRI data. Specifically, we use an MRI volume of size $181 \times 217 \times 181$ from the Simulated Brain Database.[2] This dataset is generated by a realistic MRI simulator and is widely used in the neuroimaging literature. Due to memory limitations, we downsample each frame to a spatial resolution of 60×75.

We treat the recovery of each MRI frame, i.e., each 60×75 matrix slice, as one compressed sensing task. This yields 181 related compressed sensing tasks, which are jointly used to recover

[2] http://brainweb.bic.mni.mcgill.ca/brainweb/



(a) MTLR PSNR =18.16, FSIM=0.4623 SSIM = 0.2356, ERGAS = 630.66

(b) MTTPR PSNR =17.71, FSIM=0.4678 SSIM = 0.1942, ERGAS = 664.53

(c) Ours MTTLR PSNR =23.81, FSIM=0.6078 SSIM = 0.4719, ERGAS = 314.01

(d) Original

Figure 8: Visual comparison of reconstruction results on the 156-th frame of *MRI* data .

the original MRI volume from compressed measurements. We again adopt the Walsh–Hadamard sampling strategy described above and compare the proposed method with MTLR (Vaswani, 2024) and MTTPR (Wu et al., 2025). For all methods, the compression ratio is fixed at 20%. For MTLR and MTTPR, we tune the low-rank parameter over $r = 4, 16$, respectively, to achieve the best performance. For the proposed method, we use D-AltMin-P2GD and set the core tensor dimensions to $r_1 = r_2 = 19$.

Visual and quantitative comparisons on the 120-th and 156-th MRI frames are shown in Fig. 7 and 8, respectively. As can be seen from these figures, the proposed method yields the best visual recovery quality among the three approaches.

B.2.5 End-to-End Multitask Classification on CelebA

To further evaluate the proposed shared multilinear representation beyond linear tensor regression and compressed sensing, we conduct an end-to-end multitask classification experiment on the CelebA dataset (Liu et al., 2015). CelebA provides 40 binary facial attributes, which we treat as $K = 40$ related classification tasks. We use fixed subsets of 40000/4000/4000 images from the official training, validation, and test partitions, respectively, and resize all images to 64×64.

Shared CNN backbone. All compared methods use exactly the same CNN feature extractor, and differ only in how the 40 task-specific classifiers are parameterized. The backbone consists of three convolutional layers with 16, 32, and 64 channels, respectively, each followed by batch normalization and ReLU, with max pooling after the first two convolutional layers. The resulting feature tensor has size $\mathcal{Z}(\boldsymbol{x}) \in \mathbb{R}^{64\times16\times16}$. Thus, unlike a vector representation, the feature retains one channel mode and two spatial modes, providing a natural setting for evaluating the multilinear inductive bias of our model.

Compared multitask heads. We compare the following five parameterizations while keeping the CNN backbone, data, optimizer, and evaluation protocol identical.

Flattened low-rank head. Following classical low-dimensional shared-representation MTL (Argyriou et al., 2006; Thekumparampil et al., 2021), we vectorize $\mathcal{Z}(\boldsymbol{x})$ into $\boldsymbol{z} \in \mathbb{R}^D$, where $D = 64 \times 16 \times 16$, and constrain the stacked task-weight matrix as $\boldsymbol{W} = \boldsymbol{A}\boldsymbol{B}^\top, \boldsymbol{A} \in \mathbb{R}^{40\times q}, \boldsymbol{B} \in \mathbb{R}^{D\times q}$. We set $q = 9$, yielding a parameter budget comparable to that of our multilinear head. This baseline captures a shared linear subspace across tasks but discards the channel and spatial mode structure of the CNN feature tensor.

Flattened nuclear-norm head. Trace- or nuclear-norm regularization of the stacked task-parameter matrix is a standard parameter-based MTL strategy (Pong et al., 2010; Boursier et al., 2022). We use a full task-weight matrix $\boldsymbol{W} \in \mathbb{R}^{40\times D}$ and optimize $\mathcal{L}_{\mathrm{BCE}} + \lambda_{\mathrm{nuc}}\|\boldsymbol{W}\|_*$, with $\lambda_{\mathrm{nuc}} = 10^{-3}$. This baseline therefore encourages low-rank task relatedness through a convex spectral regularizer rather than an explicit factorization.

Hard-sharing head. Hard parameter sharing is one of the standard architectures for deep multitask learning (Caruana, 1997; Ruder, 2017). After flattening the CNN feature, we use a common nonlinear bottleneck of width $q = 9$, followed by 40 task outputs: $\boldsymbol{h} = \phi(\boldsymbol{B}^\top\boldsymbol{z}), \hat{\boldsymbol{y}} = \boldsymbol{A}\boldsymbol{h} + \boldsymbol{b}$, where ϕ is a LeakyReLU activation. This provides a conventional nonlinear shared-representation baseline with approximately the same head parameter budget as our method.

Full linear head. As an unstructured high-capacity control, we use 40 independent linear classifiers on top of the common CNN feature, $\hat{\boldsymbol{y}} = \boldsymbol{W}\boldsymbol{z} + \boldsymbol{b}, \boldsymbol{W} \in \mathbb{R}^{40\times D}$, without low-rank or nuclear-norm constraints. Such shared-backbone, task-specific output layers constitute a standard deep MTL construction (Caruana, 1997; Ruder, 2017).

Shared Tucker head (ours). We preserve the tensor structure of $\mathcal{Z}(\boldsymbol{x})$ and parameterize the classifier tensor for task k as

$$\boldsymbol{\mathcal{W}}_k = \boldsymbol{\mathcal{S}}_k \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \times_3 \boldsymbol{U}_3,$$

where

$$\boldsymbol{U}_1 \in \mathbb{R}^{64\times24}, \boldsymbol{U}_2, \boldsymbol{U}_3 \in \mathbb{R}^{16\times12}, \boldsymbol{\mathcal{S}}_k \in \mathbb{R}^{24\times12\times12}.$$

The mode-wise factors are shared by all 40 tasks, whereas each task retains its own core tensor. This directly instantiates the shared multilinear representation studied in this paper and is closely related to

Model	Seeds	Total Params ↓	Head Params ↓	Macro AUROC ↑	Macro AP ↑	Balanced Acc. ↑
CNN + Flattened Low-rank Head	5	171,552	147,856	0.8863 ± 0.0011	0.6417 ± 0.0029	0.8157 ± 0.0013
CNN + Flattened Nuclear-norm Head	5	679,096	655,400	0.8953 ± 0.0002	0.6596 ± 0.0007	0.8273 ± 0.0008
CNN + Hard-sharing Head	5	171,561	147,865	0.8829 ± 0.0010	0.6334 ± 0.0034	0.8134 ± 0.0005
CNN + Full Linear Head	5	679,096	655,400	0.9085 ± 0.0001	0.7016 ± 0.0035	0.8350 ± 0.0007
CNN + Shared Tucker Head (ours)	5	**163,896**	**140,200**	$\mathbf{0.9126 \pm 0.0006}$	$\mathbf{0.7115 \pm 0.0034}$	$\mathbf{0.8456 \pm 0.0018}$

Table 6: End-to-end multitask classification results on CelebA. All methods use the same CNN backbone and differ only in the multitask classification head. Results are reported as mean ± standard deviation over five seeds; the best results are highlighted in bold.

the general use of tensor factorizations in deep MTL (De Lathauwer et al., 2000; Yang & Hospedales, 2017b). The corresponding logit is $f_k(\boldsymbol{x}) = \langle \mathcal{Z}(\boldsymbol{x}), \boldsymbol{\mathcal{W}}_k \rangle + b_k$.

Training and evaluation. CelebA attributes are highly imbalanced, so we train all models using attribute-wise weighted binary cross-entropy, where the positive-class weight of each attribute is determined from its training-set class frequency. All models are trained end-to-end with AdamW using a batch size of 256, an initial learning rate of 10^{-3}, weight decay 10^{-4}, and a cosine learning-rate schedule for at most 20 epochs. The checkpoint with the highest validation Macro AUROC is used for test evaluation. We report Macro AUROC, Macro average precision (AP), and Macro balanced accuracy, averaged across the 40 tasks. Results are reported as mean ± standard deviation over five random seeds.

Results. As shown in Table 6, the proposed shared Tucker head achieves the best performance on all three evaluation metrics while using the fewest trainable parameters. In particular, compared with the strongest full linear baseline, it improves Macro AUROC, Macro AP, and balanced accuracy by 0.41, 0.99, and 1.06 percentage points, respectively, while using approximately 76% fewer total trainable parameters. It also consistently outperforms the flattened low-rank, nuclear-norm, and hard-sharing MTL baselines. These results suggest that preserving the channel and spatial modes of the CNN feature and sharing low-dimensional factors along each mode provides a useful multilinear inductive bias for modeling task relatedness. We emphasize that this experiment evaluates the proposed *shared multilinear representation* in an end-to-end deep MTL setting; all compared heads are optimized using AdamW, and therefore the experiment is not intended as a comparison of D-AltMin-PGD/P2GD with standard deep-learning optimizers.

C Global Convergence of D-AltMin-PGD for Solving Population Objective (3)

We have presented the two-stage optimization method to solve empirical loss (3) in Section 4. Now we prove that only D-AltMin-PGD without first-stage spectral initialization can achieve global convergence for solving the population loss defined as follows.

$$\min_{\substack{\{\boldsymbol{\mathcal{S}}_k\}_{k\in[K]},\\ \{\boldsymbol{U}_n\}_{n\in N}}} \mathbb{E}\left[F\left(\{\boldsymbol{\mathcal{S}}_k\}_{k\in[K]}, \{\boldsymbol{U}_n\}_{n\in[N]}\right)\right] = \sum_{k=1}^{K} \left\| \boldsymbol{\mathcal{S}}_k \times \{\boldsymbol{U}_n\}_{n\in[N]} - \boldsymbol{\mathcal{X}}_k^\star \right\|_F^2. \tag{18}$$

The expectation in (18) is taken with $\left\{\boldsymbol{\mathcal{A}}_i^k\right\}_{k\in[K],i\in[m]}$ in (3), which is assumed to have $i.i.d$ random entries with zero means and variance $\frac{1}{m}$. The population loss in (18) can be regarded as the ideal empirical loss in (3) with infinite sample size.

For t-th iteration with $\{\boldsymbol{U}_n^t\}_{n\in[N]}$, the exact minimization solution with $\boldsymbol{\mathcal{S}}_k$ for (18) in k-th task is

$$\boldsymbol{\mathcal{S}}_k^t = \boldsymbol{\mathcal{X}}_k^\star \times \left\{\boldsymbol{U}_n^{t\top}\right\}_{n\in[N]}. \tag{19}$$

The PGD for updating $\boldsymbol{U}_n$ is

$$\hat{\boldsymbol{U}}_n^{t+1} = \boldsymbol{U}_n^t - \eta \mathbb{E}\left[\nabla_{\boldsymbol{U}_n} F\left(\left\{\boldsymbol{\mathcal{S}}_k^t\right\}_{k\in[K]}, \left\{\boldsymbol{U}_n^t\right\}_{n\in[N]}\right)\right]$$

$$= \boldsymbol{U}_n^t - \eta \sum_{k=1}^{K} \left(\boldsymbol{X}_k^t - \boldsymbol{X}_k^\star\right)_{(n)} \left(\bigotimes_{j\neq n}^{N} \boldsymbol{U}_j^t\right) \left(\boldsymbol{S}_k^t\right)_{(n)}^\top$$

$$\boldsymbol{U}_n^{t+1} = \mathrm{QR}\left(\hat{\boldsymbol{U}}_n^{t+1}\right). \tag{20}$$

The following theorem shows that even for the highly non-convex population loss (18), AltMin-PGD can achieve global convergence from the random initialization.

Theorem 5. *For initializations $\left\{\boldsymbol{U}_n^0\right\}_{n\in[N]}$ that each $\boldsymbol{U}_n^0 = \tilde{\boldsymbol{U}}_n^0 \left(\tilde{\boldsymbol{U}}_n^{0\top} \tilde{\boldsymbol{U}}_n^0\right)^{-\frac{1}{2}}$, where $\tilde{\boldsymbol{U}}_n^0 \in \mathbb{R}^{d_n \times r_n}$ is random with $i.i.d$ standard Gaussian entries, if the step size $\eta \leq \frac{1}{2\sigma_{\max}^{\star 2}}$, then to achieve subspace alignment $\delta_t \leq \epsilon$, AltMin-PGD only needs at most*

$$T \gtrsim \kappa^2 \left(\frac{d_{\min} r_{\min}}{\xi^2}\right)^{N-1} + \kappa^2 \log \frac{1}{\epsilon} \tag{21}$$

iterations with probability at least $1 - NC\xi - 2N\exp\left(-\frac{d_{\min}}{C}\right)$, where C and ξ are sufficiently small universal constants.

The iteration number in (21) has two terms that come from two stages of convergence results. The order of $\left(\frac{d_{\min} r_{\min}}{\xi^2}\right)^{N-1}$ comes from the first stage, which dominates the number of convergence. This is because in the high-dimensional setting, random initialization is nearly orthogonal to the true subspace, which requires substantial iterations of AltMin-PGD to achieve non-trivial alignment with the true subspace. After achieving such alignment, the second term is the order of $\log \frac{1}{\epsilon}$, which means that the second stage uses a linear convergence rate to achieve exact subspace alignment. Subsequent experimental results also validate the two-stage convergence result.

Remark 2. *The population loss (18) is the coupled tensor factorization problem, which is not difficult to solve. For example, we can use the distributed power method to obtain the shared principal subspaces along each mode and then obtain each core tensor based on (19). The reason why we study AltMin-PGD is that (18) is the population loss of (3). Thus, the analysis here is useful to analyse AltMin-PGD for solving the MTTLR problem (3).*

Proof. Because of symmetry for shared factors, we suffice to present the convergence for the first factor.

$$\begin{aligned}
\hat{\boldsymbol{U}}_1^t &= \boldsymbol{U}_1^{t-1} - \eta \sum_{k=1}^{K} \left(\boldsymbol{U}_1^{t-1} \left(\boldsymbol{S}_k^t\right)_{(1)} \left(\bigotimes_{n\neq 1}^{N} \boldsymbol{U}_n^{t-1}\right)^\top - \left(\boldsymbol{X}_k^\star\right)_{(1)}\right) \left(\bigotimes_{n\neq 1}^{N} \boldsymbol{U}_n^{t-1}\right) \left(\boldsymbol{S}_k^t\right)_{(1)}^\top \\
&\overset{(i)}{=} \boldsymbol{U}_1^{t-1} - \eta \boldsymbol{U}_1^{t-1} \sum_{k=1}^{K} \left(\boldsymbol{S}_k^t\right)_{(1)} \left(\boldsymbol{S}_k^t\right)_{(1)}^\top + \eta \sum_{k=1}^{K} \left(\boldsymbol{X}_k^\star\right)_{(1)} \left(\bigotimes_{n\neq 1}^{N} \boldsymbol{U}_n^{t-1}\right) \left(\boldsymbol{S}_k^t\right)_{(1)}^\top \\
&\overset{(ii)}{=} \boldsymbol{U}_1^{t-1} + \eta \left(\boldsymbol{I} - \boldsymbol{U}_1^{t-1} \left(\boldsymbol{U}_1^{t-1}\right)^\top\right) \sum_{k=1} \left(\boldsymbol{X}_k^\star\right)_{(1)} \left(\bigotimes_{n\neq 1}^{N} \boldsymbol{U}_n^{t-1}\right) \left(\bigotimes_{n\neq 1}^{N} \boldsymbol{U}_n^{t-1}\right)^\top \left(\boldsymbol{X}_k^\star\right)_{(1)}^\top \boldsymbol{U}_1^{t-1} \\
&= \boldsymbol{U}_1^{t-1} + \eta \left(\boldsymbol{I} - \boldsymbol{U}_1^{t-1} \left(\boldsymbol{U}_1^{t-1}\right)^\top\right) \boldsymbol{X}_{(1)}^\star \left(\boldsymbol{I} \bigotimes_{n\neq 1}^{N} \boldsymbol{U}_n^{t-1}\right) \left(\bigotimes_{n\neq 1}^{N} \boldsymbol{U}_n^{t-1}\right)^\top \left(\boldsymbol{X}^\star\right)_{(1)}^\top \boldsymbol{U}_1^{t-1} \\
&\mathrm{QR}(\hat{\boldsymbol{U}}_1^t) = \boldsymbol{U}_1^t \boldsymbol{R}_1^t.
\end{aligned} \tag{22}$$

In (i), we use the property of the Kronecker product, and (ii) is because we substitute the closed solution of $\boldsymbol{\mathcal{S}}_k^t$ in (19).

In addition, we also have

$$\sigma_{\min}\left(\hat{\boldsymbol{U}}_1^t\right) = \sigma_{\min}\left(\left(\boldsymbol{I} + \eta \left(\boldsymbol{I} - \boldsymbol{U}_1^{t-1} \left(\boldsymbol{U}_1^{t-1}\right)^\top\right) \boldsymbol{X}_{(1)}^\star \left(\boldsymbol{I} \bigotimes_{n\neq 1}^{N} \boldsymbol{U}_n^{t-1}\right) \left(\boldsymbol{I} \bigotimes_{n\neq 1}^{N} \boldsymbol{U}_n^{t-1}\right)^\top \left(\boldsymbol{X}^\star\right)_{(1)}^\top\right) \boldsymbol{U}_1^{t-1}\right)$$

$$
\begin{aligned}
&\geq \sigma_{\min}\left(\boldsymbol{I}+\eta\left(\boldsymbol{I}-\boldsymbol{U}_1^{t-1}\left(\boldsymbol{U}_1^{t-1}\right)^\top\right)\boldsymbol{X}_{(1)}^\star\left(\boldsymbol{I}\bigotimes_{n\neq 1}^{N}\boldsymbol{U}_n^{t-1}\right)\left(\boldsymbol{I}\bigotimes_{n\neq 1}^{N}\boldsymbol{U}_n^{t-1}\right)^\top\left(\boldsymbol{X}^\star\right)_{(1)}^\top\right)\sigma_{\min}\left(\boldsymbol{U}_1^{t-1}\right)\\
&\geq 1-\eta\left\|\left(\boldsymbol{I}-\boldsymbol{U}_1^{t-1}\left(\boldsymbol{U}_1^{t-1}\right)^\top\right)\boldsymbol{X}_{(1)}^\star\left(\boldsymbol{I}\bigotimes_{n\neq 1}^{N}\boldsymbol{U}_n^{t-1}\right)\left(\boldsymbol{I}\bigotimes_{n\neq 1}^{N}\boldsymbol{U}_n^{t-1}\right)^\top\left(\boldsymbol{X}^\star\right)_{(1)}^\top\right\|\\
&\geq 1-\eta\left\|\boldsymbol{X}_{(1)}^\star\right\|^2\\
&>0,
\end{aligned}
\tag{23}
$$

where the first inequality is due to the property that if $\boldsymbol{A}=\boldsymbol{B}\boldsymbol{C}$ and $\boldsymbol{A}$ and $\boldsymbol{B}$ are tall or square, then $\sigma_{\min}(\boldsymbol{A})\geq\sigma_{\min}(\boldsymbol{B})\,\sigma_{\min}(\boldsymbol{C})$. The last inequality is due to the step size condition $\eta\leq\frac{c}{\sigma_{\max}^2}$. Thus, there is

$$
\begin{aligned}
&\delta\left(\boldsymbol{U}_1^t,\boldsymbol{U}_1^\star\right)=\left\|\boldsymbol{U}_{1,\perp}^{\star\top}\hat{\boldsymbol{U}}_1^t\left(\boldsymbol{R}_1^t\right)^{-1}\right\|\\
&\leq\frac{\left\|\boldsymbol{U}_{1,\perp}^{\star\top}\hat{\boldsymbol{U}}_1^t\right\|}{\sigma_{\min}\left(\hat{\boldsymbol{U}}_1^t\right)}\\
&=\frac{\left\|\boldsymbol{U}_{1,\perp}^{\star\top}\boldsymbol{U}_1^{t-1}+\eta\boldsymbol{U}_{1,\perp}^{\star\top}\left(\boldsymbol{I}-\boldsymbol{U}_1^{t-1}\left(\boldsymbol{U}_1^{t-1}\right)^\top\right)\boldsymbol{X}_{(1)}^\star\left(\boldsymbol{I}\bigotimes_{n\neq 1}^{N}\boldsymbol{U}_n^{t-1}\right)\left(\boldsymbol{I}\bigotimes_{n\neq 1}^{N}\boldsymbol{U}_n^{t-1}\right)^\top\left(\boldsymbol{X}^\star\right)_{(1)}^\top\boldsymbol{U}_1^{t-1}\right\|}{\sqrt{\sigma_{\min}\left(\left(\hat{\boldsymbol{U}}_1^t\right)^\top\hat{\boldsymbol{U}}_1^t\right)}}\\
&=\frac{\left\|\boldsymbol{U}_{1,\perp}^{\star\top}\boldsymbol{U}_1^{t-1}\left(\boldsymbol{I}-\eta\left(\boldsymbol{U}_1^{t-1}\right)^\top\boldsymbol{X}_{(1)}^\star\left(\boldsymbol{I}\bigotimes_{n\neq 1}^{N}\boldsymbol{U}_n^{t-1}\right)\left(\boldsymbol{I}\bigotimes_{n\neq 1}^{N}\boldsymbol{U}_n^{t-1}\right)^\top\boldsymbol{X}_{(1)}^{\star\top}\boldsymbol{U}_1^{t-1}\right)\right\|}{\sqrt{1+\eta^2\sigma_{\min}\left(\left(\boldsymbol{I}-\boldsymbol{U}_1^{t-1}\left(\boldsymbol{U}_1^{t-1}\right)^\top\right)\boldsymbol{X}_{(1)}^\star\left(\boldsymbol{I}\bigotimes_{n\neq 1}^{N}\boldsymbol{U}_n^{t-1}\right)\left(\boldsymbol{I}\bigotimes_{n\neq 1}^{N}\boldsymbol{U}_n^{t-1}\right)^\top\boldsymbol{X}_{(1)}^{\star\top}\boldsymbol{U}_1^{t-1}\right)}}\\
&\leq\left\|\boldsymbol{U}_{1,\perp}^{\star\top}\boldsymbol{U}_1^{t-1}\right\|\left\|\boldsymbol{I}-\eta\left(\boldsymbol{U}_1^{t-1}\right)^\top\boldsymbol{X}_{(1)}^\star\left(\boldsymbol{I}\bigotimes_{n\neq 1}^{N}\boldsymbol{U}_n^{t-1}\right)\left(\boldsymbol{I}\bigotimes_{n\neq 1}^{N}\boldsymbol{U}_n^{t-1}\right)^\top\left(\boldsymbol{X}^\star\right)_{(1)}^\top\boldsymbol{U}_1^{t-1}\right\|\\
&\overset{(i)}{\leq}\left\|\boldsymbol{U}_{1,\perp}^{\star\top}\boldsymbol{U}_1^{t-1}\right\|\left(1-\eta\sigma_{\min}^2\left(\left(\boldsymbol{U}_1^{t-1}\right)^\top\boldsymbol{X}_{(1)}^\star\left(\boldsymbol{I}\bigotimes_{n\neq 1}^{N}\boldsymbol{U}_n^{t-1}\right)\right)\right)\\
&\overset{(ii)}{=}\left\|\boldsymbol{U}_{1,\perp}^{\star\top}\boldsymbol{U}_1^{t-1}\right\|\left(1-\eta\sigma_{\min}^2\left(\boldsymbol{I}\bigotimes_{n\neq 1}^{N}\left(\left(\boldsymbol{U}_n^{t-1}\right)^\top\boldsymbol{U}_n^\star\right)\boldsymbol{S}_{(1)}^{\star\top}\boldsymbol{U}_1^{\star\top}\boldsymbol{U}_1^{t-1}\right)\right)\\
&\overset{(iii)}{\leq}\left\|\boldsymbol{U}_{1,\perp}^{\star\top}\boldsymbol{U}_1^{t-1}\right\|\left(1-\eta\sigma_{\min}^2\left(\boldsymbol{I}\bigotimes_{n\neq 1}^{N}\left(\left(\boldsymbol{U}_n^{t-1}\right)^\top\boldsymbol{U}_n^\star\right)\boldsymbol{S}_{(1)}^{\star\top}\right)\sigma_{\min}^2\left(\boldsymbol{U}_1^{\star\top}\boldsymbol{U}_1^{t-1}\right)\right)\\
&\overset{(iv)}{\leq}\left\|\boldsymbol{U}_{1,\perp}^{\star\top}\boldsymbol{U}_1^{t-1}\right\|\left(1-\eta\sigma_{\min}^2\sigma_{\min}^2\left(\boldsymbol{I}\bigotimes_{n\neq 1}^{N}\left(\left(\boldsymbol{U}_n^{t-1}\right)^\top\boldsymbol{U}_n^\star\right)\right)\sigma_{\min}^2\left(\boldsymbol{U}_1^{\star\top}\boldsymbol{U}_1^{t-1}\right)\right)\\
&\overset{(v)}{=}\left\|\boldsymbol{U}_{1,\perp}^{\star\top}\boldsymbol{U}_1^{t-1}\right\|\left(1-\eta\sigma_{\min}^2\prod_{j=1}^{N}\sigma_{\min}^2\left(\boldsymbol{U}_j^{\star\top}\boldsymbol{U}_j^{t-1}\right)\right)\\
&\overset{(vi)}{=}\delta\left(\boldsymbol{U}_1^{t-1},\boldsymbol{U}_1^\star\right)\left(1-\eta\sigma_{\min}^2\prod_{j=1}^{N}\left(1-\delta^2\left(\boldsymbol{U}_j^{t-1},\boldsymbol{U}_j^\star\right)\right)\right)\\
&\leq\delta_{t-1}\left(1-\eta\sigma_{\min}^2\left(1-\delta_{t-1}^2\right)^N\right).
\end{aligned}
\tag{24}
$$

(i) is because the step condition $\eta \leq \frac{c}{\sigma_{\max}^2}$. (ii) is based on formula of $\boldsymbol{\mathcal{X}}_{(1)}^{\star}$ and Kronecker product property. (iii) and (iv) use the property that $\sigma_{\min}(\boldsymbol{AB}) \geq \sigma_{\min}(\boldsymbol{A})\sigma_{\min}(\boldsymbol{B})$ if $\boldsymbol{AB}$ and $\boldsymbol{A}$ are tall or square. (v) is due to the Kronecker product property. (vi) is because of the fact that $\sigma_{\min}^2\left(\boldsymbol{U}_j^{\star\top}\boldsymbol{U}_j^{t-1}\right) + \left\|\boldsymbol{U}_{j,\perp}^{\star\top}\boldsymbol{U}_j^{t-1}\right\|^2 = \cos^2\theta + \sin^2\theta = 1$. The last inequality is because we define $\delta_t := \max_j \delta\left(\boldsymbol{U}_j^t, \boldsymbol{U}_j^{\star}\right)$.

Thus, the dynamic of subspace distance is as

$$\delta_{t+1} \leq \delta_t\left(1 - \eta\sigma_{\min}^2\left(1-\delta_t^2\right)^N\right). \tag{25}$$

It is apparent that $\{\delta_t\}$ is a non-increasing sequence and $\delta_t \in (0,1)$ due to $\eta \leq \frac{c}{\sigma_{\max}^2}$. Then there is

$$\begin{aligned}
\epsilon_{t+1} := 1-\delta_{t+1}^2 &\geq 1 - \delta_t^2\left(1-\eta\sigma_{\min}^2\epsilon_t^N\right)^2 \\
&= 1 - \delta_t^2\left(1 - 2\eta\sigma_{\min}^2\epsilon_t^N + \eta^2\sigma_{\min}^4\epsilon_t^{2N}\right) \\
&= \left(1-\delta_t^2\right) + \delta_t^2\eta\sigma_{\min}^2\epsilon_t^N\left(2-\eta\sigma_{\min}^2\epsilon_t^N\right) \\
&\geq \epsilon_t + \delta_t^2\eta\sigma_{\min}^2\epsilon_t^N,
\end{aligned} \tag{26}$$

where the last inequality is due to the step size condition and $\epsilon_t \in (0,1)$.

Based on the above dynamic, the convergence of $\{\delta_t\}$ can be divided into two phases. This is because the random initialization $\{\boldsymbol{U}_j^0\}$ are uniformly sampled from the Stiefel manifold, which are almost orthogonal to $\{\boldsymbol{U}_j^{\star}\}$ based on Lemma 6, which shows that δ_0 is almost 1 in high dimension setting where $\max_n r_n \ll \min_n d_n$.

(i) First Phase. $\{\boldsymbol{U}_n^t\}_{n\in[N]}$ spend time leaving the plateau to attain a non-trivial alignment with $\{\boldsymbol{U}_n^{\star}\}_{n\in[N]}$ such that $\delta_t^2 \leq 1-c_0$, where c_0 is the constant that $c_0 \in \left(1-\delta_0^2, 1\right)$. Because the monotonic property of $\{\delta_t\}$, there is $\delta_t^2 \geq 1-c_0$ in the first phase. Thus, there is

$$\begin{aligned}
\xi_t := \epsilon_t^{-(N-1)} &\leq \epsilon_{t-1}^{-(N-1)}\left(1+(1-c_0)\,\eta\sigma_{\min}^2\epsilon_{t-1}^{N-1}\right)^{-(N-1)} \\
&= \xi_{t-1}\left(1 + \frac{(1-c_0)\eta\sigma_{\min}^2}{\xi_{t-1}}\right)^{-(N-1)} \\
&\overset{(i)}{\leq} \frac{\xi_{t-1}}{1+\frac{(N-1)(1-c_0)\eta\sigma_{\min}^2}{\xi_{t-1}}} \\
&= \xi_{t-1} - \frac{(N-1)(1-c_0)\eta\sigma_{\min}^2}{1+\frac{(N-1)(1-c_0)\eta\sigma_{\min}^2}{\xi_{t-1}}} \\
&\leq \xi_{t-1} - \frac{(N-1)(1-c_0)\eta\sigma_{\min}^2}{1+(N-1)(1-c_0)\eta\sigma_{\min}^2 c_0^{N-1}} \\
&\leq \xi_0 - \frac{(N-1)(1-c_0)\eta\sigma_{\min}^2 t}{1+(N-1)(1-c_0)\eta\sigma_{\min}^2 c_0^{N-1}},
\end{aligned} \tag{27}$$

where (i) uses the Bernoulli inequality that $(1+x)^r \geq 1+rx$ for $r \geq 1, x \geq -1$ and last inequality is unrolling the recursion to $t=0$. Thus, the first time that $\delta_t^2 \leq 1-c_0$ denoted as

$$t_1 := \min\left\{t : \delta_t^2 \leq 1-c_0\right\}, \quad c_0 \in \left(1-\delta_0^2, 1\right) \tag{28}$$

can be upper-bounded as

$$t_1 \leq \frac{\left(\left(1-\delta_0^2\right)^{-(N-1)} - c_0^{-(N-1)}\right)\left(\kappa^2 + c\,(N-1)\,(1-c_0)\,c_0^{N-1}\right)}{c(N-1)(1-c_0)} \tag{29}$$

based on the (27).

(ii) Second Phase. The second phase is that δ_t has a linear convergence rate to zero when the iterates finish the first phase, where $\left\{\boldsymbol{U}_j^t\right\}$ have a non-trivial alignment with $\left\{\boldsymbol{U}_j^\star\right\}$ as $\delta_t^2 \leq 1 - c_0$. Based on one step contraction in (25) for $\forall t \geq t_1$ as

$$\delta_t \leq \delta_{t-1}\left(1 - \eta\sigma_{\min}^2 c_0^N\right). \tag{30}$$

Combining with the above linear convergence rate with the first phase, for arbitrary small $\epsilon > 0$ alignment with the true subspace, the total running steps should be

$$t \lesssim \frac{\left(\left(1-\delta_0^2\right)^{-(N-1)} - c_0^{-(N-1)}\right)\left(\kappa^2 + c\left(N-1\right)\left(1-c_0\right)c_0^{N-1}\right)}{c(N-1)(1-c_0)} + \frac{\kappa^2}{cc_0^N}\log\frac{1}{\epsilon}. \tag{31}$$

Substituting the result in Lemma 6 can finish the proof. □

D Proof of Theorem 1

Firstly, we have the following well-known lemma in (Arbenz et al., 2012; Li et al., 2021; Saad, 2011) to bound the iteration number of Algorithm 1 needed to achieve the subspace distance $\delta\left(\left\{\boldsymbol{U}_n^0, \tilde{\boldsymbol{U}}_n\right\}_{n\in N}\right) \leq \epsilon$.

Lemma 1. *Denote the top-r_n eigen-space matrix of $\hat{\boldsymbol{W}}_n$ as $\hat{\boldsymbol{U}}_n$, to obtain the multiple subspaces distance $\delta\left(\left\{\boldsymbol{U}_n^0, \tilde{\boldsymbol{U}}_n\right\}_{n\in[N]}\right) \leq \epsilon$ with high probability, the iterations needed in Algorithm 1 are*

$$\Omega\left(\max_{n\in[N]}\frac{\sigma_{r_n}\left(\hat{\boldsymbol{W}}_n\right)}{\sigma_{r_n}\left(\hat{\boldsymbol{W}}_n\right) - \sigma_{r_{n+1}}\left(\hat{\boldsymbol{W}}_n\right)}\log\left(\frac{d_{\max}}{\epsilon}\right)\right), \tag{32}$$

To show the explicit formula of the above computational complexity, we need to bound the $\sigma_{r_n}\left(\hat{\boldsymbol{W}}_n\right)$ and $\sigma_{r_{n+1}}\left(\hat{\boldsymbol{W}}_n\right)$, which can be given based on the following theorem.

Theorem 6. *For the MTTLR in (1) under Assumption 1, suppose $\{\boldsymbol{\mathcal{A}}_{k,i}^0\}$ have $i.i.d$. standard Gaussian entries and noise $\left\{\xi_{k,i}^0\right\}$ are $i.i.d$ with $\xi_{k,i}^0 \sim \mathcal{N}\left(0, \sigma_v^2\right)$. If the per-task sample size m_0 satisfies*

$$m_0 \gtrsim \frac{N\mu^2\kappa^4 r_{\max}^{\frac{3}{2}}\left(\sqrt{\prod_{n=1}^N d_n} + d_{\max}\right)\log\left(m_0 N K d_{\max}\right)}{\sqrt{K}} + \frac{N\mu^2\kappa^8 d_{\max} r_{\max}^2}{K}, \tag{33}$$

ans SNR is large enough that satisfies

$$SNR \gtrsim \max\left\{\frac{N\kappa^6 d_{\max} r_{\max}}{m_0 K}, \frac{N\kappa^2 r_{\max}^{\frac{1}{2}}\left(\sqrt{\prod_{n=1}^N d_n} + d_{\max}\right)}{m_0\sqrt{K}}\right\}, \tag{34}$$

then (5) has

$$\left\|\hat{\boldsymbol{W}}_n - \boldsymbol{X}_{(n)}^\star \boldsymbol{X}_{(n)}^{\star\top}\right\| \leq \frac{3\sigma_{\min}^2}{800\sqrt{N r_{\max}}\kappa^2} \tag{35}$$

holds with probability at least $1 - \frac{1}{d_{\max}^{10}}$.

Proof. Because $\left\{\boldsymbol{\mathcal{A}}_{k,i}^0\right\}$ have $i.i.d$ normal Gaussian entries and noise is additive Gaussian, there is

$$\mathbb{E}\left[\hat{\boldsymbol{W}}_n\right] = \boldsymbol{W}_n^\star := \sum_{k=1}^K \left(\boldsymbol{X}_k^\star\right)_{(n)}\left(\boldsymbol{X}_k^\star\right)_{(n)}^\top = \boldsymbol{X}_{(n)}^\star \boldsymbol{X}_{(n)}^{\star\top}. \tag{36}$$

Using a similar decoupling technique (Xia & Yuan, 2019; Li et al., 2023; Xia et al., 2021) for 2-order U-statistics $\hat{\boldsymbol{W}}_n - \boldsymbol{W}_n^\star$, we introduce the a new 2-order concentration U-statistics estimator

$$\tilde{\boldsymbol{W}}_n = \sum_{k=1}^{K} \frac{1}{2m_0(m_0-1)} \sum_{1\le i<i'\le m_0} y_{k,i}^0 \tilde{y}_{k,i'}^0 \left(\left(\boldsymbol{A}_{k,i}^0\right)_{(n)} \left(\tilde{\boldsymbol{A}}_{k,i'}^0\right)_{(n)}^\top + \left(\tilde{\boldsymbol{A}}_{k,i'}^0\right)_{(n)} \left(\boldsymbol{A}_{k,i}^0\right)_{(n)}^\top \right), \tag{37}$$

where the $\tilde{\boldsymbol{A}}_{k,i'}^0$ is standard Gaussian that is dependent with $\boldsymbol{A}_{k,i}^0$ and

$$\tilde{y}_{k,i'}^0 = \left\langle \tilde{\boldsymbol{\mathcal{A}}}_{k,i'}^0, \boldsymbol{\mathcal{X}}_k^\star \right\rangle + \tilde{\xi}_{k,i'}, \quad \forall k \in [K], i' \in [m_0], \tag{38}$$

Based on Theorem 3.4.1 in (De la Pena & Giné, 2012), we have

$$\mathbb{P}\left(\|\hat{\boldsymbol{W}}_n - \boldsymbol{W}_n^\star\| \ge t \right) \le 15\mathbb{P}\left(\|\tilde{\boldsymbol{W}}_n - \boldsymbol{W}_n^\star\| \ge t \right). \tag{39}$$

Because there is decomposition

$$\begin{aligned}
\tilde{\boldsymbol{W}}_n - \boldsymbol{W}_n^\star = {} & \frac{m_0}{2(m_0-1)} \underbrace{\sum_{k=1}^{K} \left(\boldsymbol{\Delta}_{k,n} \tilde{\boldsymbol{\Delta}}_{k,n}^\top + \tilde{\boldsymbol{\Delta}}_{k,n} \boldsymbol{\Delta}_{k,n}^\top \right)}_{\Pi_1} + \frac{m_0}{2(m_0-1)} \underbrace{\sum_{k=1}^{K} \left(\boldsymbol{\Delta}_{k,n} + \tilde{\boldsymbol{\Delta}}_{k,n} \right) \left(\boldsymbol{X}_k^\star\right)_{(n)}^\top}_{\Pi_2} \\
& + \frac{1}{m_0 - 1} \underbrace{\sum_{k=1}^{K} \left(\frac{1}{2m_0} \sum_{i=1}^{m} y_{k,i}^0 \tilde{y}_{k,i}^0 \left(\left(\boldsymbol{A}_{k,i}^0\right)_{(n)} \left(\tilde{\boldsymbol{A}}_{k,i}^0\right)_{(n)}^\top + \left(\tilde{\boldsymbol{A}}_{k,i}^0\right)_{(n)} \left(\boldsymbol{A}_{k,i}^0\right)_{(n)}^\top \right) - \left(\boldsymbol{X}_k^\star\right)_{(n)} \left(\boldsymbol{X}_k^\star\right)_{(n)}^\top \right)}_{\Pi_3} \\
& + \frac{m_0}{2(m_0-1)} \sum_{k=1}^{K} \left(\boldsymbol{X}_k^\star\right)_{(n)} \left(\boldsymbol{\Delta}_{k,n} + \tilde{\boldsymbol{\Delta}}_{k,n} \right)^\top,
\end{aligned} \tag{40}$$

where the $\boldsymbol{\Delta}_{k,n}$ and $\tilde{\boldsymbol{\Delta}}_{k,n}$ are defined as

$$\boldsymbol{\Delta}_{k,n} := \frac{1}{m_0} \sum_{i=1}^{m_0} y_{k,i}^0 \left(\boldsymbol{A}_{k,i}^0\right)_{(n)} - \left(\boldsymbol{X}_k^\star\right)_{(n)}, \quad \tilde{\boldsymbol{\Delta}}_{k,n} := \frac{1}{m_0} \sum_{i=1}^{m_0} \tilde{y}_{k,i}^0 \left(\tilde{\boldsymbol{A}}_{k,i}^0\right)_{(n)} - \left(\boldsymbol{X}_k^\star\right)_{(n)}. \tag{41}$$

(i) Upper bound for Π_1. Because $\boldsymbol{\Delta}_{k,n}, \tilde{\boldsymbol{\Delta}}_{k,n}$ are independent for $\forall k \in [K]$, $\boldsymbol{\Delta}_{p,n}(\tilde{\boldsymbol{\Delta}}_{p,n})$ and $\boldsymbol{\Delta}_{q,n}(\tilde{\boldsymbol{\Delta}}_{q,n})$ are independent for $\forall p \ne q, \forall p, q \in [K]$, we can handle the concentration bound conditioned on $\tilde{\boldsymbol{\Delta}}_{k,n}, k = 1, \cdots, K$. Firstly, we have the following event

$$\mathcal{E}_n^1 := \left\{ \forall k \in [K], \max\left\{ \left\| \tilde{\boldsymbol{\Delta}}_{k,n} \right\|, \|\boldsymbol{\Delta}_{k,n}\| \right\} \lesssim \max_k \nu_k \sqrt{\frac{\hat{d}_n \log(K N \hat{d}_n)}{m_0}} \right\} \tag{42}$$

holds with probability at least $1 - \frac{1}{10N d_{\max}^{10}}$ based on the matrix Bernstein inequality (Koltchinskii et al., 2011) and union bound, where $\nu_k^2 := \|\boldsymbol{\mathcal{X}}_k^\star\|_F^2 + \sigma_v^2$ is the variance of Gaussian $y_{k,i}^0, \tilde{y}_{k,i}^0$ and $\hat{d}_n := \max(d_n, d_n^-)$.

Then, based on the variational formulation of the spectral norm, there is

$$\left\| \sum_{k=1}^{K} \boldsymbol{\Delta}_{k,n} \tilde{\boldsymbol{\Delta}}_{k,n}^\top \right\| = \max_{\substack{\boldsymbol{w}\in\mathbb{R}^{d_n}, \|\boldsymbol{w}\|=1 \\ \boldsymbol{z}\in\mathbb{R}^{d_n}, \|\boldsymbol{z}\|=1}} \sum_{k=1}^{K} \sum_{i=1}^{m_0} \boldsymbol{w}^\top \left(y_{k,i}^0 \left(\boldsymbol{A}_{k,i}^0\right)_{(n)} \tilde{\boldsymbol{\Delta}}_{k,n}^\top - \left(\boldsymbol{X}_k^\star\right)_{(n)} \tilde{\boldsymbol{\Delta}}_{k,n}^\top \right) \boldsymbol{z} / m_0. \tag{43}$$

The summands in the above equality are independent zero-mean, sub-exponential $r.v.s$ with sub-exponential norm

$$\psi_{ki} \lesssim \left\| y_{k,i}^0 \right\|_{\psi_2} \left\| \boldsymbol{w}\boldsymbol{z}^\top \tilde{\boldsymbol{\Delta}}_{k,n} \right\|_F / m_0$$

$$\lesssim \nu_k \left\| \boldsymbol{w}\boldsymbol{z}^\top \tilde{\boldsymbol{\Delta}}_{k,n} \right\|_F / m_0 \tag{44}$$

Therefore

$$\begin{aligned}
\frac{t^2}{\sum_{ki}\psi_{ki}^2} &\gtrsim \frac{m_0 t^2}{\sum_{k=1}^K \nu_k^2 \left\| \boldsymbol{w}\boldsymbol{z}^\top \tilde{\boldsymbol{\Delta}}_{k,n} \right\|_F^2} \\
&\geq \frac{m_0 t^2}{\max_k \nu_k^2 \left\| \boldsymbol{w}\boldsymbol{z}^\top \right\|_F^2 \sum_{k=1}^K \left\| \tilde{\boldsymbol{\Delta}}_{k,n} \right\|^2} \\
&\overset{(i)}{\gtrsim} \frac{m_0^2 t^2}{\hat{d}_n \log\left(NK\hat{d}_n\right) \max_k \nu_k^2 \sum_k \nu_k^2} \\
&\overset{(ii)}{\geq} \frac{K m_0^2 t^2}{\sigma_{\max}^4 \hat{d}_n \log\left(NK\hat{d}_n\right) \left(\mu^2 r_{\max} + (\kappa^2 \mathsf{SNR})^{-1}\right) \left(r_{\max} + (\kappa^2 \mathsf{SNR})^{-1}\right)} \\
&\overset{(iii)}{\geq} \frac{K m_0^2 t^2}{3\sigma_{\max}^4 \hat{d}_n \log\left(NK\hat{d}_n\right)} \min\left\{ \frac{1}{\mu^2 r_{\max}^2}, \frac{\kappa^2 \mathsf{SNR}}{(\mu^2+1) r_{\max}}, \kappa^4 \mathsf{SNR}^2 \right\} \\
&\gtrsim d_n
\end{aligned} \tag{45}$$

where (i) is due to (42). (ii) comes from $\max_k \nu_k^2 \leq \frac{\sigma_{\max}^2}{K}\left(\mu^2 r_{\max} + \frac{1}{\kappa^2 \mathsf{SNR}}\right), \sum_k \nu_k^2 \leq \sigma_{\max}^2\left(r_{\max} + \frac{1}{\kappa^2 \mathsf{SNR}}\right)$, which is due to definition of $\nu_k^2 = \|\boldsymbol{\mathcal{X}}_k^\star\|_F^2 + \sigma_v^2$, SNR and Assumption 1. The (iii) uses $\frac{1}{a+b+c} \geq \frac{1}{3}\min\left\{\frac{1}{a}, \frac{1}{b}, \frac{1}{c}\right\}$. The last inequality is because we can set $t = \frac{\sigma_{\min}^2}{2000\sqrt{N r_{\max}}\kappa^2}$ such that the first term in sample size condition (33) is sufficiently to guarantee $\frac{K m_0^2 t^2}{3\mu^2 r_{\max}^2 \sigma_{\max}^4 \hat{d}_n \log\left(NK\hat{d}_n\right)} \gtrsim d_n$. The second term of SNR condition in (34) and first term of sample size condition in (33) are sufficiently to induce $\frac{K\kappa^2 m_0^2 t^2 \mathsf{SNR}}{3(\mu^2+1) r_{\max} \sigma_{\max}^4 \hat{d}_n \log\left(NK\hat{d}_n\right)} \gtrsim d_n$ and $\frac{K\kappa^4 \mathsf{SNR}^2 m_0^2 t^2}{3\sigma_{\max}^4 \hat{d}_n \log\left(NK\hat{d}_n\right)} \gtrsim d_n$.

Similarly,

$$\begin{aligned}
\frac{t}{\max_{ki}\psi_{ki}} &\gtrsim \frac{m_0^{\frac{3}{2}} K t}{2\sigma_{\max}^2 \sqrt{\hat{d}_n \log\left(NK\hat{d}_n\right)}} \min\left\{ \frac{1}{\mu^2 r_{\max}}, \kappa^2 \mathsf{SNR} \right\} \\
&\gtrsim d_n,
\end{aligned} \tag{46}$$

where the last inequality is due to both first term and second term of sample complexity in (33) and SNR condition in (34). Thus, we can conclude

$$\mathbb{P}\left(\|\Pi_1\| \leq \frac{\sigma_{\min}^2}{2000\sqrt{N r_{\max}}\kappa^2} \right) \geq 1 - \frac{1}{10 N d_{\max}^{10}}, \tag{47}$$

based on sub-exponential Bernstein inequality, ϵ-net argument and $\min_n d_n \gg \max(\log d_{\max}, \log N)$.

(2) Upper bound for Π_2. Also, based on the variational formulation of the matrix spectral norm, there is

$$\left\| \sum_{k=1}^K \boldsymbol{\Delta}_{k,n} \left(\boldsymbol{X}_k^\star\right)_{(n)}^\top \right\| = \max_{\substack{\boldsymbol{w}\in\mathbb{R}^{d_n}, \|\boldsymbol{w}\|=1 \\ \boldsymbol{z}\in\mathbb{R}^{d_n}, \|\boldsymbol{z}\|=1}} \sum_{k=1}^K \sum_{i=1}^{m_0} \boldsymbol{w}^\top \left(y_{k,i}^0 \left(\boldsymbol{A}_{k,i}^0\right)_{(n)} \left(\boldsymbol{X}_k^\star\right)_{(n)}^\top - \left(\boldsymbol{X}_k^\star\right)_{(n)} \left(\boldsymbol{X}_k^\star\right)_{(n)}^\top \right) \boldsymbol{z} / m_0. \tag{48}$$

The summands in the above equality are also independent zero-mean, sub-exponential $r.v.s$ with sub-exponential norm

$$\psi_{ki} \lesssim \left\| y_{k,i}^0 \right\|_{\psi_2} \left\| \boldsymbol{w}\boldsymbol{z}^\top \left(\boldsymbol{X}_k^\star\right)_{(n)} \right\|_F / m_0$$

$$\lesssim \nu_k \left\| \boldsymbol{w}\boldsymbol{z}^\top \left(\boldsymbol{X}_k^\star\right)_{(n)} \right\|_F / m_0 \tag{49}$$

Thus, there is

$$\begin{aligned}
\frac{t^2}{\sum_{ki}\psi_{ki}^2} &\gtrsim \frac{m_0 t^2}{\sum_k \nu_k^2 \left\| \boldsymbol{w}\boldsymbol{z}^\top \left(\boldsymbol{X}_k^\star\right)_{(n)} \right\|_F^2} \\
&\overset{(i)}{\geq} \frac{m_0 t^2}{\max_k \nu_k^2 \sigma_{\max}^2} \\
&\overset{(ii)}{\geq\geq} \frac{m_0 K t^2}{\sigma_{\max}^4 \left(\mu^2 r_{\max} + \left(\kappa^2 \mathrm{SNR}\right)^{-1}\right)} \\
&\geq \frac{m_0 K t^2}{2\sigma_{\max}^4} \min\left\{ \frac{1}{\mu^2 r_{\max}}, \kappa^2 \mathrm{SNR} \right\}
\end{aligned} \tag{50}$$

where (i) follows from $\sum_{k=1}^K \left\| \boldsymbol{w}\boldsymbol{z}^\top \left(\boldsymbol{X}_k^\star\right)_{(n)} \right\|_F^2 = \sum_{k=1}^K \left\| \left(\boldsymbol{X}_k^\star\right)_{(n)} \boldsymbol{z} \right\|_2^2 = \boldsymbol{z}\left(\boldsymbol{X}_k^\star\right)_{(n)} \left(\boldsymbol{X}_k^\star\right)_{(n)}^\top \boldsymbol{z} \leq \sigma_{\max}^2$. (ii) is due to $\max_k \nu_k^2 \leq \frac{\sigma_{\max}^2}{K}\left(\mu^2 r_{\max} + \frac{1}{\kappa^2 \mathrm{SNR}}\right)$.

In addition, there is

$$\begin{aligned}
\frac{t}{\max_{k,i}\psi_{ki}} &\gtrsim \frac{m_0 t}{\max_k \nu_k \max_k \left\|\boldsymbol{\mathcal{X}}_k^\star\right\|_F} \\
&\geq \frac{K m_0 t}{\mu \sigma_{\max}^2 \sqrt{r_{\max}} \sqrt{\mu^2 r_{\max} + \left(\kappa^2 \mathrm{SNR}\right)^{-1}}} \\
&\geq \frac{K m_0 t}{\mu \sigma_{\max}^2 \sqrt{r_{\max}} \left(\mu \sqrt{r_{\max}} + \kappa^{-1} \mathrm{SNR}^{-\frac{1}{2}}\right)} \\
&\geq \frac{K m_0 t}{2\mu \sigma_{\max}^2 \sqrt{r_{\max}}} \min\left\{ \frac{1}{\mu\sqrt{r_{\max}}}, \kappa\sqrt{\mathrm{SNR}} \right\}
\end{aligned} \tag{51}$$

where the last inequality also follows from bound of ν_k and Assumption 1.

Set $t = \frac{\sigma_{\min}^2}{2000\sqrt{N r_{\max}}\kappa^2}$ and substituting it into (50) and (51) has

$$\begin{aligned}
\min\left\{ \frac{t^2}{\sum_{ki}\psi_{ki}^2}, \frac{t}{\max_{ki}\psi_{ik}} \right\} &\gtrsim \min\left\{ \frac{m_0 K t^2 \min\left\{\frac{1}{\mu^2 r_{\max}}, \kappa^2 \mathrm{SNR}\right\}}{2\sigma_{\max}^4}, \frac{K m_0 t \min\left\{\frac{1}{\mu\sqrt{r_{\max}}}, \kappa\sqrt{\mathrm{SNR}}\right\}}{2\mu\sigma_{\max}^2\sqrt{r_{\max}}} \right\} \\
&\gtrsim d_n,
\end{aligned} \tag{52}$$

where the second term of sample complexity in (33) suficiently induces $\frac{m_0 K t^2}{2\sigma_{\max}^4 \mu^2 r_{\max}} \gtrsim d_n$ and $\frac{K m_0 t}{2\mu^2 r_{\max}\sigma_{\max}^2}$. The first term of SNR condition (34) is sufficiently to guarantee $\frac{m_0 K t^2 \kappa^2 \mathrm{SNR}}{2\sigma_{\max}^4} \gtrsim d_n$ and $\frac{K\kappa m_0 t\sqrt{\mathrm{SNR}}}{2\mu\sigma_{\max}^2\sqrt{r_{\max}}} \gtrsim d_n$. Thus, we can conclude

$$\mathbb{P}\left(\|\Pi_2\| \leq \frac{\sigma_{\min}^2}{2000\sqrt{N r_{\max}}\kappa^2} \right) \geq 1 - \frac{1}{10 N d_{\max}^{10}}, \tag{53}$$

based on sub-exponential Bernstein inequality, ϵ-net argument and $\min_n d_n \gg \max(\log d_{\max}, \log N)$.

(iii) Upper bound for Π_3. There is decomposition

$$\begin{aligned}
\Pi_3 = &\frac{1}{m_0}\sum_{k=1}^K \sum_{i=1}^{m_0} y_{k,i}^0 \left(\boldsymbol{A}_{k,i}^0\right)_{(n)} \left(\tilde{y}_{k,i}^0 \left(\tilde{\boldsymbol{A}}_{k,i}^0\right)_{(n)}^\top - \left(\boldsymbol{X}_k^\star\right)_{(n)}^\top \right) \\
&+ \frac{1}{m_0}\sum_{k=1}^K \sum_{i=1}^{m_0} \left(y_{k,i}^0 \left(\boldsymbol{A}_{k,i}^0\right)_{(n)} - \left(\boldsymbol{X}_k^\star\right)_{(n)} \right) \left(\boldsymbol{X}_k^\star\right)_{(n)}^\top .
\end{aligned} \tag{54}$$

The spectral norm bound for the second term has been given in the last step. Thus, the key step is to provide a high-probability bound for the spectral norm of the first summation term, which has

$$\max_{\substack{\boldsymbol{w}\in\mathbb{R}^{d_n},\|\boldsymbol{w}\|=1\\ \boldsymbol{z}\in\mathbb{R}^{d_n},\|\boldsymbol{z}\|=1}} \sum_{k=1}^{K}\sum_{i=1}^{m_0}\frac{1}{m_0}\left\langle y_{k,i}^0\left(\boldsymbol{A}_{k,i}^0\right)_{(n)}^\top \boldsymbol{w}\boldsymbol{z}^\top, \tilde{y}_{k,i}^0\left(\tilde{\boldsymbol{A}}_{k,i}^0\right)^\top - \left(\boldsymbol{X}_k^\star\right)_{(n)}^\top\right\rangle. \tag{55}$$

The summands in above are also independent zero-mean, sub-exponential r.v.s conditioned on the following event.

$$\mathcal{E}_n^2 := \left\{\forall k\in K, \max_{k\in[K],i\in[m_0]}\left\|y_{k,i}^0\left(\boldsymbol{A}_{k,i}^0\right)_{(n)}\right\| \lesssim \sqrt{\hat{d}_n}\max_k \nu_k \log\left(m_0 NKd_{\max}\right)\right\}, \tag{56}$$

which holds with probability at least $1-\frac{1}{10Nd_{\max}^{10}}$ based on matrix Bernstein inequality in Lemma 4. Thus, we have

$$\begin{aligned}\psi_{ki} &\lesssim \left\|\tilde{y}_{k,i}^0\right\|_{\psi_2}\left\|y_{k,i}^0\left(\boldsymbol{A}_{k,i}^0\right)_{(n)}^\top \boldsymbol{w}\boldsymbol{z}^\top\right\|_F/m_0\\ &\leq \frac{\nu_k^2\sqrt{\hat{d}_n}\log\left(m_0NKd_{\max}\right)}{m_0}.\end{aligned} \tag{57}$$

Setting $t := \mathcal{O}\left(\nu_{\max}^2\sqrt{\frac{K\hat{d}_n d_n}{m_0}}\log\left(m_0NKd_{\max}\right)\right)$ can obtain

$$\begin{aligned}\min\left\{\frac{t^2}{\sum_{ki}\psi_{ki}^2}, \frac{t}{\max_{ki}\psi_{ik}}\right\} &\gtrsim \min\left\{\mathcal{O}\left(d_n\right), \sqrt{m_0Kd_n}\right\}\\ &\gtrsim d_n,\end{aligned} \tag{58}$$

where the last inequality is due to sample complexity in (33). Thus, we have

$$\begin{aligned}\frac{\|\Pi_3\|}{m_0-1} &\lesssim \frac{\sigma_{\max}^2}{m_0^{\frac{3}{2}}\sqrt{K}}\left(\mu^2 r_{\max}+\frac{1}{\kappa^2\mathrm{SNR}}\right)\sqrt{\hat{d}_n d_n}\log\left(m_0NKd_{\max}\right)+\frac{\sigma_{\min}^2}{2000\sqrt{Nr_{\max}}\kappa^2}\\ &\leq \frac{\sigma_{\min}^2}{1000\sqrt{Nr_{\max}}\kappa^2}\end{aligned} \tag{59}$$

with probability at least $1-\frac{1}{10Nd_{\max}^{10}}$ by using sub-exponential Bernstein inequality, ϵ-net argument and $\min_n d_n \gg \max\left(\log d_{\max}, \log N\right)$. The last inequality uses the first term of sample size complexity and SNR condition in Theorem 1.

Combing above bounds for Π_1, Π_1, Π_3 and union bound can finish the proof of Theorem 5. □

Now we can prove the Theorem 1 based on Lemma 1 and Theorem 5.

Proof.

$$\begin{aligned}\delta\left(\left\{\boldsymbol{U}_n^0, \boldsymbol{U}_n^\star\right\}_{n\in[N]}\right) &= \max_{n\in[N]}\left\|\boldsymbol{U}_n^0\boldsymbol{U}_n^{0^\top}-\boldsymbol{U}_n^\star\boldsymbol{U}_n^{\star^\top}\right\|\\ &\leq \max_{n\in[N]}\left\|\boldsymbol{U}_n^0\boldsymbol{U}_n^{0^\top}-\tilde{\boldsymbol{U}}_n\tilde{\boldsymbol{U}}_n^\top\right\|+\max_{n\in[N]}\left\|\tilde{\boldsymbol{U}}_n\tilde{\boldsymbol{U}}_n^\top-\boldsymbol{U}_n^\star\boldsymbol{U}_n^{\star^\top}\right\|\\ &\overset{(i)}{\leq} \max_{n\in[N]}\left\|\boldsymbol{U}_n^0\boldsymbol{U}_n^{0^\top}-\tilde{\boldsymbol{U}}_n\tilde{\boldsymbol{U}}_n^\top\right\|+\frac{\left\|\hat{\boldsymbol{W}}_n-\boldsymbol{W}_n^\star\right\|}{\sigma_{r_n}^2-0-\left\|\hat{\boldsymbol{W}}_n-\boldsymbol{W}_n^\star\right\|}\\ &\leq \max_{n\in[N]}\left\|\boldsymbol{U}_n^0\boldsymbol{U}_n^{0^\top}-\tilde{\boldsymbol{U}}_n\tilde{\boldsymbol{U}}_n^\top\right\|+\frac{\left\|\hat{\boldsymbol{W}}_n-\boldsymbol{W}_n^\star\right\|}{\sigma_{\min}^2-\left\|\hat{\boldsymbol{W}}_n-\boldsymbol{W}_n^\star\right\|}\\ &\overset{(ii)}{\leq} \max_{n\in[N]}\left\|\boldsymbol{U}_n^0\boldsymbol{U}_n^{0^\top}-\tilde{\boldsymbol{U}}_n\tilde{\boldsymbol{U}}_n^\top\right\|+\frac{\frac{3\sigma_{\min}^2}{800\sqrt{Nr_{\max}}\kappa^2}}{\sigma_{\min}^2-\frac{3\sigma_{\min}^2}{800\sqrt{Nr_{\max}}\kappa^2}}\end{aligned}$$

$$\leq \max_{n\in[N]} \left\| \boldsymbol{U}_n^0 \boldsymbol{U}_n^{0\top} - \tilde{\boldsymbol{U}}_n \tilde{\boldsymbol{U}}_n^\top \right\| + \frac{3}{400\sqrt{N r_{\max}}\kappa^2}, \tag{60}$$

where (i) uses the Davis-Kahan $\sin\Theta$ theorem and (ii) uses the Theorem 5.

To achieve $\delta\left(\left\{\boldsymbol{U}_n^0, \boldsymbol{U}_n^\star\right\}_{n\in[N]}\right) \leq \frac{1}{100\sqrt{N r_{\max}}\kappa^2}$, we should run distributed power method to make $\max_{n\in[N]} \left\| \boldsymbol{U}_n^0 \boldsymbol{U}_n^{0\top} - \tilde{\boldsymbol{U}}_n \tilde{\boldsymbol{U}}_n^\top \right\| \leq \frac{1}{400\sqrt{N r_{\max}}\kappa^2}$, which requires the $\epsilon = \frac{1}{400\sqrt{N r_{\max}}\kappa^2}$ in Lemma 1. To obtain the required number of iteration of power method, we need to bound the $\sigma_{r_n}\left(\hat{\boldsymbol{W}}_n\right), \sigma_{r_{n+1}}\left(\hat{\boldsymbol{W}}_n\right)$ based on Wely's inequality and Theorem 5 as follows

$$\begin{aligned}
\sigma_{r_n}\left(\hat{\boldsymbol{W}}_n\right) - \sigma_{r_n}\left(\boldsymbol{W}_n^\star\right) &\geq -\left\| \hat{\boldsymbol{W}}_n - \boldsymbol{X}_{(n)}^\star \boldsymbol{X}_{(n)}^{\star\top} \right\| \geq -\frac{3\sigma_{\min}^2}{800\sqrt{N r_{\max}}\kappa^2} \\
\sigma_{r_{n+1}}\left(\hat{\boldsymbol{W}}_n\right) - \sigma_{r_{n+1}}\left(\boldsymbol{W}_n^\star\right) &\leq \left\| \hat{\boldsymbol{W}}_n - \boldsymbol{X}_{(n)}^\star \boldsymbol{X}_{(n)}^{\star\top} \right\| \leq \frac{3\sigma_{\min}^2}{800\sqrt{N r_{\max}}\kappa^2}
\end{aligned} \tag{61}$$

Thus, substituting above bound and $\epsilon = \frac{1}{400\sqrt{N r_{\max}}\kappa^2}$ will finish the proof of Theorem 1. □

E Proof of Theorem 2

E.1 Exact Minimization Part

For the exact minimization problem for the core tensor, which is the solution

$$\begin{aligned}
\boldsymbol{\mathcal{S}}_k^t &= \underset{\boldsymbol{\mathcal{S}}\in\mathbb{R}^{r_1\times\cdots\times r_N}}{\arg\min} \sum_{i=1}^m \left(\left\langle \boldsymbol{\mathcal{A}}_{k,i}^t, \boldsymbol{\mathcal{S}} \times \left\{\boldsymbol{U}_n^t\right\}_{n\in[N]} \right\rangle - \left\langle \boldsymbol{\mathcal{A}}_{k,i}^t, \boldsymbol{\mathcal{S}}_k^\star \times \left\{\boldsymbol{U}_n^\star\right\}_{n\in[N]} \right\rangle - \xi_{k,i} \right)^2 \\
&= \underset{\boldsymbol{\mathcal{S}}\in\mathbb{R}^{r_1\times\cdots\times r_N}}{\arg\min} \sum_{i=1}^m \left(\left\langle \boldsymbol{\mathcal{A}}_{k,i}^t \times \left\{ \left(\boldsymbol{U}_n^t\right)^\top \right\}_{n\in[N]}, \boldsymbol{\mathcal{S}} \right\rangle - \left\langle \boldsymbol{\mathcal{A}}_{k,i}^t \times \left\{ \boldsymbol{U}_n^{\star\top} \right\}_{n\in[N]}, \boldsymbol{\mathcal{S}}_k^\star \right\rangle - \xi_{k,i} \right)^2,
\end{aligned} \tag{62}$$

which is the tensor format of the vector $\boldsymbol{s}_k^t \in \mathbb{R}^{r_1\cdots r_N}$ defined as the solution of least square problem

$$\begin{aligned}
\boldsymbol{s}_k^t &= \underset{\boldsymbol{s}\in\mathbb{R}^{r_1\cdots r_N}}{\arg\min} \sum_{i=1}^m \left(\left\langle \left(\bigotimes_{n=1}^N \boldsymbol{U}_n^t \right)^\top \mathrm{Vec}\left(\boldsymbol{\mathcal{A}}_{k,i}^t\right), \boldsymbol{s} \right\rangle - \left\langle \left(\bigotimes_{n=1}^N \boldsymbol{U}_n^\star \right)^\top \mathrm{Vec}\left(\boldsymbol{\mathcal{A}}_{k,i}^t\right), \boldsymbol{s}_k^\star \right\rangle - \xi_{k,i} \right)^2 \\
&= \underset{\boldsymbol{s}\in\mathbb{R}^{r_1\cdots r_N}}{\arg\min} \left\| \boldsymbol{A}_k^t \left(\bigotimes_{n=1}^N \boldsymbol{U}_n^t \right) \boldsymbol{s} - \boldsymbol{A}_k^t \left(\bigotimes_{n=1}^N \boldsymbol{U}_n^\star \right) \boldsymbol{s}_k^\star - \boldsymbol{\xi}_k \right\|_F^2 \\
&= \left(\left(\boldsymbol{A}_k^t \left(\bigotimes_{n=1}^N \boldsymbol{U}_n^t \right) \right)^\top \boldsymbol{A}_k^t \left(\bigotimes_{n=1}^N \boldsymbol{U}_n^t \right) \right)^{-1} \left(\boldsymbol{A}_k^t \left(\bigotimes_{n=1}^N \boldsymbol{U}_n^t \right) \right)^\top \left(\boldsymbol{A}_k^t \left(\bigotimes_{n=1}^N \boldsymbol{U}_n^\star \right) \boldsymbol{s}_k^\star + \boldsymbol{\xi}_k \right),
\end{aligned} \tag{63}$$

where the $\boldsymbol{A}_k^t := \left[\mathrm{Vec}\left(\boldsymbol{\mathcal{A}}_{k,i}^t\right)^\top; \cdots; \mathrm{Vec}\left(\boldsymbol{\mathcal{A}}_{k,m}^t\right)^\top \right] \in \mathbb{R}^{m\times n_1\cdots n_N}$ is the stacked the vectorization of $\boldsymbol{\mathcal{A}}_{k,i}^t$ and it has $i.i.d$ normal Gaussian entries.

Then for $\boldsymbol{\mathcal{G}}_k^t := \boldsymbol{\mathcal{X}}_k^\star \times \left\{ \left(\boldsymbol{U}_n^t\right)^\top \right\}_{n\in[N]} \in \mathbb{R}^{r_1\times\cdots\times r_N}$, we can have

$$\begin{aligned}
\left\| \boldsymbol{\mathcal{S}}_k^t - \boldsymbol{\mathcal{G}}_k^t \right\|_F = \Bigg\| &\left(\left(\boldsymbol{A}_k^t \left(\bigotimes_{n=1}^N \boldsymbol{U}_n^t \right) \right)^\top \boldsymbol{A}_k^t \left(\bigotimes_{n=1}^N \boldsymbol{U}_n^t \right) \right)^{-1} \left(\boldsymbol{A}_k^t \left(\bigotimes_{n=1}^N \boldsymbol{U}_n^t \right) \right)^\top \left(\boldsymbol{A}_k^t \left(\bigotimes_{n=1}^N \boldsymbol{U}_n^\star \right) \boldsymbol{s}_k^\star \right. \\
&\left. + \boldsymbol{\xi}_k \right) - \left(\bigotimes_{n=1}^N \left(\left(\boldsymbol{U}_n^t\right)^\top \boldsymbol{U}_n^\star \right) \right) \boldsymbol{s}_k^\star \Bigg\|_F
\end{aligned}$$

$$
\begin{aligned}
&\leq \left\| \left(\left(\boldsymbol{A}_k^t \left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t \right) \right)^\top \boldsymbol{A}_k^t \left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t \right) \right)^{-1} \left(\boldsymbol{A}_k^t \left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t \right) \right)^\top \boldsymbol{\xi}_k \right\|_F \\
&\quad + \left\| \left(\left(\boldsymbol{A}_k^t \left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t \right) \right)^\top \boldsymbol{A}_k^t \left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t \right) \right)^{-1} \left(\boldsymbol{A}_k^t \left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t \right) \right)^\top \boldsymbol{A}_k^t \left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^\star \right) \boldsymbol{s}_k^\star \right. \\
&\quad - \left(\left(\boldsymbol{A}_k^t \left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t \right) \right)^\top \boldsymbol{A}_k^t \left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t \right) \right)^{-1} \left(\boldsymbol{A}_k^t \left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t \right) \right)^\top \boldsymbol{A}_k^t \left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t \right) \\
&\quad \left. \cdot \left(\bigotimes_{n=1}^{N} \left(\left(\boldsymbol{U}_n^t \right)^\top \boldsymbol{U}_n^\star \right) \right) \boldsymbol{s}_k^\star \right\|_F \\
&\overset{(i)}{=} \left\| \left(\left(\boldsymbol{A}_k^t \left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t \right) \right)^\top \boldsymbol{A}_k^t \left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t \right) \right)^{-1} \left(\boldsymbol{A}_k^t \left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t \right) \right)^\top \boldsymbol{A}_k^t \left(\boldsymbol{I} - \bigotimes_{n=1}^{N} \left(\boldsymbol{U}_n^t \left(\boldsymbol{U}_n^t \right)^\top \right) \right) \right. \\
&\quad \left. \cdot \left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^\star \right) \boldsymbol{s}_k^\star \right\|_F + \left\| \left(\left(\boldsymbol{A}_k^t \left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t \right) \right)^\top \boldsymbol{A}_k^t \left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t \right) \right)^{-1} \left(\boldsymbol{A}_k^t \left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t \right) \right)^\top \boldsymbol{\xi}_k \right\|_F \\
&\leq \sigma_{\min}^{-1} \left(\left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t \right)^\top \boldsymbol{A}_k^{t\top} \boldsymbol{A}_k^t \left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t \right) \right) \left\| \left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t \right)^\top \boldsymbol{A}_k^{t\top} \boldsymbol{A}_k^t \left(\boldsymbol{I} - \bigotimes_{n=1}^{N} \left(\boldsymbol{U}_n^t \left(\boldsymbol{U}_n^t \right)^\top \right) \right) \right. \\
&\quad \left. \cdot \left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^\star \right) \boldsymbol{s}_k^\star \right\| + \left\| \left(\left(\boldsymbol{A}_k^t \left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t \right) \right)^\top \boldsymbol{A}_k^t \left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t \right) \right)^{-1} \left(\boldsymbol{A}_k^t \left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t \right) \right)^\top \boldsymbol{\xi}_k \right\|_F .
\end{aligned} \tag{64}
$$

Now we can bound the first noiseless term and second noisy term above, respectively.

First, we have $\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t$ is also orthogonal due to $\left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t \right)^\top \bigotimes_{n=1}^{N} \boldsymbol{U}_n^t = \bigotimes_{n=1}^{N} \left(\left(\boldsymbol{U}_n^t \right)^\top \boldsymbol{U}_n^t \right) = \boldsymbol{I}$, which is based on orthogonality of $\boldsymbol{U}_n^t$. Thus, there is

$$
\mathbb{E} \left[\left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t \right)^\top \boldsymbol{A}_k^{t\top} \boldsymbol{A}_k^t \left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t \right) \right] = m \boldsymbol{I}. \tag{65}
$$

Combining the sub-exponential Bernstein inequality, variational formula of spectral norm, and ϵ-net argument, there is

$$
\mathbb{P} \left\{ \left\| \left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t \right)^\top \boldsymbol{A}_k^{t\top} \boldsymbol{A}_k^t \left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t \right) - m \boldsymbol{I} \right\| \leq 1.1 \epsilon m \right\} \geq 1 - \exp \left(\prod_{n=1}^{N} r_n - \epsilon m \right). \tag{66}
$$

Thus, the lower bound of the minimal singular value of $\left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t \right)^\top \boldsymbol{A}_k^{t\top} \boldsymbol{A}_k^t \left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t \right)$ is

$$
\sigma_{\min} \left(\left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t \right)^\top \boldsymbol{A}_k^{t\top} \boldsymbol{A}_k^t \left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t \right) \right) \geq (1 - 1.1\epsilon)\, m. \tag{67}
$$

In addition, there is

$$
\mathbb{E} \left[\left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^t \right)^\top \boldsymbol{A}_k^{t\top} \boldsymbol{A}_k^t \left(\boldsymbol{I} - \bigotimes_{n=1}^{N} \left(\boldsymbol{U}_n^t \left(\boldsymbol{U}_n^t \right)^\top \right) \right) \left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^\star \right) \boldsymbol{s}_k^\star \right] = \boldsymbol{0}_{\prod_{n=1}^{N} r_n}. \tag{68}
$$

Using the sub-exponential Bernstein inequality, the variational formula of Frobenius norm, union bound and ϵ-net argument, $\forall k \in [K]$, the

$$\begin{aligned}&\left\|\left(\bigotimes_{n=1}^{N}\boldsymbol{U}_n^t\right)^\top \boldsymbol{A}_k^{t\top}\boldsymbol{A}_k^t\left(\boldsymbol{I}-\bigotimes_{n=1}^{N}\left(\boldsymbol{U}_n^t\left(\boldsymbol{U}_n^t\right)^\top\right)\right)\left(\bigotimes_{n=1}^{N}\boldsymbol{U}_n^\star\right)\boldsymbol{s}_k^\star\right\| \\ &\qquad\le 1.1\epsilon m\left\|\left(\boldsymbol{I}-\bigotimes_{n=1}^{N}\left(\boldsymbol{U}_n^t\left(\boldsymbol{U}_n^t\right)^\top\right)\right)\left(\bigotimes_{n=1}^{N}\boldsymbol{U}_n^\star\right)\boldsymbol{s}_k^\star\right\| \end{aligned} \tag{69}$$

holds with probability at least $1-K\exp\left(\prod_{n=1}^{N}r_n-\epsilon m\right)$.

For the noisy term, we denote $\boldsymbol{P}_t$ as $\boldsymbol{A}_k^t\left(\bigotimes_{n=1}^{N}\boldsymbol{U}_n^t\right)$ to simply notation. Thus, conditioned $\boldsymbol{P}_t$, $\left(\boldsymbol{P}_t^\top\boldsymbol{P}_t\right)^{-1}\boldsymbol{P}_t\boldsymbol{\xi}_k\sim\boldsymbol{\mathcal{N}}\left(\boldsymbol{0},\sigma_v^2\left(\boldsymbol{P}_t^\top\boldsymbol{P}_t\right)^{-1}\right)$, which can be rewritten as $\sigma_v\left(\boldsymbol{P}_t^\top\boldsymbol{P}_t\right)^{-\frac{1}{2}}\boldsymbol{g}$ that $\boldsymbol{g}\sim\boldsymbol{\mathcal{N}}\left(\boldsymbol{0},\boldsymbol{I}_{\prod_{n=1}^{N}r_n}\right)$. We can conclude that

$$\left\|\left(\left(\boldsymbol{A}_k^t\left(\bigotimes_{n=1}^{N}\boldsymbol{U}_n^t\right)\right)^\top\boldsymbol{A}_k^t\left(\bigotimes_{n=1}^{N}\boldsymbol{U}_n^t\right)\right)^{-1}\left(\boldsymbol{A}_k^t\left(\bigotimes_{n=1}^{N}\boldsymbol{U}_n^t\right)\right)^\top\boldsymbol{\xi}_k\right\|_F\le 40\sigma_v\sqrt{\frac{\prod_{n=1}^{N}r_n+\log\left(Kd_{\max}\right)}{m(1-1.1\epsilon)}} \tag{70}$$

holds with probability at least $1-\frac{1}{10d_{\max}^{10}}$ conditioned $\boldsymbol{P}_t$ based on (67) and Lemma 3.

Combing result in (66), (69) and (70), we can have

$$\begin{aligned}\left\|\boldsymbol{\mathcal{S}}_k^t-\boldsymbol{\mathcal{G}}_k^t\right\|_F&\le 1.2\epsilon\left\|\left(\boldsymbol{I}-\bigotimes_{n=1}^{N}\left(\boldsymbol{U}_n^t\left(\boldsymbol{U}_n^t\right)^\top\right)\right)\left(\bigotimes_{n=1}^{N}\boldsymbol{U}_n^\star\right)\boldsymbol{s}_k^\star\right\|+40\sigma_v\sqrt{\frac{\prod_{n=1}^{N}r_n+\log\left(Kd_{\max}\right)}{m(1-1.1\epsilon)}}\\ &\le 1.2\epsilon\sqrt{1-\left(1-\delta_t^2\right)^N}\left\|\boldsymbol{\mathcal{S}}_k^\star\right\|_F+\frac{40\sigma_v}{\sqrt{1-1.1\epsilon}}\sqrt{\frac{\prod_{n=1}^{N}r_n+\log\left(Kd_{\max}\right)}{m}},\end{aligned} \tag{71}$$

where the last inequality is due to Lemma 5.

E.2 Projected Gradient Descent Part

Because $\boldsymbol{U}_1^t,\cdots,\boldsymbol{U}_N^t$ are symmetric, we only need to consider the projected gradient descent with $\boldsymbol{U}_1^t$ and omit the analysis for the other factors. The gradient with $\boldsymbol{U}_1^t$ is

$$\begin{aligned}\nabla_{\boldsymbol{U}_1^t}F=&\sum_{k=1}^{K}\sum_{i=1}^{m}\left\langle\left(\boldsymbol{A}_{k,i}^t\right)_{(1)},\boldsymbol{U}_1^t\left(\boldsymbol{S}_k^t\right)_{(1)}\left(\bigotimes_{n\neq1}^{N}\boldsymbol{U}_n^t\right)^\top-\boldsymbol{X}_{(1)}^\star\right\rangle\left(\boldsymbol{A}_{k,i}^t\right)_{(1)}\left(\bigotimes_{n\neq1}^{N}\boldsymbol{U}_n^t\right)\left(\boldsymbol{S}_k^t\right)_{(1)}^\top\\ &-\sum_{k=1}^{K}\sum_{i=1}^{m}\xi_{k,i}\left(\boldsymbol{A}_{k,i}^t\right)_{(1)}\left(\bigotimes_{n\neq1}^{N}\boldsymbol{U}_n^t\right)\left(\boldsymbol{S}_k^t\right)_{(1)}^\top,\end{aligned} \tag{72}$$

where the first line is noiseless gradient and second line comes from noise. Its expectation is

$$\begin{aligned}\mathbb{E}\left[\nabla_{\boldsymbol{U}_1^t}F\right]=&m\boldsymbol{U}_1^t\sum_{k=1}^{K}\left(\boldsymbol{S}_k^t\right)_{(1)}\left(\bigotimes_{n\neq1}^{N}\boldsymbol{U}_n^t\right)^\top\left(\bigotimes_{n\neq1}^{N}\boldsymbol{U}_n^t\right)\left(\boldsymbol{S}_k^t\right)_{(1)}^\top\\ &-m\boldsymbol{U}_1^\star\sum_{k=1}^{K}\left(\boldsymbol{S}_k^\star\right)_{(1)}\left(\bigotimes_{n\neq1}^{N}\boldsymbol{U}_n^\star\right)^\top\left(\bigotimes_{n\neq1}^{N}\boldsymbol{U}_n^t\right)\left(\boldsymbol{S}_k^t\right)_{(1)}^\top\\ =&m\boldsymbol{U}_1^t\sum_{k=1}^{K}\left(\boldsymbol{S}_k^t\right)_{(1)}\left(\boldsymbol{S}_k^t\right)_{(1)}^\top-m\boldsymbol{U}_1^\star\sum_{k=1}^{K}\left(\boldsymbol{S}_k^\star\right)_{(1)}\left(\bigotimes_{n\neq1}^{N}\boldsymbol{U}_n^\star\right)^\top\left(\bigotimes_{n\neq1}^{N}\boldsymbol{U}_n^t\right)\left(\boldsymbol{S}_k^t\right)_{(1)}^\top.\end{aligned} \tag{73}$$

Then, one step of projected GD by QR decomposition is

$$\hat{\boldsymbol{U}}_1^{t+1} = \boldsymbol{U}_1^t - \eta\mathbb{E}\left[\nabla_{\boldsymbol{U}_1^t} F\right] + \eta\left(\mathbb{E}\left[\nabla_{\boldsymbol{U}_1^t} F\right] - \nabla_{\boldsymbol{U}_1^t} F\right), \quad \text{QR}\left(\hat{\boldsymbol{U}}_1^{t+1}\right) = \boldsymbol{U}_1^{t+1}\boldsymbol{R}_1^{t+1}. \tag{74}$$

Thus, we have

$$\begin{aligned}
\delta\left(\boldsymbol{U}_1^{t+1}, \boldsymbol{U}_1^\star\right) &= \left\|\boldsymbol{U}_{1,\perp}^\star \boldsymbol{U}_1^{t+1}\right\| \\
&= \left\|\boldsymbol{U}_{1,\perp}^\star \hat{\boldsymbol{U}}_1^{t+1}\left(\boldsymbol{R}_1^{t+1}\right)^{-1}\right\| \\
&\leq \left\|\boldsymbol{U}_{1,\perp}^\star \hat{\boldsymbol{U}}_1^{t+1}\right\| \sigma_{\min}\left(\left(\boldsymbol{R}_1^{t+1}\right)^{-1}\right) \\
&= \frac{\left\|\boldsymbol{U}_{1,\perp}^\star \boldsymbol{U}_1^t\left(\boldsymbol{I} - m\eta\sum_{k=1}^K \left(\boldsymbol{S}_k^t\right)_{(1)}\left(\boldsymbol{S}_k^t\right)_{(1)}^\top\right) + \boldsymbol{U}_{1,\perp}^\star \eta\left(\mathbb{E}\left[\nabla_{\boldsymbol{U}_1^t} F\right] - \nabla_{\boldsymbol{U}_1^t} F\right)\right\|}{\sigma_{\min}\left(\boldsymbol{R}_1^{t+1}\right)} \\
&\leq \frac{\delta\left(\boldsymbol{U}_1^t, \boldsymbol{U}_1^\star\right)\left\|\boldsymbol{I}_{r_1} - m\eta\sum_{k=1}^K \left(\boldsymbol{S}_k^t\right)_{(1)}\left(\boldsymbol{S}_k^t\right)_{(1)}^\top\right\| + \eta\left\|\mathbb{E}\left[\nabla_{\boldsymbol{U}_1^t} F\right] - \nabla_{\boldsymbol{U}_1^t} F\right\|}{\sigma_{\min}\left(\hat{\boldsymbol{U}}_1^{t+1}\right)}.
\end{aligned} \tag{75}$$

Now comes to upper bound terms $\left\|\boldsymbol{I}_{r_1} - m\eta\sum_{k=1}^K \left(\boldsymbol{S}_k^t\right)_{(1)}\left(\boldsymbol{S}_k^t\right)_{(1)}^\top\right\|, \left\|\mathbb{E}\left[\nabla_{\boldsymbol{U}_1^t} F\right] - \nabla_{\boldsymbol{U}_1^t} F\right\|$ and lower bound $\sigma_{\min}\left(\hat{\boldsymbol{U}}_1^{t+1}\right)$, respectively.

(I) Upper bound for $\left\|\boldsymbol{I}_{r_1} - m\eta\sum_{k=1}^K \left(\boldsymbol{S}_k^t\right)_{(1)}\left(\boldsymbol{S}_k^t\right)_{(1)}^\top\right\|$.

To achieve this, we need the lower bound for $\sigma_{\min}\left(\boldsymbol{S}_{(1)}^t\right)$. For convenience, denote $\boldsymbol{S}_{(1)}^t := \left[\left(\boldsymbol{S}_1^t\right)_{(1)}, \cdots, \left(\boldsymbol{S}_K^t\right)_{(1)}\right], \boldsymbol{G}_{(1)}^t := \left[\left(\boldsymbol{G}_1^t\right)_{(1)}, \cdots, \left(\boldsymbol{G}_K^t\right)_{(1)}\right], \boldsymbol{S}_{(1)}^\star := \left[\left(\boldsymbol{S}_1^\star\right)_{(1)}, \cdots, \left(\boldsymbol{S}_K^\star\right)_{(1)}\right]$, then it is equivalent to bound the $\left\|\boldsymbol{I}_{r_1} - m\eta\boldsymbol{S}_{(1)}^t\left(\boldsymbol{S}_{(1)}^t\right)^\top\right\|$. Thus, there is

$$\begin{aligned}
\sigma_{\min}\left(\boldsymbol{S}_{(1)}^t\right) - \sigma_{\min}\left(\boldsymbol{G}_{(1)}^t\right) &= \sigma_{\min}\left(\boldsymbol{S}_{(1)}^t\right) - \sigma_{\min}\left(\left(\boldsymbol{U}_1^t\right)^\top \boldsymbol{U}_1^\star \boldsymbol{S}_{(1)}^\star\left(\boldsymbol{I}\bigotimes_{n\neq 1}^N\left(\left(\boldsymbol{U}_n^t\right)^\top \boldsymbol{U}_j^\star\right)\right)^\top\right) \\
&\geq -\left\|\boldsymbol{S}_{(1)}^t - \boldsymbol{G}_{(1)}^t\right\|_F.
\end{aligned} \tag{76}$$

In addition,

$$\begin{aligned}
\sigma_{\min}\left(\left(\boldsymbol{U}_1^t\right)^\top \boldsymbol{U}_1^\star \boldsymbol{S}_{(1)}^\star\left(\boldsymbol{I}\bigotimes_{n\neq 1}^N\left(\left(\boldsymbol{U}_n^t\right)^\top \boldsymbol{U}_j^\star\right)\right)^\top\right) &= \sigma_{\min}\left(\boldsymbol{I}\bigotimes_{n\neq 1}^N\left(\left(\boldsymbol{U}_n^t\right)^\top \boldsymbol{U}_j^\star\right)\boldsymbol{S}_{(1)}^{\star\top}\boldsymbol{U}_1^{\star\top}\boldsymbol{U}_1^t\right) \\
&\geq \sigma_{\min}\left(\boldsymbol{I}\bigotimes_{n\neq 1}^N\left(\left(\boldsymbol{U}_n^t\right)^\top \boldsymbol{U}_n^\star\right)\right)\sigma_{\min}\left(\left(\boldsymbol{U}_1^t\right)^\top \boldsymbol{U}_1^\star\right)\sigma_{\min}\left(\boldsymbol{S}_{(1)}^\star\right) \\
&\geq \left(1-\delta_t^2\right)^{\frac{N}{2}}\sigma_{\min}.
\end{aligned} \tag{77}$$

Based on (71), we can conclude

$$\begin{aligned}
\sigma_{\min}\left(\boldsymbol{S}_{(1)}^t\right) &\geq \left(1-\delta_t^2\right)^{\frac{N}{2}}\sigma_{\min} - 1.2\epsilon\sqrt{1-\left(1-\delta_t^2\right)^N}\sqrt{\sum_{k=1}^K\left\|\boldsymbol{\mathcal{S}}_k^\star\right\|_F^2} \\
&\quad - \frac{40\sqrt{K}\sigma_v}{\sqrt{1-1.1\epsilon}}\sqrt{\frac{\prod_{n=1}^N r_n + \log\left(Kd_{\max}\right)}{m}} \\
&\overset{(i)}{\geq} \left(1-\frac{N}{2}\delta_t^2\right)\sigma_{\min} - 1.2\epsilon\sqrt{Nr_{\max}}\sigma_{\max}\delta_t - \frac{40\sigma_{\min}}{\sqrt{1-1.1\epsilon}}\sqrt{\frac{\prod_{n=1}^N r_n + \log\left(Kd_{\max}\right)}{m\text{SNR}}}
\end{aligned}$$

$$\geq 0.91\sigma_{\min}, \tag{78}$$

where (i) is due to Bernoulli inequality $(1+x)^r \geq 1+rx, x > -1, r \geq 1$. The last inequality is because initialization condition $\delta_t \leq \frac{0.01}{\sqrt{N r_{\max}}\kappa}$ and SNR condition that SNR $\gtrsim \frac{\prod_{n=1}^N r_n + \log(K d_{\max})}{m}$.

Next, there is

$$\begin{aligned}
\left\|\boldsymbol{S}_{(1)}^t\right\| &\leq \left\|\boldsymbol{G}_{(1)}^t\right\| + \left\|\boldsymbol{S}_{(1)}^t - \boldsymbol{G}_{(1)}^t\right\|_F \\
&\leq \sigma_{\max} + 1.2\epsilon\sqrt{1-(1-\delta_t^2)^N}\left\|\boldsymbol{S}_{(1)}^\star\right\|_F + \frac{40\sigma_{\min}}{\sqrt{1-1.1\epsilon}}\sqrt{\frac{\prod_{n=1}^N r_n + \log\left(K d_{\max}\right)}{m\text{SNR}}} \\
&\leq 1.09\sigma_{\max}.
\end{aligned} \tag{79}$$

The last inequality is also due to Bernoulli inequality and $\delta_t \leq \frac{0.01}{\sqrt{N r_{\max}}\kappa}$.

(II) Concentration bound for $\left\|\mathbb{E}\left[\nabla_{\boldsymbol{U}_1^t}F\right] - \nabla_{\boldsymbol{U}_1^t}F\right\|$. The $\boldsymbol{\mathcal{X}}_k^t - \boldsymbol{\mathcal{X}}_k^\star$ can be decomposed as

$$\begin{aligned}
\boldsymbol{\mathcal{X}}_k^t - \boldsymbol{\mathcal{X}}_k^\star &= \left(\boldsymbol{\mathcal{S}}_k^t - \boldsymbol{\mathcal{G}}_k^t\right) \times \left\{\boldsymbol{U}_n^t\right\}_{n\in[N]} + \boldsymbol{\mathcal{G}}_k^t \times \left\{\boldsymbol{U}_n^t\right\}_{n\in[N]} - \boldsymbol{\mathcal{X}}_k^\star \\
&= \underbrace{\left(\text{Ten}\left(\left(\boldsymbol{P}_t^\top\boldsymbol{P}_t\right)^{-1}\boldsymbol{P}_t^\top\boldsymbol{A}_k^t\left(\boldsymbol{I} - \bigotimes_{n=1}^N\left(\boldsymbol{U}_n^t\left(\boldsymbol{U}_n^t\right)^\top\right)\right)\left(\bigotimes_{n=1}^N\boldsymbol{U}_n^\star\right)\boldsymbol{s}_k^\star\right) + \boldsymbol{\mathcal{G}}_k^t\right) \times \left\{\boldsymbol{U}_n^t\right\}_{n\in[N]} - \boldsymbol{\mathcal{X}}_k^\star}_{\overline{\boldsymbol{\mathcal{X}}}_k^t} \\
&\quad + \underbrace{\text{Ten}\left(\left(\boldsymbol{P}_t^\top\boldsymbol{P}_t\right)^{-1}\boldsymbol{P}_t^\top\boldsymbol{\xi}_k\right) \times \left\{\boldsymbol{U}_n^t\right\}_{n\in[N]}}_{\boldsymbol{\mathcal{N}}_k^t}.
\end{aligned} \tag{80}$$

Thus, the difference between population gradient and empirical gradient can be decomposed as

$$\begin{aligned}
\left\|\mathbb{E}\left[\nabla_{\boldsymbol{U}_1^t}F\right] - \nabla_{\boldsymbol{U}_1^t}F\right\| &\leq \max_{\substack{\boldsymbol{w}\in\mathbb{R}^{d_1},\|\boldsymbol{w}\|=1\\ \boldsymbol{z}\in\mathbb{R}^{r_1},\|\boldsymbol{z}\|=1}}\sum_{k=1}^K\sum_{i=1}^m\left\langle\boldsymbol{\mathcal{A}}_{k,i}^t, \boldsymbol{\mathcal{X}}_k^t - \boldsymbol{\mathcal{X}}_k^\star\right\rangle\boldsymbol{w}^\top\left(\boldsymbol{A}_{k,i}^t\right)_{(1)}\left(\bigotimes_{n\neq1}^N\boldsymbol{U}_n^t\right)\left(\boldsymbol{S}_k^t\right)_{(1)}^\top\boldsymbol{z} \\
&\quad - \max_{\substack{\boldsymbol{w}\in\mathbb{R}^{d_1},\|\boldsymbol{w}\|=1\\ \boldsymbol{z}\in\mathbb{R}^{r_1},\|\boldsymbol{z}\|=1}}\sum_{k=1}^K\sum_{i=1}^m\xi_{k,i}\boldsymbol{w}^\top\left(\boldsymbol{A}_{k,i}^t\right)_{(1)}\left(\bigotimes_{n\neq1}^N\boldsymbol{U}_n^t\right)\left(\boldsymbol{S}_k^t\right)_{(1)}^\top\boldsymbol{z} - \boldsymbol{w}^\top\mathbb{E}\left[\nabla_{\boldsymbol{U}_1^t}F\right]\boldsymbol{z} \\
&\leq \underbrace{\max_{\substack{\boldsymbol{w}\in\mathbb{R}^{d_1},\|\boldsymbol{w}\|=1\\ \boldsymbol{z}\in\mathbb{R}^{r_1},\|\boldsymbol{z}\|=1}}\sum_{k=1}^K\sum_{i=1}^m\left\langle\boldsymbol{\mathcal{A}}_{k,i}^t, \overline{\boldsymbol{\mathcal{X}}}_k^t - \boldsymbol{\mathcal{X}}_k^\star\right\rangle\boldsymbol{w}^\top\left(\boldsymbol{A}_{k,i}^t\right)_{(1)}\left(\bigotimes_{n\neq1}^N\boldsymbol{U}_n^t\right)\left(\boldsymbol{S}_k^t\right)_{(1)}^\top\boldsymbol{z} - \mathbb{E}\left[\Pi_1\right]}_{\Pi_1} \\
&\quad + \underbrace{\max_{\substack{\boldsymbol{w}\in\mathbb{R}^{d_1},\|\boldsymbol{w}\|=1\\ \boldsymbol{z}\in\mathbb{R}^{r_1},\|\boldsymbol{z}\|=1}}\sum_{k=1}^K\sum_{i=1}^m\left\langle\boldsymbol{\mathcal{A}}_{k,i}^t, \boldsymbol{\mathcal{N}}_k^t\right\rangle\boldsymbol{w}^\top\left(\boldsymbol{A}_{k,i}^t\right)_{(1)}\left(\bigotimes_{n\neq1}^N\boldsymbol{U}_n^t\right)\left(\boldsymbol{S}_k^t\right)_{(1)}^\top\boldsymbol{z} - \mathbb{E}\left[\Pi_2\right]}_{\Pi_2} \\
&\quad \underbrace{- \max_{\substack{\boldsymbol{w}\in\mathbb{R}^{d_1},\|\boldsymbol{w}\|=1\\ \boldsymbol{z}\in\mathbb{R}^{r_1},\|\boldsymbol{z}\|=1}}\sum_{k=1}^K\sum_{i=1}^m\xi_{k,i}\boldsymbol{w}^\top\left(\boldsymbol{A}_{k,i}^t\right)_{(1)}\left(\bigotimes_{n\neq1}^N\boldsymbol{U}_n^t\right)\left(\boldsymbol{S}_k^t\right)_{(1)}^\top\boldsymbol{z} - \mathbb{E}\left[\Pi_3\right]}_{\Pi_3},
\end{aligned} \tag{81}$$

Now can bound $\Pi_i, i = 1,2,3$ terms separately as follows.

(i). For the $\Pi_1 - \mathbb{E}\left[\Pi_1\right]$, the summands in the above equality are independent zero-mean, sub-exponential $r.v.s$ with sub-exponential norm

$$\psi_{ki} \lesssim \left\|\overline{\boldsymbol{\mathcal{X}}}_k^t - \boldsymbol{\mathcal{X}}_k^\star\right\|_F\left\|\boldsymbol{w}\boldsymbol{z}^\top\left(\boldsymbol{S}_k^t\right)_{(1)}\left(\bigotimes_{n\neq1}^N\boldsymbol{U}_n^t\right)^\top\right\|_F. \tag{82}$$

Thus, there is

$$\begin{aligned}\frac{t^2}{\sum_{ki}\psi_{ki}^2} &\gtrsim \frac{t^2}{m\sum_{k=1}^K \left\|\overline{\boldsymbol{\mathcal{X}}}_k^t - \boldsymbol{\mathcal{X}}_k^\star\right\|_F^2 \left\|\boldsymbol{w}\boldsymbol{z}^\top \left(\boldsymbol{\mathcal{S}}_k^t\right)_{(1)} \left(\bigotimes_{n\neq 1}^N \boldsymbol{U}_n^t\right)^\top\right\|_F^2} \\ &\geq \frac{t^2}{m\max_k \left\|\overline{\boldsymbol{\mathcal{X}}}_k^t - \boldsymbol{\mathcal{X}}_k^\star\right\|_F^2 \left\|\boldsymbol{w}\boldsymbol{z}^\top \boldsymbol{S}_{(1)}^t\right\|_F^2} \\ &\geq \frac{t^2}{1.09^2 m r_{\max}\max_k \left\|\overline{\boldsymbol{\mathcal{X}}}_k^t - \boldsymbol{\mathcal{X}}_k^\star\right\|_F^2 \sigma_{\max}^2} \\ &\geq \frac{t^2}{1.09^2\times 1.5^2 mN\delta_t^2\mu^2\sigma_{\max}^4\frac{r_{\max}^2}{K}},\end{aligned} \tag{83}$$

where the third inequality uses (79). In addition, there is also

$$\begin{aligned}&\frac{t}{\max_{ki}\psi_{ki}} \gtrsim \frac{t}{\max_k \left\|\overline{\boldsymbol{\mathcal{X}}}_k^t - \boldsymbol{\mathcal{X}}_k^\star\right\|_F \|\boldsymbol{w}\boldsymbol{z}^\top\|_F \max_k \left\|\boldsymbol{\mathcal{S}}_k^t\right\|_F} \\ \geq &\frac{t}{\max_k \left\|\overline{\boldsymbol{\mathcal{X}}}_k^t - \boldsymbol{\mathcal{X}}_k^\star\right\|_F \max_k \left(\left\|\boldsymbol{\mathcal{G}}_k^t\right\|_F + 1.2\epsilon\sqrt{1-(1-\delta_t^2)^N}\left\|\boldsymbol{\mathcal{S}}_k^\star\right\|_F + 40\sigma_v\sqrt{\frac{\prod_{n=1}^N r_n + \log(Kd_{\max})}{m(1-1.1\epsilon)}}\right)} \\ \geq &\frac{t}{1.09\mu\max_k \left\|\overline{\boldsymbol{\mathcal{X}}}_k^t - \boldsymbol{\mathcal{X}}_k^\star\right\|_F \sqrt{\frac{r_{\max}}{K}}\sigma_{\max}} \\ \geq &\frac{t}{1.09\times 1.5\sqrt{N}\delta_t\mu^2\sigma_{\max}^2\frac{r_{\max}}{K}},\end{aligned} \tag{84}$$

where the second inequality uses (71) and third equality uses the SNR condition SNR $\gtrsim \frac{\prod_{n=1}^N r_n + \log(Kd_{\max})}{m}$ and Assumption 1. The last steps in (83) and (84) is due to

$$\begin{aligned}\left\|\overline{\boldsymbol{\mathcal{X}}}_k^t - \boldsymbol{\mathcal{X}}_k^\star\right\|_F &\leq \left\|\left(\boldsymbol{P}_t^\top\boldsymbol{P}_t\right)^{-1}\boldsymbol{P}_t^\top\boldsymbol{A}_k^t\left(\boldsymbol{I} - \bigotimes_{n=1}^N\left(\boldsymbol{U}_n^t\left(\boldsymbol{U}_n^t\right)^\top\right)\right)\left(\bigotimes_{n=1}^N \boldsymbol{U}_n^\star\right)\boldsymbol{s}_k^\star\right\|_F \\ &\quad + \left\|\boldsymbol{\mathcal{G}}_k^t \times \left\{\boldsymbol{U}_n^t\right\}_{n\in[N]} - \boldsymbol{\mathcal{X}}_k^\star\right\|_F \\ &\leq 1.2\epsilon\sqrt{1-(1-\delta_t^2)^N}\left\|\boldsymbol{\mathcal{S}}_k^\star\right\|_F + \left\|\boldsymbol{\mathcal{S}}_k^\star \times \left\{\left(\boldsymbol{I}_n - \boldsymbol{U}_n^t\boldsymbol{U}_n^{t^\top}\right)\boldsymbol{U}_n^\star\right\}_{n\in[N]}\right\|_F \\ &\leq 1.2\epsilon\sqrt{1-(1-\delta_t^2)^N}\left\|\boldsymbol{\mathcal{S}}_k^\star\right\|_F + \sqrt{1-(1-\delta_t^2)^N}\left\|\boldsymbol{\mathcal{S}}_k^\star\right\|_F \\ &\leq 1.5\sqrt{N}\delta_t\left\|\boldsymbol{\mathcal{S}}_k^\star\right\|_F \\ &\leq 1.5\sqrt{N}\mu\delta_t\sigma_{\max}\sqrt{\frac{r_{\max}}{K}},\end{aligned} \tag{85}$$

where the second inequality comes from (71) and last inequality comes form Assumption 1.

Thus, if set $t := \epsilon_1 m\delta_t\sigma_{\min}^2$, then

$$\begin{aligned}&\min\left\{\frac{t^2}{1.09^2\times 1.5^2 mN\delta_t^2\mu^2\sigma_{\max}^4\frac{r_{\max}^2}{K}}, \frac{t}{1.09\times 1.5\sqrt{N}\delta_t\mu^2\sigma_{\max}^2\frac{r_{\max}}{K}}\right\} \\ &= \min\left\{\frac{\epsilon_1^2 mK}{1.09^2\times 1.5^2 N\mu^2\kappa^4 r_{\max}^2}, \frac{\epsilon_1 mK}{1.09\times 1.5\sqrt{N}\mu^2\kappa^2 r_{\max}}\right\} \\ &= \frac{\epsilon_1^2 mK}{1.09^2\times 1.5^2 N\mu^2\kappa^4 r_{\max}^2}.\end{aligned} \tag{86}$$

Using the standard epsilon-net argument to bound and union bound along all orders would obtain

$$\mathbb{P}\left\{\|\Pi_1 - \mathbb{E}\left[\Pi_1\right]\| \leq \epsilon_1 m \delta_t \sigma_{\min}^2\right\} \geq 1 - \mathcal{O}\left(\exp\left\{d_{\max} - \frac{\epsilon_1^2 mK}{N\mu^2\kappa^4 r_{\max}^2}\right\}\right) \tag{87}$$

for given core $\boldsymbol{\mathcal{S}}_k^t$ updated. The sample size condition in (12) induce the high probability event.

(ii) The concentration bound for Π_2 is similar to Π_1. The summands are also independent zero-mean, sub-exponential $r.v.s$ with sub-exponential norm as

$$\psi_{ki} \lesssim \left\|\boldsymbol{\mathcal{N}}_k^t\right\|_F \left\|\boldsymbol{w}\boldsymbol{z}^\top \left(\boldsymbol{\mathcal{S}}_k^t\right)_{(1)} \left(\bigotimes_{n\neq 1}^{N} \boldsymbol{U}_n^t\right)^\top\right\|_F. \tag{88}$$

With the similar derivation as above, we can have

$$\begin{aligned}\frac{t^2}{\sum_{ki}\psi_{ki}^2} &\gtrsim \frac{t^2}{1.09^2 m \max_k \|\mathcal{N}_k^t\|_F^2 \sigma_{\max}^2} \\ &\geq \frac{(1-1.1\epsilon)t^2}{1.09^2 \times 40^2 \sigma_v^2 \sigma_{\max}^2 \left(\prod_{n=1}^N r_n + \log(Kd_{\max})\right)}\end{aligned} \tag{89}$$

and

$$\begin{aligned}\frac{t}{\max_{ki}\psi_{ki}} &\gtrsim \frac{t}{1.09\,\mu\sigma_{\max}\sqrt{r_{\max}/K}\,\max_k \|\mathcal{N}_k^t\|_F} \\ &\geq \frac{\sqrt{1-1.1\epsilon}\,t}{1.09\times 40\,\mu\sigma_v\sigma_{\max}}\sqrt{\frac{mK}{r_{\max}\left(\prod_{n=1}^N r_n + \log(Kd_{\max})\right)}}.,\end{aligned} \tag{90}$$

where the last steps in above two inequalities are due to substituting (70). We set $t := \mathcal{O}\left(\sigma_v\sigma_{\max}\sqrt{\left(\prod_{n=1}^N r_n + \log(Kd_{\max})\right)d_{\max}}\right)$ to obtain

$$\begin{aligned}\min\left\{\frac{t^2}{\sum_{ki}\psi_{ki}^2}, \frac{t}{\max_{ki}\psi_{ki}}\right\} &\gtrsim \min\left\{\frac{1-1.1\epsilon}{1.09^2\times 40^2}d_{\max}, \frac{\sqrt{1-1.1\epsilon}}{1.09\times 40\mu}\sqrt{\frac{mKd_{\max})}{r_{\max}}}\right\} \\ &= \frac{1-1.1\epsilon}{1.09^2\times 40^2}d_{\max},\end{aligned} \tag{91}$$

where the last equality is because the sample size condition (12). Using ϵ-net argument can obtain

$$\mathbb{P}\left\{\|\Pi_2 - \mathbb{E}[\Pi_2]\| \lesssim \sigma_v\sigma_{\max}\sqrt{\left(\prod_{n=1}^N r_n + \log(Kd_{\max})\right)d_{\max}}\right\} \geq 1 - \frac{1}{5Nd_{\max}^{10}} \tag{92}$$

for given core $\boldsymbol{\mathcal{S}}_k^t$ updated.

(iii) The proof for the Π_3 also uses the sub-exponential Bernstein inequality. The summands are also independent zero-mean, sub-exponential $r.v.s$ with sub-exponential norm as

$$\psi_{ki} \lesssim \sigma_v \left\|\boldsymbol{w}\boldsymbol{z}^\top \left(\boldsymbol{\mathcal{S}}_k^t\right)_{(1)} \left(\bigotimes_{n\neq 1}^{N} \boldsymbol{U}_n^t\right)^\top\right\|_F. \tag{93}$$

Thus, $t := \mathcal{O}\left(\sigma_v\sigma_{\max}\sqrt{md_{\max}}\right)$ can induce

$$\begin{aligned}\min\left\{\frac{t^2}{\sum_{ki}\psi_{ki}^2}, \frac{t}{\max_{ki}\psi_{ki}}\right\} &\gtrsim \min\left\{\frac{t^2}{1.09^2\, m\sigma_v^2\sigma_{\max}^2}, \frac{t\sqrt{K}}{1.09\,\mu\sigma_v\sigma_{\max}\sqrt{r_{\max}}}\right\} \\ &\gtrsim \min\left\{d_{\max}, \sqrt{\frac{mKd_{\max}}{\mu^2 r_{\max}}}\right\}\end{aligned}$$

$$\gtrsim d_{\max}, \tag{94}$$

where the last equality comes from (12). Thus, the concentration bound of Π_3 is

$$\mathbb{P}\left\{ \|\Pi_3 - \mathbb{E}[\Pi_3]\| \lesssim \sigma_v \sigma_{\max} \sqrt{m d_{\max}} \right\} \geq 1 - \frac{1}{5N d_{\max}^{10}} \tag{95}$$

for given core $\boldsymbol{\mathcal{S}}_k^t$ updated.

Combining above concentration inequalities for $\Pi_i, i = 1, 2, 3$ can obtain

$$\begin{aligned} \left\|\nabla_{U_n^t} F - \mathbb{E}\left[\nabla_{U_n^t} F\right]\right\| &\leq \|\Pi_1 - \mathbb{E}\Pi_1\| + \|\Pi_2 - \mathbb{E}\Pi_2\| + \|\Pi_3 - \mathbb{E}\Pi_3\| \\ &\leq 0.01\, m\delta_t \sigma_{\min}^2 + C\sigma_v \sigma_{\max} \sqrt{m d_{\max}} \\ &= 0.01\, m\delta_t \sigma_{\min}^2 + Cm\sigma_{\min}^2 \kappa \sqrt{\frac{d_{\max}}{mK\,\mathrm{SNR}}}, \end{aligned} \tag{96}$$

where the second inequality is due to sample size condition $m \gtrsim \prod_{n=1}^N r_n + \log(K d_{\max})$ and last equality is due to the definition of SNR.

(III) Lower bound for $\sigma_{\min}\left(\hat{\boldsymbol{U}}_1^{t+1}\right)$.

$$\begin{aligned} \sigma_{\min}\left(\hat{\boldsymbol{U}}_1^{t+1}\right) &\geq \sigma_{\min}\left(\boldsymbol{U}_1^t\right) - \eta \left\|\nabla_{\boldsymbol{U}_1^t} F\right\| \\ &\geq 1 - \eta \left\|\mathbb{E}\left[\nabla_{\boldsymbol{U}_1^t} F\right] - \nabla_{\boldsymbol{U}_1^t} F\right\| - \eta \left\|\mathbb{E}\left[\nabla_{\boldsymbol{U}_1^t} F\right]\right\| \\ &\geq 1 - 0.01 m\eta\delta_t\sigma_{\min}^2 - Cm\eta\sigma_{\min}^2\kappa\sqrt{\frac{d_{\max}}{mK\,\mathrm{SNR}}} - \eta m \left\|\boldsymbol{X}_{(1)}^t - \boldsymbol{X}_{(1)}^\star\right\|_F \left\|\boldsymbol{S}_{(1)}^t\right\| \\ &\geq 1 - 0.01 m\eta\delta_t\sigma_{\min}^2 - Cm\eta\sigma_{\min}^2\kappa\sqrt{\frac{d_{\max}}{mK\,\mathrm{SNR}}} \\ &\quad - 1.09 m\eta\sigma_{\max}\left(1.5\sqrt{N r_{\max}}\,\delta_t\sigma_{\max} + C\sigma_{\min}\sqrt{\frac{\prod_n r_n + \log(K d_{\max})}{m\,\mathrm{SNR}}}\right) \\ &\geq 1 - 0.1 m\eta\sigma_{\min}^2, \end{aligned} \tag{97}$$

where the last inequality is due to $\delta_t \leq \frac{0.01}{\sqrt{N r_{\max}}\kappa^2}$ and sample size condition induces $\mathrm{SNR} \gtrsim \kappa^2 \max\left\{\frac{(\prod_n r_n + \log(K d_{\max}))}{m}, \frac{d_{\max}}{mK}\right\}$

Finally, substituting results in (I)-(III) into (75), there is

$$\begin{aligned} \delta_{t+1} &\leq \frac{\left(1 - 0.81 m\eta\sigma_{\min}^2\right)\delta_t + Cm\eta\sigma_{\min}^2\kappa\sqrt{\frac{d_{\max}}{mK\,\mathrm{SNR}}}}{1 - 0.1 m\eta\sigma_{\min}^2} \\ &\leq \left(1 - 0.5 m\eta\sigma_{\min}^2\right)\delta_t + Cm\eta\sigma_{\min}^2\kappa\sqrt{\frac{d_{\max}}{mK\,\mathrm{SNR}}} \\ &\leq \left(1 - 0.5 m\eta\sigma_{\min}^2\right)^{t+1}\delta_0 + C\kappa\sqrt{\frac{d_{\max}}{mK\,\mathrm{SNR}}} \end{aligned} \tag{98}$$

The second inequality is because $\eta = \frac{0.1}{m\sigma_{\max}^2}$. To maintain the each step in the local basin as $\delta_t \leq \frac{0.01}{\sqrt{N r_{\max}}\kappa^2}$, the SNR should satisfy $\mathrm{SNR} \gtrsim \frac{\kappa^6 N d_{\max} r_{\max}}{mK}$ in Theorem 2.

To obtain the statistical error rate for estimated parameter, we have

$$\begin{aligned} \left\|\boldsymbol{\mathcal{X}}_k^t - \boldsymbol{\mathcal{X}}_k^\star\right\|_F &= \left\|\left(\boldsymbol{\mathcal{S}}_k^t - \boldsymbol{\mathcal{G}}_k^t\right) \times \left\{\boldsymbol{U}_n^t\right\}_{n\in[N]} + \boldsymbol{\mathcal{G}}_k^t \times \left\{\boldsymbol{U}_n^t\right\}_{n\in[N]} - \boldsymbol{\mathcal{X}}_k^\star\right\|_F \\ &\lesssim \sqrt{1 - (1 - \delta_t^2)^N}\left\|\boldsymbol{\mathcal{S}}_k^\star\right\|_F + \sigma_v\sqrt{\frac{\prod_{n=1}^N r_n + \log(K d_{\max})}{m}} \end{aligned}$$

$$\lesssim \mu\sqrt{N}\sigma_{\max}\sqrt{\frac{r_{\max}}{K}}\cdot\left(\left(1-0.5m\eta\sigma_{\min}^2\right)^t\delta_0+\kappa\sqrt{\frac{d_{\max}}{mK\,\mathrm{SNR}}}\right)$$
$$+\sigma_v\sqrt{\frac{\prod_{n=1}^N r_n+\log\left(Kd_{\max}\right)}{m}}$$
$$\leq \frac{\mu\sigma_{\max}}{\sqrt{K}}\left(1-0.5m\eta\sigma_{\min}^2\right)^t+\sigma_v\left(\mu\kappa^2\sqrt{\frac{Nd_{\max}r_{\max}}{mK}}+\sqrt{\frac{\prod_{n=1}^N r_n+\log\left(Kd_{\max}\right)}{m}}\right), \tag{99}$$

where the second inequality is due to substituting (98) and Assumption 1. The last inequality uses the initialization condition $\delta_t \leq \frac{0.01}{\sqrt{Nr_{\max}}\kappa^2}$.

F Proof of Theorem 3

Based on the gradient formula in (72), the preconditioned gradient is

$$\nabla^p_{\boldsymbol{U}_1^t}F=\left(\sum_{k=1}^K\sum_{i=1}^m\left\langle\boldsymbol{\mathcal{A}}_{k,i}^t,\boldsymbol{\mathcal{X}}_k^t-\boldsymbol{\mathcal{X}}_k^\star\right\rangle\left(\boldsymbol{A}_{k,i}^t\right)_{(1)}\left(\bigotimes_{n\neq1}^N\boldsymbol{U}_n^t\right)\left(\boldsymbol{S}_k^t\right)_{(1)}^\top\right)\left(\sum_{k=1}^K\left(\boldsymbol{S}_k^t\right)_{(1)}\left(\boldsymbol{S}_k^t\right)_{(1)}^\top\right)^{-1}$$
$$-\sum_{k=1}^K\sum_{i=1}^m\xi_{k,i}\left(\boldsymbol{A}_{k,i}^t\right)_{(1)}\left(\bigotimes_{n\neq1}^N\boldsymbol{U}_n^t\right)\left(\boldsymbol{S}_k^t\right)_{(1)}^\top\left(\sum_{k=1}^K\left(\boldsymbol{S}_k^t\right)_{(1)}\left(\boldsymbol{S}_k^t\right)_{(1)}^\top\right)^{-1} \tag{100}$$

and its expectation is

$$\mathbb{E}\left[\nabla^p_{\boldsymbol{U}_1^t}F\right]=\left(m\boldsymbol{U}_1^t\boldsymbol{S}_{(1)}^t\boldsymbol{S}_{(1)}^{t^\top}-m\boldsymbol{X}_{(1)}^\star\left(\boldsymbol{I}\bigotimes_{n\neq1}^N\boldsymbol{U}_n^t\right)\boldsymbol{S}_{(1)}^{t^\top}\right)\left(\boldsymbol{S}_{(1)}^t\boldsymbol{S}_{(1)}^{t^\top}\right)^{-1}$$
$$=m\boldsymbol{U}_1^t-m\boldsymbol{X}_{(1)}^\star\left(\boldsymbol{I}\bigotimes_{n\neq1}^N\boldsymbol{U}_n^t\right)\boldsymbol{S}_{(1)}^{t^\top}\left(\boldsymbol{S}_{(1)}^t\boldsymbol{S}_{(1)}^{t^\top}\right)^{-1}. \tag{101}$$

Thus, we have

$$\delta\left(\boldsymbol{U}_1^{t+1},\boldsymbol{U}_1^\star\right)\leq\left\|\boldsymbol{U}_{1,\perp}^\star\left(\boldsymbol{U}_1^t-\eta\mathbb{E}\left[\nabla^p_{\boldsymbol{U}_1^t}F\right]+\eta\left(\mathbb{E}\left[\nabla^p_{\boldsymbol{U}_1^t}F\right]-\nabla^p_{\boldsymbol{U}_1^t}F\right)\right)\right\|\sigma_{\min}^{-1}\left(\boldsymbol{R}_1^{t+1}\right)$$
$$\leq\frac{\delta\left(\boldsymbol{U}_1^t,\boldsymbol{U}_1^\star\right)(1-\eta m)+\eta\left\|\mathbb{E}\left[\nabla^p_{\boldsymbol{U}_1^t}F\right]-\nabla^p_{\boldsymbol{U}_1^t}F\right\|}{\sigma_{\min}\left(\hat{\boldsymbol{U}}_1^{t+1}\right)}. \tag{102}$$

In addition, based on $\nabla^p_{\boldsymbol{U}_1^t}F=\nabla_{\boldsymbol{U}_1^t}F\left(\boldsymbol{S}_{(1)}^t\boldsymbol{S}_{(1)}^{t^\top}\right)^{-1}$, there is

$$\left\|\mathbb{E}\left[\nabla^p_{\boldsymbol{U}_1^t}F\right]-\nabla^p_{\boldsymbol{U}_1^t}F\right\|\leq\left\|\mathbb{E}\left[\nabla_{\boldsymbol{U}_1^t}F\right]-\nabla_{\boldsymbol{U}_1^t}F\right\|\left\|\left(\boldsymbol{S}_{(1)}^t\boldsymbol{S}_{(1)}^{t^\top}\right)^{-1}\right\|$$
$$\leq\left(0.01\,m\delta_t\sigma_{\min}^2+Cm\sigma_{\min}^2\kappa\sqrt{\frac{d_{\max}}{mK\,\mathrm{SNR}}}\right)(0.91\sigma_{\min})^{-2}$$
$$\leq 0.02\,m\delta_t+2Cm\kappa\sqrt{\frac{d_{\max}}{mK\,\mathrm{SNR}}}, \tag{103}$$

and

$$\left\|\mathbb{E}\left[\nabla^p_{\boldsymbol{U}_1^t}F\right]\right\|\leq\left\|\mathbb{E}\left[\nabla_{\boldsymbol{U}_1^t}F\right]\right\|\left\|\left(\boldsymbol{S}_{(1)}^t\boldsymbol{S}_{(1)}^{t^\top}\right)^{-1}\right\|$$
$$\leq 1.09m\sigma_{\max}\left(1.5\sqrt{Nr_{\max}}\,\delta_t\sigma_{\max}+C\sigma_{\min}\sqrt{\frac{\prod_n r_n+\log(Kd_{\max})}{m\,\mathrm{SNR}}}\right)(0.91\sigma_{\min})^{-2}$$

$$\leq m\left(1.98\sqrt{Nr}\,\kappa^2\delta_t + C\kappa\sqrt{\frac{\prod_n r_n + \log(Kd_{\max})}{m\,\mathrm{SNR}}}\right), \tag{104}$$

where the second inequalities in above results uses the (96) and (78). Thus, the lower bound for $\sigma_{\min}\left(\hat{\boldsymbol{U}}_1^{t+1}\right)$ is

$$\begin{aligned}\sigma_{\min}\left(\hat{\boldsymbol{U}}_1^{t+1}\right) &\geq 1-\eta\left(\left\|\mathbb{E}\left[\nabla^p_{\boldsymbol{U}_1^t}F\right]-\nabla^p_{\boldsymbol{U}_1^t}F\right\|+\left\|\mathbb{E}\left[\nabla^p_{\boldsymbol{U}_1^t}F\right]\right\|\right)\\ &\geq 1-m\eta\left(0.02\delta_t+2C\kappa\sqrt{\frac{d_{\max}}{mK\,\mathrm{SNR}}}+1.98\sqrt{Nr}\,\kappa^2\delta_t+C\kappa\sqrt{\frac{\prod_n r_n+\log(Kd_{\max})}{m\,\mathrm{SNR}}}\right)\\ &\geq 1-0.1m\eta, \end{aligned}\tag{105}$$

where the last inequality is also due to $\delta_t \leq \frac{0.01}{\sqrt{Nr_{\max}}\kappa^2}$ and sample size condition induces $\mathrm{SNR} \gtrsim \kappa^2\max\left\{\frac{(\prod_n r_n+\log(Kd_{\max}))}{m}, \frac{d_{\max}}{mK}\right\}$.

Finally, substituting (103), (104) and (105) into (102) can obtain

$$\begin{aligned}\delta_{t+1} &\leq \frac{(1-0.81m\eta)\,\delta_t + Cm\eta\kappa\sqrt{\frac{d_{\max}}{mK\,\mathrm{SNR}}}}{1-0.1m\eta}\\ &\leq (1-0.5m\eta)\,\delta_t + Cm\eta\kappa\sqrt{\frac{d_{\max}}{mK\mathrm{SNR}}}\\ &\leq (1-0.5m\eta)^{t+1}\,\delta_0 + C\kappa\sqrt{\frac{d_{\max}}{mK\,\mathrm{SNR}}}, \end{aligned}\tag{106}$$

where we use the step size condition $\eta = \frac{0.1}{m}$ and SNR condition in (12) to ensure the iterates in local basin to maintain $\delta_t \leq \frac{0.01}{\sqrt{Nr_{\max}}\kappa^2}$.

G Proof of Theorem 4

The proof idea is reducing estimation to testing on several finite subsets of $\mathcal{F}$. The guiding implication is that an estimator with sufficiently small error would identify the correct candidate parameter. We construct candidates that are separated in parameter distance but whose observation laws cannot be reliably distinguished. This forces a nonzero probability of a large estimation error for every estimator.

Fix an arbitrary estimator $\hat{\boldsymbol{\mathcal{X}}}$ and parameter $\boldsymbol{\mathcal{X}} = (\boldsymbol{\mathcal{X}}_1, \ldots, \boldsymbol{\mathcal{X}}_K)$, we denote two distances $d_\Sigma^2(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}}') = \sum_{k=1}^K \|\boldsymbol{\mathcal{X}}_k - \boldsymbol{\mathcal{X}}'_k\|_F^2, d_\infty(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}}') = \max_{k\in[K]}\|\boldsymbol{\mathcal{X}}_k - \boldsymbol{\mathcal{X}}'_k\|_F$. It is apparent that there is $d_\Sigma^2 \leq Kd_\infty^2$. We first translate parameter separation to a testing lower bound problem.

Let $\mathbb{P}_{\boldsymbol{\mathcal{X}}}$ be the joint distribution of all mK input-output pairs and $\boldsymbol{\mathcal{A}}$ denotes $\{\boldsymbol{\mathcal{A}}_{k,i}\}_{k\in[K],i\in[m]}$. Conditional on Gaussian distribution and independence give

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_{\boldsymbol{\mathcal{X}}}\|\mathbb{P}_{\boldsymbol{\mathcal{X}}'}) &= \mathbb{E}_{\boldsymbol{\mathcal{A}}}\,\mathrm{KL}\big(\mathbb{P}_{\boldsymbol{\mathcal{X}}}(\cdot\mid\boldsymbol{\mathcal{A}})\|\mathbb{P}_{\boldsymbol{\mathcal{X}}'}(\cdot\mid\boldsymbol{\mathcal{A}})\big)\\ &= \frac{1}{2\sigma_v^2}\sum_{k=1}^K\sum_{i=1}^M \mathbb{E}\langle\boldsymbol{\mathcal{A}}_{k,i}, \boldsymbol{\mathcal{X}}_k - \boldsymbol{\mathcal{X}}'_k\rangle^2\\ &= \frac{M}{2\sigma_v^2}d_\Sigma^2(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}}'). \end{aligned}\tag{107}$$

Consider the subset $\{\boldsymbol{\mathcal{X}}^\theta : \theta \in \{-1,1\}^p\} \subset \mathcal{F}$, $p \geq 1$, satisfying

$$d_\Sigma^2(\boldsymbol{\mathcal{X}}^\theta, \boldsymbol{\mathcal{X}}^{\theta'}) \geq 4a^2 d_H(\theta,\theta'), \forall\theta,\theta' \quad \text{and} \quad \mathrm{KL}(\mathbb{P}_{\boldsymbol{\mathcal{X}}^\theta}\|\mathbb{P}_{\boldsymbol{\mathcal{X}}^{\theta'}}) \leq \frac{1}{16} \qquad \text{if } d_H(\theta,\theta') = 1, \tag{108}$$

where $a := \frac{\sigma_v}{8\sqrt{m}}, d_H := \sum_{j=1}^{p} 1_{\{\theta_j \neq \theta'_j\}}$ is Hamming distance. We will construct set that can be verified to satisfy (108).

Choose a nearest candidate in $\mathcal{F}$:

$$\hat{\theta} \in \arg\min_{\vartheta \in \{-1,1\}^p} d_\Sigma(\hat{\boldsymbol{\mathcal{X}}}, \boldsymbol{\mathcal{X}}^\vartheta). \tag{109}$$

By the triangle inequality and optimality of nearest candidate, there is

$$\begin{aligned} d_\Sigma(\boldsymbol{\mathcal{X}}^{\hat{\theta}}, \boldsymbol{\mathcal{X}}^\theta) &\leq d_\Sigma(\boldsymbol{\mathcal{X}}^{\hat{\theta}}, \hat{\boldsymbol{\mathcal{X}}}) + d_\Sigma(\hat{\boldsymbol{\mathcal{X}}}, \boldsymbol{\mathcal{X}}^\theta) \leq 2d_\Sigma(\hat{\boldsymbol{\mathcal{X}}}, \boldsymbol{\mathcal{X}}^\theta), \\ K d_\infty^2(\hat{\boldsymbol{\mathcal{X}}}, \boldsymbol{\mathcal{X}}^\theta) &\geq d_\Sigma^2(\hat{\boldsymbol{\mathcal{X}}}, \boldsymbol{\mathcal{X}}^\theta) \geq \frac{1}{4} d_\Sigma^2(\mathcal{X}^{\hat{\theta}}, \boldsymbol{\mathcal{X}}^\theta) \geq a^2 d_H(\hat{\theta}, \theta). \end{aligned} \tag{110}$$

Thus many incorrect signs force a large tensor estimation error. Write $\mathbb{P}_\theta = \mathbb{P}_{\boldsymbol{\mathcal{X}}^\theta}$ and let $\theta^{(j)}$ be obtained by flipping sign j. For two laws with densities f, g, write $\mathrm{TV}(P, Q) = \frac{1}{2}\int |f - g|$. Pinsker's inequality and the neighboring-KL condition imply

$$\begin{aligned} &\mathrm{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta^{(j)}}) \leq \sqrt{\frac{1}{2}\,\mathrm{KL}(\mathbb{P}_\theta \| \mathbb{P}_{\theta^{(j)}})} \leq \frac{1}{4}, \\ &\mathbb{P}_\theta(\hat{\theta}_j \neq \theta_j) + \mathbb{P}_{\theta^{(j)}}(\hat{\theta}_j \neq \theta_j^{(j)}) \geq 1 - \mathrm{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta^{(j)}}) \geq \frac{3}{4}. \end{aligned} \tag{111}$$

Now, we can calculate the accumulated the Hamming errors. There are 2^{p-1} unordered neighboring pairs for each sign, so averaging over the signs and the data gives

$$\mathbb{E} d_H(\hat{\theta}, \theta) = \sum_{j=1}^{p} \frac{1}{2^p} \sum_{\theta} \mathbb{P}_\theta(\hat{\theta}_j \neq \theta_j) \geq p\, \frac{2^{p-1}}{2^p}\, \frac{3}{4} = \frac{3p}{8}. \tag{112}$$

Since $0 \leq d_H(\hat{\theta}, \theta) \leq p$, letting $q = \mathbb{P}\{d_H(\hat{\theta}, \theta) \geq p/8\}$ yields

$$\frac{3p}{8} \leq \mathbb{E} d_H(\hat{\theta}, \theta) \leq \frac{p}{8}(1 - q) + pq, \qquad q \geq \frac{2}{7}. \tag{113}$$

Finally, we can return to the maximum-task estimation loss. By (110), the event $d_H(\hat{\theta}, \theta) \geq p/8$ implies $d_\infty(\hat{\mathcal{X}}, \mathcal{X}^\theta) \geq a\sqrt{p/(8K)}$. The worst-case error probability is no smaller than its uniform average. Thus, there is

$$\sup_{\boldsymbol{\mathcal{X}} \in \mathcal{F}} \mathbb{P}_{\boldsymbol{\mathcal{X}}} \left\{ d_\infty(\hat{\boldsymbol{\mathcal{X}}}, \boldsymbol{\mathcal{X}}) \geq a\sqrt{\frac{p}{8K}} \right\} \geq \frac{1}{2^p} \sum_{\theta} \mathbb{P}_\theta\{d_H(\hat{\theta}, \theta) \geq p/8\} = q \geq \frac{1}{8}. \tag{114}$$

For core perturbations $\{\boldsymbol{\mathcal{E}}_k\}$, set

$$\tilde{\boldsymbol{\mathcal{X}}}_k = (\boldsymbol{\mathcal{S}}_k^\star + \boldsymbol{\mathcal{E}}_k) \times_1 \boldsymbol{U}_1^\star \cdots \times_N \boldsymbol{U}_N^\star; \qquad \boldsymbol{E}_{(n)} = [(\boldsymbol{E}_1)_{(n)}, \ldots, (\boldsymbol{E}_K)_{(n)}], \tag{115}$$

We verify that any perturbation satisfying

$$\varepsilon := \left(\sum_k \|\boldsymbol{\mathcal{E}}_k\|_F^2 \right)^{1/2} \leq \frac{\sigma_{\min}}{8}, \qquad \max_k \|\boldsymbol{\mathcal{E}}_k\|_F \leq \frac{\sigma_{\min}}{8\sqrt{K}}. \tag{116}$$

can guarantee $\tilde{\boldsymbol{\mathcal{X}}} \in \mathcal{F}$. Besides, $\forall n \in [N]$,

$$\begin{aligned} &\|\boldsymbol{E}_{(n)}\| \leq \|\boldsymbol{E}_{(n)}\|_F = \left(\sum_k \|\boldsymbol{\mathcal{E}}_k\|_F^2 \right)^{1/2} = \varepsilon \leq \frac{\sigma_{\min}}{8}, \\ &\sigma_{r_n}(\boldsymbol{S}_{(n)}^\star + \boldsymbol{E}_{(n)}) \geq \sigma_{r_n}(\boldsymbol{S}_{(n)}^\star) - \|\boldsymbol{E}_{(n)}\| \geq \frac{7}{8}\sigma_{\min} > 0. \end{aligned} \tag{117}$$

The last two lines use Weyl's inequality. The stacked perturbed core has r_n rows and positive r_n-th singular value. Orthonormal embedding then implies $\mathrm{rank}(\tilde{\mathbf{X}}_{(n)}) = \mathrm{rank}(\boldsymbol{S}_{(n)}^\star + \boldsymbol{E}_{(n)}) = r_n$.

Writing $\tilde{\lambda} = \sigma_{\min}(\tilde{\boldsymbol{\mathcal{X}}})$ and $\tilde{\Lambda} = \sigma_{\max}(\tilde{\boldsymbol{\mathcal{X}}})$, Weyl's inequality also gives

$$\begin{aligned}
\tilde{\lambda} &\geq \sigma_{\min} - \varepsilon \geq \frac{7}{8}\sigma_{\min} > \frac{1}{2}\sigma_{\min}, \\
\tilde{\Lambda} &\leq \sigma_{\max} + \varepsilon \leq \frac{9}{8}\sigma_{\max} < 2\sigma_{\max}, \\
\tilde{\Lambda} &\geq \sigma_{\max} - \varepsilon \geq \frac{7}{8}\sigma_{\max}, \\
\frac{\tilde{\Lambda}}{\tilde{\lambda}} &\leq \frac{(9/8)\sigma_{\max}}{(7/8)\sigma_{\min}} < 2\kappa.
\end{aligned} \tag{118}$$

In addition, there is

$$\max_k \|\tilde{\boldsymbol{\mathcal{X}}}_k\|_F = \max_k \|\boldsymbol{\mathcal{S}}_k^\star + \boldsymbol{\mathcal{E}}_k\|_F \leq \mu\sigma_{\max}\sqrt{\frac{r_{\max}}{K}} + \frac{\sigma_{\min}}{8\sqrt{K}} \leq \frac{9}{8}\mu\sigma_{\max}\sqrt{\frac{r_{\max}}{K}} < 2\mu\tilde{\Lambda}\sqrt{\frac{r_{\max}}{K}}. \tag{119}$$

There are total $K\prod_n r_n$ core entries by $\theta \in \{-1,1\}^{K\prod_n r_n}$ and set

$$(\boldsymbol{\mathcal{E}}_k^\theta)_{j_1,\ldots,j_N} = a\theta_{k,j_1,\ldots,j_N}, \quad \boldsymbol{\mathcal{X}}_k^\theta = (\boldsymbol{\mathcal{S}}_k^\star + \boldsymbol{\mathcal{E}}_k^\theta) \times_1 \boldsymbol{U}_1^\star \cdots \times_N \boldsymbol{U}_N^\star. \tag{120}$$

Based on SNR condition in Theorem 4, there are

$$\left(\sum_k \|\boldsymbol{\mathcal{E}}_k^\theta\|_F^2\right)^{1/2} = a\sqrt{KR} = \frac{\sigma_v\sqrt{KR}}{8\sqrt{m}} \leq \frac{\sigma_{\min}}{8}, \quad \max_k \|\boldsymbol{\mathcal{E}}_k^\theta\|_F = a\sqrt{R} \leq \frac{\sigma_{\min}}{8\sqrt{K}}. \tag{121}$$

Thus $\boldsymbol{\mathcal{X}}^\theta \in \mathcal{F}$. Orthonormality yields

$$d_\Sigma^2(\boldsymbol{\mathcal{X}}^\theta, \boldsymbol{\mathcal{X}}^{\theta'}) = \sum_k \|\boldsymbol{\mathcal{E}}_k^\theta - \boldsymbol{\mathcal{E}}_k^{\theta'}\|_F^2 = 4a^2 d_H(\theta,\theta'). \tag{122}$$

For $d_H(\theta,\theta') = 1$,

$$\mathrm{KL}(\mathbb{P}_{\boldsymbol{\mathcal{X}}^\theta} \| \mathbb{P}_{\boldsymbol{\mathcal{X}}^{\theta'}}) = \frac{m}{2\sigma_v^2}\, 4a^2 < \frac{1}{16}, \tag{123}$$

where the last inequality is due to definition of $a := \frac{\sigma_v}{8\sqrt{m}}$. Applying (114) with $p = K\prod_n r_n$ gives

$$\sup_{\boldsymbol{\mathcal{X}}\in\mathcal{F}} \mathbb{P}_{\boldsymbol{\mathcal{X}}}\left\{ d_\infty(\hat{\boldsymbol{\mathcal{X}}}, \boldsymbol{\mathcal{X}}) \geq \frac{\sigma_v}{8\sqrt{8m}}\sqrt{R} \right\} \geq \frac{1}{8}. \tag{124}$$

We now further construct the perturbation set along the shared subspace. By orthogonal invariance of the designs, norms and singular values, take $\boldsymbol{U}_n^\star = [\boldsymbol{I}_{r_n}; 0]$. Define $\boldsymbol{H}_n = \boldsymbol{S}_{(n)}^\star (\boldsymbol{S}_{(n)}^\star)^\top \succeq \sigma_{\min}^2 \boldsymbol{I}_{r_n}$. For $\theta \in \{-1,1\}^{\sum_n r_n(d_n - r_n)}$, let

$$\boldsymbol{C}_n^\theta \in \mathbb{R}^{(d_n - r_n)\times r_n},\ (\boldsymbol{C}_n^\theta)_{ij} = a\theta_{n,i,j},\ \boldsymbol{B}_n^\theta = \boldsymbol{C}_n^\theta H_n^{-1/2},\ \boldsymbol{F}_n^\theta = \left[\boldsymbol{I}_{r_n} \boldsymbol{B}_n^\theta\right],\ \boldsymbol{\mathcal{X}}_k^\theta = \boldsymbol{\mathcal{S}}_k^\star \times \{\boldsymbol{F}_n^\theta\}. \tag{125}$$

By SNR condition in Theorem 4, there is

$$\gamma_\theta := \sum_n \|\boldsymbol{B}_n^\theta\|^2 \leq \sum_n \|\boldsymbol{C}_n^\theta\|_F^2 \|\boldsymbol{H}_n^{-1/2}\|^2 \leq \frac{a^2 P}{\sigma_{\min}^2} = \frac{\sigma_v^2 P}{64 M \sigma_{\min}^2} \leq \frac{1}{64}, \tag{126}$$

and

$$\begin{aligned}
(\boldsymbol{F}_n^\theta)^\top \boldsymbol{F}_n^\theta &= \boldsymbol{I}_{r_n} + (\boldsymbol{B}_n^\theta)^\top \boldsymbol{B}_n^\theta \succeq \boldsymbol{I}_{r_n}, \\
\prod_n \|\boldsymbol{F}_n^\theta\|^2 &= \prod_n (1 + \|\boldsymbol{B}_n^\theta\|^2) \leq \exp\left(\sum_n \|\boldsymbol{B}_n^\theta\|^2\right) = e^{\gamma_\theta} \leq e^{1/64}.
\end{aligned} \tag{127}$$

Write $\boldsymbol{F}_n^\theta = \boldsymbol{Q}_n^\theta \boldsymbol{D}_n^\theta$, where $\boldsymbol{D}_n^\theta = \left(\boldsymbol{I}_{r_n} + (\boldsymbol{B}_n^\theta)^\top \boldsymbol{B}_n^\theta\right)^{1/2} \succeq \boldsymbol{I}_{r_n}, (\boldsymbol{Q}_n^\theta)^\top \boldsymbol{Q}_n^\theta = I_{r_n}$. The unfolding and its orthonormal representation are

$$\mathbf{X}_{(n)}^\theta = \boldsymbol{F}_n^\theta \boldsymbol{S}_{(n)}^\star \left(\boldsymbol{I}_K \bigotimes_{j\neq n} \boldsymbol{F}_j^\theta\right)^\top = \boldsymbol{Q}_n^\theta \left[\boldsymbol{D}_n^\theta \boldsymbol{S}_{(n)}^\star \left(\boldsymbol{I}_K \bigotimes_{j\neq n} \boldsymbol{D}_j^\theta\right)^\top\right] \left(\boldsymbol{I}_K \otimes \bigotimes_{j\neq n} \boldsymbol{Q}_j^\theta\right)^\top. \tag{128}$$

Since all $\boldsymbol{D}_j^\theta$ are invertible with minimum singular value at least one, $\operatorname{rank}(\boldsymbol{X}_{(n)}^\theta) = \operatorname{rank}(\boldsymbol{S}_{(n)}^\star) = r_n$ and

$$\sigma_{r_n}(\boldsymbol{X}_{(n)}^\theta) \geq \sigma_{r_n}(\boldsymbol{D}_n^\theta)\, \sigma_{r_n}(\boldsymbol{S}_{(n)}^\star) \prod_{j \neq n} \sigma_{r_j}(\boldsymbol{D}_j^\theta) \geq \sigma_{r_n}(\boldsymbol{S}_{(n)}^\star) \geq \sigma_{\min}. \tag{129}$$

Let $\lambda_\theta = \sigma_{\min}(\boldsymbol{\mathcal{X}}^\theta)$ and $\Lambda_\theta = \sigma_{\max}(\boldsymbol{\mathcal{X}}^\theta)$. Then

$$\lambda_\theta \geq \sigma_{\min} > \sigma_{\min}/2,\ \Lambda_\theta \leq \sigma_{\max} \prod_j \|\boldsymbol{F}_j^\theta\| \leq e^{\gamma_\theta/2}\sigma_{\max} < 2\sigma_{\max},\ \frac{\Lambda_\theta}{\lambda_\theta} \leq e^{\gamma_\theta/2}\frac{\sigma_{\max}}{\sigma_{\min}} < 2\kappa. \tag{130}$$

Restricting every mode to its first r_n coordinates recovers the cores, so $S_{(n)}^\star$ is a submatrix of $\mathbf{X}_{(n)}^\theta$. Hence $\Lambda_\theta = \max_n \|\boldsymbol{X}_{(n)}^\theta\| \geq \max_n \|\boldsymbol{S}_{(n)}^\star\| = \sigma_{\max}$ and

$$\max_k \|\boldsymbol{\mathcal{X}}_k^\theta\|_F \leq \left(\prod_n \|\boldsymbol{F}_n^\theta\|\right) \max_k \|\boldsymbol{\mathcal{S}}_k^\star\|_F \leq e^{\gamma_\theta/2}\mu\sigma_{\max}\sqrt{\frac{r_{\max}}{K}} < 2\mu\Lambda_\theta\sqrt{\frac{r_{\max}}{K}}. \tag{131}$$

Therefore $\boldsymbol{\mathcal{X}}^\theta \in \mathcal{F}$. For arbitrary θ, θ', put $\Delta\boldsymbol{B}_n = \boldsymbol{B}_n^\theta - \boldsymbol{B}_n^{\theta'}$. The coordinate blocks

$$\mathcal{I}_n = [r_1] \times \cdots \times [r_{n-1}] \times \{r_n + 1, \ldots, d_n\} \times [r_{n+1}] \times \cdots \times [r_N]$$

are pairwise disjoint. Keeping their contributions gives

$$\begin{aligned} d_\Sigma^2(\boldsymbol{\mathcal{X}}^\theta, \boldsymbol{\mathcal{X}}^{\theta'}) &\geq \sum_n \sum_k \|\boldsymbol{\mathcal{S}}_k^\star \times_n \Delta\boldsymbol{B}_n\|_F^2 \\ &= \sum_n \|\Delta\boldsymbol{B}_n \boldsymbol{S}_{(n)}^\star\|_F^2 \\ &= \sum_n \operatorname{tr}(\Delta\boldsymbol{B}_n \boldsymbol{H}_n \Delta\boldsymbol{B}_n^\top) \\ &= \sum_n \|\Delta\boldsymbol{B}_n \boldsymbol{H}_n^{1/2}\|_F^2 \\ &= \sum_n \|\boldsymbol{C}_n^\theta - \boldsymbol{C}_n^{\theta'}\|_F^2 \\ &= 4a^2 d_H(\theta, \theta'). \end{aligned} \tag{132}$$

If only one sign in mode n differs, then $\boldsymbol{F}_j^\theta = \boldsymbol{F}_j^{\theta'}$ for $j \neq n$. With $\Delta\boldsymbol{F}_n = [\mathbf{0}; \Delta\boldsymbol{B}_n]$, multilinear property gives

$$\boldsymbol{\mathcal{X}}_k^\theta - \boldsymbol{\mathcal{X}}_k^{\theta'} = \boldsymbol{\mathcal{S}}_k^\star \times_n \Delta\boldsymbol{F}_n \underset{j \neq n}{\times} \boldsymbol{F}_j^\theta. \tag{133}$$

Consequently, there are

$$\begin{aligned} d_\Sigma^2(\boldsymbol{\mathcal{X}}^\theta, \boldsymbol{\mathcal{X}}^{\theta'}) &\leq \left(\prod_{j \neq n} \|\boldsymbol{F}_j^\theta\|^2\right) \sum_k \|\boldsymbol{\mathcal{S}}_k^\star \times_n \Delta\boldsymbol{F}_n\|_F^2 \\ &= \left(\prod_{j \neq n} \|\boldsymbol{F}_j^\theta\|^2\right) \|\boldsymbol{C}_n^\theta - \boldsymbol{C}_n^{\theta'}\|_F^2 \\ &\leq 4a^2 e^{\gamma_\theta} \\ &\leq 4a^2 e^{1/64}, \\ \mathrm{KL}(\mathbb{P}_{\boldsymbol{\mathcal{X}}^\theta} \| \mathbb{P}_{\boldsymbol{\mathcal{X}}^{\theta'}}) &\leq \frac{m}{2\sigma_v^2} 4a^2 e^{1/64} = \frac{e^{1/64}}{32} < \frac{1}{16}. \end{aligned} \tag{134}$$

Applying (114) with $p = \sum_n r_n(d_n - r_n)$,

$$\sup_{\boldsymbol{\mathcal{X}} \in \mathcal{F}} \mathbb{P}_{\boldsymbol{\mathcal{X}}} \left\{ d_\infty(\hat{\boldsymbol{\mathcal{X}}}, \boldsymbol{\mathcal{X}}) \geq \frac{\sigma_v}{8\sqrt{8m}} \sqrt{\frac{\sum_n r_n(d_n - r_n)}{K}} \right\} \geq \frac{1}{8}. \tag{135}$$

For $\sum_n r_n(d_n - r_n) = 0$, this construction is unnecessary. In either case, with $D_0 = \prod_n r_n + \sum_n r_n(d_n - r_n)/K$,

$$\max\{\prod_n r_n, \sum_n r_n(d_n - r_n)/K\} \geq \frac{D_0}{2}, \ \frac{\sigma_v}{8\sqrt{8m}}\sqrt{\max\{\prod_n r_n, \sum_n r_n(d_n - r_n)/K\}} \geq \frac{\sigma_v\sqrt{D_0}}{32\sqrt{m}}. \tag{136}$$

Choosing the larger contribution in (124) and (135) therefore yields

$$\sup_{\boldsymbol{\mathcal{X}} \in \mathcal{F}} \mathbb{P}_{\boldsymbol{\mathcal{X}}} \left\{ d_\infty(\hat{\boldsymbol{\mathcal{X}}}, \boldsymbol{\mathcal{X}}) \geq \frac{\sigma_v}{32\sqrt{m}} \sqrt{\prod_n r_n + \sum_n r_n(d_n - r_n)/K} \right\} \geq \frac{1}{8}. \tag{137}$$

Choose $\boldsymbol{\mathcal{E}} \in \mathbb{R}^{r_1 \times \cdots \times r_N}$ with $\|\boldsymbol{\mathcal{E}}\|_F = 1$ and set $\boldsymbol{\mathcal{Z}} = \boldsymbol{\mathcal{E}} \times_1 \boldsymbol{U}_1^\star \cdots \times_N \boldsymbol{U}_N^\star, \qquad \|\boldsymbol{\mathcal{Z}}\|_F = 1$. For $K \geq 16$, define

$$h_K = a\sqrt{\log K}, \qquad \boldsymbol{\mathcal{X}}_k^{[j]} = \boldsymbol{\mathcal{X}}_k^\star + h_K \boldsymbol{\mathcal{Z}} \, \mathbf{1}_{\{k=j\}}, \qquad j \in [K]. \tag{138}$$

By SNR condition in Theorem 4,

$$h_K^2 = \frac{\sigma_v^2 \log K}{64m} \leq \frac{\sigma_{\min}^2}{64K} \leq \frac{\sigma_{\min}^2}{64}. \tag{139}$$

Both the total and the maximum per-task core perturbation norms equal h_K, so (116) implies $\boldsymbol{\mathcal{X}}^{[j]} \in \mathcal{F}$. For $j \neq \ell$,

$$d_\infty(\boldsymbol{\mathcal{X}}^{[j]}, \boldsymbol{\mathcal{X}}^{[\ell]}) = h_K, \ d_\Sigma^2(\boldsymbol{\mathcal{X}}^{[j]}, \boldsymbol{\mathcal{X}}^\star) = h_K^2, \ \mathrm{KL}(\mathbb{P}_{\boldsymbol{\mathcal{X}}^{[j]}} \| \mathbb{P}_{\boldsymbol{\mathcal{X}}^\star}) = \frac{m h_K^2}{2\sigma_v^2} = \frac{\log K}{128}. \tag{140}$$

Let J be uniform on $[K]$, let W denote the observations, and write $Q_j = \mathbb{P}_{\boldsymbol{\mathcal{X}}^{[j]}}, Q_0 = \mathbb{P}_{\boldsymbol{\mathcal{X}}^\star}, \overline{Q} = \frac{1}{K}\sum_{j=1}^K Q_j$. For the experiment $W \mid \{J = j\} \sim Q_j$,

$$I(J; W) = \frac{1}{K}\sum_j \mathrm{KL}(Q_j \| \overline{Q}) = \frac{1}{K}\sum_j \mathrm{KL}(Q_j \| Q_0) - \mathrm{KL}(\overline{Q} \| Q_0) \leq \frac{\log K}{128}. \tag{141}$$

For any $\hat{J}$, put $p_e = \mathbb{P}(\hat{J} \neq J)$. The conditional entropy bound gives

$$\begin{aligned} &H(J \mid W) \leq H(J \mid \hat{J}) \leq \log K, \\ &p_e \geq 1 - \frac{I(J; W) + \log 2}{\log K} \geq 1 - \frac{1}{128} - \frac{\log 2}{\log K}\frac{1}{2}. \end{aligned} \tag{142}$$

Take $\hat{J}$ to be a nearest candidate to $\hat{\boldsymbol{\mathcal{X}}}$ in d_∞, with fixed tie breaking. By the separation in (140),

$$\{\hat{J} \neq j\} \subseteq \{d_\infty(\hat{\boldsymbol{\mathcal{X}}}, \boldsymbol{\mathcal{X}}^{[j]}) \geq h_K/2\}. \tag{143}$$

Hence

$$\sup_{\boldsymbol{\mathcal{X}} \in \mathcal{F}} \mathbb{P}_{\boldsymbol{\mathcal{X}}} \left\{ d_\infty(\hat{\boldsymbol{\mathcal{X}}}, \boldsymbol{\mathcal{X}}) \geq \frac{\sigma_v}{16\sqrt{m}} \sqrt{\log K} \right\} \geq \frac{1}{K}\sum_j Q_j(\hat{J} \neq j) = p_e \geq \frac{1}{2}. \tag{144}$$

For any $K \geq 1$ and $0 < \alpha \leq 1/32$, set $L_\alpha = \log \frac{1}{8\alpha}, h_\alpha = a\sqrt{L_\alpha}, \boldsymbol{\mathcal{X}}_k^\pm = \boldsymbol{\mathcal{X}}_k^\star \pm h_\alpha \boldsymbol{\mathcal{Z}} \, \mathbf{1}_{\{k=1\}}$. Since $L_\alpha \leq \log(K/\alpha), h_\alpha^2 = \frac{\sigma_v^2 L_\alpha}{64m} \leq \frac{\sigma_{\min}^2}{64K} \leq \frac{\sigma_{\min}^2}{64}$, and (116) again implies $\boldsymbol{\mathcal{X}}^\pm \in \mathcal{F}$. Write $\mathbb{P}_\pm = \mathbb{P}_{\mathcal{X}^\pm}$. Then

$$d_\infty(\boldsymbol{\mathcal{X}}^+, \boldsymbol{\mathcal{X}}^-) = 2h_\alpha, \ d_\Sigma^2(\boldsymbol{\mathcal{X}}^+, \boldsymbol{\mathcal{X}}^-) = 4h_\alpha^2, \ \mathrm{KL}(\mathbb{P}_+ \| \mathbb{P}_-) = \frac{m}{2\sigma_v^2} 4h_\alpha^2 = \frac{L_\alpha}{32}.. \tag{145}$$

Let f_+, f_- be their densities relative to a common dominating measure ν, and put $\rho = \int \sqrt{f_+ f_-}\, d\nu$. Jensen's inequality and Cauchy–Schwarz yield

$$\log \rho = \log \mathbb{E}_+ \sqrt{f_-/f_+} \geq \mathbb{E}_+ \log \sqrt{f_-/f_+} = -\tfrac{1}{2}\,\mathrm{KL}(\mathbb{P}_+ \| \mathbb{P}_-),$$

$$\begin{aligned}\mathrm{TV}(\mathbb{P}_+,\mathbb{P}_-) &= \frac{1}{2}\int |\sqrt{f_+}-\sqrt{f_-}|(\sqrt{f_+}+\sqrt{f_-})\,d\nu \\ &\le \frac{1}{2}\left[\int(\sqrt{f_+}-\sqrt{f_-})^2 d\nu \int(\sqrt{f_+}+\sqrt{f_-})^2 d\nu\right]^{1/2} \\ &= \sqrt{1-\rho^2}. \end{aligned} \tag{146}$$

Therefore, for every binary test $\phi \in \{+,-\}$,

$$\mathbb{P}_+(\phi=-)+\mathbb{P}_-(\phi=+) \ge 1-\mathrm{TV}(\mathbb{P}_+,\mathbb{P}_-) \ge 1-\sqrt{1-\rho^2} \ge \frac{1}{2}\exp\{-\mathrm{KL}(\mathbb{P}_+\|\mathbb{P}_-)\} = \frac{1}{2}e^{-L_\alpha/32}. \tag{147}$$

Let $\hat{\phi}$ choose the nearer of $\boldsymbol{\mathcal{X}}^+$ and $\boldsymbol{\mathcal{X}}^-$ to $\hat{\boldsymbol{\mathcal{X}}}$ in d_∞. Since their distance is $2h_\alpha$,

$$\sup_{\boldsymbol{\mathcal{X}}\in\mathcal{F}} \mathbb{P}_{\boldsymbol{\mathcal{X}}}\{d_\infty(\hat{\boldsymbol{\mathcal{X}}},\boldsymbol{\mathcal{X}}) \ge h_\alpha\} \ge \frac{1}{2}\sum_{s\in\{+,-\}} \mathbb{P}_s\{d_\infty(\hat{\boldsymbol{\mathcal{X}}},\boldsymbol{\mathcal{X}}^s) \ge h_\alpha\} \ge \frac{1}{2}\big[\mathbb{P}_+(\hat{\phi}=-)+\mathbb{P}_-(\hat{\phi}=+)\big] \ge 2\alpha, \tag{148}$$

Finally, define the nonincreasing function

$$\Psi(t) = \sup_{\boldsymbol{\mathcal{X}}\in\mathcal{F}} \mathbb{P}_{\boldsymbol{\mathcal{X}}}\{d_\infty(\hat{\boldsymbol{\mathcal{X}}},\boldsymbol{\mathcal{X}}) \ge t\}.$$

Since $L_\alpha \ge \log 4$, $\log(1/\alpha) = L_\alpha + \log 8 \le \frac{5}{2}L_\alpha$. Set

$$T = \begin{cases} \max\{\prod_n r_n + \sum_n r_n(d_n-r_n)/K, \log K, L_\alpha\}, & K \ge 16, \\ \max\{\prod_n r_n + \sum_n r_n(d_n-r_n)/K, L_\alpha\}, & 1 \le K < 16. \end{cases} \tag{149}$$

For $K<16$, $\log K < 3 \le 3R \le 3D_0$. Thus, by $\log(1/\alpha) = L_\alpha + \log 8 \le \frac{5}{2}L_\alpha$ and the definition of T,

$$\begin{aligned}\prod_n r_n + \sum_n r_n(d_n-r_n)/K + \log(K/\alpha) &\le \begin{cases} \frac{9}{2}T, & K \ge 16, \\ \frac{13}{2}T, & 1 \le K < 16, \end{cases} \\ &\le 7T. \end{aligned} \tag{150}$$

Thus, there is

$$t_\star := \frac{\sigma_v}{128\sqrt{m}}\sqrt{\prod_n r_n + \sum_n r_n(d_n-r_n)/K + \log(K/\alpha)} \le \frac{\sqrt{7}\,\sigma_v}{128\sqrt{m}}\sqrt{T} \le \frac{\sigma_v}{32\sqrt{m}}\sqrt{T}. \tag{151}$$

Selecting the bound in (137), (144) or (148) corresponding to T, and using monotonicity and $2\alpha \le 1/8$, gives

$$\Psi(t_\star) \ge \Psi\left(\frac{\sigma_v}{32\sqrt{M}}\sqrt{T}\right) \ge 2\alpha. \tag{152}$$

Since $\hat{\boldsymbol{\mathcal{X}}}$ was arbitrary, taking the infimum over all estimators proves (17). The two-term expression following the theorem follows from $\sqrt{u+v} \le \sqrt{u}+\sqrt{v} \le \sqrt{2(u+v)}$, $u = \prod_n r_n(d_n - r_n)/K$, $v = \prod_n r_n + \log(K/\alpha)$.

H Technical Lemmas

Lemma 2. *(Subexponential Bernstein inequality (Vershynin, 2025)). Let $X_1,\cdots,X_N$ be independent, mean-zero, subexponential random variables, and $\boldsymbol{a} = (a_1,\cdots,a_N) \in \mathbb{R}^N$. Then, for every $t \ge 0$, we have*

$$\mathbb{P}\left(\left|\sum_{i=1}^N a_i X_i\right| \ge t\right) \le 2\exp\left(-c\min\left(\frac{t^2}{K^2\|\boldsymbol{a}\|^2}, \frac{t}{K\|\boldsymbol{a}\|_\infty}\right)\right), \tag{153}$$

where $K = \max_i \|X_i\|_{\psi_1}$.

Lemma 3. *(Gaussian norm concentration (Laurent & Massart, 2000, Lemma 1)) Let $g \sim \mathcal{N}(0, I_q)$. Then, for every $x > 0$, there is*

$$\mathbb{P}\Big(\|g\|_2 > \sqrt{q} + \sqrt{2x}\Big) \leq e^{-x}. \tag{154}$$

Lemma 4. *(Matrix Bernstein inequality (Vershynin, 2025)). Let $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_N$ be independent mean-zero, $n \times n$ symmetric matrices, such that $\|\boldsymbol{X}_i\| \leq K$ almost surely for all i. Then for every $t > 0$, we have*

$$\mathbb{P}\left(\left\|\sum_{i=1}^{N} \boldsymbol{X}_i\right\| \geq t\right) \leq 2n \exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right). \tag{155}$$

Here $\sigma^2 = \left\|\sum_{i=1}^{N} \mathbb{E}\left[\boldsymbol{X}_i^2\right]\right\|$ is the operator norm of the matrix variance of the sum.

Lemma 5. *If we have $\left\|\left(\boldsymbol{I} - \boldsymbol{U}_n \boldsymbol{U}_n^\top\right) \boldsymbol{U}_n^\star\right\| \leq \delta_t, \forall n \in [N]$, then there is*

$$\left\|\left(\boldsymbol{I} - \bigotimes_{n=1}^{N} \left(\boldsymbol{U}_n \boldsymbol{U}_n^\top\right)\right)\left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^\star\right)\right\| \leq \sqrt{1 - (1-\delta^2)^N}. \tag{156}$$

Proof. Based on the definition of subspace distance, we only need to obtain the minimum singular value of $\sigma_{\min}\left(\left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n\right)^\top \left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^\star\right)\right) = \cos\left(\Theta_{\max}\right)$. Because

$$\begin{aligned}
\sigma_{\min}\left(\left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n\right)^\top \left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^\star\right)\right) &= \sigma_{\min}\left(\bigotimes_{n=1}^{N} \left(\boldsymbol{U}_n^\top \boldsymbol{U}_n^\star\right)\right) \\
&\overset{(i)}{=} \prod_{n=1}^{N} \sigma_{\min}\left(\boldsymbol{U}_n^\top \boldsymbol{U}_n^\star\right) \\
&\geq \left(1 - \delta^2\right)^{\frac{N}{2}},
\end{aligned} \tag{157}$$

where (i) uses the property that $\sigma_{\min}\left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n\right) = \prod_{n=1}^{N} \sigma_{\min}\left(\boldsymbol{U}_n\right)$. Using $\left\|\left(\boldsymbol{I} - \bigotimes_{n=1}^{N} \left(\boldsymbol{U}_n \boldsymbol{U}_n^\top\right)\right)\left(\bigotimes_{n=1}^{N} \boldsymbol{U}_n^\star\right)\right\| = \sin\left(\Theta_{\max}\right) = \sqrt{1 - \cos^2\left(\Theta_{\max}\right)}$ and above inequality, we prove the result. □

Lemma 6. *For a given fixed column orthogonal matrix $\boldsymbol{U}^\star \in \mathbb{R}^{n \times r} (r < n)$ and random $\tilde{\boldsymbol{U}}_0$ with i.i.d. standard Gaussian entries , the random initialization $\boldsymbol{U}_0 = \tilde{\boldsymbol{U}}_0 (\tilde{\boldsymbol{U}}_0^\top \tilde{\boldsymbol{U}}_0)^{-\frac{1}{2}}$ satisfies*

$$\mathbb{P}\left(\sigma_{\min}(\boldsymbol{U}^{\star\top} \boldsymbol{U}_0) \geq \frac{\epsilon}{8\sqrt{r}(\sqrt{n} + \sqrt{r})}\right) \geq 1 - (C\epsilon) - 2\exp(-n/C) \tag{158}$$

and

$$\mathbb{P}\left(\sigma_{\min}(\boldsymbol{U}^{\star\top} \boldsymbol{U}_0) \leq \frac{2\sqrt{r} + \sqrt{2\log\frac{2}{\delta}}}{\sqrt{n} - \sqrt{r} - \sqrt{2\log\frac{2}{\delta}}}\right) \geq 1 - \delta, \tag{159}$$

where $\epsilon, \delta \in (0, 1)$ are sufficient small constants.

Proof. Based on the rotational invariant property of the Gaussian distribution, we may assume $\boldsymbol{U}^\star = \left[\boldsymbol{I}_r; \boldsymbol{0}_{(n-r)\times r}\right]$. Then $\boldsymbol{U}^{\star\top} \boldsymbol{U}_0 = \tilde{\boldsymbol{U}}_{0,1} (\tilde{\boldsymbol{U}}_0^\top \tilde{\boldsymbol{U}}_0)^{-\frac{1}{2}}$ where $\tilde{\boldsymbol{U}}_{0,1}$ is top-r row of $\tilde{\boldsymbol{U}}_0$.

Thus, for (158) there is

$$\sigma_{\min}(\boldsymbol{U}^{\star\top} \boldsymbol{U}_0) = \sigma_{\min}(\tilde{\boldsymbol{U}}_{0,1} (\tilde{\boldsymbol{U}}_0^\top \tilde{\boldsymbol{U}}_0)^{-\frac{1}{2}})$$

$$\geq \frac{\sigma_{\min}(\tilde{\boldsymbol{U}}_{0,1})}{\sigma_{\max}(\tilde{\boldsymbol{U}}_0)}$$
$$\geq \frac{\epsilon}{8\sqrt{r}(\sqrt{n}+\sqrt{r})}, \tag{160}$$

where the last inequality holds with probability at least $1 - C\epsilon - 2\exp(-n/C)$ for any $\epsilon > 0$ based on the proof for Lemma 1 in (Xu et al., 2023).

For the (159), there is

$$\sigma_{\min}({\boldsymbol{U}^\star}^\top \boldsymbol{U}_0) = \sigma_{\min}(\tilde{\boldsymbol{U}}_{0,1}(\tilde{\boldsymbol{U}}_0^\top \tilde{\boldsymbol{U}}_0)^{-\frac{1}{2}})$$
$$\leq \frac{\sigma_{\max}(\tilde{\boldsymbol{U}}_{0,1})}{\sigma_{\min}(\tilde{\boldsymbol{U}}_0)}$$
$$\leq \frac{2\sqrt{r} + \sqrt{2\log\frac{2}{\delta}}}{\sqrt{n} - \sqrt{r} - \sqrt{2\log\frac{2}{\delta}}}, \tag{161}$$

where the last inequality holds with probability at least $1 - 3\delta$ for any $\delta > 0$ based on Theorem 4.4.3 and Exercise 7.13 in (Vershynin, 2025). □