Fast and Provable Tensor Robust Principal Component Analysis via Scaled Gradient Descent

Harry Dong* CMU | Tian Tong* CMU | Cong Ma† UChicago | Yuejie Chi* CMU

June 2022; Revised: February 2023

Abstract

An increasing number of data science and machine learning problems rely on computation with tensors, which better capture the multi-way relationships and interactions of data than matrices. When tapping into this critical advantage, a key challenge is to develop computationally efficient and provably correct algorithms for extracting useful information from tensor data that are simultaneously robust to corruptions and ill-conditioning. This paper tackles tensor robust principal component analysis (RPCA), which aims to recover a low-rank tensor from its observations contaminated by sparse corruptions, under the Tucker decomposition. To minimize the computation and memory footprints, we propose to directly recover the low-dimensional tensor factors—starting from a tailored spectral initialization—via scaled gradient descent (ScaledGD), coupled with an iteration-varying thresholding operation to adaptively remove the impact of corruptions. Theoretically, we establish that the proposed algorithm converges linearly to the true low-rank tensor at a constant rate that is independent with its condition number, as long as the level of corruptions is not too large. Empirically, we demonstrate that the proposed algorithm achieves better and more scalable performance than state-of-the-art tensor RPCA algorithms through synthetic experiments and real-world applications.

Keywords: low-rank tensors, Tucker decomposition, robust principal component analysis, scaled gradient descent, preconditioning.

Contents

1 Introduction **2**
- 1.1 Our approach . . . 3
- 1.2 Related works . . . 4
- 1.3 Notation and tensor preliminaries . . . 5

2 Main results **6**
- 2.1 Problem formulation . . . 6
- 2.2 Proposed algorithm . . . 7
- 2.3 Performance guarantees . . . 8

3 Outline of the analysis **9**

*Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA; Emails: {harryd,ttong1,yuejiec}@andrew.cmu.edu. The work of T. Tong was completed while he was a graduate student at CMU.

†Department of Statistics, University of Chicago, Chicago, IL 60637, USA; Email: congm@uchicago.edu.

4 Numerical experiments **11**
4.1 Experiments on synthetic data . . . 11
4.2 Image denoising and outlier detection . . . 12
4.3 Background subtraction via selective mode update . . . 13

5 Conclusions **13**

A Proof of Lemma 1 **18**
A.1 Proof of (24) . . . 19
A.2 Proof of (26) . . . 22

B Proof of Lemma 2 **24**

C Proof of Lemma 3 **27**

D Proof of Lemma 4 **28**

E Technical lemmas **31**
E.1 Tensor algebra . . . 31
E.2 Perturbation bounds . . . 33
E.3 Sparse outliers . . . 36

1 Introduction

An increasing number of data science and machine learning problems rely on computation with tensors [KB09, PFS16], which better capture the multi-way relationships and interactions of data than matrices; examples include recommendation systems [KABO10], topic modeling [AGH+14], image processing [LMWY12], anomaly detection [LWQ+15], and so on. Oftentimes the data object of interest can be represented by a much smaller number of latent factors than what its ambient dimension suggests, which induces a low-rank structure in the underlying tensor. Unlike the matrix case, the flexibility of tensor modeling allows one to decompose a tensor under several choices of popular decompositions. The particular tensor decomposition studied in this paper is the Tucker decomposition, where a third-order tensor $\boldsymbol{\mathcal{X}}_\star \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is *low-rank* if it can be decomposed as[1]

$$\boldsymbol{\mathcal{X}}_\star = \left(\boldsymbol{U}_\star^{(1)}, \boldsymbol{U}_\star^{(2)}, \boldsymbol{U}_\star^{(3)}\right) \cdot \boldsymbol{\mathcal{G}}_\star,$$

where $\boldsymbol{U}_\star^{(1)} \in \mathbb{R}^{n_1 \times r_1}$, $\boldsymbol{U}_\star^{(2)} \in \mathbb{R}^{n_2 \times r_2}$, $\boldsymbol{U}_\star^{(3)} \in \mathbb{R}^{n_3 \times r_3}$ are the factor matrices along each mode, $\boldsymbol{\mathcal{G}}_\star \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ is the core tensor, and $\{r_i\}_{i=1}^3$ are the rank of each mode; see Section 2.1 for the precise definition. If we flatten the tensor along each mode, then the obtained matrices are all correspondingly low-rank:

$$r_1 = \operatorname{rank}\left(\mathcal{M}_1(\boldsymbol{\mathcal{X}}_\star)\right), \quad r_2 = \operatorname{rank}\left(\mathcal{M}_2(\boldsymbol{\mathcal{X}}_\star)\right), \quad r_3 = \operatorname{rank}\left(\mathcal{M}_3(\boldsymbol{\mathcal{X}}_\star)\right),$$

where $\mathcal{M}_k(\cdot)$ denotes the matricization of an input tensor along the k-th mode ($k = 1, 2, 3$). Intuitively, this means that the fibers along each mode lie in the same low-dimensional subspace. In other words, the tensor $\boldsymbol{\mathcal{X}}_\star$ has a multi-linear rank $\boldsymbol{r} = (r_1, r_2, r_3)$, where typically $r_k \ll n_k$. Throughout the paper, we denote $n := \max_k n_k$ and $r := \max_k r_k$.

This paper tackles tensor robust principal component analysis (RPCA), which aims to recover a low-rank tensor $\boldsymbol{\mathcal{X}}_\star$ from its observations contaminated by sparse corruptions. Mathematically, imagine we have access to a set of measurements given as

$$\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{X}}_\star + \boldsymbol{\mathcal{S}}_\star,$$

where $\boldsymbol{\mathcal{S}}_\star \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is a sparse tensor—in which the number of nonzero entries is much smaller than its ambient dimension—modeling corruptions or gross errors in the observations due to sensor failures,

[1] Note that there are several other popular notation for denoting the Tucker decomposition; our choice is made to facilitate the presentation of the analysis.

anomalies, or adversarial perturbations. Our goal is to recover $\boldsymbol{\mathcal{X}}_\star$ from the corrupted observation $\boldsymbol{\mathcal{Y}}$ in a computationally efficient and provably correct manner.

1.1 Our approach

In this paper, we propose a novel iterative method for tensor RPCA with provable convergence guarantees. To minimize the memory footprint, we aim to directly estimate the ground truth factors, collected in $\boldsymbol{F}_\star = (\boldsymbol{U}_\star^{(1)}, \boldsymbol{U}_\star^{(2)}, \boldsymbol{U}_\star^{(3)}, \boldsymbol{\mathcal{G}}_\star)$, via optimizing the following objective function:

$$\mathcal{L}(\boldsymbol{F}, \boldsymbol{\mathcal{S}}) \coloneqq \frac{1}{2} \left\| (\boldsymbol{U}^{(1)}, \boldsymbol{U}^{(2)}, \boldsymbol{U}^{(3)}) \cdot \boldsymbol{\mathcal{G}} + \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{Y}} \right\|_{\mathsf{F}}^2, \tag{1}$$

where $\boldsymbol{F} = (\boldsymbol{U}^{(1)}, \boldsymbol{U}^{(2)}, \boldsymbol{U}^{(3)}, \boldsymbol{\mathcal{G}})$ and $\boldsymbol{\mathcal{S}}$ are the optimization variables for the tensor factors and the corruption tensor, respectively. Despite the nonconvexity of the objective function, a simple and intuitive approach is to update the tensor factors via gradient descent, which, unfortunately, converges slowly even when the problem instance is moderately ill-conditioned [HWZ20].

On a high level, our proposed method alternates between corruption pruning (i.e., updating $\boldsymbol{\mathcal{S}}$) and factor refinements (i.e., updating $\boldsymbol{F}$). At the beginning of each iteration, we update the corruption tensor $\boldsymbol{\mathcal{S}}$ via thresholding the observation residuals as

$$\boldsymbol{\mathcal{S}}_{t+1} = \mathcal{T}_{\zeta_{t+1}} \left(\boldsymbol{\mathcal{Y}} - (\boldsymbol{U}_t^{(1)}, \boldsymbol{U}_t^{(2)}, \boldsymbol{U}_t^{(3)}) \cdot \boldsymbol{\mathcal{G}}_t \right), \qquad t = 0, 1, \dots \tag{2a}$$

where $\boldsymbol{\mathcal{S}}_{t+1}$ is the update of the corruption tensor at the t-th iteration, $\mathcal{T}_{\zeta_{t+1}}(\cdot)$ trims away the entries with magnitudes smaller than an iteration-varying threshold ζ_{t+1} that is carefully orchestrated, e.g., following a geometric decaying schedule. As the estimate of the data tensor $\boldsymbol{\mathcal{X}}_t = (\boldsymbol{U}_t^{(1)}, \boldsymbol{U}_t^{(2)}, \boldsymbol{U}_t^{(3)}) \cdot \boldsymbol{\mathcal{G}}_t$ gets more accurate, the observation residual becomes more aligned with the corruptions, therefore the thresholding operator (2a) becomes more effective in identifying and removing the impact of corruptions. Turning to the low-rank tensor factors $\boldsymbol{F}$, motivated by the recent success of scaled gradient descent (ScaledGD) [TMC21a, TMC21b, TMPB+22] for accelerating ill-conditioned low-rank estimation, we propose to update the tensor factors iteratively by descending along the scaled gradient directions:

$$\begin{aligned} \boldsymbol{U}_{t+1}^{(k)} &= \boldsymbol{U}_t^{(k)} - \eta \nabla_{\boldsymbol{U}_t^{(k)}} \mathcal{L}(\boldsymbol{F}_t, \boldsymbol{\mathcal{S}}_{t+1}) \big(\breve{\boldsymbol{U}}_t^{(k)\top} \breve{\boldsymbol{U}}_t^{(k)}\big)^{-1}, \qquad k = 1, 2, 3, \qquad \text{and} \\ \boldsymbol{\mathcal{G}}_{t+1} &= \boldsymbol{\mathcal{G}}_t - \eta \Big(\big(\boldsymbol{U}_t^{(1)\top} \boldsymbol{U}_t^{(1)}\big)^{-1}, \big(\boldsymbol{U}_t^{(2)\top} \boldsymbol{U}_t^{(2)}\big)^{-1}, \big(\boldsymbol{U}_t^{(3)\top} \boldsymbol{U}_t^{(3)}\big)^{-1} \Big) \cdot \nabla_{\boldsymbol{\mathcal{G}}_t} \mathcal{L}(\boldsymbol{F}_t, \boldsymbol{\mathcal{S}}_{t+1}). \end{aligned} \tag{2b}$$

Here, $\boldsymbol{F}_t = (\boldsymbol{U}_t^{(1)}, \boldsymbol{U}_t^{(2)}, \boldsymbol{U}_t^{(3)}, \boldsymbol{\mathcal{G}}_t)$ is the estimate of the tensor factors at the t-th iteration, $\nabla_{\boldsymbol{U}^{(k)}} \mathcal{L}(\boldsymbol{F}, \boldsymbol{\mathcal{S}})$ and $\nabla_{\boldsymbol{\mathcal{G}}} \mathcal{L}(\boldsymbol{F}, \boldsymbol{\mathcal{S}})$ are the partial derivatives of $\mathcal{L}(\boldsymbol{F}, \boldsymbol{\mathcal{S}})$ with respect to the corresponding variables, $\eta > 0$ is the learning rate, and

$$\breve{\boldsymbol{U}}_t^{(1)} = (\boldsymbol{U}_t^{(3)} \otimes \boldsymbol{U}_t^{(2)}) \mathcal{M}_1(\boldsymbol{\mathcal{G}}_t)^\top, \quad \breve{\boldsymbol{U}}_t^{(2)} = (\boldsymbol{U}_t^{(3)} \otimes \boldsymbol{U}_t^{(1)}) \mathcal{M}_2(\boldsymbol{\mathcal{G}}_t)^\top, \quad \breve{\boldsymbol{U}}_t^{(3)} = (\boldsymbol{U}_t^{(2)} \otimes \boldsymbol{U}_t^{(1)}) \mathcal{M}_3(\boldsymbol{\mathcal{G}}_t)^\top$$

are used to construct the preconditioned directions of the gradients, with $\otimes$ denoting the Kronecker product. With the preconditioners, ScaledGD balances the tensor factors to find better descent directions, the benefits of which are more accentuated in ill-conditioned tensors where the convergence rate of vanilla gradient descent degenerates significantly, while ScaledGD is capable of maintaining a linear rate of convergence regardless of the condition number.

Theoretical guarantees. Coupled with a tailored spectral initialization scheme, the proposed ScaledGD method converges linearly to the true low-rank tensor in both the Frobenius norm and the entrywise ℓ_∞ norm at a constant rate that is independent of its condition number, as long as the level of corruptions—measured in terms of the fraction of nonzero entries per fiber—does not exceed the order of $\frac{1}{\mu^2 \kappa r_1 r_2 r_3}$, where μ and κ are respectively the incoherence parameter and the condition number of the ground truth tensor $\boldsymbol{\mathcal{X}}_\star$ (to be formally defined later). This not only enables fast global convergence by virtue of following the scaled gradients rather than the vanilla gradients [TMPB+22], but also lends additional robustness to finding the low-rank Tucker decomposition despite the presence of corruptions and gross errors. Moreover, our work

provides the first refined entrywise error analysis for tensor RPCA, suggesting the errors are distributed evenly across the entries when the ground low-rank truth tensor is incoherent. To corroborate the theoretical findings, we further demonstrate that the proposed ScaledGD algorithm achieves better and more scalable performance than state-of-the-art matrix and tensor RPCA algorithms through synthetic experiments and real-world applications.

Comparisons to prior art. While tensor RPCA has been previously investigated under various low-rank tensor decompositions, e.g., [LFC+16, AJSN16, DBBG19], the development of provably efficient algorithms under the Tucker decomposition remains scarce. The most closely related work is [CLX21], which proposed a Riemannian gradient descent algorithm for the same tensor RPCA model as ours. Their algorithm is proven to also achieve a constant rate of convergence—at a higher per-iteration expense—as long as the fraction of outliers per fiber does not exceed the order of $\min\left\{\frac{1}{\mu_s^4\kappa_s^{14}r^2\log^2 n}, \frac{1}{\mu_s^{12}\kappa_s^{12}r^3}\right\}$ (cf. [CLX21, Theorem 5.1]), where μ_s and κ_s are the spikiness parameter and the worst-case condition number of $\boldsymbol{\mathcal{X}}_\star$, respectively. Using the relation $\mu \le \mu_s^2\kappa_s^2$ (cf. [CLX21, Lemma 13.5]) and $\kappa \le \kappa_s$ (cf. (14)) to conservatively translate our bound, our algorithm succeeds as long as the corruption level is below the order of $\frac{1}{\mu_s^4\kappa_s^5 r^3}$, which is still significantly higher than that allowed in [CLX21], when the outliers are evenly distributed across the fibers. See additional numerical comparisons in Section 4.

1.2 Related works

Broadly speaking, our work falls under the recent surge of developing both computationally efficient and provably correct algorithms for high-dimensional signal estimation via nonconvex optimization, which has been particularly fruitful for problems with inherent low-rank structures; we refer interested readers to the recent overviews [CLC19, CC18] for further pointers. In the rest of this section, we focus on works that are most closely related to our paper.

Provable algorithms for matrix RPCA. The matrix RPCA problem, which aims to decompose a low-rank matrix and a sparse matrix from their sum, has been heavily investigated since its introduction in the seminar papers [CLMW11, CSPW11]. Convex relaxation based approaches, which minimize a weighted sum of the nuclear norm of the data matrix and the ℓ_1 norm of the corruption matrix, have been demonstrated to achieve near-optimal performance guarantees [CLMW11, WGR+09, CSPW11, LCM10, CFMY21, CC14]. However, their computational and memory complexities are prohibitive when applied to large-scale problem instances; for example, solving the resulting semidefinite programs via accelerated proximal gradient descent [TY10] only results in a sublinear rate of convergence with a per-iteration complexity that scales cubically with the matrix dimension. To address the computational bottleneck, nonconvex methods have been developed to achieve both statistical and computational efficiencies simultaneously [NNS+14, CCW19, GWL16, YPCC16, TMC21a, CLY21]. Our tensor RPCA algorithm draws inspiration from [TMC21a, TMPB+22], which adopt a factored representation of the low-rank object and invoke scaled gradient updates to bypass the dependence of the convergence rate on the condition number. The matrix RPCA method in [CLY21] differs from [TMC21a] by using a threshold-based trimming procedure—which we also adopt—rather than a sorting-based one to identify the sparse matrix, for further computational savings.

Provable algorithms for tensor RPCA. Moving onto tensors, although one could unfold a tensor and feed the resulting matrices into a matrix RPCA algorithm [GQ14, ZWZM19], destroying the tensor structure through matricizations can result in suboptimal performance because it ignores the higher-order interactions [YZ16]. Therefore, it is desirable to directly operate in the tensor space. However, tensor algorithms encounter unique issues not present for matrices. For instance, while it appears straightforward to generalize the convex relaxation approach to tensors, it has been shown that computing the tensor nuclear norm is in fact NP-hard [FL18]; a similar drawback is applicable to the atomic norm formulation studied in [DBBG19]. Tensor RPCA has also been studied under different low-rank tensor decompositions, a small number of samples including the tubal rank [LFC+16, LFC+19] and the CP-rank [AJSN16, DBBG19]. These algorithms are not directly comparable with ours which uses the multilinear rank.

Robust low-rank tensor recovery. Broadly speaking, tensor RPCA concerns with reconstructing a high-dimensional tensor with certain low-dimensional structures from incomplete and corrupted observations. Pertaining to works that deal with the Tucker decomposition, [XY19] proposed a gradient descent based algorithm for tensor completion, [TMPB+22, TMC22] proposed scaled gradient descent algorithms for tensor regression and tensor completion (which our algorithm also adopts), [LZ21] proposed a Gauss-Newton algorithm for tensor regression that achieves quadratic convergence, [WCW21] proposed a Riemannian gradient method with entrywise convergence guarantees, and [ARB20] studied tensor regression assuming the underlying tensor is simultaneously low-rank and sparse.

1.3 Notation and tensor preliminaries

Throughout this paper, we use boldface calligraphic letters (e.g. $\boldsymbol{\mathcal{X}}$) to denote tensors, and boldface capitalized letters (e.g. $\boldsymbol{X}$) to denote matrices. For any matrix $\boldsymbol{X}$, let $\sigma_i(\boldsymbol{X})$ be its i-th largest singular value, and $\sigma_{\max}(\boldsymbol{X})$ (resp. $\sigma_{\min}(\boldsymbol{X})$) to denote its largest (resp. smallest) nonzero singular value. Let $\|\boldsymbol{X}\|$, $\|\boldsymbol{X}\|_{\mathsf{F}}$, $\|\boldsymbol{X}\|_{2,\infty}$, and $\|\boldsymbol{X}\|_{\infty}$ be the spectral norm, the Frobenius norm, the $\ell_{2,\infty}$ norm (largest ℓ_2 norm of the rows), and the entrywise ℓ_∞ norm of a matrix $\boldsymbol{X}$, respectively. The $r \times r$ identity matrix is denoted by $\boldsymbol{I}_r$. The set of invertible matrices in $\mathbb{R}^{r\times r}$ is denoted by $\mathrm{GL}(r)$.

We now describe some preliminaries on tensor algebra that are used throughout this paper. For a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1\times n_2\times n_3}$, let $[\boldsymbol{\mathcal{X}}]_{i,j,k}$ be its (i,j,k)-th entry. For a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1\times n_2\times n_3}$, suppose it can be represented via the multilinear multiplication

$$\boldsymbol{\mathcal{X}} = \big(\boldsymbol{U}^{(1)}, \boldsymbol{U}^{(2)}, \boldsymbol{U}^{(3)}\big)\cdot \boldsymbol{\mathcal{G}},$$

where $\boldsymbol{U}^{(k)} \in \mathbb{R}^{n_k\times r_k}$, $k = 1, 2, 3,$, and $\boldsymbol{\mathcal{G}} \in \mathbb{R}^{r_1\times r_2\times r_3}$. Equivalently, the entries of $\boldsymbol{\mathcal{X}}$ can be expressed as

$$\big[\boldsymbol{\mathcal{X}}\big]_{i_1,i_2,i_3} = \sum_{j_1=1}^{r_1}\sum_{j_2=1}^{r_2}\sum_{j_3=1}^{r_3}\big[\boldsymbol{U}^{(1)}\big]_{i_1,j_1}\big[\boldsymbol{U}^{(2)}\big]_{i_2,j_2}\big[\boldsymbol{U}^{(3)}\big]_{i_3,j_3}\big[\boldsymbol{\mathcal{G}}\big]_{j_1,j_2,j_3}.$$

The multilinear multiplication possesses several nice properties. A crucial one is that for any $\boldsymbol{B}^{(k)} \in \mathbb{R}^{r_k\times r_k}$, $k = 1, 2, 3$, it holds that

$$\big(\boldsymbol{U}^{(1)}\boldsymbol{B}^{(1)}, \boldsymbol{U}^{(2)}\boldsymbol{B}^{(2)}, \boldsymbol{U}^{(3)}\boldsymbol{B}^{(3)}\big)\cdot\boldsymbol{\mathcal{G}} = \big(\boldsymbol{U}^{(1)}, \boldsymbol{U}^{(2)}, \boldsymbol{U}^{(3)}\big)\cdot\Big(\big(\boldsymbol{B}^{(1)}, \boldsymbol{B}^{(2)}, \boldsymbol{B}^{(3)}\big)\cdot\boldsymbol{\mathcal{G}}\Big). \tag{3}$$

In addition, if we flatten the tensor $\boldsymbol{\mathcal{X}}$ along different modes, the obtained matrices obey the following low-rank decompositions:

$$\mathcal{M}_1(\boldsymbol{\mathcal{X}}) = \boldsymbol{U}^{(1)}\mathcal{M}_1(\boldsymbol{\mathcal{G}})\big(\boldsymbol{U}^{(3)}\otimes\boldsymbol{U}^{(2)}\big)^\top = \boldsymbol{U}^{(1)}\breve{\boldsymbol{U}}^{(1)\top}, \qquad \breve{\boldsymbol{U}}^{(1)} := \big(\boldsymbol{U}^{(3)}\otimes\boldsymbol{U}^{(2)}\big)\mathcal{M}_1(\boldsymbol{\mathcal{G}})^\top, \tag{4a}$$

$$\mathcal{M}_2(\boldsymbol{\mathcal{X}}) = \boldsymbol{U}^{(2)}\mathcal{M}_2(\boldsymbol{\mathcal{G}})(\boldsymbol{U}^{(3)}\otimes\boldsymbol{U}^{(1)})^\top = \boldsymbol{U}^{(2)}\breve{\boldsymbol{U}}^{(2)\top}, \qquad \breve{\boldsymbol{U}}^{(2)} := \big(\boldsymbol{U}^{(3)}\otimes\boldsymbol{U}^{(1)}\big)\mathcal{M}_2(\boldsymbol{\mathcal{G}})^\top, \tag{4b}$$

$$\mathcal{M}_3(\boldsymbol{\mathcal{X}}) = \boldsymbol{U}^{(3)}\mathcal{M}_3(\boldsymbol{\mathcal{G}})(\boldsymbol{U}^{(2)}\otimes\boldsymbol{U}^{(1)})^\top = \boldsymbol{U}^{(3)}\breve{\boldsymbol{U}}^{(3)\top}, \qquad \breve{\boldsymbol{U}}^{(3)} := \big(\boldsymbol{U}^{(2)}\otimes\boldsymbol{U}^{(1)}\big)\mathcal{M}_3(\boldsymbol{\mathcal{G}})^\top. \tag{4c}$$

Given two tensors $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\mathcal{B}}$, their inner product is defined as $\langle\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}}\rangle = \sum_{i_1,i_2,i_3}\boldsymbol{\mathcal{A}}_{i_1,i_2,i_3}\boldsymbol{\mathcal{B}}_{i_1,i_2,i_3}$. The inner product satisfies the following property:

$$\Big\langle\big(\boldsymbol{U}^{(1)}, \boldsymbol{U}^{(2)}, \boldsymbol{U}^{(3)}\big)\cdot\boldsymbol{\mathcal{G}}, \boldsymbol{\mathcal{X}}\Big\rangle = \Big\langle\boldsymbol{\mathcal{G}}, \big(\boldsymbol{U}^{(1)\top}, \boldsymbol{U}^{(2)\top}, \boldsymbol{U}^{(3)\top}\big)\cdot\boldsymbol{\mathcal{X}}\Big\rangle. \tag{5}$$

Denote the Frobenius norm and the ℓ_∞ norm of $\boldsymbol{\mathcal{X}}$ as $\|\boldsymbol{\mathcal{X}}\|_{\mathsf{F}} = \sqrt{\langle\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}}\rangle}$ and $\|\boldsymbol{\mathcal{X}}\|_\infty = \max_{i_1,i_2,i_3}|\boldsymbol{\mathcal{X}}_{i_1,i_2,i_3}|$, respectively. It follows that for $\boldsymbol{Q}_k \in \mathbb{R}^{r_k\times r_k}$, $k = 1, 2, 3$:

$$\big\|(\boldsymbol{Q}_1, \boldsymbol{Q}_2, \boldsymbol{Q}_3)\cdot\boldsymbol{\mathcal{G}}\big\|_{\mathsf{F}} \le \|\boldsymbol{Q}_1\|\|\boldsymbol{Q}_2\|\|\boldsymbol{Q}_3\|\|\boldsymbol{\mathcal{G}}\|_{\mathsf{F}}. \tag{6}$$

Let $\boldsymbol{r} = (r_1, r_2, r_3)$. For a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1\times n_2\times n_3}$, let its rank-$\boldsymbol{r}$ higher-order singular value decomposition (HOSVD) $\mathcal{H}_{\boldsymbol{r}}(\boldsymbol{\mathcal{X}})$ be

$$\mathcal{H}_{\boldsymbol{r}}(\boldsymbol{\mathcal{X}}) = (\boldsymbol{U}^{(1)}, \boldsymbol{U}^{(2)}, \boldsymbol{U}^{(3)}, \boldsymbol{\mathcal{G}}), \tag{7}$$

where $\boldsymbol{U}^{(k)}$ is the top r_k left singular vectors of $\mathcal{M}_k(\boldsymbol{\mathcal{X}})$, $k = 1, 2, 3$, and $\boldsymbol{\mathcal{G}} = (\boldsymbol{U}^{(1)\top}, \boldsymbol{U}^{(2)\top}, \boldsymbol{U}^{(3)\top}) \cdot \boldsymbol{\mathcal{X}}$ is the core tensor. The HOSVD is an extension of the matrix SVD and can be seen as a special case of the Tucker decomposition; see [BL10] for an exposition. Although there are faster methods—such as [VVM12]—available, one straightforward way of computing the HOSVD is to obtain the singular vectors from performing matrix SVD on each matricization of $\boldsymbol{\mathcal{X}}$. With these vectors, we can construct each $\boldsymbol{U}^{(k)}$, followed by finding $\boldsymbol{\mathcal{G}} = (\boldsymbol{U}^{(1)\top}, \boldsymbol{U}^{(2)\top}, \boldsymbol{U}^{(3)\top}) \cdot \boldsymbol{\mathcal{X}}$ to complete the process. In contrast to its matrix counterpart, the core tensor $\boldsymbol{\mathcal{G}}$ will not necessarily be diagonal.

2 Main results

2.1 Problem formulation

Suppose that the ground truth tensor $\boldsymbol{\mathcal{X}}_\star \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ with multilinear rank $\boldsymbol{r} = (r_1, r_2, r_3)$ admits the following Tucker decomposition

$$\boldsymbol{\mathcal{X}}_\star = \left(\boldsymbol{U}_\star^{(1)}, \boldsymbol{U}_\star^{(2)}, \boldsymbol{U}_\star^{(3)}\right) \cdot \boldsymbol{\mathcal{G}}_\star, \tag{8}$$

where $\boldsymbol{U}_\star^{(1)} \in \mathbb{R}^{n_1 \times r_1}$, $\boldsymbol{U}_\star^{(2)} \in \mathbb{R}^{n_2 \times r_2}$, $\boldsymbol{U}_\star^{(3)} \in \mathbb{R}^{n_3 \times r_3}$ are the factor matrices along each mode, and $\boldsymbol{\mathcal{G}}_\star \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ is the core tensor. The Tucker decomposition is not unique since for any $\boldsymbol{Q}^{(k)} \in \mathrm{GL}(r_k)$, $k = 1, 2, 3$, in view of (3), we have

$$\left(\boldsymbol{U}^{(1)}, \boldsymbol{U}^{(2)}, \boldsymbol{U}^{(3)}\right) \cdot \boldsymbol{\mathcal{G}} = \left(\boldsymbol{U}^{(1)}\boldsymbol{Q}^{(1)}, \boldsymbol{U}^{(2)}\boldsymbol{Q}^{(2)}, \boldsymbol{U}^{(3)}\boldsymbol{Q}^{(3)}\right) \cdot \boldsymbol{\mathcal{G}}_{\boldsymbol{Q}}$$

where $\boldsymbol{\mathcal{G}}_{\boldsymbol{Q}} = \left((\boldsymbol{Q}^{(1)})^{-1}, (\boldsymbol{Q}^{(2)})^{-1}, (\boldsymbol{Q}^{(3)})^{-1}\right) \cdot \boldsymbol{\mathcal{G}}$. Without loss of generality, to address ambiguity, we set the ground truth $\boldsymbol{F}_\star = (\boldsymbol{U}_\star^{(1)}, \boldsymbol{U}_\star^{(2)}, \boldsymbol{U}_\star^{(3)}, \boldsymbol{\mathcal{G}}_\star)$ to satisfy that for each mode, $\boldsymbol{U}_\star^{(k)} \in \mathbb{R}^{n_k \times r_k}$ to have orthonormal columns, and

$$\mathcal{M}_k(\boldsymbol{\mathcal{G}}_\star)\, \mathcal{M}_k(\boldsymbol{\mathcal{G}}_\star)^\top = \left(\boldsymbol{\Sigma}_\star^{(k)}\right)^2, \tag{9}$$

the squared singular value matrix $\boldsymbol{\Sigma}_\star^{(k)}$ of $\mathcal{M}_k(\boldsymbol{\mathcal{X}}_\star)$. This can be easily met, for example, by taking the tensor factors $\boldsymbol{F}_\star$ as the HOSVD of $\boldsymbol{\mathcal{X}}_\star$.

Observation model and goal. Suppose that we collect a set of corrupted observations of $\boldsymbol{\mathcal{X}}_\star$ as

$$\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{X}}_\star + \boldsymbol{\mathcal{S}}_\star, \tag{10}$$

where $\boldsymbol{\mathcal{S}}_\star$ is the corruption tensor. The problem of tensor RPCA seeks to separate $\boldsymbol{\mathcal{X}}_\star$ and $\boldsymbol{\mathcal{S}}_\star$ from their sum $\boldsymbol{\mathcal{Y}}$ as efficiently and accurately as possible.

Key quantities. Obviously, the tensor RPCA problem is ill-posed without imposing additional constraints on the low-rank tensor $\boldsymbol{\mathcal{X}}_\star$ and the corruption tensor $\boldsymbol{\mathcal{S}}_\star$, which are crucial in determining the performance of the proposed algorithm. We first introduce the incoherence parameter of the tensor $\boldsymbol{\mathcal{X}}_\star$.

Definition 1 (Incoherence)**.** The incoherence parameter μ of $\boldsymbol{\mathcal{X}}_\star$ is defined as

$$\mu := \max_k \left\{ \frac{n_k}{r_k} \left\| \boldsymbol{U}_\star^{(k)} \right\|_{2,\infty}^2 \right\}, \tag{11}$$

where $\boldsymbol{\mathcal{X}}_\star = \left(\boldsymbol{U}_\star^{(1)}, \boldsymbol{U}_\star^{(2)}, \boldsymbol{U}_\star^{(3)}\right) \cdot \boldsymbol{\mathcal{G}}_\star$ is its Tucker decomposition.

The incoherence parameter roughly measures how spread the energy of $\boldsymbol{\mathcal{X}}_\star$ is over its entries—the energy is more spread as μ gets smaller. Moreover, we define a new notion of condition number that measures the conditioning of the ground truth tensor $\boldsymbol{\mathcal{X}}_\star$ as follows, which is weaker than previously used notions.

Definition 2 (Condition number)**.** The condition number κ of $\boldsymbol{\mathcal{X}}_\star$ is defined as

$$\kappa := \frac{\min_k \sigma_{\max}(\mathcal{M}_k(\boldsymbol{\mathcal{X}}_\star))}{\min_k \sigma_{\min}(\mathcal{M}_k(\boldsymbol{\mathcal{X}}_\star))}. \tag{12}$$

With slight abuse of terminology, denote

$$\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star) = \min_k \sigma_{\min}(\mathcal{M}_k(\boldsymbol{\mathcal{X}}_\star)) \tag{13}$$

as the minimum nonzero singular value of $\boldsymbol{\mathcal{X}}_\star$.

Remark 1. The above-defined condition number can be much smaller than the *worst-case* condition number κ_s used in prior analyses [TMPB+22, CLX21, HWZ20], which is defined as

$$\kappa_s := \frac{\max_k \sigma_{\max}(\mathcal{M}_k(\boldsymbol{\mathcal{X}}_\star))}{\min_k \sigma_{\min}(\mathcal{M}_k(\boldsymbol{\mathcal{X}}_\star))} \geq \frac{\min_k \sigma_{\max}(\mathcal{M}_k(\boldsymbol{\mathcal{X}}_\star))}{\min_k \sigma_{\min}(\mathcal{M}_k(\boldsymbol{\mathcal{X}}_\star))} = \kappa. \tag{14}$$

Furthermore, the condition number κ is also upper bounded by the largest condition number of the matricization along different modes, i.e., $\kappa \leq \max_k \kappa_k = \max_k \frac{\sigma_{\max}(\mathcal{M}_k(\boldsymbol{\mathcal{X}}_\star))}{\sigma_{\min}(\mathcal{M}_k(\boldsymbol{\mathcal{X}}_\star))}$.

Turning to the corruption tensor, we consider a deterministic sparsity model following the matrix case [CSPW11, NNS+14, YPCC16], where $\boldsymbol{\mathcal{S}}_\star$ contains at most a small fraction of nonzero entries per fiber. This is captured in the following definition.

Definition 3 (α-fraction sparsity)**.** The corruption tensor $\boldsymbol{\mathcal{S}}_\star$ is α-fraction sparse, i.e., $\boldsymbol{\mathcal{S}}_\star \in \boldsymbol{\mathcal{S}}_\alpha$, where

$$\boldsymbol{\mathcal{S}}_\alpha := \Big\{ \boldsymbol{\mathcal{S}} \in \mathbb{R}^{n_1 \times n_2 \times n_3} : \ \|\boldsymbol{\mathcal{S}}_{i_1,i_2,:}\|_0 \leq \alpha n_3, \ \|\boldsymbol{\mathcal{S}}_{i_1,:,i_3}\|_0 \leq \alpha n_2, \ \|\boldsymbol{\mathcal{S}}_{:,i_2,i_3}\|_0 \leq \alpha n_1, \\ \text{for all } 1 \leq i_k \leq n_k, \quad k = 1,2,3 \Big\}. \tag{15}$$

With this setup in hand, we are now ready to describe the proposed algorithm.

2.2 Proposed algorithm

Our algorithm alternates between corruption removal and factor refinements. To remove the corruption, we use the following soft-shrinkage operator that trims the magnitudes of the entries by the amount of some carefully pre-set threshold.

Definition 4 (Soft-shrinkage operator)**.** For an order-3 tensor $\boldsymbol{\mathcal{X}}$, the soft-shrinkage operator $\mathcal{T}_\zeta(\cdot) : \mathbb{R}^{n_1 \times n_2 \times n_3} \mapsto \mathbb{R}^{n_1 \times n_2 \times n_3}$ with threshold $\zeta > 0$ is defined as

$$\big[\mathcal{T}_\zeta(\boldsymbol{\mathcal{X}})\big]_{i_1,i_2,i_3} := \operatorname{sgn}\big([\boldsymbol{\mathcal{X}}]_{i_1,i_2,i_3}\big) \cdot \max\big(0, \big|[\boldsymbol{\mathcal{X}}]_{i_1,i_2,i_3}\big| - \zeta\big).$$

The soft-shrinkage operator $\mathcal{T}_\zeta()$ sets entries with magnitudes smaller than ζ to 0, while uniformly shrinking the magnitudes of the other entries by ζ. At the beginning of each iteration, the corruption tensor is updated via

$$\boldsymbol{\mathcal{S}}_{t+1} = \mathcal{T}_{\zeta_{t+1}}\Big(\boldsymbol{\mathcal{Y}} - (\boldsymbol{U}_t^{(1)}, \boldsymbol{U}_t^{(2)}, \boldsymbol{U}_t^{(3)}) \cdot \boldsymbol{\mathcal{G}}_t\Big), \tag{16a}$$

with the schedule ζ_t to be specified shortly. With the newly updated estimate of the corruption tensor, the tensor factors are then updated by scaled gradient descent [TMPB+22], for which they they are computed according to (2b) with respect to $\mathcal{L}(\boldsymbol{F}_t, \boldsymbol{\mathcal{S}}_{t+1})$ in (1):

$$\begin{aligned} \boldsymbol{U}_{t+1}^{(k)} &= \boldsymbol{U}_t^{(k)} - \eta \nabla_{\boldsymbol{U}_t^{(k)}} \mathcal{L}(\boldsymbol{F}_t, \boldsymbol{\mathcal{S}}_{t+1}) \big(\breve{\boldsymbol{U}}_t^{(k)\top} \breve{\boldsymbol{U}}_t^{(k)}\big)^{-1} \\ &= (1-\eta)\boldsymbol{U}_t^{(k)} - \eta\big(\mathcal{M}_k(\boldsymbol{\mathcal{S}}_{t+1}) - \mathcal{M}_k(\boldsymbol{\mathcal{Y}})\big)\breve{\boldsymbol{U}}_t^{(k)} \big(\breve{\boldsymbol{U}}_t^{(k)\top} \breve{\boldsymbol{U}}_t^{(k)}\big)^{-1} \end{aligned} \tag{16b}$$

for $k = 1, 2, 3$ and

$$\begin{aligned} \boldsymbol{\mathcal{G}}_{t+1} &= \boldsymbol{\mathcal{G}}_t - \eta \Big(\big(\boldsymbol{U}_t^{(1)\top} \boldsymbol{U}_t^{(1)}\big)^{-1}, \big(\boldsymbol{U}_t^{(2)\top} \boldsymbol{U}_t^{(2)}\big)^{-1}, \big(\boldsymbol{U}_t^{(3)\top} \boldsymbol{U}_t^{(3)}\big)^{-1} \Big) \cdot \nabla_{\boldsymbol{\mathcal{G}}_t} \mathcal{L}(\boldsymbol{F}_t, \boldsymbol{\mathcal{S}}_{t+1}) \\ &= (1-\eta)\boldsymbol{\mathcal{G}}_t - \eta \Big(\big(\boldsymbol{U}_t^{(1)\top} \boldsymbol{U}_t^{(1)}\big)^{-1} \boldsymbol{U}_t^{(1)\top}, \big(\boldsymbol{U}_t^{(2)\top} \boldsymbol{U}_t^{(2)}\big)^{-1} \boldsymbol{U}_t^{(2)\top}, \big(\boldsymbol{U}_t^{(3)\top} \boldsymbol{U}_t^{(3)}\big)^{-1} \boldsymbol{U}_t^{(3)\top} \Big) \cdot (\boldsymbol{\mathcal{S}}_{t+1} - \boldsymbol{\mathcal{Y}}). \end{aligned} \tag{16c}$$

Algorithm 1 ScaledGD for tensor robust principal component analysis

Input: the observed tensor $\boldsymbol{\mathcal{Y}}$, the multilinear rank $\boldsymbol{r}$, learning rate η, and threshold schedule $\{\zeta_t\}_{t=0}^{T}$.
Initialization: $\boldsymbol{\mathcal{S}}_0 = \mathcal{T}_{\zeta_0}(\boldsymbol{\mathcal{Y}})$ and $\left(\boldsymbol{U}_0^{(1)}, \boldsymbol{U}_0^{(2)}, \boldsymbol{U}_0^{(3)}, \boldsymbol{\mathcal{G}}_0\right) = \mathcal{H}_{\boldsymbol{r}}(\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{S}}_0)$.
for $t = 0, 1, \ldots, T-1$ **do**
  Update the corruption tensor $\boldsymbol{\mathcal{S}}_{t+1}$ via (16a);
  Update the tensor factors $\boldsymbol{F}_{t+1} = \left(\boldsymbol{U}_{t+1}^{(1)}, \boldsymbol{U}_{t+1}^{(2)}, \boldsymbol{U}_{t+1}^{(3)}, \boldsymbol{\mathcal{G}}_{t+1}\right)$ via (16b) and (16c);
end for
Output: the tensor factors $\boldsymbol{F}_T = \left(\boldsymbol{U}_T^{(1)}, \boldsymbol{U}_T^{(2)}, \boldsymbol{U}_T^{(3)}, \boldsymbol{\mathcal{G}}_T\right)$.

Here, $\eta > 0$ is the learning rate, and

$$\breve{\boldsymbol{U}}_t^{(1)} := \left(\boldsymbol{U}_t^{(3)} \otimes \boldsymbol{U}_t^{(2)}\right)\mathcal{M}_1(\boldsymbol{\mathcal{G}}_t)^\top, \quad \breve{\boldsymbol{U}}_t^{(2)} := \left(\boldsymbol{U}_t^{(3)} \otimes \boldsymbol{U}_t^{(1)}\right)\mathcal{M}_2(\boldsymbol{\mathcal{G}}_t)^\top, \text{ and } \quad \breve{\boldsymbol{U}}_t^{(3)} := \left(\boldsymbol{U}_t^{(2)} \otimes \boldsymbol{U}_t^{(1)}\right)\mathcal{M}_3(\boldsymbol{\mathcal{G}}_t)^\top.$$

To complete the algorithm description, we still need to specify how to initialize the algorithm. We will estimate the tensor factors via the spectral method, by computing the HOSVD of the observation after applying the soft-shrinkage operator:

$$\left(\boldsymbol{U}_0^{(1)}, \boldsymbol{U}_0^{(2)}, \boldsymbol{U}_0^{(3)}, \boldsymbol{\mathcal{G}}_0\right) = \mathcal{H}_{\boldsymbol{r}}(\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{S}}_0), \qquad \text{where} \quad \boldsymbol{\mathcal{S}}_0 = \mathcal{T}_{\zeta_0}(\boldsymbol{\mathcal{Y}}).$$

Altogether, we arrive at Algorithm 1, which we still dub as ScaledGD for simplicity.

Computational benefits. It is worth highlighting that the proposed tensor RPCA algorithm possesses several computational benefits which might be of interest in applications.

- *Advantages over matrix RPCA algorithms.* While it is possible to matricize the input tensor and then apply the matrix RPCA algorithms, they can only exploit the low-rank structure along the mode that the tensor is unfolded, rather than along multiple rows simultaneously as in the tensor RPCA algorithm. In addition, the space complexity of storing and computing the factors is much higher for the matrix RPCA algorithms, where the size of the factors become multiplicative in terms of the tensor dimensions due to unfolding, rather than linear as in the tensor RPCA algorithm.

- *Generalization to N-th order tensors.* Although the description of Algorithm 1 is tailored to an order-3 tensor, our algorithm is easily generalizable to any N-th order tensor; in fact, Algorithm 1 can be applied almost verbatim by redefining

  $$\breve{\boldsymbol{U}}_t^{(k)} = \left(\boldsymbol{U}_t^{(N)} \otimes \cdots \otimes \boldsymbol{U}_t^{(k+1)} \otimes \boldsymbol{U}_t^{(k-1)} \otimes \cdots \otimes \boldsymbol{U}_t^{(1)}\right)\mathcal{M}_k(\boldsymbol{\mathcal{G}}_t)^\top, \qquad k = 1, \ldots, N$$

  to its natural high-order counterpart. This extension is numerically evaluated in our experiments in Section 4.

- *Parallelizability.* At each iteration of the proposed algorithm, each tensor factor is updated independently as done in (16b) and (16c), therefore we can update them in a parallel manner. This improvement becomes more apparent as the order of the tensor increases.

- *Selective modes to update*: If we know the underlying ground truth tensor is only low-rank along certain mode, we can choose to skip the iterative updates of the rest of the modes after initialization to reduce computational costs, which we demonstrate empirically in Section 4.3.

2.3 Performance guarantees

Motivated by the analysis in [TMPB+22], we consider the following distance metric, which not only resolves the ambiguity in the Tucker decomposition, but also takes the preconditioning factor into consideration.

Definition 5 (Distance metric). Letting $\boldsymbol{F} := \big(\boldsymbol{U}^{(1)}, \boldsymbol{U}^{(2)}, \boldsymbol{U}^{(3)}, \boldsymbol{\mathcal{G}}\big)$ and $\boldsymbol{F}_\star := \big(\boldsymbol{U}_\star^{(1)}, \boldsymbol{U}_\star^{(2)}, \boldsymbol{U}_\star^{(3)}, \boldsymbol{\mathcal{G}}_\star\big)$, denote

$$\mathrm{dist}^2(\boldsymbol{F}, \boldsymbol{F}_\star) := \inf_{\boldsymbol{Q}^{(k)} \in \mathrm{GL}(r_k)} \sum_{k=1}^{3} \Big\| \big(\boldsymbol{U}^{(k)}\boldsymbol{Q}^{(k)} - \boldsymbol{U}_\star^{(k)}\big)\boldsymbol{\Sigma}_\star^{(k)} \Big\|_{\mathsf{F}}^2 + \Big\| \big((\boldsymbol{Q}^{(1)})^{-1}, (\boldsymbol{Q}^{(2)})^{-1}, (\boldsymbol{Q}^{(3)})^{-1}\big) \cdot \boldsymbol{\mathcal{G}} - \boldsymbol{\mathcal{G}}_\star \Big\|_{\mathsf{F}}^2, \tag{17}$$

where we recall $\boldsymbol{\Sigma}_\star^{(k)}$ is the singular value matrix of $\mathcal{M}_k(\boldsymbol{\mathcal{X}}_\star)$, $k = 1, 2, 3$. Moreover, if the infimum is attained at the arguments $\{\boldsymbol{Q}^{(k)}\}_{k=1}^3$, they are called the optimal alignment matrices between $\boldsymbol{F}$ and $\boldsymbol{F}_\star$.

Fortunately, the proposed ScaledGD algorithm (cf. Algorithm 1) provably recovers the ground truth tensor—as long as the fraction of corruptions is not too large—with proper choices of the tuning parameters, as captured in following theorem.

Theorem 1. *Let $\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{X}}_\star + \boldsymbol{\mathcal{S}}_\star \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, where $\boldsymbol{\mathcal{X}}_\star$ is μ-incoherent with multilinear rank $\boldsymbol{r} = (r_1, r_2, r_3)$, and $\boldsymbol{\mathcal{S}}_\star$ is α-sparse. Suppose that the threshold values $\{\zeta_k\}_{k=0}^{\infty}$ obey that $\|\boldsymbol{\mathcal{X}}_\star\|_\infty \le \zeta_0 \le 2\|\boldsymbol{\mathcal{X}}_\star\|_\infty$ and $\zeta_{t+1} = \rho\zeta_t$, $t \ge 1$, for some properly tuned $\zeta_1 := 8\sqrt{\frac{\mu^3 r_1 r_2 r_3}{n_1 n_2 n_3}}\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$ and $\frac{1}{7} \le \eta \le \frac{1}{4}$, where $\rho = 1 - 0.45\eta$. Then, the iterates $\boldsymbol{\mathcal{X}}_t = \big(\boldsymbol{U}_t^{(1)}, \boldsymbol{U}_t^{(2)}, \boldsymbol{U}_t^{(3)}\big) \cdot \boldsymbol{\mathcal{G}}_t$ satisfy*

$$\|\boldsymbol{\mathcal{X}}_t - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \le 0.03\rho^t \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star), \tag{18a}$$

$$\|\boldsymbol{\mathcal{X}}_t - \boldsymbol{\mathcal{X}}_\star\|_\infty \le 8\rho^t \sqrt{\frac{\mu^3 r_1 r_2 r_3}{n_1 n_2 n_3}}\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star), \tag{18b}$$

$$\|\boldsymbol{\mathcal{S}}_t - \boldsymbol{\mathcal{S}}_\star\|_\infty \le 16\rho^{t-1} \sqrt{\frac{\mu^3 r_1 r_2 r_3}{n_1 n_2 n_3}}\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star) \tag{18c}$$

for all $t \ge 0$, as long as the level of corruptions obeys $\alpha \le \frac{c_0}{\mu^2 r_1 r_2 r_3 \kappa}$ for some sufficiently small $c_0 > 0$.

The value of ρ was selected to simplify the proof and should not be taken as an optimal convergence rate. In a nutshell, Theorem 1 has the following immediate consequences:

- **Exact recovery.** Upon appropriate choices of the parameters, if the level of corruptions α is small enough, i.e. not exceeding the order of $\frac{1}{\mu^2 r_1 r_2 r_3 \kappa}$, we can ensure that the proposed Algorithm 1 exactly recovers the ground truth tensor $\boldsymbol{\mathcal{X}}_\star$ even when the gross corruptions are arbitrary and adversarial. As mentioned earlier, our result significantly enlarges the range of allowable corruption levels for exact recovery when the outliers are evenly distributed across the fibers, compared with the prior art established in [CLX21].
- **Constant linear rate of convergence.** The proposed ScaledGD algorithm (cf. Algorithm 1) finds the ground truth tensor at a *constant* linear rate, which is independent of the condition number, from a carefully designed spectral initialization. Consequently, the proposed ScaledGD algorithm inherits the computational robustness against ill-conditioning as [TMPB+22], even in the presence of gross outliers, as long as the thresholding operations are properly carried out.
- **Refined entrywise error guarantees.** Furthermore, when $\mu = O(1)$ and $r = O(1)$, the entrywise error bound (18b)—which is smaller than the Frobenius error (18a) by a factor of $\sqrt{\frac{1}{n_1 n_2 n_3}}$—suggests the errors are distributed in an evenly manner across the entries for incoherent and low-rank tensors. The same applies to the entrywise error bound of the sparse tensor (18c) which exhibits similar behavior as (18b). To the best of our knowledge, this is the first time such a refined entrywise error analysis is established for tensor RPCA.

3 Outline of the analysis

In this section, we outline the proof of Theorem 1. The proof is inductive in nature, where we aim to establish the following induction hypothesis at all the iterations:

$$\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \le \epsilon_0 \rho^t \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star), \tag{19a}$$

$$\max_k \left\{ \sqrt{\frac{n_k}{r_k}} \left\| (\boldsymbol{U}_t^{(k)} \boldsymbol{Q}_t^{(k)} - \boldsymbol{U}_\star^{(k)}) \boldsymbol{\Sigma}_\star^{(k)} \right\|_{2,\infty} \right\} \le \rho^t \sqrt{\mu} \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star), \tag{19b}$$

where $\rho = 1 - 0.45\eta$, $\epsilon_0 < 0.01$ is some sufficiently small constant, and $\{\boldsymbol{Q}_t^{(k)}\}_{k=1}^3$ are the optimal alignment matrices between $\boldsymbol{F}_t$ and $\boldsymbol{F}_\star$. The claims (18) in Theorem 1 follow immediately with the aid of Lemma 10 and Lemma 8 (see Appendix E). The following set of lemmas, whose proofs are deferred to Appendix A, establishes the induction hypothesis (19) for both the induction case and the base case.

Induction: local contraction. We start by outlining the local contraction of the proposed Algorithm 1, by establishing the induction hypothesis (19) continues to hold at the $(t+1)$-th iteration, assuming it holds at the t-th iteration, as long as the corruption level is not too large.

Lemma 1 (Distance contraction). *Let $\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{X}}_\star + \boldsymbol{\mathcal{S}}_\star \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, where $\boldsymbol{\mathcal{X}}_\star$ is μ-incoherent with multilinear rank $\boldsymbol{r} = (r_1, r_2, r_3)$, and $\boldsymbol{\mathcal{S}}_\star$ is α-sparse. Let $\boldsymbol{F}_t := \left(\boldsymbol{U}_t^{(1)}, \boldsymbol{U}_t^{(2)}, \boldsymbol{U}_t^{(3)}, \boldsymbol{\mathcal{G}}_t\right)$ be the t-th iterate of Algorithm 1. Suppose that the induction hypothesis (19) holds at the t-th iteration. Under the assumption $\alpha \le \frac{c_0 \epsilon_0}{\sqrt{\mu^3 r_1 r_2 r_3 r}}$ for some sufficiently small constant c_0 and the choice of ζ_{t+1} in Theorem 1, the $(t+1)$-th iterate $\boldsymbol{F}_{t+1}$ satisfies*

$$\operatorname{dist}(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) \le \epsilon_0 \rho^{t+1} \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$$

as long as $\eta \le 1/4$.

While Lemma 1 guarantees the contraction of the distance metric, the next Lemma 2 establishes the contraction of the incoherence metric, so that we can repeatedly apply Lemma 1 and Lemma 2 for induction.

Lemma 2 (Incoherence contraction). *Let $\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{X}}_\star + \boldsymbol{\mathcal{S}}_\star \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, where $\boldsymbol{\mathcal{X}}_\star$ is μ-incoherent with multilinear rank $\boldsymbol{r} = (r_1, r_2, r_3)$, and $\boldsymbol{\mathcal{S}}_\star$ is α-sparse. Let $\boldsymbol{F}_t := \left(\boldsymbol{U}_t^{(1)}, \boldsymbol{U}_t^{(2)}, \boldsymbol{U}_t^{(3)}, \boldsymbol{\mathcal{G}}_t\right)$ be the t-th iterate of Algorithm 1. Suppose that the induction hypothesis (19) holds at the t-th iteration. Under the assumption that $\alpha \le \frac{c_1}{\mu^2 r_1 r_2 r_3}$ for some sufficiently small constant c_1 and the choice of ζ_{t+1} in Theorem 1, the $(t+1)$-th iterate $\boldsymbol{F}_{t+1}$ satisfies*

$$\max_k \left\{ \sqrt{\frac{n_k}{r_k}} \left\| (\boldsymbol{U}_{t+1}^{(k)} \boldsymbol{Q}_{t+1}^{(k)} - \boldsymbol{U}_\star^{(k)}) \boldsymbol{\Sigma}_\star^{(k)} \right\|_{2,\infty} \right\} \le \rho^{t+1} \sqrt{\mu} \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$$

as long as $1/7 \le \eta \le 1/4$, where $\{\boldsymbol{Q}_t^{(k)}\}_{k=1}^3$ are the optimal alignment matrices between $\boldsymbol{F}_t$ and $\boldsymbol{F}_\star$.

It is worthwhile to note that, the local linear convergence of the proposed Algorithm 1, as ensured by the above two lemmas collectively, require the corruption level to not exceed the order of $\frac{1}{\mu^2 r_1 r_2 r_3}$, which is also independent of the condition number. Indeed, the range of the corruption level is mainly constrained by the spectral initialization, as demonstrated next.

Base case: spectral initialization. To establish the induction hypothesis, we still need to check the spectral initialization. The following lemmas state that the spectral initialization satisfies the induction hypothesis (19) at the base case $t = 0$, allowing us to invoke local contraction.

Lemma 3 (Distance at initialization). *Let $\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{X}}_\star + \boldsymbol{\mathcal{S}}_\star \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, where $\boldsymbol{\mathcal{X}}_\star$ is μ-incoherent with rank $\boldsymbol{r} = (r_1, r_2, r_3)$, and $\boldsymbol{\mathcal{S}}_\star$ is α-sparse. Let $\boldsymbol{F}_0 := \left(\boldsymbol{U}_0^{(1)}, \boldsymbol{U}_0^{(2)}, \boldsymbol{U}_0^{(3)}, \boldsymbol{\mathcal{G}}_0\right)$ be the output of spectral initialization with the threshold obeying $\|\boldsymbol{\mathcal{X}}_\star\|_\infty \le \zeta_0 \le 2\|\boldsymbol{\mathcal{X}}_\star\|_\infty$. If $\alpha \le \frac{c_0}{\sqrt{\mu^3 r_1 r_2 r_3 r \kappa}}$ for some constant $c_0 > 0$, we have*

$$\operatorname{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \le 54.1 c_0 \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star).$$

Lastly, the next lemma ensures that our initialization satisfies the incoherence condition, which requires nontrivial efforts to exploit the algebraic structures of the Tucker decomposition.

Lemma 4 (Incoherence at initialization). *Let $\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{X}}_\star + \boldsymbol{\mathcal{S}}_\star \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, where $\boldsymbol{\mathcal{X}}_\star$ is μ-incoherent with rank $\boldsymbol{r} = (r_1, r_2, r_3)$, and $\boldsymbol{\mathcal{S}}_\star$ is α-sparse. Let $\boldsymbol{F}_0 := (\boldsymbol{U}_0^{(1)}, \boldsymbol{U}_0^{(2)}, \boldsymbol{U}_0^{(3)}, \boldsymbol{\mathcal{G}}_0)$ be the output of spectral*

initialization with the threshold obeying $\|\boldsymbol{\mathcal{X}}_\star\|_\infty \le \zeta_0 \le 2\|\boldsymbol{\mathcal{X}}_\star\|_\infty$. *If* $\alpha \le \frac{c_0}{\mu^2 r_1 r_2 r_3 \kappa}$ *for some sufficiently small constant* c_0*, then the spectral initialization satisfies the incoherence condition*

$$\max_k \left\{ \sqrt{\frac{n_k}{r_k}} \left\| (\boldsymbol{U}_0^{(k)} \boldsymbol{Q}_0^{(k)} - \boldsymbol{U}_\star^{(k)}) \boldsymbol{\Sigma}_\star^{(k)} \right\|_{2,\infty} \right\} \le \sqrt{\mu} \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star),$$

where $\{\boldsymbol{Q}_0^{(k)}\}_{k=1}^3$ *are the optimal alignment matrices between* $\boldsymbol{F}_0$ *and* $\boldsymbol{F}_\star$.

4 Numerical experiments

4.1 Experiments on synthetic data

We begin with evaluating the phase transition performance of ScaledGD (cf. Algorithm 1) with respect to the multilinear rank and the level of corruption. For each κ, we randomly generate an $n \times n \times n$ tensor, with $n = 100$, multilinear rank $\boldsymbol{r} = (r, r, r)$, $r \in \{2, 5, 10, 20, \ldots, 80\}$, and level of corruption $\alpha \in \{0.1, 0.2, \ldots, 0.9, 1\}$. The factor matrices are generated uniformly at random with orthonormal columns, and a diagonal core tensor $\boldsymbol{\mathcal{G}}_\star$ is generated such that $[\boldsymbol{\mathcal{G}}_\star]_{i,i,i} = \kappa^{-(i-1)/(r-1)}$ for $i = 1, 2, \ldots, r$. We further randomly corrupt α-fraction of the entries, by adding uniformly sampled numbers from the range $[-\sum_{i,j,k} |[\boldsymbol{\mathcal{X}}_\star]_{i,j,k}|/n^3, \sum_{i,j,k} |[\boldsymbol{\mathcal{X}}_\star]_{i,j,k}|/n^3]$ to the selected entries, where $\sum_{i,j,k} |[\boldsymbol{\mathcal{X}}_\star]_{i,j,k}|/n^3$ is the mean of the entry-wise magnitudes of $\boldsymbol{\mathcal{X}}_\star$. To tune the constant step size η, and the hyperparameters ζ_0, ζ_1, and the decay rate ρ of the thresholding parameter for each tensor automatically, we used the Bayesian optimization method described in [ASY+19]. Specifically, we run the toolbox [ASY+19] for 200 trials or until the tuned parameters satisfy $\frac{\|\boldsymbol{\mathcal{X}}_T - \boldsymbol{\mathcal{X}}_\star\|_{\mathrm{F}}}{\|\boldsymbol{\mathcal{X}}_\star\|_{\mathrm{F}}} < 10^{-6}$ for $T = 200$, whichever happened first.

Figure 1 shows the log median of the relative reconstruction error $\frac{\|\boldsymbol{\mathcal{X}}_T - \boldsymbol{\mathcal{X}}_\star\|_{\mathrm{F}}}{\|\boldsymbol{\mathcal{X}}_\star\|_{\mathrm{F}}}$ when $T = 200$, over 20 random tensor realizations for $\kappa = 1, 5, 10$. Our results show a distinct negative linear relationship between the corruption level and the multilinear rank with regards to the final relative loss of ScaledGD. In particular, the performance is almost independent of the condition number κ, suggesting the performance of ScaledGD is indeed quite insensitive to the condition number.

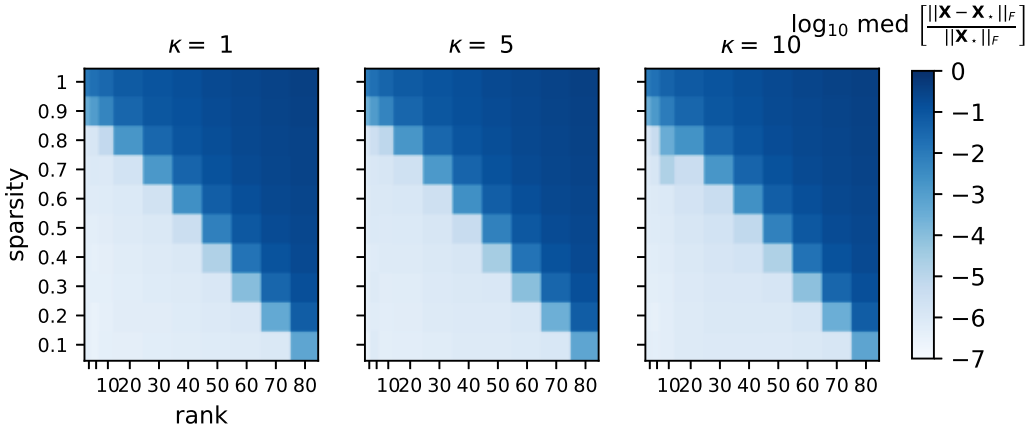


Figure 1: Log median of the relative reconstruction error of $\frac{\|\boldsymbol{\mathcal{X}}_T - \boldsymbol{\mathcal{X}}_\star\|_{\mathrm{F}}}{\|\boldsymbol{\mathcal{X}}_\star\|_{\mathrm{F}}}$ across 20 randomly generated tensors with varying ranks and levels of corruption when the condition number is set as $\kappa = 1, 5, 10$.

We further investigate the effect of the decay rate ρ of the thresholding parameters while fixing the other hyperparameters tuned as earlier. Using the same method, we generate a $100 \times 100 \times 100$ tensor with $\kappa = 5$ and 20% of the entries corrupted. Figure 2 shows the relative reconstruction error versus the iteration count using different decay rates ρ. It can be seen that ScaledGD enables exact recovery over a wide range of ρ as long as it is not too small. Moreover, within the range of decay rates that still admits exact recovery, the smaller ρ is, the faster ScaledGD converges. Note that the tuned decay rate $\rho \approx 0.931$ does not achieve the

fastest convergence rate since the stopping criteria for hyperparameter tuning were not set to optimize the convergence rate but some accuracy-speed trade-off.

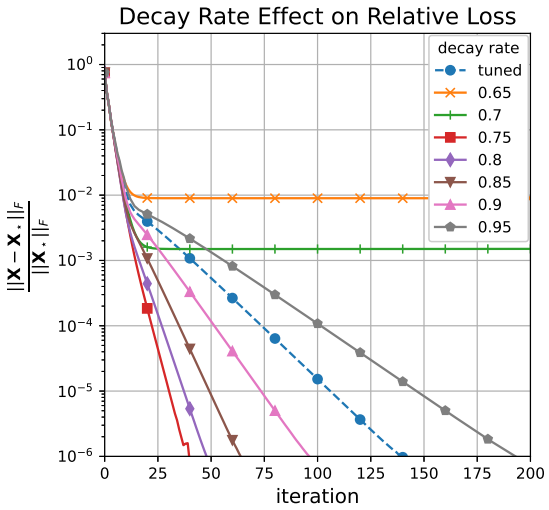


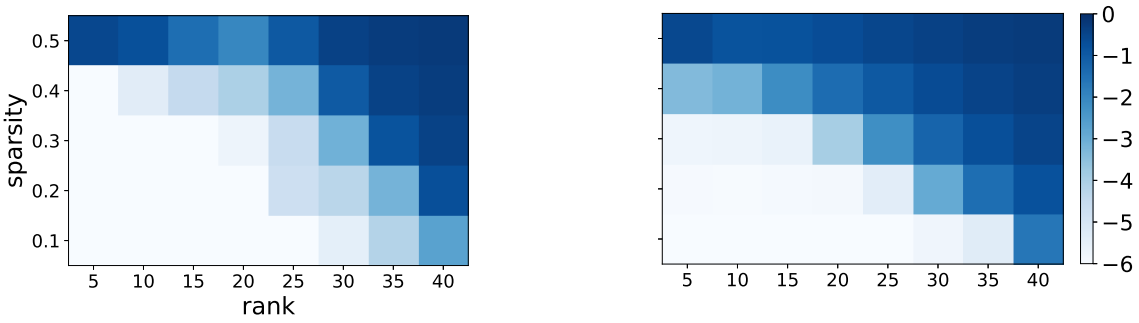
Figure 2: The relative reconstruction error $\frac{\|\boldsymbol{\mathcal{X}}_T-\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}}{\|\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}}$ with respect to the iteration count, when varying the decay rate ρ with other hyperparameters fixed.

Next, we also examine the performance of ScaledGD with shot noise, corruptions drawn from a Poisson distribution, with comparisons to the Riemannian gradient descent (RiemannianGD) algorithm in [CLX21]. More specifically, if corrupted, a ground truth entry $\boldsymbol{\mathcal{X}}_{i,j,k}$ has noise drawn from $10^{-5}\mathsf{Poisson}(10^5|\boldsymbol{\mathcal{X}}_{i,j,k}|)$ added to itself, where the 10^5 scaling is to encourage draws that are nonzero. This type of noise is non-negative and perturbs higher magnitude entries more. Note that the per-iteration cost of RiemannianGD is significantly higher than ours, due to the fact that it requires an evaluation of a rank-$\boldsymbol{r}$ HOSVD. Therefore, for this experiment, we generate tensors of a smaller size $50 \times 50 \times 50$ to accommodate the high computation need of RiemannianGD. We similarly tune the hyperparameters of RiemannianGD using the same method mentioned earlier for 100 trials. Figure 3 shows the log median of the relative reconstruction error $\frac{\|\boldsymbol{\mathcal{X}}_T-\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}}{\|\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}}$ when $T = 100$, over 20 random tensor realizations when $\kappa = 5$. It can be observed that the empirical performance of the two methods, indicated by the phase transition curves, are similar when tuned properly. However, the ScaledGD method is considerably faster, the difference of which is accentuated for larger and lower rank tensors, due to the fact that it works in the factor space and does not need to perform rank-$\boldsymbol{r}$ HOSVD at every iteration.

4.2 Image denoising and outlier detection

In this experiment, we examine the performance of ScaledGD for imaging denoising and outlier detection, with comparisons to the tensor RPCA algorithm proposed in [LFC+19] called TNN for their use of a newly defined tensor nuclear norm (TNN). We consider a sequence of handwritten digits "2" from the MNIST database [LCB10] containing 5958 images of size 28×28, leading to a 3-way tensor. We assume the tensor is low-rank along the image sequence, but not within the image for simplicity; in other words, the multilinear rank is assumed as $\boldsymbol{r} = (5, 28, 28)$. For both algorithms, the hyperparameters are best tuned by hands.

We examine the performance of ScaledGD and TNN when the image sequence is contaminated in the following scenarios: 1) 70% salt and pepper noise; 2) 500 out of the total images are randomly selected and swapped by random images from the entire MNIST training set; and 3) 50% salt and pepper noise and 500 randomly swapped images. Figure 4 demonstrates the performance of the compared algorithms on the first 100 instances for each corruption scenario. In all situations, ScaledGD recovers the low-rank component corresponding to the correct digit more accurately than TNN from a visual inspection. Furthermore, ScaledGD corrected the oddly-shaped or outlying digits to make the low-rank component be more homogeneous, but



(a) RiemannianGD (b) ScaledGD

Figure 3: Comparison between RiemannianGD [CLX21] and ScaledGD on the log median of the relative recovery error, $\frac{\|\boldsymbol{\mathcal{X}}_\star - \boldsymbol{\mathcal{X}}_T\|_{\mathrm{F}}}{\|\boldsymbol{\mathcal{X}}_\star\|_{\mathrm{F}}}$, across 20 randomly generated tensors with varying ranks and levels of shot noise corruption when the condition number is set as $\kappa = 5$.

TNN mostly preserved these cases in the low-rank output. More importantly, ScaledGD runs much faster as a scalable nonconvex approach, while TNN is more computationally expensive using convex optimization.

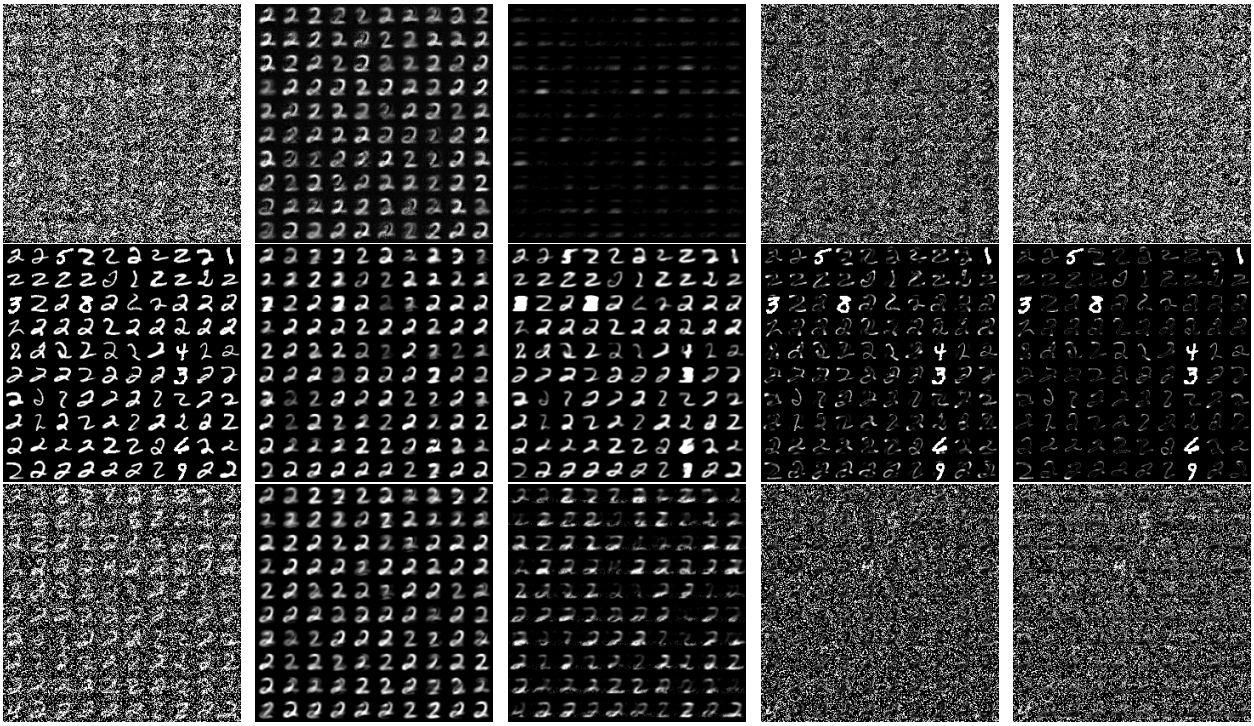
4.3 Background subtraction via selective mode update

We now apply ScaledGD to the task of background subtraction using videos from the VIRAT dataset [OHP+11], where the height and width of the videos are reduced by a factor of 4 due to hardware limitations. The video data can be thought as a multi-way tensor spanning across the height, width, frames, as well as different color channels of the scene. Here, the low-rank tensor corresponds to the background in the video which is fairly static over the frames, and the sparse tensor corresponds to the foreground containing moving objects which takes a small number of pixels. In particular, it is reasonable to assume that the background tensor is low-rank for the mode corresponding to the frames, but full rank in other modes. Motivated by this observation, one might be tempted to selectively only update the core tensor and the factor matrix corresponding to the mode for frames while keeping the other factor matrices fixed after the spectral initialization. We compare the results using this selective mode update strategy with the original ScaledGD algorithm in Figure 5, where the same hyperparameters are used in both. It can be seen that skipping updates of the full-rank factor matrices produced qualitatively similar results while gaining a significant per-iteration speed-up of about 4.6 to 5 times. We expect the speed improvement to be greater for larger tensors, as more computation can be bypassed.

5 Conclusions

In this paper, we proposed a new algorithm for tensor RPCA, which applies scaled gradient descent with an iteration-varying thresholding operation to adaptively remove the impact of corruptions. The algorithm is demonstrated to be fast, provably accurate, and achieve competitive performance over real-world applications. It opens several interesting directions for future research.

- *Dependency with the condition number from spectral initialization.* As seen from the analysis, the local linear convergence of Algorithm 1 succeeds under a larger range of the sparsity level α independent of the condition number κ. The constraint on α with respect to the condition number κ mainly stems from the spectral initialization, and it is of great interest to see if it is possible to refine the analysis in terms of the dependency with κ, which likely will require new tools.



(a) (b) (c) (d) (e)

Figure 4: Imaging denoising and outlier removal on an image sequence of handwritten digits with various corruption scenarios, using tensor RPCA via ScaledGD and TNN [LFC+19]. From top to bottom, the rows show results on the first 100 images when 1) corrupted with 70% salt and pepper noise; 2) 500 randomly swapped images; and 3) 50% salt and pepper noise and 500 randomly swapped images. From left to right, the columns show (a) the corrupted input, (b) the low-rank output of ScaledGD, (c) the low-rank output of TNN, (d) the sparse output of ScaledGD, and (e) the sparse output of TNN.

- *Missing data.* An important extension is to handle missing data in tensor RPCA, which seeks to recover a low-rank ground truth tensor from its partially observed entries possibly corrupted by gross errors. Our proposed algorithm can be adapted to this case in a straightforward fashion by considering the loss function defined only over the observed entries, and understanding its performance guarantees is a natural step.
- *Streaming data.* An equally interesting direction is to perform tensor RPCA over online and streaming data, where the fibers or slices of the tensor arrive sequentially over time, a situation that is common in data analytics [BCL18, VN18]. It is of great interest to develop low-complexity algorithms that can estimate and track the low-rank tensor factors as quickly as possible.
- *Hyperparameters.* The proposed algorithm contains several hyperparameters that need to be tuned carefully to fully unleash its potential. A recent follow-up [DSDC22] examined a learned approach based on algorithm unfolding and self-supervised learning to enable automatic hyperparameter tuning. In addition, understanding the performance when the rank is only imperfectly specified is also of great importance, which is closely related to [XSCM23].

Acknowledgements

The work of H. Dong, T. Tong and Y. Chi is supported in part by Office of Naval Research under N00014-19-1-2404, by Air Force Research Laboratory under FA8750-20-2-0504, by National Science Foundation under CAREER ECCS-1818571, CCF-1901199 and ECCS-2126634, and by Department of Transportation under

Figure 5: Examples of extracted background and foreground in video surveillance with and without selective mode updates. The first two columns use the same street surveillance video, and the last two columns use the same parking lot video, where the same frame is shown for each pair. In a pair, the left column is the result when using the original ScaledGD algorithm, and the right column is the result employing selective mode updates. The first row is the original frame, and second row is the sparse foreground, and the third row is the background.

693JJ321C000013. The work of H. Dong is also supported by CIT Dean's Fellowship, Liang Ji-Dian Graduate Fellowship, and Michel and Kathy Doreau Graduate Fellowship in Electrical and Computer Engineering at Carnegie Mellon University. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

References

[AGH+14] A. Anandkumar, R. Ge, D. Hsu, S. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832, 2014.

[AJSN16] A. Anandkumar, P. Jain, Y. Shi, and U. N. Niranjan. Tensor vs. matrix methods: Robust tensor decomposition under block sparse perturbations. In *Artificial Intelligence and Statistics*, pages 268–276. PMLR, 2016.

[ARB20] T. Ahmed, H. Raja, and W. U. Bajwa. Tensor regression using low-rank and sparse Tucker decompositions. *SIAM Journal on Mathematics of Data Science*, 2(4):944–966, 2020.

[ASY+19] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.

[BCL18] L. Balzano, Y. Chi, and Y. M. Lu. Streaming PCA and subspace tracking: The missing data case. *Proceedings of the IEEE*, 106(8):1293–1310, 2018.

[BL10] G. Bergqvist and E. G. Larsson. The higher-order singular value decomposition: Theory and an application [lecture notes]. *IEEE Signal Processing Magazine*, 27(3):151–154, 2010.

[CC14] Y. Chen and Y. Chi. Robust spectral compressed sensing via structured matrix completion. *IEEE Transactions on Information Theory*, 60(10):6576 – 6601, Oct. 2014.

[CC18] Y. Chen and Y. Chi. Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization. *IEEE Signal Processing Magazine*, 35(4):14 – 31, 2018.

[CCW19] H. Cai, J.-F. Cai, and K. Wei. Accelerated alternating projections for robust principal component analysis. *The Journal of Machine Learning Research*, 20(1):685–717, 2019.

[CFMY21] Y. Chen, J. Fan, C. Ma, and Y. Yan. Bridging convex and nonconvex optimization in robust PCA: Noise, outliers, and missing data. *The Annals of Statistics*, 49(5):2948–2971, 2021.

[CLC19] Y. Chi, Y. M. Lu, and Y. Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

[CLMW11] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):11:1–11:37, 2011.

[CLX21] J.-F. Cai, J. Li, and D. Xia. Generalized low-rank plus sparse tensor estimation by fast Riemannian optimization. *arXiv preprint arXiv:2103.08895*, 2021.

[CLY21] H. Cai, J. Liu, and W. Yin. Learned robust PCA: A scalable deep unfolding approach for high-dimensional outlier detection. *Advances in Neural Information Processing Systems*, 34, 2021.

[CSPW11] V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

[DBBG19] D. Driggs, S. Becker, and J. Boyd-Graber. Tensor robust principal component analysis: Better recovery with atomic norm regularization. *arXiv preprint arXiv:1901.10991*, 2019.

[DLDMV00] L. De Lathauwer, B. De Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.

[DSDC22] H. Dong, M. Shah, S. Donegan, and Y. Chi. Deep unfolded tensor robust PCA with self-supervised learning. *arXiv preprint arXiv:2212.11346*, 2022.

[FL18] S. Friedland and L.-H. Lim. Nuclear norm of higher-order tensors. *Mathematics of Computation*, 87(311):1255–1281, 2018.

[GGH14] Q. Gu, H. Gui, and J. Han. Robust tensor decomposition with gross corruption. *Advances in Neural Information Processing Systems*, 27, 2014.

[GQ14] D. Goldfarb and Z. Qin. Robust low-rank tensor recovery: Models and algorithms. *SIAM Journal on Matrix Analysis and Applications*, 35(1):225–253, 2014.

[GWL16] Q. Gu, Z. Wang, and H. Liu. Low-rank and sparse structure pursuit via alternating minimization. In *Artificial Intelligence and Statistics*, pages 600–609, 2016.

[HWZ20] R. Han, R. Willett, and A. Zhang. An optimal statistical and computational framework for generalized tensor estimation. *arXiv preprint arXiv:2002.11255*, 2020.

[KABO10] A. Karatzoglou, X. Amatriain, L. Baltrunas, and N. Oliver. Multiverse recommendation: n-dimensional tensor factorization for context-aware collaborative filtering. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 79–86, 2010.

[KB09] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[LCB10] Y. LeCun, C. Cortes, and C. Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[LCM10] Z. Lin, M. Chen, and Y. Ma. The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.

[LFC+16] C. Lu, J. Feng, Y. Chen, W. Liu, Z. Lin, and S. Yan. Tensor robust principal component analysis: Exact recovery of corrupted low-rank tensors via convex optimization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5249–5257, 2016.

[LFC+19] C. Lu, J. Feng, Y. Chen, W. Liu, Z. Lin, and S. Yan. Tensor robust principal component analysis with a new tensor nuclear norm. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):925–938, 2019.

[LMWY12] J. Liu, P. Musialski, P. Wonka, and J. Ye. Tensor completion for estimating missing values in visual data. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 35(1):208–220, 2012.

[LWQ+15] S. Li, W. Wang, H. Qi, B. Ayhan, C. Kwan, and S. Vance. Low-rank tensor decomposition based anomaly detection for hyperspectral imagery. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 4525–4529. IEEE, 2015.

[LZ21] Y. Luo and A. R. Zhang. Low-rank tensor estimation via Riemannian Gauss-Newton: Statistical optimality and second-order convergence. *arXiv preprint arXiv:2104.12031*, 2021.

[NNS+14] P. Netrapalli, U. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain. Non-convex robust PCA. In *Advances in Neural Information Processing Systems*, pages 1107–1115, 2014.

[OHP+11] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis, et al. A large-scale benchmark dataset for event recognition in surveillance video. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3153–3160, 2011.

[PFS16] E. E. Papalexakis, C. Faloutsos, and N. D. Sidiropoulos. Tensors for data mining and data fusion: Models, applications, and scalable algorithms. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(2):1–44, 2016.

[TMC21a] T. Tong, C. Ma, and Y. Chi. Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *Journal of Machine Learning Research*, 22(150):1–63, 2021.

[TMC21b] T. Tong, C. Ma, and Y. Chi. Low-rank matrix recovery with scaled subgradient methods: Fast and robust convergence without the condition number. *IEEE Transactions on Signal Processing*, 69:2396–2409, 2021.

[TMC22] T. Tong, C. Ma, and Y. Chi. Accelerating ill-conditioned robust low-rank tensor regression. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2022.

[TMPB+22] T. Tong, C. Ma, A. Prater-Bennette, E. Tripp, and Y. Chi. Scaling and scalability: Provable nonconvex low-rank tensor estimation from incomplete measurements. *Journal of Machine Learning Research*, 23(163):1–77, 2022.

[TY10] K.-C. Toh and S. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of optimization*, 6(615-640):15, 2010.

[VN18] N. Vaswani and P. Narayanamurthy. Static and dynamic robust PCA and matrix completion: A review. *Proceedings of the IEEE*, 106(8):1359–1379, 2018.

[VVM12] N. Vannieuwenhoven, R. Vandebril, and K. Meerbergen. A new truncation strategy for the higher-order singular value decomposition. *SIAM Journal on Scientific Computing*, 34(2):A1027–A1052, 2012.

[WCW21] H. Wang, J. Chen, and K. Wei. Entrywise convergence of riemannian gradient method for low rank tensor completion via tucker decomposition. *arXiv preprint arXiv:2108.07899*, 2021.

[WGR+09] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. *Advances in neural information processing systems*, 22, 2009.

[XSCM23] X. Xu, Y. Shen, Y. Chi, and C. Ma. The power of preconditioning in overparameterized low-rank matrix sensing. *arXiv preprint arXiv:2302.01186*, 2023.

[XY19] D. Xia and M. Yuan. On polynomial time methods for exact low-rank tensor completion. *Foundations of Computational Mathematics*, 19(6):1265–1313, 2019.

[YPCC16] X. Yi, D. Park, Y. Chen, and C. Caramanis. Fast algorithms for robust PCA via gradient descent. In *Advances in Neural Information Processing Systems*, pages 4152–4160, 2016.

[YZ16] M. Yuan and C.-H. Zhang. On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics*, 16(4):1031–1068, 2016.

[ZWZM19] X. Zhang, D. Wang, Z. Zhou, and Y. Ma. Robust low-rank tensor recovery with rectification and alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):238–255, 2019.

A Proof of Lemma 1

Since $\operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) < \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$, [TMPB+22, Lemma 6] ensures that the optimal alignment matrices $\{\boldsymbol{Q}_t^{(k)}\}_{k=1}^3$ between $\boldsymbol{F}_t$ and $\boldsymbol{F}_\star$ exist. Since $\{\boldsymbol{Q}_t^{(k)}\}_{k=1}^3$ may be a suboptimal alignment between $\boldsymbol{F}_{t+1}$ and $\boldsymbol{F}_\star$, we therefore have

$$\operatorname{dist}^2(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) \le \sum_{k=1}^3 \left\| (\boldsymbol{U}_{t+1}^{(k)} \boldsymbol{Q}_t^{(k)} - \boldsymbol{U}_\star^{(k)}) \boldsymbol{\Sigma}_\star^{(k)} \right\|_{\mathsf{F}}^2 + \left\| ((\boldsymbol{Q}_t^{(1)})^{-1}, (\boldsymbol{Q}_t^{(2)})^{-1}, (\boldsymbol{Q}_t^{(3)})^{-1}) \cdot \boldsymbol{\mathcal{G}}_{t+1} - \boldsymbol{\mathcal{G}}_\star \right\|_{\mathsf{F}}^2. \tag{20}$$

Before we embark on the control of the terms on the right hand side of (20), we introduce the following short-hand notations:

$$\begin{aligned} \boldsymbol{U}^{(k)} &:= \boldsymbol{U}_t^{(k)} \boldsymbol{Q}_t^{(k)}, & \breve{\boldsymbol{U}}^{(k)} &:= \breve{\boldsymbol{U}}_t^{(k)} (\boldsymbol{Q}_t^{(k)})^{-\top}, & \boldsymbol{\mathcal{S}} &:= \boldsymbol{\mathcal{S}}_{t+1}, \\ \boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}} &:= \boldsymbol{U}^{(k)} - \boldsymbol{U}_\star^{(k)}, & \boldsymbol{\Delta}_{\breve{\boldsymbol{U}}^{(k)}} &:= \breve{\boldsymbol{U}}^{(k)} - \breve{\boldsymbol{U}}_\star^{(k)}, & \boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}} &:= \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{S}}_\star, \\ \boldsymbol{\mathcal{G}} &:= ((\boldsymbol{Q}_t^{(1)})^{-1}, (\boldsymbol{Q}_t^{(2)})^{-1}, (\boldsymbol{Q}_t^{(3)})^{-1}) \cdot \boldsymbol{\mathcal{G}}_t, & \boldsymbol{\Delta}_{\boldsymbol{\mathcal{G}}} &:= \boldsymbol{\mathcal{G}} - \boldsymbol{\mathcal{G}}_\star. \end{aligned} \tag{21}$$

In addition, Lemma 10 in conjunction with the induction hypothesis (19) tells us that

$$\|\boldsymbol{\mathcal{X}}_t - \boldsymbol{\mathcal{X}}_\star\|_\infty \le 8\sqrt{\frac{\mu^3 r_1 r_2 r_3}{n_1 n_2 n_3}} \rho^t \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star) =: \zeta_{t+1}. \tag{22}$$

Step 1: bounding the first term of (20). Using the update rule (16b) for $\boldsymbol{U}_{t+1}^{(k)}$, we have

$$\begin{aligned} &(\boldsymbol{U}_{t+1}^{(k)} \boldsymbol{Q}_t^{(k)} - \boldsymbol{U}_\star^{(k)}) \boldsymbol{\Sigma}_\star^{(k)} \\ &= \left[\left((1-\eta) \boldsymbol{U}_t^{(k)} - \eta \big(\mathcal{M}_k(\boldsymbol{\mathcal{S}}_{t+1}) - \mathcal{M}_k(\boldsymbol{\mathcal{Y}}) \big) \breve{\boldsymbol{U}}_t^{(k)} \big(\breve{\boldsymbol{U}}_t^{(k)\top} \breve{\boldsymbol{U}}_t^{(k)} \big)^{-1} \right) \boldsymbol{Q}_t^{(k)} - \boldsymbol{U}_\star^{(k)} \right] \boldsymbol{\Sigma}_\star^{(k)} \\ &= \left[\left((1-\eta) \boldsymbol{U}_t^{(k)} - \eta \big(\mathcal{M}_k(\boldsymbol{\mathcal{S}}_{t+1} - \boldsymbol{\mathcal{S}}_\star) - \boldsymbol{U}_\star^{(k)} \breve{\boldsymbol{U}}_\star^{(k)\top} \big) \breve{\boldsymbol{U}}_t^{(k)} \big(\breve{\boldsymbol{U}}_t^{(k)\top} \breve{\boldsymbol{U}}_t^{(k)} \big)^{-1} \right) \boldsymbol{Q}_t^{(k)} - \boldsymbol{U}_\star^{(k)} \right] \boldsymbol{\Sigma}_\star^{(k)}, \end{aligned}$$

where the second equality follows from $\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{X}}_\star + \boldsymbol{\mathcal{S}}_\star$ as well as the matricization property (4). With the set of notation (21) in place, simple algebraic simplifications yield

$$\begin{aligned} (\boldsymbol{U}_{t+1}^{(k)} \boldsymbol{Q}_t^{(k)} - \boldsymbol{U}_\star^{(k)}) \boldsymbol{\Sigma}_\star^{(k)} &= \left[(1-\eta) \boldsymbol{U}^{(k)} - \eta \left(\mathcal{M}_k(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}) - \boldsymbol{U}_\star^{(k)} \breve{\boldsymbol{U}}_\star^{(k)\top} \right) \breve{\boldsymbol{U}}^{(k)} \big(\breve{\boldsymbol{U}}^{(k)\top} \breve{\boldsymbol{U}}^{(k)} \big)^{-1} - \boldsymbol{U}_\star^{(k)} \right] \boldsymbol{\Sigma}_\star^{(k)} \\ &= \left[(1-\eta) \boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}} - \eta \left(\mathcal{M}_k(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}) + \boldsymbol{U}_\star^{(k)} \boldsymbol{\Delta}_{\breve{\boldsymbol{U}}^{(k)}}^\top \right) \breve{\boldsymbol{U}}^{(k)} \big(\breve{\boldsymbol{U}}^{(k)\top} \breve{\boldsymbol{U}}^{(k)} \big)^{-1} \right] \boldsymbol{\Sigma}_\star^{(k)} \end{aligned}$$

$$= (1-\eta)\boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}}\boldsymbol{\Sigma}_\star^{(k)} - \eta\left(\mathcal{M}_k\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right) + \boldsymbol{U}_\star^{(k)}\boldsymbol{\Delta}_{\breve{\boldsymbol{U}}^{(k)}}^\top\right)\breve{\boldsymbol{U}}^{(k)}\big(\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}\big)^{-1}\boldsymbol{\Sigma}_\star^{(k)}. \quad (23)$$

Detailed in Appendix A.1, we claim the following bound holds:

$$\sum_{k=1}^3\left\|(\boldsymbol{U}_{t+1}^{(k)}\boldsymbol{Q}_t^{(k)} - \boldsymbol{U}_\star^{(k)})\boldsymbol{\Sigma}_\star^{(k)}\right\|_{\mathsf{F}}^2 \le (1-\eta)^2\sum_{k=1}^3\left\|\boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}}\boldsymbol{\Sigma}_\star^{(k)}\right\|_{\mathsf{F}}^2 + 0.15\eta(1-\eta)\sum_{k=1}^3\left\|\boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}}\boldsymbol{\Sigma}_\star^{(k)}\right\|_{\mathsf{F}}\epsilon_0\rho^t\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$$
$$+ 2\eta^2\epsilon_0^2\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star) + 0.06\eta(1-\eta)\epsilon_0^2\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star). \quad (24)$$

Step 2: bounding the second term of (20). Using the update rule (16c) for $\boldsymbol{\mathcal{G}}_{t+1}$, we have

$$\begin{aligned}
&\big((\boldsymbol{Q}_t^{(1)})^{-1},(\boldsymbol{Q}_t^{(2)})^{-1},(\boldsymbol{Q}_t^{(3)})^{-1}\big)\cdot\boldsymbol{\mathcal{G}}_{t+1} - \boldsymbol{\mathcal{G}}_\star\\
&= (1-\eta)\big((\boldsymbol{Q}_t^{(1)})^{-1},(\boldsymbol{Q}_t^{(2)})^{-1},(\boldsymbol{Q}_t^{(3)})^{-1}\big)\cdot\boldsymbol{\mathcal{G}}_t\\
&\quad - \eta\left(\big(\boldsymbol{U}_t^{(1)\top}\boldsymbol{U}_t^{(1)}\boldsymbol{Q}_t^{(1)}\big)^{-1}\boldsymbol{U}_t^{(1)\top},\big(\boldsymbol{U}_t^{(2)\top}\boldsymbol{U}_t^{(2)}\boldsymbol{Q}_t^{(2)}\big)^{-1}\boldsymbol{U}_t^{(2)\top},\big(\boldsymbol{U}_t^{(3)\top}\boldsymbol{U}_t^{(3)}\boldsymbol{Q}_t^{(3)}\big)^{-1}\boldsymbol{U}_t^{(3)\top}\right)\cdot(\boldsymbol{\mathcal{S}}_{t+1}-\boldsymbol{\mathcal{Y}}) - \boldsymbol{\mathcal{G}}_\star\\
&= (1-\eta)\boldsymbol{\mathcal{G}} - \eta\left(\big(\boldsymbol{U}^{(1)\top}\boldsymbol{U}^{(1)}\big)^{-1}\boldsymbol{U}^{(1)\top},\big(\boldsymbol{U}^{(2)\top}\boldsymbol{U}^{(2)}\big)^{-1}\boldsymbol{U}^{(2)\top},\big(\boldsymbol{U}^{(3)\top}\boldsymbol{U}^{(3)}\big)^{-1}\boldsymbol{U}^{(3)\top}\right)\cdot(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}-\boldsymbol{\mathcal{X}}_\star) - \boldsymbol{\mathcal{G}}_\star\\
&= (1-\eta)\boldsymbol{\Delta}_{\boldsymbol{\mathcal{G}}} - \eta\left(\big(\boldsymbol{U}^{(1)\top}\boldsymbol{U}^{(1)}\big)^{-1}\boldsymbol{U}^{(1)\top},\big(\boldsymbol{U}^{(2)\top}\boldsymbol{U}^{(2)}\big)^{-1}\boldsymbol{U}^{(2)\top},\big(\boldsymbol{U}^{(3)\top}\boldsymbol{U}^{(3)}\big)^{-1}\boldsymbol{U}^{(3)\top}\right)\\
&\qquad\cdot\left(\big(\boldsymbol{U}^{(1)},\boldsymbol{U}^{(2)},\boldsymbol{U}^{(3)}\big)\cdot\boldsymbol{\mathcal{G}}_\star - \boldsymbol{\mathcal{X}}_\star + \boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right), \quad (25)
\end{aligned}$$

where the last two lines make use of the short-hand notation in (21), as well as the multilinear property (3). Detailed in Appendix A.2, we claim the following bound holds:

$$\left\|\big((\boldsymbol{Q}_t^{(1)})^{-1},(\boldsymbol{Q}_t^{(2)})^{-1},(\boldsymbol{Q}_t^{(3)})^{-1}\big)\cdot\boldsymbol{\mathcal{G}}_{t+1} - \boldsymbol{\mathcal{G}}_\star\right\|_{\mathsf{F}}^2 \le (1-\eta)^2\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{G}}}\right\|_{\mathsf{F}}^2 + 2\cdot 0.15\eta(1-\eta)\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{G}}}\right\|_{\mathsf{F}}\epsilon_0\rho^t\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$$
$$+ 0.02\eta(1-\eta)\epsilon_0^2\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star) + 0.06\eta^2\epsilon_0^2\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star). \quad (26)$$

Step 3: putting the bounds together. Plugging the bounds (24) and (26) into (20) yields

$$\begin{aligned}
&\mathrm{dist}^2(\boldsymbol{F}_{t+1},\boldsymbol{F}_\star)\\
&\le (1-\eta)^2\left(\sum_{k=1}^3\left\|\boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}}\boldsymbol{\Sigma}_\star^{(k)}\right\|_{\mathsf{F}}^2 + \left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{G}}}\right\|_{\mathsf{F}}^2\right) + 0.15\eta(1-\eta)\left(\sum_{k=1}^3\left\|\boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}}\boldsymbol{\Sigma}_\star^{(k)}\right\|_{\mathsf{F}} + 2\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{G}}}\right\|_{\mathsf{F}}\right)\epsilon_0\rho^t\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)\\
&\quad + 2.1\eta^2\epsilon_0^2\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star) + 0.08\eta(1-\eta)\epsilon_0^2\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star)\\
&\overset{\text{(i)}}{\le} (1-\eta)^2\,\mathrm{dist}^2(\boldsymbol{F}_t,\boldsymbol{F}_\star) + 0.4\eta(1-\eta)\,\mathrm{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star)\epsilon_0\rho^t\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star) + 2.1\eta^2\epsilon_0^2\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star) \quad (27)\\
&\quad + 0.08\eta(1-\eta)\epsilon_0^2\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star)\\
&\overset{\text{(ii)}}{\le} ((1-\eta)^2 + 0.4\eta(1-\eta) + 2.1\eta^2 + 0.08\eta(1-\eta))\epsilon_0^2\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star)\\
&= ((1-\eta)^2 + 0.5\eta(1-\eta) + 2.1\eta^2)\epsilon_0^2\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star), \quad (28)
\end{aligned}$$

where (i) follows from the definition of $\mathrm{dist}^2(\boldsymbol{F}_t,\boldsymbol{F}_\star)$ and Cauchy-Schwarz, and (ii) follows from the induction hypothesis $\mathrm{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star)\le\epsilon_0\rho^t\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$. For $0<\eta\le 1/4$ and $\rho = 1-0.45\eta$, this simplifies to the claimed bound

$$\mathrm{dist}^2(\boldsymbol{F}_{t+1},\boldsymbol{F}_\star) \le (1-0.45\eta)^2\epsilon_0^2\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star) = \epsilon_0^2\rho^{2t+2}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star).$$

A.1 Proof of (24)

Taking the squared norm on both sides of (23), we obtain

$$\left\|(\boldsymbol{U}_{t+1}^{(k)}\boldsymbol{Q}_{t,k} - \boldsymbol{U}_\star^{(k)})\boldsymbol{\Sigma}_\star^{(k)}\right\|_{\mathsf{F}}^2 = (1-\eta)^2\left\|\boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}}\boldsymbol{\Sigma}_\star^{(k)}\right\|_{\mathsf{F}}^2 + \eta^2\underbrace{\left\|\mathcal{M}_k\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right)\breve{\boldsymbol{U}}^{(k)}\big(\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}\big)^{-1}\boldsymbol{\Sigma}_\star^{(k)}\right\|_{\mathsf{F}}^2}_{=:\mathfrak{A}_{1,k}}$$

$$
\begin{aligned}
&- 2\eta(1-\eta)\underbrace{\left\langle \boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}}\boldsymbol{\Sigma}_\star^{(k)}, \boldsymbol{U}_\star^{(k)}\boldsymbol{\Delta}_{\breve{\boldsymbol{U}}^{(k)}}^\top \breve{\boldsymbol{U}}^{(k)}\big(\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}\big)^{-1}\boldsymbol{\Sigma}_\star^{(k)}\right\rangle}_{=:\mathfrak{A}_{2,k}} \\
&+ \eta^2\underbrace{\left\|\boldsymbol{U}_\star^{(k)}\boldsymbol{\Delta}_{\breve{\boldsymbol{U}}^{(k)}}^\top \breve{\boldsymbol{U}}^{(k)}\big(\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}\big)^{-1}\boldsymbol{\Sigma}_\star^{(k)}\right\|_{\mathsf{F}}^2}_{=:\mathfrak{A}_{3,k}} \\
&- 2\eta(1-\eta)\underbrace{\left\langle \boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}}\boldsymbol{\Sigma}_\star^{(k)}, \mathcal{M}_k\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right)\breve{\boldsymbol{U}}^{(k)}\big(\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}\big)^{-1}\boldsymbol{\Sigma}_\star^{(k)}\right\rangle}_{=:\mathfrak{A}_{4,k}} \\
&+ 2\eta^2\underbrace{\left\langle \mathcal{M}_k\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right)\breve{\boldsymbol{U}}^{(k)}\big(\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}\big)^{-1}\boldsymbol{\Sigma}_\star^{(k)}, \boldsymbol{U}_\star^{(k)}\boldsymbol{\Delta}_{\breve{\boldsymbol{U}}^{(k)}}^\top \breve{\boldsymbol{U}}^{(k)}\big(\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}\big)^{-1}\boldsymbol{\Sigma}_\star^{(k)}\right\rangle}_{=:\mathfrak{A}_{5,k}}.
\end{aligned}
\tag{29}
$$

In the sequel, we shall bound each term separately.

- **Bounding $\mathfrak{A}_{1,k}$.** Since the quantity inside the norm is of rank r_k, we have

$$
\begin{aligned}
\mathfrak{A}_{1,k} &= \left\|\mathcal{M}_k\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right)\breve{\boldsymbol{U}}^{(k)}\big(\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}\big)^{-1}\boldsymbol{\Sigma}_\star^{(k)}\right\|_{\mathsf{F}}^2 \\
&\le r_k\left\|\mathcal{M}_k\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right)\breve{\boldsymbol{U}}^{(k)}\big(\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}\big)^{-1}\boldsymbol{\Sigma}_\star^{(k)}\right\|^2 \\
&\le r_k\left\|\mathcal{M}_k\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right)\right\|^2\left\|\breve{\boldsymbol{U}}^{(k)}\big(\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}\big)^{-1}\boldsymbol{\Sigma}_\star^{(k)}\right\|^2 \le \frac{r_k}{(1-\epsilon_0)^6}\left\|\mathcal{M}_k\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right)\right\|^2,
\end{aligned}
$$

  where the last inequality follows from Lemma 8 (cf. (70c)). To continue, notice that the choice of ζ_{t+1} (cf. (22)) guarantees that $\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}$ (and hence $\mathcal{M}_k\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right)$) is α-sparse (cf. Lemma 12). This allows us to invoke Lemma 11 and obtain

$$
\left\|\mathcal{M}_k\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right)\right\| \le \alpha\sqrt{n_1n_2n_3}\left\|\mathcal{M}_k\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right)\right\|_\infty = \alpha\sqrt{n_1n_2n_3}\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right\|_\infty \le 2\alpha\sqrt{n_1n_2n_3}\zeta_{t+1}. \tag{30}
$$

  Plugging this into the previous inequality, we arrive at

$$
\mathfrak{A}_{1,k} \le \frac{4\alpha^2 n_1n_2n_3r_k}{(1-\epsilon_0)^6}\zeta_{t+1}^2 \le \frac{256\alpha^2\mu^3r_1r_2r_3r_k}{(1-\epsilon_0)^6}\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star),
$$

  where the second inequality follows from the choice of ζ_{t+1} (cf. (22)). Finally, with the assumption on the sparsity level $\alpha \le \frac{c_0\epsilon_0}{\sqrt{\mu^3r_1r_2r_3r}}$, we have

$$
\mathfrak{A}_{1,k} \le \frac{256c_0^2}{(1-\epsilon_0)^6}\epsilon_0^2\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star) \le 0.02\epsilon_0^2\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star) \tag{31}
$$

  for sufficiently small c_0 and $\epsilon_0 < 0.01$.

- **Bounding $\mathfrak{A}_{2,k}$.** This term is identical to the term that is bounded in [TMPB+22, Section B.1], which obeys

$$
\mathfrak{A}_{2,k} \ge \left\langle \boldsymbol{\mathcal{K}}^{(k)}, \boldsymbol{\mathcal{K}}^{(1)}+\boldsymbol{\mathcal{K}}^{(2)}+\boldsymbol{\mathcal{K}}^{(3)}\right\rangle - C_1\epsilon_0\rho^t\,\mathrm{dist}^2(\boldsymbol{F}_t,\boldsymbol{F}_\star)
$$

  for some constant $C_1 > 1$ with

$$
\begin{aligned}
\boldsymbol{\mathcal{K}}^{(1)} &:= \big(\boldsymbol{U}_\star^{(1)\top}\boldsymbol{\Delta}_{\boldsymbol{U}^{(1)}}, \boldsymbol{I}_{r_2}, \boldsymbol{I}_{r_3}\big)\cdot\boldsymbol{\mathcal{G}}_\star, \\
\boldsymbol{\mathcal{K}}^{(2)} &:= \big(\boldsymbol{I}_{r_1}, \boldsymbol{U}_\star^{(2)\top}\boldsymbol{\Delta}_{\boldsymbol{U}^{(2)}}, \boldsymbol{I}_{r_3}\big)\cdot\boldsymbol{\mathcal{G}}_\star, \\
\boldsymbol{\mathcal{K}}^{(3)} &:= \big(\boldsymbol{I}_{r_1}, \boldsymbol{I}_{r_2}, \boldsymbol{U}_\star^{(3)\top}\boldsymbol{\Delta}_{\boldsymbol{U}^{(3)}}\big)\cdot\boldsymbol{\mathcal{G}}_\star.
\end{aligned}
$$

  As long as the choice of ϵ_0 is small enough such that $C_1\epsilon_0\rho^t \le C_1\epsilon_0 < 0.01$, the induction hypothesis (19a) tells us that.

$$
\mathfrak{A}_{2,k} \ge \left\langle \boldsymbol{\mathcal{K}}^{(k)}, \boldsymbol{\mathcal{K}}^{(1)}+\boldsymbol{\mathcal{K}}^{(2)}+\boldsymbol{\mathcal{K}}^{(3)}\right\rangle - 0.01\epsilon_0^2\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star). \tag{32}
$$

- **Bounding $\mathfrak{A}_3$.** This term is identical to the term that is bounded in [TMPB+22, Section B.2], which obeys

$$\mathfrak{A}_{3,k} \le \left\|\boldsymbol{\mathcal{K}}^{(1)} + \boldsymbol{\mathcal{K}}^{(2)} + \boldsymbol{\mathcal{K}}^{(3)}\right\|_{\mathsf{F}}^2 + C_2\epsilon_0\rho^t \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star)$$

for some constant $C_2 > 1$. As long as the choice of ϵ_0 is small enough such that $C_2\epsilon_0\rho^t \le C_2\epsilon_0 < 0.01$, the induction hypothesis (19a) results in

$$\mathfrak{A}_{3,k} \le \left\|\boldsymbol{\mathcal{K}}^{(1)} + \boldsymbol{\mathcal{K}}^{(2)} + \boldsymbol{\mathcal{K}}^{(3)}\right\|_{\mathsf{F}}^2 + 0.01\epsilon_0^2\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star). \tag{33}$$

- **Bounding $\mathfrak{A}_{4,k}$.** To control $|\mathfrak{A}_{4,k}|$, we apply the definition of the matrix inner product to rewrite it as

$$\begin{aligned}
|\mathfrak{A}_{4,k}| &= \left|\operatorname{tr}\left(\mathcal{M}_k\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right)\breve{\boldsymbol{U}}^{(k)}\left(\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}\right)^{-1}\left(\boldsymbol{\Sigma}_\star^{(k)}\right)^2\boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}}^\top\right)\right| \\
&\le \left\|\mathcal{M}_k\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right)\right\|\left\|\breve{\boldsymbol{U}}^{(k)}\left(\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}\right)^{-1}\left(\boldsymbol{\Sigma}_\star^{(k)}\right)^2\boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}}^\top\right\|_* \\
&\le 2\alpha\sqrt{n_1n_2n_3}\zeta_{t+1}\left\|\breve{\boldsymbol{U}}^{(k)}\left(\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}\right)^{-1}\left(\boldsymbol{\Sigma}_\star^{(k)}\right)^2\boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}}^\top\right\|_*,
\end{aligned}$$

where the second line follows from Hölder's inequality, and the third line follows from the bound of $\|\mathcal{M}_k\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right)\|$ from (30). To bound the remaining term, we have

$$\begin{aligned}
\left\|\breve{\boldsymbol{U}}^{(k)}\left(\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}\right)^{-1}\left(\boldsymbol{\Sigma}_\star^{(k)}\right)^2\boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}}^\top\right\|_* &\le \sqrt{r_k}\left\|\breve{\boldsymbol{U}}^{(k)}\left(\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}\right)^{-1}\left(\boldsymbol{\Sigma}_\star^{(k)}\right)^2\boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}}^\top\right\|_{\mathsf{F}} \\
&\le \sqrt{r_k}\left\|\breve{\boldsymbol{U}}^{(k)}\left(\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}\right)^{-1}\boldsymbol{\Sigma}_\star^{(k)}\right\|\left\|\boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}}\boldsymbol{\Sigma}_\star^{(k)}\right\|_{\mathsf{F}} \\
&\le \frac{\sqrt{r_k}}{(1-\epsilon_0)^3}\left\|\boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}}\boldsymbol{\Sigma}_\star^{(k)}\right\|_{\mathsf{F}},
\end{aligned}$$

where the last line follows from Lemma 8 (cf. (70c)). Plugging this into the previous inequality, we reach

$$|\mathfrak{A}_{4,k}| \le \frac{2\alpha\zeta_{t+1}\sqrt{n_1n_2n_3r_k}}{(1-\epsilon_0)^3}\left\|\boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}}\boldsymbol{\Sigma}_\star^{(k)}\right\|_{\mathsf{F}} \le \frac{16\alpha\sqrt{\mu^3r_1r_2r_3r_k}}{(1-\epsilon_0)^3}\rho^t\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)\left\|\boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}}\boldsymbol{\Sigma}_\star^{(k)}\right\|_{\mathsf{F}},$$

where the second inequality follows from the choice of ζ_{t+1} (cf. (22)). Finally, with the assumption on the sparsity level $\alpha \le \frac{c_0\epsilon_0}{\sqrt{\mu^3r_1r_2r_3r}}$ for sufficiently small c_0 and $\epsilon_0 < 0.01$, we have

$$|\mathfrak{A}_{4,k}| \le \frac{16c_0}{(1-\epsilon_0)^3}\left\|\boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}}\boldsymbol{\Sigma}_\star^{(k)}\right\|_{\mathsf{F}}\epsilon_0\rho^t\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star) \le 0.15\left\|\boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}}\boldsymbol{\Sigma}_\star^{(k)}\right\|_{\mathsf{F}}\epsilon_0\rho^t\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star). \tag{34}$$

- **Bounding $\mathfrak{A}_{5,k}$.** Similar to $\mathfrak{A}_{4,k}$, we first apply the definition of the matrix inner product and then Hölder's inequality, leading to

$$\begin{aligned}
|\mathfrak{A}_{5,k}| &= \left|\operatorname{tr}\left(\mathcal{M}_k\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right)\breve{\boldsymbol{U}}^{(k)}\left(\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}\right)^{-1}\left(\boldsymbol{\Sigma}_\star^{(k)}\right)^2\left(\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}\right)^{-1}\breve{\boldsymbol{U}}^{(k)\top}\boldsymbol{\Delta}_{\breve{\boldsymbol{U}}^{(k)}}\boldsymbol{U}_\star^{(k)\top}\right)\right| \\
&\le \left\|\mathcal{M}_k\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right)\right\|\left\|\breve{\boldsymbol{U}}^{(k)}\left(\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}\right)^{-1}\left(\boldsymbol{\Sigma}_\star^{(k)}\right)^2\left(\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}\right)^{-1}\breve{\boldsymbol{U}}^{(k)\top}\boldsymbol{\Delta}_{\breve{\boldsymbol{U}}^{(k)}}\boldsymbol{U}_\star^{(k)\top}\right\|_* \\
&\le 2\zeta_{t+1}\alpha\sqrt{n_1n_2n_3r_k}\left\|\breve{\boldsymbol{U}}^{(k)}\left(\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}\right)^{-1}\left(\boldsymbol{\Sigma}_\star^{(k)}\right)^2\left(\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}\right)^{-1}\breve{\boldsymbol{U}}^{(k)\top}\boldsymbol{\Delta}_{\breve{\boldsymbol{U}}^{(k)}}\boldsymbol{U}_\star^{(k)\top}\right\|_{\mathsf{F}},
\end{aligned}$$

where the last line follows from (30), as well as the norm relation $\|\boldsymbol{A}\|_* \le \sqrt{r_k}\|\boldsymbol{A}\|_{\mathsf{F}}$ for a matrix of rank at most r_k. To continue, noting that $\boldsymbol{U}_\star^{(k)}$ has orthonormal columns, we have

$$\begin{aligned}
&\left\|\breve{\boldsymbol{U}}^{(k)}\left(\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}\right)^{-1}\left(\boldsymbol{\Sigma}_\star^{(k)}\right)^2\left(\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}\right)^{-1}\breve{\boldsymbol{U}}^{(k)\top}\boldsymbol{\Delta}_{\breve{\boldsymbol{U}}^{(k)}}\boldsymbol{U}_\star^{(k)\top}\right\|_{\mathsf{F}} \\
&\le \left\|\breve{\boldsymbol{U}}^{(k)}\left(\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}\right)^{-1}\boldsymbol{\Sigma}_\star^{(k)}\right\|^2\left\|\boldsymbol{\Delta}_{\breve{\boldsymbol{U}}^{(k)}}\right\|_{\mathsf{F}}
\end{aligned}$$

$$\leq \frac{\left(1+\epsilon_0+\epsilon_0^2/3\right)}{(1-\epsilon_0)^6}\left(\left\|\boldsymbol{\Delta}_{\boldsymbol{U}^{(2)}}\boldsymbol{\Sigma}_\star^{(2)}\right\|_{\mathsf{F}}+\left\|\boldsymbol{\Delta}_{\boldsymbol{U}^{(3)}}\boldsymbol{\Sigma}_\star^{(3)}\right\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{G}}}\|_{\mathsf{F}}\right)\leq \frac{2\left(1+\epsilon_0+\epsilon_0^2/3\right)}{(1-\epsilon_0)^6}\operatorname{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star),$$

where the penultimate line follows from Lemma 8 (cf. (70c) and (70e)), and the last line follows from Cauchy-Schwarz inequality. Plug this into the previous bound to arrive at

$$|\mathfrak{A}_{5,k}|\leq \frac{4\zeta_{t+1}\alpha\sqrt{n_1n_2n_3r_k}\left(1+\epsilon_0+\epsilon_0^2/3\right)}{(1-\epsilon_0)^6}\operatorname{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star)\leq \frac{32\alpha\sqrt{\mu^3r_1r_2r_3r_k}\left(1+\epsilon_0+\epsilon_0^2/3\right)}{(1-\epsilon_0)^6}\epsilon_0\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star),$$

where we have used $\operatorname{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star)\leq\epsilon_0\rho^t\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$ and the choice of ζ_{t+1} (cf. (22)). Finally, with the assumption on the sparsity level $\alpha\leq\frac{c_0\epsilon_0}{\sqrt{\mu^3r_1r_2r_3r}}$ for sufficiently small c_0 and $\epsilon_0<0.01$, we have

$$|\mathfrak{A}_{5,k}|\leq \frac{32c_0\left(1+\epsilon_0+\epsilon_0^2/3\right)}{(1-\epsilon_0)^6}\epsilon_0^2\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star)\leq 0.3\epsilon_0^2\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star). \tag{35}$$

Putting things together. Summing (29) over all k, we obtain

$$\begin{aligned}&\sum_{k=1}^3\left\|(\boldsymbol{U}_{t+1}^{(k)}\boldsymbol{Q}_t^{(k)}-\boldsymbol{U}_\star^{(k)})\boldsymbol{\Sigma}_\star^{(k)}\right\|_{\mathsf{F}}^2\\&=(1-\eta)^2\sum_{k=1}^3\left\|\boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}}\boldsymbol{\Sigma}_\star^{(k)}\right\|_{\mathsf{F}}^2+\eta^2\sum_{k=1}^3(\mathfrak{A}_{1,k}+\mathfrak{A}_{3,k}+2\mathfrak{A}_{5,k})-2\eta(1-\eta)\sum_{k=1}^3(\mathfrak{A}_{2,k}+\mathfrak{A}_{4,k}).\end{aligned}$$

Plugging in our bounds in (31)-(35), we have

$$\begin{aligned}\sum_{k=1}^3\left\|(\boldsymbol{U}_{t+1}^{(k)}\boldsymbol{Q}_t^{(k)}-\boldsymbol{U}_\star^{(k)})\boldsymbol{\Sigma}_\star^{(k)}\right\|_{\mathsf{F}}^2&\leq(1-\eta)^2\sum_{k=1}^3\left\|\boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}}\boldsymbol{\Sigma}_\star^{(k)}\right\|_{\mathsf{F}}^2\\&+2\eta^2\epsilon_0^2\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star)-2\eta(1-\eta)\left\|\boldsymbol{\mathcal{K}}^{(1)}+\boldsymbol{\mathcal{K}}^{(2)}+\boldsymbol{\mathcal{K}}^{(3)}\right\|_{\mathsf{F}}^2+0.06\eta(1-\eta)\epsilon_0^2\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star)\\&+3\eta^2\left\|\boldsymbol{\mathcal{K}}^{(1)}+\boldsymbol{\mathcal{K}}^{(2)}+\boldsymbol{\mathcal{K}}^{(3)}\right\|_{\mathsf{F}}^2+0.15\eta(1-\eta)\sum_{k=1}^3\left\|\boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}}\boldsymbol{\Sigma}_\star^{(k)}\right\|_{\mathsf{F}}\epsilon_0\rho^t\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star).\end{aligned}$$

Note that when $0<\eta<2/5$,

$$-2\eta(1-\eta)\left\|\boldsymbol{\mathcal{K}}^{(1)}+\boldsymbol{\mathcal{K}}^{(2)}+\boldsymbol{\mathcal{K}}^{(3)}\right\|_{\mathsf{F}}^2+3\eta^2\left\|\boldsymbol{\mathcal{K}}^{(1)}+\boldsymbol{\mathcal{K}}^{(2)}+\boldsymbol{\mathcal{K}}^{(3)}\right\|_{\mathsf{F}}^2=-\eta(2-5\eta)\left\|\boldsymbol{\mathcal{K}}^{(1)}+\boldsymbol{\mathcal{K}}^{(2)}+\boldsymbol{\mathcal{K}}^{(3)}\right\|_{\mathsf{F}}^2<0.$$

Therefore, the previous bound can be simplified to

$$\begin{aligned}\sum_{k=1}^3\left\|(\boldsymbol{U}_{t+1}^{(k)}\boldsymbol{Q}_t^{(k)}-\boldsymbol{U}_\star^{(k)})\boldsymbol{\Sigma}_\star^{(k)}\right\|_{\mathsf{F}}^2&\leq(1-\eta)^2\sum_{k=1}^3\left\|\boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}}\boldsymbol{\Sigma}_\star^{(k)}\right\|_{\mathsf{F}}^2+0.15\eta(1-\eta)\sum_{k=1}^3\left\|\boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}}\boldsymbol{\Sigma}_\star^{(k)}\right\|_{\mathsf{F}}\epsilon_0\rho^t\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)\\&\quad+2\eta^2\epsilon_0^2\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star)+0.06\eta(1-\eta)\epsilon_0^2\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star).\end{aligned}$$

A.2 Proof of (26)

Taking the squared Frobenius norm of (25), it follows

$$\left\|\left((\boldsymbol{Q}_t^{(1)})^{-1},(\boldsymbol{Q}_t^{(2)})^{-1},(\boldsymbol{Q}_t^{(3)})^{-1}\right)\cdot\boldsymbol{\mathcal{G}}_{t+1}-\boldsymbol{\mathcal{G}}_\star\right\|_{\mathsf{F}}^2=(1-\eta)^2\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{G}}}\|_{\mathsf{F}}^2-2\eta(1-\eta)\mathfrak{B}_1+\eta^2\mathfrak{B}_2, \tag{36}$$

where

$$\begin{aligned}\mathfrak{B}_1&=\left\langle\boldsymbol{\Delta}_{\boldsymbol{\mathcal{G}}},\left(\left(\boldsymbol{U}^{(1)\top}\boldsymbol{U}^{(1)}\right)^{-1}\boldsymbol{U}^{(1)\top},\ldots,\left(\boldsymbol{U}^{(3)\top}\boldsymbol{U}^{(3)}\right)^{-1}\boldsymbol{U}^{(3)\top}\right)\cdot\left((\boldsymbol{U}^{(1)},\boldsymbol{U}^{(2)},\boldsymbol{U}^{(3)})\cdot\boldsymbol{\mathcal{G}}_\star-\boldsymbol{\mathcal{X}}_\star+\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right)\right\rangle,\\\mathfrak{B}_2&=\left\|\left(\left(\boldsymbol{U}^{(1)\top}\boldsymbol{U}^{(1)}\right)^{-1}\boldsymbol{U}^{(1)\top},\ldots,\left(\boldsymbol{U}^{(3)\top}\boldsymbol{U}^{(3)}\right)^{-1}\boldsymbol{U}^{(3)\top}\right)\cdot\left((\boldsymbol{U}^{(1)},\boldsymbol{U}^{(2)},\boldsymbol{U}^{(3)})\cdot\boldsymbol{\mathcal{G}}_\star-\boldsymbol{\mathcal{X}}_\star+\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right)\right\|_{\mathsf{F}}^2.\end{aligned}$$

We will now bound $\mathfrak{B}_1$ and $\mathfrak{B}_2$ separately.

Bounding $\mathfrak{B}_1$. We start by breaking up the inner product into

$$\mathfrak{B}_1 = \underbrace{\left\langle \boldsymbol{\Delta}_{\boldsymbol{\mathcal{G}}}, \left(\left(\boldsymbol{U}^{(1)\top}\boldsymbol{U}^{(1)}\right)^{-1}\boldsymbol{U}^{(1)\top}, \ldots, \left(\boldsymbol{U}^{(3)\top}\boldsymbol{U}^{(3)}\right)^{-1}\boldsymbol{U}^{(3)\top} \right) \cdot \left(\left(\boldsymbol{U}^{(1)}, \boldsymbol{U}^{(2)}, \boldsymbol{U}^{(3)}\right) \cdot \boldsymbol{\mathcal{G}}_\star - \boldsymbol{\mathcal{X}}_\star \right) \right\rangle}_{=:\mathfrak{B}_{1,1}}$$
$$+ \underbrace{\left\langle \boldsymbol{\Delta}_{\boldsymbol{\mathcal{G}}}, \left(\left(\boldsymbol{U}^{(1)\top}\boldsymbol{U}^{(1)}\right)^{-1}\boldsymbol{U}^{(1)\top}, \ldots, \left(\boldsymbol{U}^{(3)\top}\boldsymbol{U}^{(3)}\right)^{-1}\boldsymbol{U}^{(3)\top} \right) \cdot \boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}} \right\rangle}_{=:\mathfrak{B}_{1,2}}.$$

Note that $\mathfrak{B}_{1,1}$ is identical to the term that is bounded in [TMPB+22, Section B.3], which obeys

$$\mathfrak{B}_{1,1} \geq \sum_{k=1}^{3} \left\| \boldsymbol{D}^{(k)} \right\|_{\mathsf{F}}^2 - C_1\epsilon_0\rho^t \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star) \geq \sum_{k=1}^{3} \left\| \boldsymbol{D}^{(k)} \right\|_{\mathsf{F}}^2 - 0.01\epsilon_0^2\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star), \tag{37}$$

where we have used the induction hypothesis (19a) and $C_1\epsilon_0\rho^t \leq C_1\epsilon_0 < 0.01$ as long as ϵ_0 is small enough. Here,

$$\boldsymbol{D}^{(k)} = \left(\boldsymbol{U}^{(k)\top}\boldsymbol{U}^{(k)}\right)^{-1/2}\boldsymbol{U}^{(k)\top}\boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}}\boldsymbol{\Sigma}_\star^{(k)}, \qquad k = 1, 2, 3.$$

Turning to $\mathfrak{B}_{1,2}$, since the inner product is invariant to matricization, we flatten the tensor along the first mode to bound it as

$$\begin{aligned}
|\mathfrak{B}_{1,2}| &= \left| \left\langle \mathcal{M}_1\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{G}}}\right), \left(\boldsymbol{U}^{(1)\top}\boldsymbol{U}^{(1)}\right)^{-1}\boldsymbol{U}^{(1)\top}\mathcal{M}_1\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right)\left(\left(\boldsymbol{U}^{(3)\top}\boldsymbol{U}^{(3)}\right)^{-1}\left(\boldsymbol{U}^{(3)}\right)^{\top} \otimes \left(\boldsymbol{U}^{(2)\top}\boldsymbol{U}^{(2)}\right)^{-1}\boldsymbol{U}^{(2)\top} \right)^{\top} \right\rangle \right| \\
&\leq \left\|\mathcal{M}_1\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{G}}}\right)\right\|_* \left\| \left(\boldsymbol{U}^{(1)\top}\boldsymbol{U}^{(1)}\right)^{-1}\boldsymbol{U}^{(1)\top}\mathcal{M}_1\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right)\left(\left(\boldsymbol{U}^{(3)\top}\boldsymbol{U}^{(3)}\right)^{-1}\boldsymbol{U}^{(3)\top} \otimes \left(\boldsymbol{U}^{(2)\top}\boldsymbol{U}^{(2)}\right)^{-1}\boldsymbol{U}^{(2)\top} \right)^{\top} \right\| \\
&\leq \sqrt{r_1}\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{G}}}\right\|_{\mathsf{F}} \prod_{k=1}^{3} \left\| \boldsymbol{U}^{(k)}\left(\boldsymbol{U}^{(k)\top}\boldsymbol{U}^{(k)}\right)^{-1} \right\| \left\|\mathcal{M}_1\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right)\right\|,
\end{aligned}$$

where the second line uses Hölder's inequality, and the last line uses $\left\|\mathcal{M}_1\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{G}}}\right)\right\|_* \leq \sqrt{r_1}\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{G}}}\right\|_{\mathsf{F}}$ with the fact that $\mathcal{M}_1\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{G}}}\right)$ is of rank r_1. To continue, invoke Lemma 8 (cf. (70b)) as well as (30) to further obtain

$$|\mathfrak{B}_{1,2}| \leq \frac{2\zeta_{t+1}\alpha\sqrt{n_1n_2n_3r_1}}{(1-\epsilon_0)^3}\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{G}}}\right\|_{\mathsf{F}} = \frac{16\alpha\sqrt{\mu^3r_1^2r_2r_3}}{(1-\epsilon_0)^3}\rho^t\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{G}}}\right\|_{\mathsf{F}},$$

where the second equality follows from the choice of ζ_{t+1} (cf. (22)). Finally, with the assumption on the sparsity level $\alpha \leq \frac{c_0\epsilon_0}{\sqrt{\mu^3r_1r_2r_3r}}$ for sufficiently small c_0 and $\epsilon_0 < 0.01$, we have

$$|\mathfrak{B}_{1,2}| \leq \frac{16c_0}{(1-\epsilon_0)^3}\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{G}}}\right\|_{\mathsf{F}}\epsilon_0\rho^t\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star) \leq 0.15\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{G}}}\right\|_{\mathsf{F}}\epsilon_0\rho^t\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star). \tag{38}$$

Put (37) and (38) together to see

$$\mathfrak{B}_1 \geq \sum_{k=1}^{3}\left\|\boldsymbol{D}^{(k)}\right\|_{\mathsf{F}}^2 - 0.01\epsilon_0^2\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star) - 0.15\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{G}}}\right\|_{\mathsf{F}}\epsilon_0\rho^t\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star). \tag{39}$$

Bounding $\mathfrak{B}_2$. Expanding out the square and applying Cauchy-Schwarz, we can upper bound $\mathfrak{B}_2$ by

$$\mathfrak{B}_2 \leq 2\underbrace{\left\| \left(\left(\boldsymbol{U}^{(1)\top}\boldsymbol{U}^{(1)}\right)^{-1}\boldsymbol{U}^{(1)\top}, \ldots, \left(\boldsymbol{U}^{(3)\top}\boldsymbol{U}^{(3)}\right)^{-1}\boldsymbol{U}^{(3)\top} \right) \cdot \left(\left(\boldsymbol{U}^{(1)}, \boldsymbol{U}^{(2)}, \boldsymbol{U}^{(3)}\right) \cdot \boldsymbol{\mathcal{G}}_\star - \boldsymbol{\mathcal{X}}_\star \right) \right\|_{\mathsf{F}}^2}_{=:\mathfrak{B}_{2,1}}$$
$$+ 2\underbrace{\left\| \left(\left(\boldsymbol{U}^{(1)\top}\boldsymbol{U}^{(1)}\right)^{-1}\boldsymbol{U}^{(1)\top}, \ldots, \left(\boldsymbol{U}^{(3)\top}\boldsymbol{U}^{(3)}\right)^{-1}\boldsymbol{U}^{(3)\top} \right) \cdot \boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}} \right\|_{\mathsf{F}}^2}_{=:\mathfrak{B}_{2,2}}.$$

$\mathfrak{B}_{2,1}$ is identical to the term that is bounded in [TMPB+22, Section B.4], which obeys

$$\mathfrak{B}_{2,1} \le 3\sum_{k=1}^{3}\left\|\boldsymbol{D}^{(k)}\right\|_{\mathsf{F}}^2 + C_2\epsilon_0\rho^t \operatorname{dist}^2(\boldsymbol{F}_t,\boldsymbol{F}_\star) \le 3\sum_{k=1}^{3}\left\|\boldsymbol{D}^{(k)}\right\|_{\mathsf{F}}^2 + 0.01\epsilon_0^2\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star), \tag{40}$$

where we make use of the induction hypothesis (19a) and $C_2\epsilon_0\rho^t \le C_2\epsilon_0 < 0.01$ as long as ϵ_0 is small enough.

For $\mathfrak{B}_{2,2}$, the matricization of the term in the norm along the first mode is of rank at most r_1, so

$$\begin{aligned}
\mathfrak{B}_{2,2} &= \left\|\left(((\boldsymbol{U}^{(1)})^\top\boldsymbol{U}^{(1)})^{-1}(\boldsymbol{U}^{(1)})^\top, ((\boldsymbol{U}^{(2)})^\top\boldsymbol{U}^{(2)})^{-1}(\boldsymbol{U}^{(2)})^\top, ((\boldsymbol{U}^{(3)})^\top\boldsymbol{U}^{(3)})^{-1}(\boldsymbol{U}^{(3)})^\top\right)\cdot\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right\|_{\mathsf{F}}^2 \\
&\le r_1\left\|((\boldsymbol{U}^{(1)})^\top\boldsymbol{U}^{(1)})^{-1}(\boldsymbol{U}^{(1)})^\top\mathcal{M}_1\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right)\left(((\boldsymbol{U}^{(3)})^\top\boldsymbol{U}^{(3)})^{-1}(\boldsymbol{U}^{(3)})^\top\otimes((\boldsymbol{U}^{(2)})^\top\boldsymbol{U}^{(2)})^{-1}(\boldsymbol{U}^{(2)})^\top\right)^\top\right\|^2 \\
&\le r_1\left\|((\boldsymbol{U}^{(1)})^\top\boldsymbol{U}^{(1)})^{-1}(\boldsymbol{U}^{(1)})^\top\right\|^2\left\|((\boldsymbol{U}^{(2)})^\top\boldsymbol{U}^{(2)})^{-1}(\boldsymbol{U}^{(2)})^\top\right\|^2\left\|((\boldsymbol{U}^{(3)})^\top\boldsymbol{U}^{(3)})^{-1}(\boldsymbol{U}^{(3)})^\top\right\|^2\left\|\mathcal{M}_1\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right)\right\|^2 \\
&\le \frac{r_1}{(1-\epsilon_0)^6}\left\|\mathcal{M}_1\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right)\right\|^2,
\end{aligned}$$

where the last line follows from Lemma 8 (cf. (70b)). To continue, we use (30) to obtain

$$\mathfrak{B}_{2,2} \le \frac{4\zeta_{t+1}^2\alpha^2 n_1n_2n_3r_1}{(1-\epsilon_0)^6} = \frac{128\alpha^2\mu^3r_1^2r_2r_3}{(1-\epsilon_0)^6}\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star),$$

where the second relation follows by the choice of ζ_{t+1} (cf. (22)). Lastly, with the assumption on the sparsity level $\alpha \le \frac{c_0\epsilon_0}{\sqrt{\mu^3r_1r_2r_3r}}$ for sufficiently small c_0 and $\epsilon_0 < 0.01$, we have

$$\mathfrak{B}_{2,2} \le \frac{128c_0^2}{(1-\epsilon_0)^6}\epsilon_0^2\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star) \le 0.02\epsilon_0^2\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star). \tag{41}$$

Combining (40) and (41), we get

$$\mathfrak{B}_2 \le 6\sum_{k=1}^{3}\left\|\boldsymbol{D}^{(k)}\right\|_{\mathsf{F}}^2 + 0.06\epsilon_0^2\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star). \tag{42}$$

Sum up. Going back to (36), we can substitute in our bounds for $\mathfrak{B}_1$ (cf. (39)) and $\mathfrak{B}_2$ (cf. (42)) to get

$$\begin{aligned}
&\left\|(\boldsymbol{Q}_t^{(1)})^{-1},(\boldsymbol{Q}_t^{(2)})^{-1},(\boldsymbol{Q}_t^{(3)})^{-1})\cdot\boldsymbol{\mathcal{G}}_{t+1}-\boldsymbol{\mathcal{G}}_\star\right\|_{\mathsf{F}}^2 \\
&\le (1-\eta)^2\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{G}}}\right\|_{\mathsf{F}}^2 - 2\eta(1-\eta)\left(\sum_{k=1}^{3}\left\|\boldsymbol{D}^{(k)}\right\|_{\mathsf{F}}^2 - 0.01\epsilon_0^2\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star) - 0.15\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{G}}}\right\|_{\mathsf{F}}\epsilon_0\rho^t\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)\right) \\
&\quad + \eta^2\left(6\sum_{k=1}^{3}\left\|\boldsymbol{D}^{(k)}\right\|_{\mathsf{F}}^2 + 0.06\epsilon_0^2\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star)\right).
\end{aligned}$$

Notice that $-2\eta(1-\eta)\left\|\boldsymbol{D}^{(k)}\right\|_{\mathsf{F}}^2 + 6\eta^2\left\|\boldsymbol{D}^{(k)}\right\|_{\mathsf{F}}^2 = -2\eta(1-4\eta)\left\|\boldsymbol{D}^{(k)}\right\|_{\mathsf{F}}^2 \le 0$ whenever $0 < \eta \le 1/4$, leading to the conclusion that

$$\begin{aligned}
\left\|(\boldsymbol{Q}_t^{(1)})^{-1},(\boldsymbol{Q}_t^{(2)})^{-1},(\boldsymbol{Q}_t^{(3)})^{-1})\cdot\boldsymbol{\mathcal{G}}_{t+1}-\boldsymbol{\mathcal{G}}_\star\right\|_{\mathsf{F}}^2 &\le (1-\eta)^2\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{G}}}\right\|_{\mathsf{F}}^2 + 2\cdot 0.15\eta(1-\eta)\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{G}}}\right\|_{\mathsf{F}}\epsilon_0\rho^t\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star) \\
&\quad + 0.02\eta(1-\eta)\epsilon_0^2\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star) + 0.06\eta^2\epsilon_0^2\rho^{2t}\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star).
\end{aligned}$$

B Proof of Lemma 2

In view of Lemma 1 and [TMPB+22, Lemma 6], the optimal alignment matrices $\{\boldsymbol{Q}_t^{(k)}\}_{k=1}^3$ (resp. $\{\boldsymbol{Q}_{(t+1)}^{(k)}\}_{k=1}^3$)) between $\boldsymbol{F}_t$ (resp. $\boldsymbol{F}_{t+1}$) and $\boldsymbol{F}_\star$ exist. Fix any $k = 1,2,3$. By the triangle inequality, we have

$$\left\|(\boldsymbol{U}_{t+1}^{(k)}\boldsymbol{Q}_{t+1}^{(k)} - \boldsymbol{U}_\star^{(k)})\boldsymbol{\Sigma}_\star^{(k)}\right\|_{2,\infty} \le \left\|(\boldsymbol{U}_{t+1}^{(k)}\boldsymbol{Q}_t^{(k)} - \boldsymbol{U}_\star^{(k)})\boldsymbol{\Sigma}_\star^{(k)}\right\|_{2,\infty} + \left\|\boldsymbol{U}_{t+1}^{(k)}(\boldsymbol{Q}_{t+1}^{(k)} - \boldsymbol{Q}_t^{(k)})\boldsymbol{\Sigma}_\star^{(k)}\right\|_{2,\infty}. \tag{43}$$

Below we control the two terms in turn.

Step 1: controlling $\left\|(\boldsymbol{U}_{t+1}^{(k)}\boldsymbol{Q}_t^{(k)}-\boldsymbol{U}_\star^{(k)})\boldsymbol{\Sigma}_\star^{(k)}\right\|_{2,\infty}$. By the update rule, we have

$$(\boldsymbol{U}_{t+1}^{(k)}\boldsymbol{Q}_t^{(k)}-\boldsymbol{U}_\star^{(k)})\boldsymbol{\Sigma}_\star^{(k)}=(1-\eta)\boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}}\boldsymbol{\Sigma}_\star^{(k)}-\eta\left(\mathcal{M}_k\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right)+\boldsymbol{U}_\star^{(k)}\boldsymbol{\Delta}_{\breve{\boldsymbol{U}}^{(k)}}^\top\right)\breve{\boldsymbol{U}}^{(k)}\big(\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}\big)^{-1}\boldsymbol{\Sigma}_\star^{(k)}.$$

Take the $\ell_{2,\infty}$-norm of both sides and apply the triangle inequality to see that

$$\begin{aligned}\left\|(\boldsymbol{U}_{t+1}^{(k)}\boldsymbol{Q}_t^{(k)}-\boldsymbol{U}_\star^{(k)})\boldsymbol{\Sigma}_\star^{(k)}\right\|_{2,\infty}\le(1-\eta)\underbrace{\left\|\boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}}\boldsymbol{\Sigma}_\star^{(k)}\right\|_{2,\infty}}_{=:\mathfrak{C}_{1,k}}+\eta\underbrace{\left\|\mathcal{M}_k\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right)\breve{\boldsymbol{U}}^{(k)}\big(\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}\big)^{-1}\boldsymbol{\Sigma}_\star^{(k)}\right\|_{2,\infty}}_{=:\mathfrak{C}_{2,k}}\\+\eta\underbrace{\left\|\boldsymbol{U}_\star^{(k)}\boldsymbol{\Delta}_{\breve{\boldsymbol{U}}^{(k)}}^\top\breve{\boldsymbol{U}}^{(k)}\big(\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}\big)^{-1}\boldsymbol{\Sigma}_\star^{(k)}\right\|_{2,\infty}}_{=:\mathfrak{C}_{3,k}}.\end{aligned}$$

We then proceed to bound each term separately.

- $\mathfrak{C}_{1,k}$ is captured by the induction hypothesis (19b), which directly implies

$$\mathfrak{C}_{1,k}\le\rho^t\sqrt{\frac{\mu r_k}{n_k}}\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star). \tag{44}$$

- We now move on to $\mathfrak{C}_{2,k}$, which can be bounded by

$$\mathfrak{C}_{2,k}\le\left\|\mathcal{M}_k\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right)\right\|_{2,\infty}\left\|\breve{\boldsymbol{U}}^{(k)}\big(\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}\big)^{-1}\boldsymbol{\Sigma}_\star^{(k)}\right\|\le\frac{1}{(1-\epsilon_0)^3}\left\|\mathcal{M}_k\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right)\right\|_{2,\infty},$$

  where the second inequality follows from Lemma 8 (cf. (70c)). Recall that $\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}$ is α-sparse following the choice of ζ_{t+1}, which gives us

$$\left\|\mathcal{M}_k\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right)\right\|_{2,\infty}\le\sqrt{\frac{\alpha n_1n_2n_3}{n_k}}\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{S}}}\right\|_\infty\le2\sqrt{\frac{\alpha n_1n_2n_3}{n_k}}\zeta_{t+1}=16\sqrt{\frac{\alpha\mu^3r_1r_2r_3}{n_k}}\rho^t\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$$

  due to Lemma 11, (30), and the choice of ζ_{t+1} (cf. (22)). Plug this into the previous bound to obtain

$$\mathfrak{C}_{2,k}\le\frac{16}{(1-\epsilon_0)^3}\sqrt{\frac{\alpha\mu^3r_1r_2r_3}{n_k}}\rho^t\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)\le0.15\sqrt{\frac{\mu r_k}{n_k}}\rho^t\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star), \tag{45}$$

  where the last inequality follows from the assumption on the sparsity level $\alpha\le\frac{c_1}{\mu^2r_1r_2r_3}$ with a sufficiently small constant c_1.

- Finally, for $\mathfrak{C}_{3,k}$, we have the upper bound

$$\begin{aligned}\mathfrak{C}_{3,k}&\le\left\|\boldsymbol{U}_\star^{(k)}\right\|_{2,\infty}\left\|\boldsymbol{\Delta}_{\breve{\boldsymbol{U}}^{(k)}}\right\|_{\mathsf{F}}\left\|\breve{\boldsymbol{U}}^{(k)}\big(\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}\big)^{-1}\boldsymbol{\Sigma}_\star^{(k)}\right\|\\&\le\sqrt{\frac{3\mu r_k}{n_k}}\left(1+\epsilon_0+\frac{\epsilon_0^2}{3}\right)\frac{1}{(1-\epsilon_0)^3}\operatorname{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star),\end{aligned}$$

  where the second inequality follows from the incoherence assumption $\left\|\boldsymbol{U}_\star^{(k)}\right\|_{2,\infty}\le\sqrt{\frac{\mu r_k}{n_k}}$, and Lemma 8 (cf. (70c) and (70e)). Since $\operatorname{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star)\le\epsilon_0\rho^t\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$, we arrive at

$$\mathfrak{C}_{3,k}\le\frac{1}{(1-\epsilon_0)^3}\sqrt{\frac{3\mu r_k}{n_k}}\left(1+\epsilon_0+\frac{\epsilon_0^2}{3}\right)\epsilon_0\rho^t\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)\le0.02\sqrt{\frac{\mu r_k}{n_k}}\rho^t\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star) \tag{46}$$

  as long as $\epsilon_0<0.01$.

Combining (44), (45), and (46) together, we reach the conclusion that

$$\left\|(\boldsymbol{U}_{t+1}^{(k)}\boldsymbol{Q}_t^{(k)}-\boldsymbol{U}_\star^{(k)})\boldsymbol{\Sigma}_\star^{(k)}\right\|_{2,\infty}\le(1-0.83\eta)\sqrt{\frac{\mu r_k}{n_k}}\rho^t\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star). \tag{47}$$

In view of the basic relation

$$\left\|(\boldsymbol{U}_{t+1}^{(k)}\boldsymbol{Q}_t^{(k)}-\boldsymbol{U}_\star^{(k)})\boldsymbol{\Sigma}_\star^{(k)}\right\|_{2,\infty}\ge\left\|\boldsymbol{U}_{t+1}^{(k)}\boldsymbol{Q}_t^{(k)}-\boldsymbol{U}_\star^{(k)}\right\|_{2,\infty}\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star),$$

this also implies

$$\left\|\boldsymbol{U}_{t+1}^{(k)}\boldsymbol{Q}_t^{(k)}-\boldsymbol{U}_\star^{(k)}\right\|_{2,\infty}\le(1-0.83\eta)\sqrt{\frac{\mu r_k}{n_k}}\rho^t. \tag{48}$$

Step 2: controlling $\left\|\boldsymbol{U}_{t+1}^{(k)}(\boldsymbol{Q}_{t+1}^{(k)}-\boldsymbol{Q}_t^{(k)})\boldsymbol{\Sigma}_\star^{(k)}\right\|_{2,\infty}$. Observe that

$$\begin{aligned}\left\|\boldsymbol{U}_{t+1}^{(k)}(\boldsymbol{Q}_{t+1}^{(k)}-\boldsymbol{Q}_t^{(k)})\boldsymbol{\Sigma}_\star^{(k)}\right\|_{2,\infty}&=\left\|\boldsymbol{U}_{t+1}^{(k)}\boldsymbol{Q}_t^{(k)}(\boldsymbol{Q}_t^{(k)})^{-1}(\boldsymbol{Q}_{t+1}^{(k)}-\boldsymbol{Q}_t^{(k)})\boldsymbol{\Sigma}_\star^{(k)}\right\|_{2,\infty}\\&\le\left\|\boldsymbol{U}_{t+1}^{(k)}\boldsymbol{Q}_t^{(k)}\right\|_{2,\infty}\left\|(\boldsymbol{Q}_t^{(k)})^{-1}(\boldsymbol{Q}_{t+1}^{(k)}-\boldsymbol{Q}_t^{(k)})\boldsymbol{\Sigma}_\star^{(k)}\right\|.\end{aligned}$$

For the first term $\left\|\boldsymbol{U}_{t+1}^{(k)}\boldsymbol{Q}_t^{(k)}\right\|_{2,\infty}$, we have

$$\begin{aligned}\left\|\boldsymbol{U}_{t+1}^{(k)}\boldsymbol{Q}_t^{(k)}\right\|_{2,\infty}&\le\left\|\boldsymbol{U}_{t+1}^{(k)}\boldsymbol{Q}_t^{(k)}-\boldsymbol{U}_\star^{(k)}\right\|_{2,\infty}+\left\|\boldsymbol{U}_\star^{(k)}\right\|_{2,\infty}\\&\le\left((1-0.83\eta)\rho^t+1\right)\sqrt{\frac{\mu r_k}{n_k}}\le(2-0.83\eta)\sqrt{\frac{\mu r_k}{n_k}},\end{aligned}$$

where the second line follows from (48) and the incoherence assumption $\left\|\boldsymbol{U}_\star^{(k)}\right\|_{2,\infty}\le\sqrt{\frac{\mu r_k}{n_k}}$.

Moving on to the second term $\left\|(\boldsymbol{Q}_t^{(k)})^{-1}(\boldsymbol{Q}_{t+1}^{(k)}-\boldsymbol{Q}_t^{(k)})\boldsymbol{\Sigma}_\star^{(k)}\right\|$, we plan to invoke Lemma 7. Given the assumption of the sparsity level α satisfies the requirement of Lemma 1, following Lemma 1 as well as its proof (cf. (20)), we have

$$\begin{aligned}\left\|(\boldsymbol{U}_{t+1}^{(k)}\boldsymbol{Q}_{t+1}^{(k)}-\boldsymbol{U}_\star^{(k)})\boldsymbol{\Sigma}_\star^{(k)}\right\|_{\mathsf{F}}&\le\operatorname{dist}(\boldsymbol{F}_{t+1},\boldsymbol{F}_\star)\le\epsilon_0\rho^{t+1}\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star),\\\left\|(\boldsymbol{U}_{t+1}^{(k)}\boldsymbol{Q}_t^{(k)}-\boldsymbol{U}_\star^{(k)})\boldsymbol{\Sigma}_\star^{(k)}\right\|_{\mathsf{F}}&\le\epsilon_0\rho^{t+1}\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star),\end{aligned}$$

which in turn implies

$$\max\left\{\left\|\boldsymbol{U}_{t+1}^{(k)}\boldsymbol{Q}_t^{(k)}-\boldsymbol{U}_\star^{(k)}\right\|,\frac{\left\|(\boldsymbol{U}_{t+1}^{(k)}\boldsymbol{Q}_{t+1}^{(k)}-\boldsymbol{U}_\star^{(k)})\boldsymbol{\Sigma}_\star^{(k)}\right\|}{\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)},\frac{\left\|(\boldsymbol{U}_{t+1}^{(k)}\boldsymbol{Q}_t^{(k)}-\boldsymbol{U}_\star^{(k)})\boldsymbol{\Sigma}_\star^{(k)}\right\|}{\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)}\right\}\le\epsilon_0\rho^{t+1}<1. \tag{49}$$

Setting $\boldsymbol{U}:=\boldsymbol{U}_{t+1}^{(k)}$, $\boldsymbol{U}_\star:=\boldsymbol{U}_\star^{(k)}$, $\boldsymbol{Q}:=\boldsymbol{Q}_{t+1}^{(k)}$, $\bar{\boldsymbol{Q}}:=\boldsymbol{Q}_t^{(k)}$ and $\boldsymbol{\Sigma}:=\boldsymbol{\Sigma}_\star^{(k)}$, (49) demonstrates that Lemma 7 is applicable, leading to

$$\begin{aligned}\left\|(\boldsymbol{Q}_t^{(k)})^{-1}(\boldsymbol{Q}_{t+1}^{(k)}-\boldsymbol{Q}_t^{(k)})\boldsymbol{\Sigma}_\star^{(k)}\right\|&\le\frac{\left\|\boldsymbol{U}_{t+1}^{(k)}(\boldsymbol{Q}_{t+1}^{(k)}-\boldsymbol{Q}_t^{(k)})\boldsymbol{\Sigma}_\star^{(k)}\right\|}{\sigma_{\min}(\boldsymbol{U}_\star)-\left\|\boldsymbol{U}_{t+1}^{(k)}\boldsymbol{Q}_t^{(k)}-\boldsymbol{U}_\star^{(k)}\right\|}\\&\le\frac{\left\|(\boldsymbol{U}_{t+1}^{(k)}\boldsymbol{Q}_{t+1}^{(k)}-\boldsymbol{U}_\star^{(k)})\boldsymbol{\Sigma}_\star^{(k)}\right\|+\left\|(\boldsymbol{U}_{t+1}^{(k)}\boldsymbol{Q}_t^{(k)}-\boldsymbol{Q}_\star^{(k)})\boldsymbol{\Sigma}_\star^{(k)}\right\|}{1-\epsilon_0}\\&\le\frac{2\epsilon_0}{1-\epsilon_0}\rho^{t+1}\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star),\end{aligned}$$

where we used (49) in the second and third inequalities. Combining the above two bounds, we have

$$\left\|\boldsymbol{U}_{t+1}^{(k)}(\boldsymbol{Q}_{t+1}^{(k)}-\boldsymbol{Q}_t^{(k)})\boldsymbol{\Sigma}_\star^{(k)}\right\|_{2,\infty}\le(2-0.83\eta)\frac{2\epsilon_0}{1-\epsilon_0}\sqrt{\frac{\mu r_k}{n_k}}\rho^{t+1}\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star). \tag{50}$$

Step 3: combining the bounds. Plug (47) and (50) into (43) to get

$$\begin{aligned}\left\|(\boldsymbol{U}_{t+1}^{(k)}\boldsymbol{Q}_{t+1}^{(k)}-\boldsymbol{U}_{\star}^{(k)})\boldsymbol{\Sigma}_{\star}^{(k)}\right\|_{2,\infty} &\le \left[(1-0.83\eta)+(2-0.83\eta)\frac{2\rho\epsilon_0}{1-\epsilon_0}\right]\sqrt{\frac{\mu r_k}{n_k}}\rho^t\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)\\ &\le (1.05-0.84\eta)\sqrt{\frac{\mu r_k}{n_k}}\rho^t\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)\\ &\le (1-0.45\eta)\sqrt{\frac{\mu r_k}{n_k}}\rho^t\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)=\sqrt{\frac{\mu r_k}{n_k}}\rho^{t+1}\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)\end{aligned}$$

where the last line follows from $\eta\ge\frac{1}{7}$ and $\rho=1-0.45\eta$.

C Proof of Lemma 3

In view of [TMPB+22, Lemma 8], one has

$$\operatorname{dist}(\boldsymbol{F}_0,\boldsymbol{F}_\star)\le(\sqrt{2}+1)^{\frac{3}{2}}\left\|(\boldsymbol{U}_0^{(1)},\boldsymbol{U}_0^{(2)},\boldsymbol{U}_0^{(3)})\cdot\boldsymbol{\mathcal{G}}_0-\boldsymbol{\mathcal{X}}_\star\right\|_{\mathsf{F}}, \tag{51}$$

where we have used the definition $\boldsymbol{F}_0=(\boldsymbol{U}_0^{(1)},\boldsymbol{U}_0^{(2)},\boldsymbol{U}_0^{(3)})\cdot\boldsymbol{\mathcal{G}}_0$. As a result, we focus on bounding the term $\left\|(\boldsymbol{U}_0^{(1)},\boldsymbol{U}_0^{(2)},\boldsymbol{U}_0^{(3)})\cdot\boldsymbol{\mathcal{G}}_0-\boldsymbol{\mathcal{X}}_\star\right\|_{\mathsf{F}}$ below.

Recall from the definition of the HOSVD that $\boldsymbol{\mathcal{G}}_0=((\boldsymbol{U}_0^{(1)})^\top,(\boldsymbol{U}_0^{(2)})^\top,(\boldsymbol{U}_0^{(3)})^\top)\cdot(\boldsymbol{\mathcal{Y}}-\mathcal{T}_{\zeta_0}(\boldsymbol{\mathcal{Y}}))$, which in turn implies

$$(\boldsymbol{U}_0^{(1)},\boldsymbol{U}_0^{(2)},\boldsymbol{U}_0^{(3)})\cdot\boldsymbol{\mathcal{G}}_0=\left(\boldsymbol{U}_0^{(1)}\boldsymbol{U}_0^{(1)\top},\boldsymbol{U}_0^{(2)}\boldsymbol{U}_0^{(2)\top},\boldsymbol{U}_0^{(3)}\boldsymbol{U}_0^{(3)\top}\right)\cdot(\boldsymbol{\mathcal{Y}}-\mathcal{T}_{\zeta_0}(\boldsymbol{\mathcal{Y}})). \tag{52}$$

Note that $\boldsymbol{U}_0^{(k)}$ has orthonormal columns. We thus define $\boldsymbol{P}^{(k)}:=\boldsymbol{U}_0^{(k)}\boldsymbol{U}_0^{(k)\top}$, the orthogonal projection onto the column space of $\boldsymbol{U}_0^{(k)}$. This allows us to rewrite the squared Frobenius norm in (51) as

$$\left\|(\boldsymbol{U}_0^{(1)},\boldsymbol{U}_0^{(2)},\boldsymbol{U}_0^{(3)})\cdot\boldsymbol{\mathcal{G}}_0-\boldsymbol{\mathcal{X}}_\star\right\|_{\mathsf{F}}^2=\left\|(\boldsymbol{P}^{(1)},\boldsymbol{P}^{(2)},\boldsymbol{P}^{(3)})\cdot(\boldsymbol{\mathcal{Y}}-\mathcal{T}_{\zeta_0}(\boldsymbol{\mathcal{Y}}))-\boldsymbol{\mathcal{X}}_\star\right\|_{\mathsf{F}}^2.$$

Since $\boldsymbol{\mathcal{X}}_\star$ can be decomposed into a sum of orthogonal projections and its orthogonal complements, namely,

$$\begin{aligned}\boldsymbol{\mathcal{X}}_\star=&(\boldsymbol{P}^{(1)},\boldsymbol{P}^{(2)},\boldsymbol{P}^{(3)})\cdot\boldsymbol{\mathcal{X}}_\star+(\boldsymbol{I}_{n_1}-\boldsymbol{P}^{(1)},\boldsymbol{P}^{(2)},\boldsymbol{P}^{(3)})\cdot\boldsymbol{\mathcal{X}}_\star\\&+(\boldsymbol{I}_{n_1},\boldsymbol{I}_{n_2}-\boldsymbol{P}^{(2)},\boldsymbol{P}^{(3)})\cdot\boldsymbol{\mathcal{X}}_\star+(\boldsymbol{I}_{n_1},\boldsymbol{I}_{n_2},\boldsymbol{I}_{n_3}-\boldsymbol{P}^{(3)})\cdot\boldsymbol{\mathcal{X}}_\star,\end{aligned}$$

we have the following identity

$$\begin{aligned}&\left\|(\boldsymbol{U}_0^{(1)},\boldsymbol{U}_0^{(2)},\boldsymbol{U}_0^{(3)})\cdot\boldsymbol{\mathcal{G}}_0-\boldsymbol{\mathcal{X}}_\star\right\|_{\mathsf{F}}^2=\left\|(\boldsymbol{P}^{(1)},\boldsymbol{P}^{(2)},\boldsymbol{P}^{(3)})\cdot(\boldsymbol{\mathcal{Y}}-\mathcal{T}_{\zeta_0}(\boldsymbol{\mathcal{Y}})-\boldsymbol{\mathcal{X}}_\star)\right\|_{\mathsf{F}}^2\\&+\left\|(\boldsymbol{I}_{n_1}-\boldsymbol{P}^{(1)},\boldsymbol{P}^{(2)},\boldsymbol{P}^{(3)})\cdot\boldsymbol{\mathcal{X}}_\star\right\|_{\mathsf{F}}^2+\left\|(\boldsymbol{I}_{n_1},\boldsymbol{I}_{n_2}-\boldsymbol{P}^{(2)},\boldsymbol{P}^{(3)})\cdot\boldsymbol{\mathcal{X}}_\star\right\|_{\mathsf{F}}^2+\left\|(\boldsymbol{I}_{n_1},\boldsymbol{I}_{n_2},\boldsymbol{I}_{n_3}-\boldsymbol{P}^{(3)})\cdot\boldsymbol{\mathcal{X}}_\star\right\|_{\mathsf{F}}^2.\end{aligned} \tag{53}$$

In what follows, we bound each term respectively.

Bounding the first term. Matricize along the first mode and change to the operator norm to obtain

$$\begin{aligned}\left\|(\boldsymbol{P}^{(1)},\boldsymbol{P}^{(2)},\boldsymbol{P}^{(3)})\cdot(\boldsymbol{\mathcal{Y}}-\mathcal{T}_{\zeta_0}(\boldsymbol{\mathcal{Y}})-\boldsymbol{\mathcal{X}}_\star)\right\|_{\mathsf{F}}&\le\sqrt{r_1}\left\|\boldsymbol{P}^{(1)}\mathcal{M}_1(\boldsymbol{\mathcal{Y}}-\mathcal{T}_{\zeta_0}(\boldsymbol{\mathcal{Y}})-\boldsymbol{\mathcal{X}}_\star)(\boldsymbol{P}^{(3)}\otimes\boldsymbol{P}^{(2)})^\top\right\|\\&\le\sqrt{r_1}\left\|\mathcal{M}_1(\boldsymbol{\mathcal{Y}}-\mathcal{T}_{\zeta_0}(\boldsymbol{\mathcal{Y}})-\boldsymbol{\mathcal{X}}_\star)\right\|\\&=\sqrt{r_1}\left\|\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star-\mathcal{T}_{\zeta_0}(\boldsymbol{\mathcal{Y}}))\right\|,\end{aligned} \tag{54}$$

where the last relation holds due to the definition of $\boldsymbol{\mathcal{Y}}$.

Bounding the remaining three terms. We present the bound on the second term, while the remaining two can be bounded in a similar fashion. Matricize along the first mode, and in view of the fact that $\|\boldsymbol{P}^{(k)}\| \leq 1$, we have

$$\begin{aligned}\left\|\left(\boldsymbol{I}_{n_1}-\boldsymbol{P}^{(1)},\boldsymbol{P}^{(2)},\boldsymbol{P}^{(3)}\right)\cdot\boldsymbol{\mathcal{X}}_\star\right\|_{\mathsf{F}} &\leq \left\|\left(\boldsymbol{I}_{n_1}-\boldsymbol{P}^{(1)}\right)\mathcal{M}_1\left(\boldsymbol{\mathcal{X}}_\star\right)\right\|_{\mathsf{F}} \\ &\leq \sqrt{r_1}\left\|\left(\boldsymbol{I}_{n_1}-\boldsymbol{P}^{(1)}\right)\mathcal{M}_1\left(\boldsymbol{\mathcal{X}}_\star\right)\right\| \\ &= \sqrt{r_1}\left\|\left(\boldsymbol{I}_{n_1}-\boldsymbol{P}^{(1)}\right)\mathcal{M}_1\left(\boldsymbol{\mathcal{Y}}-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)-\boldsymbol{\mathcal{S}}_\star+\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)\right\| \\ &\leq \sqrt{r_1}\left\|\mathcal{M}_1\left(\boldsymbol{\mathcal{S}}_\star-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)\right\|+\sqrt{r_1}\left\|\left(\boldsymbol{I}_{n_1}-\boldsymbol{P}^{(1)}\right)\mathcal{M}_1\left(\boldsymbol{\mathcal{Y}}-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)\right\|.\end{aligned}$$

To continue, note that $\boldsymbol{P}^{(1)}\mathcal{M}_1\left(\boldsymbol{\mathcal{Y}}-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)$ is the best rank-r_1 approximation to $\mathcal{M}_1\left(\boldsymbol{\mathcal{Y}}-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)$, which implies

$$\begin{aligned}\left\|\left(\boldsymbol{I}_{n_1}-\boldsymbol{P}^{(1)}\right)\mathcal{M}_1\left(\boldsymbol{\mathcal{Y}}-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)\right\| &\leq \sigma_{r_1+1}(\mathcal{M}_1\left(\boldsymbol{\mathcal{Y}}-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)) \\ &\leq \sigma_{r_1+1}(\mathcal{M}_1\left(\boldsymbol{\mathcal{X}}_\star\right))+\left\|\mathcal{M}_1\left(\boldsymbol{\mathcal{S}}_\star-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)\right\| = \left\|\mathcal{M}_1\left(\boldsymbol{\mathcal{S}}_\star-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)\right\|,\end{aligned}$$

where the last line follows from Weyl's inequality and the fact that $\mathcal{M}_1\left(\boldsymbol{\mathcal{X}}_\star\right)$ has rank r_1. Plug this into the previous inequality to obtain

$$\left\|\left(\boldsymbol{I}_{n_1}-\boldsymbol{P}^{(1)},\boldsymbol{P}^{(2)},\boldsymbol{P}^{(3)}\right)\cdot\boldsymbol{\mathcal{X}}_\star\right\|_{\mathsf{F}} \leq 2\sqrt{r_1}\left\|\mathcal{M}_1\left(\boldsymbol{\mathcal{S}}_\star-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)\right\|. \tag{55}$$

Plug our bounds (54) and (55) into (53) to obtain

$$\begin{aligned}\left\|\left(\boldsymbol{U}_0^{(1)},\boldsymbol{U}_0^{(2)},\boldsymbol{U}_0^{(3)}\right)\cdot\boldsymbol{\mathcal{G}}_0-\boldsymbol{\mathcal{X}}_\star\right\|_{\mathsf{F}}^2 &\leq r_1\left\|\mathcal{M}_1\left(\boldsymbol{\mathcal{S}}_\star-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)\right\|^2+4r_1\left\|\mathcal{M}_1\left(\boldsymbol{\mathcal{S}}_\star-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)\right\|^2 \\ &\quad+4r_2\left\|\mathcal{M}_2\left(\boldsymbol{\mathcal{S}}_\star-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)\right\|^2+4r_3\left\|\mathcal{M}_3\left(\boldsymbol{\mathcal{S}}_\star-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)\right\|^2.\end{aligned} \tag{56}$$

It then boils down to controlling $\left\|\mathcal{M}_k\left(\boldsymbol{\mathcal{S}}_\star-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)\right\|$ for $k=1,2,3$. With our choice of ζ_0 (i.e. $\|\boldsymbol{\mathcal{X}}_\star\|_\infty \leq \zeta_0 \leq 2\|\boldsymbol{\mathcal{X}}_\star\|_\infty$), setting $\boldsymbol{\mathcal{X}}=\boldsymbol{0}$ in Lemma 12 guarantees that $\mathcal{M}_k\left(\boldsymbol{\mathcal{S}}_\star-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)$ is α-sparse for all k. Hence, we can apply Lemma 11 to arrive at

$$\left\|\mathcal{M}_k\left(\boldsymbol{\mathcal{S}}_\star-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)\right\| \leq \alpha\sqrt{n_1n_2n_3}\left\|\boldsymbol{\mathcal{S}}_\star-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right\|_\infty \leq 2\alpha\sqrt{n_1n_2n_3}\zeta_0 \leq 4\alpha\sqrt{n_1n_2n_3}\|\boldsymbol{\mathcal{X}}_\star\|_\infty, \tag{57}$$

where the penultimate inequality follows from Lemma 12 (cf. (76)). Plug (57) into (56) to obtain

$$\left\|\left(\boldsymbol{U}_0^{(1)},\boldsymbol{U}_0^{(2)},\boldsymbol{U}_0^{(3)}\right)\cdot\boldsymbol{\mathcal{G}}_0-\boldsymbol{\mathcal{X}}_\star\right\|_{\mathsf{F}}^2 \leq 208\alpha^2n_1n_2n_3r\|\boldsymbol{\mathcal{X}}_\star\|_\infty^2 \leq 208\alpha^2\mu^3r_1r_2r_3r\kappa^2\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star),$$

where the second inequality follows from Lemma 5. Under the assumption that $\alpha \leq \frac{c_0}{\sqrt{\mu^3r_1r_2r_3r}\kappa}$, it follows

$$\left\|\left(\boldsymbol{U}_0^{(1)},\boldsymbol{U}_0^{(2)},\boldsymbol{U}_0^{(3)}\right)\cdot\boldsymbol{\mathcal{G}}_0-\boldsymbol{\mathcal{X}}_\star\right\|_{\mathsf{F}}^2 \leq 208c_0^2\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star).$$

This combined with (51) finishes the proof.

D Proof of Lemma 4

We provide the control on the first mode, as the other two modes can be bounded using the same arguments.

We begin with a useful decomposition of the quantity we care about, whose proof will be supplied in the end of this section:

$$\begin{aligned}\boldsymbol{\Delta}_{\boldsymbol{U}^{(1)}}\boldsymbol{\Sigma}_\star^{(1)} &= \mathcal{M}_1\left(\boldsymbol{\mathcal{S}}_\star-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)\mathcal{M}_1\left(\boldsymbol{\mathcal{Y}}-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)^\top\boldsymbol{U}_0^{(1)}(\boldsymbol{Q}_0^{(1)})^{-\top}(\breve{\boldsymbol{U}}^{(1)\top}\breve{\boldsymbol{U}}^{(1)})^{-1}\boldsymbol{\Sigma}_\star^{(1)} \\ &\quad+\boldsymbol{U}_\star^{(1)}\boldsymbol{\Delta}_{\breve{\boldsymbol{U}}^{(1)}}^\top\mathcal{M}_1\left(\boldsymbol{\mathcal{Y}}-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)^\top\boldsymbol{U}_0^{(1)}(\boldsymbol{Q}_0^{(1)})^{-\top}(\breve{\boldsymbol{U}}^{(1)\top}\breve{\boldsymbol{U}}^{(1)})^{-1}\boldsymbol{\Sigma}_\star^{(1)}.\end{aligned} \tag{58}$$

Taking the $\ell_{2,\infty}$-norm and using the triangle inequality, we obtain

$$\left\|\boldsymbol{\Delta}_{\boldsymbol{U}^{(1)}}\boldsymbol{\Sigma}_\star^{(1)}\right\|_{2,\infty} \le \underbrace{\left\|\boldsymbol{U}_\star^{(1)}\boldsymbol{\Delta}_{\breve{\boldsymbol{U}}^{(1)}}^\top \mathcal{M}_1\left(\boldsymbol{\mathcal{Y}}-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)^\top \boldsymbol{U}_0^{(1)}(\boldsymbol{Q}_0^{(1)})^{-\top}(\breve{\boldsymbol{U}}^{(1)\top}\breve{\boldsymbol{U}}^{(1)})^{-1}\boldsymbol{\Sigma}_\star^{(1)}\right\|_{2,\infty}}_{=:\mathfrak{A}_1}$$
$$+\underbrace{\left\|\mathcal{M}_1\left(\boldsymbol{\mathcal{S}}_\star-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)\mathcal{M}_1\left(\boldsymbol{\mathcal{Y}}-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)^\top \boldsymbol{U}_0^{(1)}(\boldsymbol{Q}_0^{(1)})^{-\top}(\breve{\boldsymbol{U}}^{(1)\top}\breve{\boldsymbol{U}}^{(1)})^{-1}\boldsymbol{\Sigma}_\star^{(1)}\right\|_{2,\infty}}_{=:\mathfrak{A}_2}. \quad (59)$$

We now proceed to bound these two terms separately.

Step 1: bounding $\mathfrak{A}_1$. To begin, note that

$$\begin{aligned}\mathfrak{A}_1 &\le \left\|\boldsymbol{U}_\star^{(1)}\right\|_{2,\infty}\left\|\boldsymbol{\Delta}_{\breve{\boldsymbol{U}}^{(1)}}\right\|\left\|\mathcal{M}_1\left(\boldsymbol{\mathcal{Y}}-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)^\top \boldsymbol{U}_0^{(1)}(\boldsymbol{Q}_0^{(1)})^{-\top}(\breve{\boldsymbol{U}}^{(1)\top}\breve{\boldsymbol{U}}^{(1)})^{-1}\boldsymbol{\Sigma}_\star^{(1)}\right\| \\ &= \left\|\boldsymbol{U}_\star^{(1)}\right\|_{2,\infty}\left\|\boldsymbol{\Delta}_{\breve{\boldsymbol{U}}^{(1)}}\right\|\left\|\breve{\boldsymbol{U}}^{(1)}(\breve{\boldsymbol{U}}^{(1)\top}\breve{\boldsymbol{U}}^{(1)})^{-1}\boldsymbol{\Sigma}_\star^{(1)}\right\|,\end{aligned}$$

where in the second line we have used the relation

$$\breve{\boldsymbol{U}}_0^{(1)\top}\breve{\boldsymbol{U}}_0^{(1)} = \boldsymbol{U}_0^{(1)\top}\mathcal{M}_1\left(\boldsymbol{\mathcal{Y}}-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)\mathcal{M}_1\left(\boldsymbol{\mathcal{Y}}-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)^\top \boldsymbol{U}_0^{(1)}$$

given by Lemma 6 (cf. (66)), and the short-hand notation in (21). Invoking Lemma 8 (cf. (70c)) and the incoherence assumption $\left\|\boldsymbol{U}_\star^{(1)}\right\|_{2,\infty} \le \sqrt{\frac{\mu r_1}{n_1}}$, we arrive at

$$\mathfrak{A}_1 \le \frac{1}{(1-\epsilon_0)^3}\sqrt{\frac{\mu r_1}{n_1}}\left\|\boldsymbol{\Delta}_{\breve{\boldsymbol{U}}^{(1)}}\right\|.$$

Furthermore, by Lemma 8 (cf. (70e)), it holds that

$$\left\|\boldsymbol{\Delta}_{\breve{\boldsymbol{U}}^{(1)}}\right\| \le \left\|\boldsymbol{\Delta}_{\breve{\boldsymbol{U}}^{(1)}}\right\|_{\mathsf{F}} \le 2\left(1+\epsilon_0+\frac{\epsilon_0^2}{3}\right)\mathrm{dist}(\boldsymbol{F}_0,\boldsymbol{F}_\star) \le 2\left(1+\epsilon_0+\frac{\epsilon_0^2}{3}\right)\epsilon_0\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star),$$

leading to

$$\mathfrak{A}_1 \le \frac{2\epsilon_0}{(1-\epsilon_0)^3}\left(1+\epsilon_0+\frac{\epsilon_0^2}{3}\right)\sqrt{\frac{\mu r_1}{n_1}}\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star) \le 0.57\sqrt{\frac{\mu r_1}{n_1}}\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star), \quad (60)$$

where the last inequality holds as long as $\epsilon_0 \le 0.15$.

Step 2: bounding $\mathfrak{A}_2$. For $\mathfrak{A}_2$, we have

$$\begin{aligned}\mathfrak{A}_2 &\le \left\|\mathcal{M}_1\left(\boldsymbol{\mathcal{S}}_\star-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)\right\|_{1,\infty}\left\|\mathcal{M}_1\left(\boldsymbol{\mathcal{Y}}-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)^\top \boldsymbol{U}_0^{(1)}(\boldsymbol{Q}_0^{(1)})^{-\top}(\boldsymbol{\Sigma}_\star^{(1)})^{-1}\right\|_{2,\infty}\left\|\boldsymbol{\Sigma}_\star^{(1)}(\breve{\boldsymbol{U}}^{(1)\top}\breve{\boldsymbol{U}}^{(1)})^{-1}\boldsymbol{\Sigma}_\star^{(1)}\right\| \\ &\le \frac{\alpha n_2 n_3}{(1-\epsilon_0)^6}\left\|\boldsymbol{\mathcal{S}}_\star-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{X}}_\star+\boldsymbol{\mathcal{S}}_\star\right)\right\|_\infty\left\|\mathcal{M}_1\left(\boldsymbol{\mathcal{Y}}-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)^\top \boldsymbol{U}_0^{(1)}(\boldsymbol{Q}_0^{(1)})^{-\top}(\boldsymbol{\Sigma}_\star^{(1)})^{-1}\right\|_{2,\infty}.\end{aligned}$$

where we have used the α-sparsity of $\mathcal{M}_1\left(\boldsymbol{\mathcal{S}}_\star-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)$ given by Lemma 12 and Lemma 8 (cf. (70d)) in the second line. Apply Lemma 12 with $\boldsymbol{\mathcal{X}}=\boldsymbol{0}$ to get

$$\begin{aligned}\mathfrak{A}_2 &\le \frac{2\alpha n_2 n_3}{(1-\epsilon_0)^6}\zeta_0\left\|\mathcal{M}_1\left(\boldsymbol{\mathcal{Y}}-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)^\top \boldsymbol{U}_0^{(1)}(\boldsymbol{Q}_0^{(1)})^{-\top}(\boldsymbol{\Sigma}_\star^{(1)})^{-1}\right\|_{2,\infty} \\ &\le \frac{4c_0}{(1-\epsilon_0)^6}\sqrt{\frac{n_2 n_3}{\mu n_1 r_1 r_2 r_3}}\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)\left\|\mathcal{M}_1\left(\boldsymbol{\mathcal{Y}}-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)^\top \boldsymbol{U}_0^{(1)}(\boldsymbol{Q}_0^{(1)})^{-\top}(\boldsymbol{\Sigma}_\star^{(1)})^{-1}\right\|_{2,\infty}, \quad (61)\end{aligned}$$

where the second line follows from $\zeta_0 \le 2\|\boldsymbol{\mathcal{X}}_\star\|_\infty \le 2\sqrt{\frac{\mu^3 r_1 r_2 r_3}{n_1 n_2 n_3}}\kappa\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$ (cf. Lemma 5), as well as the assumption $\alpha \le \frac{c_0}{\mu^2 r_1 r_2 r_3 \kappa}$. To continue,

$$\left\|\mathcal{M}_1\left(\boldsymbol{\mathcal{Y}}-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)^\top \boldsymbol{U}_0^{(1)}(\boldsymbol{Q}_0^{(1)})^{-\top}(\boldsymbol{\Sigma}_\star^{(1)})^{-1}\right\|_{2,\infty} \le r_1\left\|\mathcal{M}_1\left(\boldsymbol{\mathcal{Y}}-\mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)^\top \boldsymbol{U}_0^{(1)}(\boldsymbol{Q}_0^{(1)})^{-\top}(\boldsymbol{\Sigma}_\star^{(1)})^{-1}\right\|_\infty$$

$$
\begin{aligned}
&\leq r_1 \left\| (\boldsymbol{\Sigma}_\star^{(1)})^{-1} (\boldsymbol{Q}_0^{(1)})^{-1} \boldsymbol{U}_0^{(1)\top} \mathcal{M}_1 \left(\boldsymbol{\mathcal{Y}} - \mathcal{T}_{\zeta_0} \left(\boldsymbol{\mathcal{Y}} \right) \right) \right\|_{2,\infty} \\
&= r_1 \left\| (\boldsymbol{\Sigma}_\star^{(1)})^{-1} \breve{\boldsymbol{U}}^{(1)\top} \breve{\boldsymbol{U}}^{(1)} (\boldsymbol{\Sigma}_\star^{(1)})^{-1} \right\|_{\infty}^{1/2} \\
&= r_1 \left\| \breve{\boldsymbol{U}}^{(1)} (\boldsymbol{\Sigma}_\star^{(1)})^{-1} \right\|_{2,\infty},
\end{aligned}
$$

where we have used Lemma 6 (cf. (66)) and the relation $\|\boldsymbol{A}\|_{2,\infty}^2 = \left\|\boldsymbol{A}\boldsymbol{A}^\top\right\|_\infty$ in the first equality. Using the definition of $\breve{\boldsymbol{U}}^{(k)}$ from (4), we have

$$
\begin{aligned}
\left\| \breve{\boldsymbol{U}}^{(1)} (\boldsymbol{\Sigma}_\star^{(1)})^{-1} \right\|_{2,\infty} &= \left\| (\boldsymbol{U}^{(3)} \otimes \boldsymbol{U}^{(2)}) \mathcal{M}_1 \left(\boldsymbol{\mathcal{G}} \right)^\top \left(\boldsymbol{\Sigma}_\star^{(1)} \right)^{-1} \right\|_{2,\infty} \leq \left\| \boldsymbol{U}^{(3)} \otimes \boldsymbol{U}^{(2)} \right\|_{2,\infty} \left\| \mathcal{M}_1 \left(\boldsymbol{\mathcal{G}} \right)^\top \left(\boldsymbol{\Sigma}_\star^{(1)} \right)^{-1} \right\| \\
&\leq \left\| \boldsymbol{U}^{(3)} \right\|_{2,\infty} \left\| \boldsymbol{U}^{(2)} \right\|_{2,\infty} \left(\left\| \mathcal{M}_1 \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{G}}} \right)^\top (\boldsymbol{\Sigma}_\star^{(1)})^{-1} \right\| + \left\| \mathcal{M}_1 \left(\boldsymbol{\mathcal{G}}_\star \right)^\top (\boldsymbol{\Sigma}_\star^{(1)})^{-1} \right\| \right).
\end{aligned}
$$

Applying the triangle inequality on the decompositions $\boldsymbol{U}^{(k)} = \boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}} + \boldsymbol{U}_\star^{(k)}$ and $\boldsymbol{\mathcal{G}} = \boldsymbol{\Delta}_{\boldsymbol{\mathcal{G}}} + \boldsymbol{\mathcal{G}}_\star$, and with Lemma 8 (cf. (70a)) and $\left\| \mathcal{M}_1 \left(\boldsymbol{\mathcal{G}}_\star \right)^\top (\boldsymbol{\Sigma}_\star^{(1)})^{-1} \right\| = 1$, it follows from the above inequalities that

$$
\begin{aligned}
&\left\| \mathcal{M}_1 \left(\boldsymbol{\mathcal{Y}} - \mathcal{T}_{\zeta_0} \left(\boldsymbol{\mathcal{Y}} \right) \right)^\top \boldsymbol{U}_0^{(1)} (\boldsymbol{Q}_0^{(1)})^{-\top} (\boldsymbol{\Sigma}_\star^{(1)})^{-1} \right\|_{2,\infty} \\
&\leq (1+\epsilon_0) r_1 \left(\|\boldsymbol{\Delta}_{\boldsymbol{U}^{(3)}}\|_{2,\infty} + \left\| \boldsymbol{U}_\star^{(3)} \right\|_{2,\infty} \right) \left(\|\boldsymbol{\Delta}_{\boldsymbol{U}^{(2)}}\|_{2,\infty} + \left\| \boldsymbol{U}_\star^{(2)} \right\|_{2,\infty} \right) \\
&\leq (1+\epsilon_0) r_1 \left(\frac{\left\| \boldsymbol{\Delta}_{\boldsymbol{U}^{(3)}} \boldsymbol{\Sigma}_\star^{(3)} \right\|_{2,\infty}}{\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)} + \sqrt{\frac{\mu r_3}{n_3}} \right) \left(\frac{\left\| \boldsymbol{\Delta}_{\boldsymbol{U}^{(2)}} \boldsymbol{\Sigma}_\star^{(2)} \right\|_{2,\infty}}{\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)} + \sqrt{\frac{\mu r_2}{n_2}} \right),
\end{aligned}
$$

where the last line uses the relationship $\|\boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}}\|_{2,\infty} \leq \frac{\|\boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}} \boldsymbol{\Sigma}_\star^{(k)}\|_{2,\infty}}{\sigma_{\min}(\boldsymbol{\Sigma}_\star^{(k)})}$ and the inequality $\left\| \boldsymbol{U}_\star^{(k)} \right\|_{2,\infty} \leq \sqrt{\frac{\mu r_k}{n_k}}$. Plug this back into (61) to arrive at

$$
\mathfrak{A}_2 \leq \frac{0.02}{\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)} \sqrt{\frac{r_1}{\mu n_1}} \left(\sqrt{\frac{n_3}{r_3}} \left\| \boldsymbol{\Delta}_{\boldsymbol{U}^{(3)}} \boldsymbol{\Sigma}_\star^{(3)} \right\|_{2,\infty} + \sqrt{\mu} \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star) \right) \left(\sqrt{\frac{n_2}{r_2}} \left\| \boldsymbol{\Delta}_{\boldsymbol{U}^{(2)}} \boldsymbol{\Sigma}_\star^{(2)} \right\|_{2,\infty} + \sqrt{\mu} \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star) \right), \tag{62}
$$

where we simplified the constants using the assumption $\epsilon_0 = 54.1 c_0 \leq 0.15$.

Step 3: Putting things together. Combining (60), (62), and (59), we have

$$
\begin{aligned}
&\frac{1}{\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)} \sqrt{\frac{n_1}{\mu r_1}} \left\| \boldsymbol{\Delta}_{\boldsymbol{U}^{(1)}} \boldsymbol{\Sigma}_\star^{(1)} \right\|_{2,\infty} \\
&\leq 0.57 + 0.02 \left(\frac{1}{\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)} \sqrt{\frac{n_3}{\mu r_3}} \left\| \boldsymbol{\Delta}_{\boldsymbol{U}^{(3)}} \boldsymbol{\Sigma}_\star^{(3)} \right\|_{2,\infty} + 1 \right) \left(\frac{1}{\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)} \sqrt{\frac{n_2}{\mu r_2}} \left\| \boldsymbol{\Delta}_{\boldsymbol{U}^{(2)}} \boldsymbol{\Sigma}_\star^{(2)} \right\|_{2,\infty} + 1 \right).
\end{aligned}
$$

Similar inequalities hold for $\sqrt{\frac{n_2}{r_2}} \left\| \boldsymbol{\Delta}_{\boldsymbol{U}^{(2)}} \boldsymbol{\Sigma}_\star^{(2)} \right\|_{2,\infty}$ and $\sqrt{\frac{n_3}{r_3}} \left\| \boldsymbol{\Delta}_{\boldsymbol{U}^{(3)}} \boldsymbol{\Sigma}_\star^{(3)} \right\|_{2,\infty}$. Taking the maximum as $\mathcal{I} := \max_k \frac{1}{\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)} \sqrt{\frac{n_k}{\mu r_k}} \left\| \boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}} \boldsymbol{\Sigma}_\star^{(k)} \right\|_{2,\infty}$, we have

$$
\mathcal{I} \leq 0.57 + 0.02(\mathcal{I}+1)^2 \qquad \Longrightarrow \qquad \mathcal{I} \leq 0.62,
$$

and consequently, $\max_k \sqrt{\frac{n_k}{r_k}} \left\| \boldsymbol{\Delta}_{\boldsymbol{U}^{(k)}} \boldsymbol{\Sigma}_\star^{(k)} \right\|_{2,\infty} < \sqrt{\mu} \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$ as claimed.

We are left with proving the decomposition (58).

Proof of (58) . By the assumption

$$\alpha \le \frac{c_0}{\mu^2 r_1 r_2 r_3 \kappa} \le \frac{c_0}{\sqrt{\mu^3 r_1 r_2 r_3 r \kappa}},$$

in view of Lemma 3, we have that

$$\operatorname{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) < 54.1 c_0 \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star) =: \epsilon_0 \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star) \le 0.15 \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star), \tag{63}$$

where $\epsilon_0 = 54.1 c_0 \le 0.15$ as long as $c_0 > 0$ is small enough. Given $\operatorname{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) < \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$, we know that $\{\boldsymbol{Q}_0^{(k)}\}_{k=1}^3$, the optimal alignment matrices between $\boldsymbol{F}_0$ and $\boldsymbol{F}_\star$ exist by [TMPB+22, Lemma 6].

We now aim to control the incoherence. We begin with the equality guaranteed by Lemma 6 (cf. (66)),

$$\boldsymbol{U}_0^{(k)} \breve{\boldsymbol{U}}_0^{(k)\top} \breve{\boldsymbol{U}}_0^{(k)} = \mathcal{M}_k\left(\boldsymbol{\mathcal{Y}} - \mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right) \mathcal{M}_k\left(\boldsymbol{\mathcal{Y}} - \mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)^\top \boldsymbol{U}_0^{(k)}.$$

Again, we will focus on the case with $k = 1$; the other modes will follow from the same arguments. Given that $\left(\breve{\boldsymbol{U}}_0^{(1)\top} \breve{\boldsymbol{U}}_0^{(1)}\right)^{-1}$ exists since $\breve{\boldsymbol{U}}_0^{(1)\top} \breve{\boldsymbol{U}}_0^{(1)} = \mathcal{M}_1\left(\boldsymbol{\mathcal{G}}_0\right) \mathcal{M}_1\left(\boldsymbol{\mathcal{G}}_0\right)^\top$ is positive definite, right-multiplying $\left(\breve{\boldsymbol{U}}_0^{(1)\top} \breve{\boldsymbol{U}}_0^{(1)}\right)^{-1}$ on both sides of the above equation yields

$$\boldsymbol{U}_0^{(1)} = \mathcal{M}_1\left(\boldsymbol{\mathcal{Y}} - \mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right) \mathcal{M}_1\left(\boldsymbol{\mathcal{Y}} - \mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)^\top \boldsymbol{U}_0^{(1)} (\breve{\boldsymbol{U}}_0^{(1)\top} \breve{\boldsymbol{U}}_0^{(1)})^{-1}.$$

Plug in the relation $\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{X}}_\star + \boldsymbol{\mathcal{S}}_\star$ to get

$$\begin{aligned}
\boldsymbol{U}_0^{(1)} &= \mathcal{M}_1\left(\boldsymbol{\mathcal{S}}_\star - \mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right) \mathcal{M}_1\left(\boldsymbol{\mathcal{Y}} - \mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)^\top \boldsymbol{U}_0^{(1)} (\breve{\boldsymbol{U}}_0^{(1)\top} \breve{\boldsymbol{U}}_0^{(1)})^{-1} \\
&\quad + \mathcal{M}_1\left(\boldsymbol{\mathcal{X}}_\star\right) \mathcal{M}_1\left(\boldsymbol{\mathcal{Y}} - \mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)^\top \boldsymbol{U}_0^{(1)} (\breve{\boldsymbol{U}}_0^{(1)\top} \breve{\boldsymbol{U}}_0^{(1)})^{-1} \\
&= \mathcal{M}_1\left(\boldsymbol{\mathcal{S}}_\star - \mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right) \mathcal{M}_1\left(\boldsymbol{\mathcal{Y}} - \mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)^\top \boldsymbol{U}_0^{(1)} (\breve{\boldsymbol{U}}_0^{(1)\top} \breve{\boldsymbol{U}}_0^{(1)})^{-1} \\
&\quad + \boldsymbol{U}_\star^{(1)} \breve{\boldsymbol{U}}_\star^{(1)\top} \mathcal{M}_1\left(\boldsymbol{\mathcal{Y}} - \mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)^\top \boldsymbol{U}_0^{(1)} (\breve{\boldsymbol{U}}_0^{(1)\top} \breve{\boldsymbol{U}}_0^{(1)})^{-1}.
\end{aligned}$$

Subtracting $\boldsymbol{U}_\star^{(1)} (\boldsymbol{Q}_0^{(1)})^{-1}$ on both sides gets us

$$\begin{aligned}
\boldsymbol{U}_0^{(1)} - \boldsymbol{U}_\star^{(1)} (\boldsymbol{Q}_0^{(1)})^{-1} &= \mathcal{M}_1\left(\boldsymbol{\mathcal{S}}_\star - \mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right) \mathcal{M}_1\left(\boldsymbol{\mathcal{Y}} - \mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)^\top \boldsymbol{U}_0^{(1)} (\breve{\boldsymbol{U}}_0^{(1)\top} \breve{\boldsymbol{U}}_0^{(1)})^{-1} \\
&\quad + \boldsymbol{U}_\star^{(1)} \breve{\boldsymbol{U}}_\star^{(1)\top} \mathcal{M}_1\left(\boldsymbol{\mathcal{Y}} - \mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)^\top \boldsymbol{U}_0^{(1)} (\breve{\boldsymbol{U}}_0^{(1)\top} \breve{\boldsymbol{U}}_0^{(1)})^{-1} - \boldsymbol{U}_\star^{(1)} (\boldsymbol{Q}_0^{(1)})^{-1}.
\end{aligned} \tag{64}$$

Observe that

$$\begin{aligned}
\boldsymbol{U}_\star^{(1)} (\boldsymbol{Q}_0^{(1)})^{-1} &= \boldsymbol{U}_\star^{(1)} (\boldsymbol{Q}_0^{(1)})^{-1} \breve{\boldsymbol{U}}_0^{(1)\top} \breve{\boldsymbol{U}}_0^{(1)} (\breve{\boldsymbol{U}}_0^{(1)\top} \breve{\boldsymbol{U}}_0^{(1)})^{-1} \\
&= \boldsymbol{U}_\star^{(1)} (\boldsymbol{Q}_0^{(1)})^{-1} \breve{\boldsymbol{U}}_0^{(1)\top} \mathcal{M}_1\left(\boldsymbol{\mathcal{Y}} - \mathcal{T}_{\zeta_0}\left(\boldsymbol{\mathcal{Y}}\right)\right)^\top \boldsymbol{U}_0^{(1)} (\breve{\boldsymbol{U}}_0^{(1)\top} \breve{\boldsymbol{U}}_0^{(1)})^{-1},
\end{aligned}$$

where we have used Lemma 6 (cf. (65)) in the second step. Plug this result into (64), multiply both sides with $\boldsymbol{Q}_0^{(1)} \boldsymbol{\Sigma}_\star^{(1)}$, and recall the short-hand notation in (21) initiated at $t = 0$ to arrive at the claimed decomposition.

E Technical lemmas

This section collects several technical lemmas that are useful in the main proofs.

E.1 Tensor algebra

We start with a simple bound on the element-wise maximum norm of an incoherent tensor.

Lemma 5. *Suppose that* $\boldsymbol{\mathcal{X}}_\star = (\boldsymbol{U}_\star^{(1)}, \boldsymbol{U}_\star^{(2)}, \boldsymbol{U}_\star^{(3)}) \cdot \boldsymbol{\mathcal{G}}_\star \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ *have multilinear rank* $\boldsymbol{r} = (r_1, r_2, r_3)$ *and is* μ*-incoherent. Then one has* $\|\boldsymbol{\mathcal{X}}_\star\|_\infty \le \sqrt{\frac{\mu^3 r_1 r_2 r_3}{n_1 n_2 n_3}} \kappa \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$.

Proof. By the property of matricizations (cf. (4)), we have $\mathcal{M}_k\left(\boldsymbol{\mathcal{X}}_\star\right) = \boldsymbol{U}_\star^{(k)}\breve{\boldsymbol{U}}_\star^{(k)\top}$ for any $k = 1, 2, 3$. Furthermore, $\|\cdot\|_\infty$ is invariant to matricizations, so

$$\|\boldsymbol{\mathcal{X}}_\star\|_\infty = \|\mathcal{M}_k\left(\boldsymbol{\mathcal{X}}_\star\right)\|_\infty = \left\|\boldsymbol{U}_\star^{(k)}\breve{\boldsymbol{U}}_\star^{(k)\top}\right\|_\infty.$$

Without loss of generality, we choose $k = 1$. It then follows that

$$\begin{aligned}\|\boldsymbol{\mathcal{X}}_\star\|_\infty &\le \left\|\boldsymbol{U}_\star^{(1)}\right\|_{2,\infty}\left\|\breve{\boldsymbol{U}}_\star^{(1)}\right\|_{2,\infty} \le \left\|\boldsymbol{U}_\star^{(1)}\right\|_{2,\infty}\left\|\boldsymbol{U}_\star^{(3)}\right\|_{2,\infty}\left\|\boldsymbol{U}_\star^{(2)}\right\|_{2,\infty}\|\mathcal{M}_1\left(\boldsymbol{\mathcal{G}}_\star\right)\| \\ &\le \sqrt{\frac{\mu^3 r_1 r_2 r_3}{n_1 n_2 n_3}}\sigma_{\max}(\mathcal{M}_1\left(\boldsymbol{\mathcal{X}}_\star\right)),\end{aligned}$$

where the second line follows from the definition of incoherence of $\boldsymbol{\mathcal{X}}_\star$, i.e., $\left\|\boldsymbol{U}_\star^{(k)}\right\|_{2,\infty} \le \sqrt{\frac{\mu r_k}{n_k}}$, and $\|\mathcal{M}_1\left(\boldsymbol{\mathcal{G}}_\star\right)\| = \sigma_{\max}(\mathcal{M}_1\left(\boldsymbol{\mathcal{X}}_\star\right))$. Applying the above bound for any k and taking the tightest bound, we have

$$\|\boldsymbol{\mathcal{X}}_\star\|_\infty \le \sqrt{\frac{\mu^3 r_1 r_2 r_3}{n_1 n_2 n_3}}\min_k \sigma_{\max}(\mathcal{M}_k\left(\boldsymbol{\mathcal{X}}_\star\right)) = \sqrt{\frac{\mu^3 r_1 r_2 r_3}{n_1 n_2 n_3}}\kappa\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star),$$

where we have used the definition of κ. □

We next show a key tensor algebraic result that is crucial in establishing the incoherence property of the spectral initialization.

Lemma 6. *Given a tensor $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{n_1\times n_2\times n_3}$, suppose its rank-$\boldsymbol{r}$ truncated HOSVD is $(\boldsymbol{U}^{(1)}, \boldsymbol{U}^{(2)}, \boldsymbol{U}^{(3)})\cdot\boldsymbol{\mathcal{G}}$ with $\boldsymbol{U}^{(k)} \in \mathbb{R}^{n_k\times r_k}$ and $\boldsymbol{\mathcal{G}} \in \mathbb{R}^{r_1\times r_2\times r_3}$, and $r_k \le n_k$ for $k = 1, 2, 3$. Then,*

$$\boldsymbol{U}^{(k)\top}\mathcal{M}_k\left(\boldsymbol{\mathcal{T}}\right)\breve{\boldsymbol{U}}^{(k)} = \breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}, \tag{65}$$

$$\mathcal{M}_k\left(\boldsymbol{\mathcal{T}}\right)\mathcal{M}_k\left(\boldsymbol{\mathcal{T}}\right)^\top\boldsymbol{U}^{(k)} = \boldsymbol{U}^{(k)}\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}, \tag{66}$$

where $\breve{\boldsymbol{U}}^{(k)}$ is defined in (4).

Proof. Set the full HOSVD of $\boldsymbol{\mathcal{T}}$ to be $\left(\boldsymbol{U}_{\boldsymbol{\mathcal{T}}}^{(1)}, \boldsymbol{U}_{\boldsymbol{\mathcal{T}}}^{(2)}, \boldsymbol{U}_{\boldsymbol{\mathcal{T}}}^{(3)}\right)\cdot\boldsymbol{\mathcal{G}}_{\boldsymbol{\mathcal{T}}}$. Since $\boldsymbol{U}_{\boldsymbol{\mathcal{T}}}^{(k)}$ contains all the left singular vectors of $\mathcal{M}_k\left(\boldsymbol{\mathcal{T}}\right)$, it can be decomposed into the following block structure:

$$\boldsymbol{U}_{\boldsymbol{\mathcal{T}}}^{(k)} = \begin{bmatrix}\boldsymbol{U}^{(k)} & \bar{\boldsymbol{U}}^{(k)}\end{bmatrix}, \tag{67}$$

where $\bar{\boldsymbol{U}}^{(k)} \in \mathbb{R}^{n_k\times(n_k-r_k)}$ contains the bottom $(n_k - r_k)$ left singular vectors. The rest of the proof focuses on the first mode (i.e., $k = 1$), while other modes follow from similar arguments.

Let us begin with proving (65). Plugging in the definition of $\breve{\boldsymbol{U}}^{(1)}$ from (4a), we see that

$$\begin{aligned}\boldsymbol{U}^{(1)\top}\mathcal{M}_1\left(\boldsymbol{\mathcal{T}}\right)\breve{\boldsymbol{U}}^{(1)} &= \boldsymbol{U}^{(1)\top}\boldsymbol{U}_{\boldsymbol{\mathcal{T}}}^{(1)}\mathcal{M}_1\left(\boldsymbol{\mathcal{G}}_{\boldsymbol{\mathcal{T}}}\right)\left(\boldsymbol{U}_{\boldsymbol{\mathcal{T}}}^{(3)}\otimes\boldsymbol{U}_{\boldsymbol{\mathcal{T}}}^{(2)}\right)^\top\left(\boldsymbol{U}^{(3)}\otimes\boldsymbol{U}^{(2)}\right)\mathcal{M}_1\left(\boldsymbol{\mathcal{G}}\right)^\top \\ &= \boldsymbol{U}^{(1)\top}\begin{bmatrix}\boldsymbol{U}^{(1)} & \bar{\boldsymbol{U}}^{(1)}\end{bmatrix}\mathcal{M}_1\left(\boldsymbol{\mathcal{G}}_{\boldsymbol{\mathcal{T}}}\right)\left(\begin{bmatrix}\boldsymbol{U}^{(3)\top} \\ \bar{\boldsymbol{U}}^{(3)\top}\end{bmatrix}\otimes\begin{bmatrix}\boldsymbol{U}^{(2)\top} \\ \bar{\boldsymbol{U}}^{(2)\top}\end{bmatrix}\right)\left(\boldsymbol{U}^{(3)}\otimes\boldsymbol{U}^{(2)}\right)\mathcal{M}_1\left(\boldsymbol{\mathcal{G}}\right)^\top,\end{aligned}$$

where the second line uses the block structure (67). By the mixed product property of Kronecker products, we have

$$\begin{aligned}\boldsymbol{U}^{(1)\top}\mathcal{M}_1\left(\boldsymbol{\mathcal{T}}\right)\breve{\boldsymbol{U}}^{(1)} &= \begin{bmatrix}\boldsymbol{I}_{r_1} & \boldsymbol{0}\end{bmatrix}\mathcal{M}_1\left(\boldsymbol{\mathcal{G}}_{\boldsymbol{\mathcal{T}}}\right)\left(\begin{bmatrix}\boldsymbol{U}^{(3)\top}\boldsymbol{U}^{(3)} \\ \bar{\boldsymbol{U}}^{(3)\top}\boldsymbol{U}^{(3)}\end{bmatrix}\otimes\begin{bmatrix}\boldsymbol{U}^{(2)\top}\boldsymbol{U}^{(2)} \\ \bar{\boldsymbol{U}}^{(2)\top}\boldsymbol{U}^{(2)}\end{bmatrix}\right)\mathcal{M}_1\left(\boldsymbol{\mathcal{G}}\right)^\top \\ &= \begin{bmatrix}\boldsymbol{I}_{r_1} & \boldsymbol{0}\end{bmatrix}\mathcal{M}_1\left(\boldsymbol{\mathcal{G}}_{\boldsymbol{\mathcal{T}}}\right)\left(\begin{bmatrix}\boldsymbol{I}_{r_3} \\ \boldsymbol{0}\end{bmatrix}\otimes\begin{bmatrix}\boldsymbol{I}_{r_2} \\ \boldsymbol{0}\end{bmatrix}\right)\mathcal{M}_1\left(\boldsymbol{\mathcal{G}}\right)^\top,\end{aligned} \tag{68}$$

where we used the fact that the singular vectors are orthonormal. Note that

$$\begin{bmatrix} \boldsymbol{I}_{r_1} & \boldsymbol{0} \end{bmatrix} \mathcal{M}_1\left(\boldsymbol{\mathcal{G}}_{\boldsymbol{\mathcal{T}}}\right)\left(\begin{bmatrix} \boldsymbol{I}_{r_3} \\ \boldsymbol{0} \end{bmatrix} \otimes \begin{bmatrix} \boldsymbol{I}_{r_2} \\ \boldsymbol{0} \end{bmatrix}\right) = \mathcal{M}_1\left(\left(\begin{bmatrix} \boldsymbol{I}_{r_1} & \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{I}_{r_2} & \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{I}_{r_3} & \boldsymbol{0} \end{bmatrix}\right) \cdot \boldsymbol{\mathcal{G}}_{\boldsymbol{\mathcal{T}}}\right) \in \mathbb{R}^{r_1 \times r_2 r_3},$$

where $\left(\begin{bmatrix} \boldsymbol{I}_{r_1} & \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{I}_{r_2} & \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{I}_{r_3} & \boldsymbol{0} \end{bmatrix}\right) \cdot \boldsymbol{\mathcal{G}}_{\boldsymbol{\mathcal{T}}}$ is equivalent to trimming off the entries $[\boldsymbol{\mathcal{G}}_{\boldsymbol{\mathcal{T}}}]_{i_1,i_2,i_3}$ for all $i_1 > r_1$, $i_2 > r_2$, or $i_3 > r_3$. Since $\boldsymbol{\mathcal{G}}$ is the section of $[\boldsymbol{\mathcal{G}}_{\boldsymbol{\mathcal{T}}}]_{i_1,i_2,i_3}$ where $1 \le i_1 \le r_1, 1 \le i_2 \le r_2, 1 \le i_3 \le r_3$ [VVM12], we have

$$\left(\begin{bmatrix} \boldsymbol{I}_{r_1} & \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{I}_{r_2} & \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{I}_{r_3} & \boldsymbol{0} \end{bmatrix}\right) \cdot \boldsymbol{\mathcal{G}}_{\boldsymbol{\mathcal{T}}} = \boldsymbol{\mathcal{G}}.$$

This allows us to simplify (68) as

$$\boldsymbol{U}^{(1)\top} \mathcal{M}_1\left(\boldsymbol{\mathcal{T}}\right) \breve{\boldsymbol{U}}^{(1)} = \mathcal{M}_1\left(\boldsymbol{\mathcal{G}}\right) \mathcal{M}_1\left(\boldsymbol{\mathcal{G}}\right)^\top = \breve{\boldsymbol{U}}^{(1)\top} \breve{\boldsymbol{U}}^{(1)},$$

where the last equality follows from the definition of $\breve{\boldsymbol{U}}^{(1)}$ from (4) and $(\boldsymbol{U}^{(3)} \otimes \boldsymbol{U}^{(2)})^\top (\boldsymbol{U}^{(3)} \otimes \boldsymbol{U}^{(2)}) = \boldsymbol{I}$ by construction of the HOSVD. This completes the proof of (65).

Turning to (66), we begin with the observation

$$\begin{aligned} \mathcal{M}_1\left(\boldsymbol{\mathcal{T}}\right) \mathcal{M}_1\left(\boldsymbol{\mathcal{T}}\right)^\top &= \boldsymbol{U}_{\boldsymbol{\mathcal{T}}}^{(1)} \mathcal{M}_1\left(\boldsymbol{\mathcal{G}}_{\boldsymbol{\mathcal{T}}}\right) \left(\boldsymbol{U}_{\boldsymbol{\mathcal{T}}}^{(3)} \otimes \boldsymbol{U}_{\boldsymbol{\mathcal{T}}}^{(2)}\right)^\top \left(\boldsymbol{U}_{\boldsymbol{\mathcal{T}}}^{(3)} \otimes \boldsymbol{U}_{\boldsymbol{\mathcal{T}}}^{(2)}\right) \mathcal{M}_1\left(\boldsymbol{\mathcal{G}}_{\boldsymbol{\mathcal{T}}}\right)^\top \boldsymbol{U}_{\boldsymbol{\mathcal{T}}}^{(1)\top} \\ &= \boldsymbol{U}_{\boldsymbol{\mathcal{T}}}^{(1)} \mathcal{M}_1\left(\boldsymbol{\mathcal{G}}_{\boldsymbol{\mathcal{T}}}\right) \mathcal{M}_1\left(\boldsymbol{\mathcal{G}}_{\boldsymbol{\mathcal{T}}}\right)^\top \boldsymbol{U}_{\boldsymbol{\mathcal{T}}}^{(1)\top}. \end{aligned} \tag{69}$$

By the "all-orthogonal" property in [DLDMV00], $\mathcal{M}_1\left(\boldsymbol{\mathcal{G}}_{\boldsymbol{\mathcal{T}}}\right) \mathcal{M}_1\left(\boldsymbol{\mathcal{G}}_{\boldsymbol{\mathcal{T}}}\right)^\top = (\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{T}}}^{(1)})^2$, the squared singular value matrix of $\mathcal{M}_1\left(\boldsymbol{\mathcal{T}}\right)$. By assigning $\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{T}}}^{(1)}$ with the block representation

$$\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{T}}}^{(1)} = \begin{bmatrix} \boldsymbol{\Sigma}^{(1)} & \boldsymbol{0} \\ \boldsymbol{0} & \bar{\boldsymbol{\Sigma}}^{(1)} \end{bmatrix},$$

where $\boldsymbol{\Sigma}^{(1)} = \sqrt{\mathcal{M}_1\left(\boldsymbol{\mathcal{G}}\right) \mathcal{M}_1\left(\boldsymbol{\mathcal{G}}\right)^\top}$ and $\bar{\boldsymbol{\Sigma}}^{(1)}$ contain the top r_1 singular values and bottom $n_1 - r_1$ singular values, respectively, of $\mathcal{M}_1\left(\boldsymbol{\mathcal{T}}\right)$. Coupled with same block structure as in (67), (69) becomes

$$\mathcal{M}_1\left(\boldsymbol{\mathcal{T}}\right) \mathcal{M}_1\left(\boldsymbol{\mathcal{T}}\right)^\top = \begin{bmatrix} \boldsymbol{U}^{(1)} & \bar{\boldsymbol{U}}^{(1)} \end{bmatrix} \begin{bmatrix} (\boldsymbol{\Sigma}^{(1)})^2 & \boldsymbol{0} \\ \boldsymbol{0} & (\bar{\boldsymbol{\Sigma}}^{(1)})^2 \end{bmatrix} \begin{bmatrix} \boldsymbol{U}^{(1)\top} \\ \bar{\boldsymbol{U}}^{(1)\top} \end{bmatrix}.$$

Multiply by $\boldsymbol{U}^{(1)}$ on the right to arrive at $\mathcal{M}_1\left(\boldsymbol{\mathcal{T}}\right) \mathcal{M}_1\left(\boldsymbol{\mathcal{T}}\right)^\top \boldsymbol{U}^{(1)} = \boldsymbol{U}^{(1)} (\boldsymbol{\Sigma}^{(1)})^2 = \boldsymbol{U}^{(1)} \breve{\boldsymbol{U}}^{(1)\top} \breve{\boldsymbol{U}}^{(1)}$. □

E.2 Perturbation bounds

Below is a useful perturbation bound for matrices.

Lemma 7. *Given two matrices $\boldsymbol{U}, \boldsymbol{U}_\star \in \mathbb{R}^{n \times r}$ that have full column rank, two invertible matrices $\bar{\boldsymbol{Q}}, \boldsymbol{Q} \in \mathbb{R}^{r \times r}$, and a positive definite matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{r \times r}$. Suppose that $\sigma_{\min}(\boldsymbol{U}_\star) > \left\|\boldsymbol{U}\bar{\boldsymbol{Q}} - \boldsymbol{U}_\star\right\|$. Then the following holds true*

$$\left\|\bar{\boldsymbol{Q}}^{-1}(\boldsymbol{Q} - \bar{\boldsymbol{Q}})\boldsymbol{\Sigma}\right\| \le \frac{\left\|\boldsymbol{U}(\boldsymbol{Q} - \bar{\boldsymbol{Q}})\boldsymbol{\Sigma}\right\|}{\sigma_{\min}(\boldsymbol{U}_\star) - \left\|\boldsymbol{U}\bar{\boldsymbol{Q}} - \boldsymbol{U}_\star\right\|}.$$

Proof. It follows that

$$\begin{aligned} \left\|\bar{\boldsymbol{Q}}^{-1}(\boldsymbol{Q} - \bar{\boldsymbol{Q}})\boldsymbol{\Sigma}\right\| &= \left\|\bar{\boldsymbol{Q}}^{-1}(\boldsymbol{U}^\top \boldsymbol{U})^{-1} \boldsymbol{U}^\top \boldsymbol{U}(\boldsymbol{Q} - \bar{\boldsymbol{Q}})\boldsymbol{\Sigma}\right\| \\ &\le \left\|\bar{\boldsymbol{Q}}^{-1}(\boldsymbol{U}^\top \boldsymbol{U})^{-1} \boldsymbol{U}^\top\right\| \left\|\boldsymbol{U}(\boldsymbol{Q} - \bar{\boldsymbol{Q}})\boldsymbol{\Sigma}\right\| \\ &= \frac{\left\|\boldsymbol{U}(\boldsymbol{Q} - \bar{\boldsymbol{Q}})\boldsymbol{\Sigma}\right\|}{\sigma_{\min}(\boldsymbol{U}\bar{\boldsymbol{Q}})}, \end{aligned}$$

where the last equality comes from the fact that $\bar{\boldsymbol{Q}}^{-1}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\boldsymbol{U}^\top$ and $\boldsymbol{U}\bar{\boldsymbol{Q}}$ are pseudoinverses. By Weyl's inequality, we know $|\sigma_i(\boldsymbol{M}) - \sigma_i(\boldsymbol{M}+\boldsymbol{\Delta}_{\boldsymbol{M}})| \le \|\boldsymbol{\Delta}_{\boldsymbol{M}}\|$. Taking $\boldsymbol{M} := \boldsymbol{U}_\star$ and $\boldsymbol{\Delta}_{\boldsymbol{M}} := \boldsymbol{U}\bar{\boldsymbol{Q}} - \boldsymbol{U}_\star$, we have

$$\|\bar{\boldsymbol{Q}}^{-1}(\boldsymbol{Q}-\bar{\boldsymbol{Q}})\boldsymbol{\Sigma}\| \le \frac{\|\boldsymbol{U}(\boldsymbol{Q}-\bar{\boldsymbol{Q}})\boldsymbol{\Sigma}\|}{\sigma_{\min}(\boldsymbol{U}_\star) - \|\boldsymbol{U}\bar{\boldsymbol{Q}}-\boldsymbol{U}_\star\|}$$

as long as $\sigma_{\min}(\boldsymbol{U}_\star) > \|\boldsymbol{U}\bar{\boldsymbol{Q}}-\boldsymbol{U}_\star\|$. □

We also collect a useful lemma regarding perturbation bounds for tensors from [TMPB+22].

Lemma 8 ([TMPB+22, Lemma 10]). *Suppose $\boldsymbol{F} = (\boldsymbol{U}^{(1)},\boldsymbol{U}^{(2)},\boldsymbol{U}^{(3)},\boldsymbol{\mathcal{G}})$ and $\boldsymbol{F} = (\boldsymbol{U}_\star^{(1)},\boldsymbol{U}_\star^{(2)},\boldsymbol{U}_\star^{(3)},\boldsymbol{\mathcal{G}}_\star)$ are aligned, and $\operatorname{dist}(\boldsymbol{F},\boldsymbol{F}_\star) \le \epsilon\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$ for some $0<\epsilon<1$. Then, the following bounds are true:*

$$\left\|\mathcal{M}_k\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{G}}}\right)^\top(\boldsymbol{\Sigma}_\star^{(k)})^{-1}\right\| \le \epsilon; \tag{70a}$$

$$\left\|\boldsymbol{U}^{(k)}\left(\boldsymbol{U}^{(k)\top}\boldsymbol{U}^{(k)}\right)^{-1}\right\| \le \frac{1}{1-\epsilon}; \tag{70b}$$

$$\left\|\breve{\boldsymbol{U}}^{(k)}\left(\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}\right)^{-1}\boldsymbol{\Sigma}_\star^{(k)}\right\| \le \frac{1}{(1-\epsilon)^3}; \tag{70c}$$

$$\left\|\boldsymbol{\Sigma}_\star^{(k)}\left(\breve{\boldsymbol{U}}^{(k)\top}\breve{\boldsymbol{U}}^{(k)}\right)^{-1}\boldsymbol{\Sigma}_\star^{(k)}\right\| \le \frac{1}{(1-\epsilon)^6}; \tag{70d}$$

$$\left\|\breve{\boldsymbol{U}}^{(1)}-\breve{\boldsymbol{U}}_\star^{(1)}\right\|_{\mathsf{F}} \le \left(1+\epsilon+\frac{\epsilon^2}{3}\right)\left(\left\|(\boldsymbol{U}^{(2)}-\boldsymbol{U}_\star^{2)})\boldsymbol{\Sigma}_\star^{(2)}\right\|_{\mathsf{F}} + \left\|(\boldsymbol{U}^{(3)}-\boldsymbol{U}_\star^{3)})\boldsymbol{\Sigma}_\star^{(3)}\right\|_{\mathsf{F}} + \|\boldsymbol{\mathcal{G}}-\boldsymbol{\mathcal{G}}_\star\|_{\mathsf{F}}\right). \tag{70e}$$

For (70e), similar bounds exist for the other modes. Furthermore, if $0<\epsilon\le 0.2$,

$$\left\|(\boldsymbol{U}^{(1)},\boldsymbol{U}^{(2)},\boldsymbol{U}^{(3)})\cdot\boldsymbol{\mathcal{G}} - \boldsymbol{\mathcal{X}}_\star\right\|_{\mathsf{F}} \le 3\operatorname{dist}(\boldsymbol{F},\boldsymbol{F}_\star). \tag{71}$$

The next set of lemmas, which is crucial in our analysis, deals with perturbation bounds when relating a tensor $\boldsymbol{\mathcal{X}} = (\boldsymbol{U}^{(1)},\boldsymbol{U}^{(2)},\boldsymbol{U}^{(3)})\cdot\boldsymbol{\mathcal{G}}$ to the ground truth $\boldsymbol{\mathcal{X}}_\star$, where the tensor tuples $\boldsymbol{F} = (\boldsymbol{U}^{(1)},\boldsymbol{U}^{(2)},\boldsymbol{U}^{(3)},\boldsymbol{\mathcal{G}})$ and $\boldsymbol{F}_\star = (\boldsymbol{U}_\star^{(1)},\boldsymbol{U}_\star^{(2)},\boldsymbol{U}_\star^{(3)},\boldsymbol{\mathcal{G}}_\star)$ are aligned.

Lemma 9. *Let $\boldsymbol{\mathcal{X}}_\star \in \mathbb{R}^{n_1\times n_2\times n_3}$ be μ-incoherent with the Tucker decomposition $\boldsymbol{\mathcal{X}}_\star = (\boldsymbol{U}_\star^{(1)},\boldsymbol{U}_\star^{(2)},\boldsymbol{U}_\star^{(3)})\cdot\boldsymbol{\mathcal{G}}_\star$ of rank $\boldsymbol{r} = (r_1,r_2,r_3)$, and $\{\boldsymbol{\Sigma}_\star^{(k)}\}_{k=1,2,3}$ be the set of singular value matrices of different matricizations of $\boldsymbol{\mathcal{X}}_\star$. In addition, let $\boldsymbol{F} := (\boldsymbol{U}^{(1)},\boldsymbol{U}^{(2)},\boldsymbol{U}^{(3)},\boldsymbol{\mathcal{G}})$ and $\boldsymbol{F}_\star := (\boldsymbol{U}_\star^{(1)},\boldsymbol{U}_\star^{(2)},\boldsymbol{U}_\star^{(3)},\boldsymbol{\mathcal{G}}_\star)$ be aligned, where $\boldsymbol{\mathcal{X}} = (\boldsymbol{U}^{(1)},\boldsymbol{U}^{(2)},\boldsymbol{U}^{(3)})\cdot\boldsymbol{\mathcal{G}}$. Suppose*

$$\max_k\left\{\sqrt{\frac{n_k}{r_k}}\left\|(\boldsymbol{U}^{(k)}-\boldsymbol{U}_\star^{(k)})\boldsymbol{\Sigma}_\star^{(k)}\right\|_{2,\infty}\right\} \le c\sqrt{\mu}\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star) \tag{72}$$

for some $0<c\le 1$. Then for $k=1,2,3$,

$$\left\|\boldsymbol{U}^{(k)}-\boldsymbol{U}_\star^{(k)}\right\|_{2,\infty} \le c\sqrt{\frac{\mu r_k}{n_k}}, \qquad \text{and} \qquad \left\|\boldsymbol{U}^{(k)}\right\|_{2,\infty} \le 2\sqrt{\frac{\mu r_k}{n_k}}. \tag{73}$$

Proof. It follows that for all k,

$$\left\|\boldsymbol{U}^{(k)}-\boldsymbol{U}_\star^{(k)}\right\|_{2,\infty} \le \frac{1}{\sigma_{\min}(\boldsymbol{\Sigma}_\star^{(k)})}\left\|(\boldsymbol{U}^{(k)}-\boldsymbol{U}_\star^{(k)})\boldsymbol{\Sigma}_\star^{(k)}\right\|_{2,\infty} \le c\sqrt{\frac{\mu r_k}{n_k}},$$

where the second inequality follows from (72) and $\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star) \le \sigma_{\min}(\boldsymbol{\Sigma}_\star^{(k)})$. This completes the proof for the first part of (73). With this and the incoherence assumption $\left\|\boldsymbol{U}_\star^{(k)}\right\|_{2,\infty} \le \sqrt{\frac{\mu r_k}{n_k}}$, after applying triangle inequality, we arrive at

$$\left\|\boldsymbol{U}^{(k)}\right\|_{2,\infty} \le \left\|\boldsymbol{U}^{(k)}-\boldsymbol{U}_\star^{(k)}\right\|_{2,\infty} + \left\|\boldsymbol{U}_\star^{(k)}\right\|_{2,\infty} \le 2\sqrt{\frac{\mu r_k}{n_k}},$$

which completes the proof. □

Lemma 10. *Let $\boldsymbol{\mathcal{X}}_\star \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ be μ-incoherent with the Tucker decomposition $\boldsymbol{\mathcal{X}}_\star = \left(\boldsymbol{U}_\star^{(1)}, \boldsymbol{U}_\star^{(2)}, \boldsymbol{U}_\star^{(3)}\right) \cdot \boldsymbol{\mathcal{G}}_\star$ of rank $\boldsymbol{r} = (r_1, r_2, r_3)$. In addition, let $\boldsymbol{F} := (\boldsymbol{U}^{(1)}, \boldsymbol{U}^{(2)}, \boldsymbol{U}^{(3)}, \boldsymbol{\mathcal{G}})$ and $\boldsymbol{F}_\star := (\boldsymbol{U}_\star^{(1)}, \boldsymbol{U}_\star^{(2)}, \boldsymbol{U}_\star^{(3)}, \boldsymbol{\mathcal{G}}_\star)$ be aligned, where $\boldsymbol{\mathcal{X}} = \left(\boldsymbol{U}^{(1)}, \boldsymbol{U}^{(2)}, \boldsymbol{U}^{(3)}\right) \cdot \boldsymbol{\mathcal{G}}$. For $0 < \epsilon_0 < 0.1$ and $0 < c \leq 1$, if*

$$\operatorname{dist}(\boldsymbol{F}, \boldsymbol{F}_\star) \leq \epsilon_0 c \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star), \tag{74a}$$

$$\max_k \left\{ \sqrt{\frac{n_k}{r_k}} \left\| (\boldsymbol{U}^{(k)} - \boldsymbol{U}_\star^{(k)}) \boldsymbol{\Sigma}_\star^{(k)} \right\|_{2,\infty} \right\} \leq c\sqrt{\mu} \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star) \tag{74b}$$

are satisfied, then

$$\|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{X}}_\star\|_\infty \leq \sqrt{\frac{\mu^3 r_1 r_2 r_3}{n_1 n_2 n_3}} \left(8\epsilon_0 + 7\right) c \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star) \leq 8 \sqrt{\frac{\mu^3 r_1 r_2 r_3}{n_1 n_2 n_3}} c \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star).$$

Proof. We can decompose $\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{X}}_\star$ into

$$\begin{aligned} \boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{X}}_\star = &\left(\boldsymbol{U}^{(1)}, \boldsymbol{U}^{(2)}, \boldsymbol{U}^{(3)}\right) \cdot (\boldsymbol{\mathcal{G}} - \boldsymbol{\mathcal{G}}_\star) + \left(\boldsymbol{U}^{(1)} - \boldsymbol{U}_\star^{(1)}, \boldsymbol{U}^{(2)}, \boldsymbol{U}^{(3)}\right) \cdot \boldsymbol{\mathcal{G}}_\star \\ &+ \left(\boldsymbol{U}_\star^{(1)}, \boldsymbol{U}^{(2)} - \boldsymbol{U}_\star^{(2)}, \boldsymbol{U}^{(3)}\right) \cdot \boldsymbol{\mathcal{G}}_\star + \left(\boldsymbol{U}_\star^{(1)}, \boldsymbol{U}_\star^{(2)}, \boldsymbol{U}^{(3)} - \boldsymbol{U}_\star^{(3)}\right) \cdot \boldsymbol{\mathcal{G}}_\star. \end{aligned} \tag{75}$$

Then by the triangle inequality,

$$\begin{aligned} \|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{X}}_\star\|_\infty \leq &\left\| \left(\boldsymbol{U}^{(1)}, \boldsymbol{U}^{(2)}, \boldsymbol{U}^{(3)}\right) \cdot (\boldsymbol{\mathcal{G}} - \boldsymbol{\mathcal{G}}_\star) \right\|_\infty + \left\| \left(\boldsymbol{U}^{(1)} - \boldsymbol{U}_\star^{(1)}, \boldsymbol{U}^{(2)}, \boldsymbol{U}^{(3)}\right) \cdot \boldsymbol{\mathcal{G}}_\star \right\|_\infty \\ &+ \left\| \left(\boldsymbol{U}_\star^{(1)}, \boldsymbol{U}^{(2)} - \boldsymbol{U}_\star^{(2)}, \boldsymbol{U}^{(3)}\right) \cdot \boldsymbol{\mathcal{G}}_\star \right\|_\infty + \left\| \left(\boldsymbol{U}_\star^{(1)}, \boldsymbol{U}_\star^{(2)}, \boldsymbol{U}^{(3)} - \boldsymbol{U}_\star^{(3)}\right) \cdot \boldsymbol{\mathcal{G}}_\star \right\|_\infty \\ = &\underbrace{\left\| \boldsymbol{U}^{(1)} \mathcal{M}_1\left(\boldsymbol{\mathcal{G}} - \boldsymbol{\mathcal{G}}_\star\right) \left(\boldsymbol{U}^{(3)} \otimes \boldsymbol{U}^{(2)}\right)^\top \right\|_\infty}_{=:\mathfrak{A}_{\text{core}}} + \underbrace{\left\| \left(\boldsymbol{U}^{(1)} - \boldsymbol{U}_\star^{(1)}\right) \mathcal{M}_1\left(\boldsymbol{\mathcal{G}}_\star\right) \left(\boldsymbol{U}^{(3)} \otimes \boldsymbol{U}^{(2)}\right)^\top \right\|_\infty}_{=:\mathfrak{A}_1} \\ &+ \underbrace{\left\| \left(\boldsymbol{U}^{(2)} - \boldsymbol{U}_\star^{(2)}\right) \mathcal{M}_2\left(\boldsymbol{\mathcal{G}}_\star\right) \left(\boldsymbol{U}^{(3)} \otimes \boldsymbol{U}_\star^{(1)}\right)^\top \right\|_\infty}_{=:\mathfrak{A}_2} + \underbrace{\left\| \left(\boldsymbol{U}^{(3)} - \boldsymbol{U}_\star^{(3)}\right) \mathcal{M}_3\left(\boldsymbol{\mathcal{G}}_\star\right) \left(\boldsymbol{U}_\star^{(2)} \otimes \boldsymbol{U}_\star^{(1)}\right)^\top \right\|_\infty}_{=:\mathfrak{A}_3}, \end{aligned}$$

where the second inequality follows from the invariance of ℓ_∞ norm to matricizations. We will bound each term separately.

- For $\mathfrak{A}_{\text{core}}$, it follows from basic norm relations that

$$\begin{aligned} \mathfrak{A}_{\text{core}} &\leq \left\| \boldsymbol{U}^{(1)} \right\|_{2,\infty} \left\| \mathcal{M}_1\left(\boldsymbol{\mathcal{G}} - \boldsymbol{\mathcal{G}}_\star\right) \right\| \left\| \boldsymbol{U}^{(3)} \otimes \boldsymbol{U}^{(2)} \right\|_{2,\infty} \\ &\leq \left\| \boldsymbol{U}^{(1)} \right\|_{2,\infty} \left\| \boldsymbol{U}^{(2)} \right\|_{2,\infty} \left\| \boldsymbol{U}^{(3)} \right\|_{2,\infty} \|\boldsymbol{\mathcal{G}} - \boldsymbol{\mathcal{G}}_\star\|_{\mathsf{F}} \leq 8 \sqrt{\frac{\mu^3 r_1 r_2 r_3}{n_1 n_2 n_3}} \epsilon_0 c \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star), \end{aligned}$$

  where the last inequality follows from $\|\boldsymbol{\mathcal{G}} - \boldsymbol{\mathcal{G}}_\star\|_{\mathsf{F}} \leq \operatorname{dist}(\boldsymbol{F}, \boldsymbol{F}_\star) \leq \epsilon_0 c \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$ by assumption (74a) and Lemma 9 by assumption (74b).

- Next, for $\mathfrak{A}_1$,

$$\begin{aligned} \mathfrak{A}_1 &\leq \left\| (\boldsymbol{U}^{(1)} - \boldsymbol{U}_\star^{(1)}) \boldsymbol{\Sigma}_\star^{(1)} \right\|_{2,\infty} \left\| \boldsymbol{U}^{(2)} \right\|_{2,\infty} \left\| \boldsymbol{U}^{(3)} \right\|_{2,\infty} \left\| \mathcal{M}_1\left(\boldsymbol{\mathcal{G}}_\star\right)^\top \left(\boldsymbol{\Sigma}_\star^{(1)}\right)^{-1} \right\| \\ &= \left\| (\boldsymbol{U}^{(1)} - \boldsymbol{U}_\star^{(1)}) \boldsymbol{\Sigma}_\star^{(1)} \right\|_{2,\infty} \left\| \boldsymbol{U}^{(2)} \right\|_{2,\infty} \left\| \boldsymbol{U}^{(3)} \right\|_{2,\infty} \leq 4 \sqrt{\frac{\mu^3 r_1 r_2 r_3}{n_1 n_2 n_3}} c \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star), \end{aligned}$$

  where the equality follows from $\left\| \mathcal{M}_k\left(\boldsymbol{\mathcal{G}}_\star\right)^\top \left(\boldsymbol{\Sigma}_\star^{(k)}\right)^{-1} \right\| = 1$ since $\mathcal{M}_k\left(\boldsymbol{\mathcal{G}}_\star\right) \mathcal{M}_k\left(\boldsymbol{\mathcal{G}}_\star\right)^\top = \left(\boldsymbol{\Sigma}_\star^{(k)}\right)^2$, and the last inequality follows from the assumption (74b) and Lemma 9 by assumption (74b).

- Similarly, for $\mathfrak{A}_2$, it follows

$$\begin{aligned}\mathfrak{A}_2 &\le \left\|(\boldsymbol{U}^{(2)} - \boldsymbol{U}_\star^{(2)})\boldsymbol{\Sigma}_\star^{(2)}\right\|_{2,\infty} \left\|\boldsymbol{U}^{(3)}\right\|_{2,\infty} \left\|\boldsymbol{U}_\star^{(1)}\right\|_{2,\infty} \left\|\mathcal{M}_2\left(\boldsymbol{\mathcal{G}}_\star\right)^\top \left(\boldsymbol{\Sigma}_\star^{(2)}\right)^{-1}\right\| \\ &\le 2\sqrt{\frac{\mu^3 r_1 r_2 r_3}{n_1 n_2 n_3}} c\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star).\end{aligned}$$

- Finally, repeat the same approach for $\mathfrak{A}_3$ to get

$$\begin{aligned}\mathfrak{A}_3 &\le \left\|(\boldsymbol{U}^{(3)} - \boldsymbol{U}_\star^{(3)})\boldsymbol{\Sigma}_\star^{(3)}\right\|_{2,\infty} \left\|\boldsymbol{U}_\star^{(2)}\right\|_{2,\infty} \left\|\boldsymbol{U}_\star^{(1)}\right\|_{2,\infty} \left\|\mathcal{M}_3\left(\boldsymbol{\mathcal{G}}_\star\right)^\top \left(\boldsymbol{\Sigma}_\star^{(3)}\right)^{-1}\right\| \\ &\le \sqrt{\frac{\mu^3 r_1 r_2 r_3}{n_1 n_2 n_3}} c\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star).\end{aligned}$$

Putting these together, we have the advertised bound. □

E.3 Sparse outliers

The following two lemmas are useful to control the sparse corruption term, of which the second lemma follows directly from translating [CLY21, Lemma 5] to the tensor case.

Lemma 11 ([YPCC16, Lemma 1] [CLY21, Lemma 6]). *Suppose that $\boldsymbol{S} \in \mathbb{R}^{m\times n}$ is α-sparse. Then one has*

$$\|\boldsymbol{S}\| \le \alpha\sqrt{mn}\,\|\boldsymbol{S}\|_\infty, \qquad \|\boldsymbol{S}\|_{2,\infty} \le \sqrt{\alpha n}\,\|\boldsymbol{S}\|_\infty, \qquad \text{and} \qquad \|\boldsymbol{S}\|_{1,\infty} \le \alpha n\,\|\boldsymbol{S}\|_\infty.$$

Lemma 12 ([CLY21, Lemma 5]). *Suppose that $\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{X}}_\star + \boldsymbol{\mathcal{S}}_\star$ for some α-sparse $\boldsymbol{\mathcal{S}}_\star$. Fix a tensor $\boldsymbol{\mathcal{X}}$, and let $\boldsymbol{\mathcal{S}} = \mathcal{T}_\zeta\left(\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}\right)$ where the threshold satisfies $\zeta \ge \|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{X}}_\star\|_\infty$. We then have*

$$\|\boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{S}}_\star\|_\infty \le \|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{X}}_\star\|_\infty + \zeta \le 2\zeta \tag{76}$$

and

$$\operatorname{supp}(\boldsymbol{\mathcal{S}}) \subseteq \operatorname{supp}(\boldsymbol{\mathcal{S}}_\star). \tag{77}$$

The relation (77) also implies that $\boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{S}}_\star$ is α-sparse.