Scaling and Scalability: Provable Nonconvex Low-Rank Tensor Estimation from Incomplete Measurements

Tian Tong* CMU
Cong Ma† UC Berkeley
Ashley Prater-Bennette‡ AFRL
Erin Tripp‡ AFRL
Yuejie Chi* CMU

April 2021

Abstract

Tensors, which provide a powerful and flexible model for representing multi-attribute data and multi-way interactions, play an indispensable role in modern data science across various fields in science and engineering. A fundamental task is to faithfully recover the tensor from highly incomplete measurements in a statistically and computationally efficient manner. Harnessing the low-rank structure of tensors in the Tucker decomposition, this paper develops a scaled gradient descent (ScaledGD) algorithm to directly recover the tensor factors with tailored spectral initializations, and shows that it provably converges at a linear rate independent of the condition number of the ground truth tensor for two canonical problems — tensor completion and tensor regression — as soon as the sample size is above the order of $n^{3/2}$ ignoring other parameter dependencies, where n is the dimension of the tensor. This leads to an extremely scalable approach to low-rank tensor estimation compared with prior art, which suffers from at least one of the following drawbacks: extreme sensitivity to ill-conditioning, high per-iteration costs in terms of memory and computation, or poor sample complexity guarantees. To the best of our knowledge, ScaledGD is the first algorithm that achieves near-optimal statistical and computational complexities simultaneously for low-rank tensor completion with the Tucker decomposition. Our algorithm highlights the power of appropriate preconditioning in accelerating nonconvex statistical estimation, where the iteration-varying preconditioners promote desirable invariance properties of the trajectory with respect to the underlying symmetry in low-rank tensor factorization.

Keywords: low-rank tensor completion, low-rank tensor regression, Tucker decomposition, scaled gradient descent, ill-conditioning.

Contents

*Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA; Emails: {ttong1,yuejiec}@andrew.cmu.edu.

†Department of Statistics and Department of Electrical Engineering and Computer Sciences, UC Berkeley, Berkeley, CA 94720, USA; Email: congm@berkeley.edu.

‡Air Force Research Laboratory, Rome, NY 13441, USA; Email: {ashley.prater-bennette,erin.tripp.4}@us.af.mil.

1 Introduction

Tensors [KB09, SDLF$^+$17], which provide a powerful and flexible model for representing multi-attribute data and multi-way interactions across various fields, play an indispensable role in modern data science with ubiquitous applications in image inpainting [LMWY12], hyperspectral imaging [DFL17], collaborative filtering [XCH$^+$10], topic modeling [AGH$^+$14], network analysis [PFS16], and many more.

1.1 Low-rank tensor estimation

In many problems across science and engineering, the central task can be regarded as tensor estimation from highly incomplete measurements, where the goal is to estimate an order-3 tensor[1] $\boldsymbol{\mathcal{X}}_\star \in \mathbb{R}^{n_1\times n_2\times n_3}$ from its observations $\boldsymbol{y} \in \mathbb{R}^m$ given by

$$\boldsymbol{y} \approx \mathcal{A}(\boldsymbol{\mathcal{X}}_\star).$$

Here, $\mathcal{A} : \mathbb{R}^{n_1\times n_2\times n_3} \mapsto \mathbb{R}^m$ represents a certain linear map modeling the data collection process. Importantly, the number m of observations is often much smaller than the ambient dimension $n_1n_2n_3$ of the tensor due to resource or physical constraints, necessitating the need of exploiting low-dimensional structures to allow for meaningful recovery.

One of the most widely adopted low-dimensional structures—which is the focus of this paper—is the low-rank structure under the *Tucker* decomposition [Tuc66]. Specifically, we assume that the ground truth

[1] For ease of presentation, we focus on 3-way tensors; our algorithm and theory can be generalized to higher-order tensors in a straightforward manner.

tensor $\boldsymbol{\mathcal{X}}_\star$ admits the following Tucker decomposition[2]

$$\boldsymbol{\mathcal{X}}_\star = (\boldsymbol{U}_\star, \boldsymbol{V}_\star, \boldsymbol{W}_\star) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_\star,$$

where $\boldsymbol{\mathcal{S}}_\star \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ is the core tensor, and $\boldsymbol{U}_\star \in \mathbb{R}^{n_1 \times r_1}$, $\boldsymbol{V}_\star \in \mathbb{R}^{n_2 \times r_2}$, $\boldsymbol{W}_\star \in \mathbb{R}^{n_3 \times r_3}$ are orthonormal matrices corresponding to the factors of each mode. The tensor $\boldsymbol{\mathcal{X}}_\star$ is said to be low-multilinear-rank, or simply low-rank, when its multilinear rank $\boldsymbol{r} = (r_1, r_2, r_3)$ satisfies $r_k \ll n_k$, for all $k = 1, 2, 3$. Compared with other tensor decompositions such as the CP decomposition [KB09] and tensor-SVD [ZEA+14], the Tucker decomposition offers several advantages: it allows flexible modeling of low-rank tensor factors with a small number of parameters, fully exploits the multi-dimensional algebraic structure of a tensor, and admits efficient and stable computation without suffering from degeneracy [Paa00].

Motivating examples. We point out two representative settings of tensor recovery that guide our work.

- *Tensor completion.* A widely encountered problem is tensor completion, where one aims to predict the entries in a tensor from only a small subset of its revealed entries. A celebrated application is collaborative filtering, where one aims to predict the users' evolving preferences from partial observations of a tensor composed of ratings for any triplet of *user, item, time* [KABO10]. Mathematically, we are given entries

  $$\boldsymbol{\mathcal{X}}_\star(i_1, i_2, i_3), \qquad (i_1, i_2, i_3) \in \Omega,$$

  in some index set Ω, where $(i_1, i_2, i_3) \in \Omega$ if and only if that entry is observed. The goal is then to recover the low-rank tensor $\boldsymbol{\mathcal{X}}_\star$ from the observed entries in Ω.

- *Tensor regression.* In machine learning and signal processing, one is often concerned with determining how the covariates relate to the response—a task known as regression. Due to advances in data acquisition, there is no shortage of scenarios where the covariates are available in the form of tensors, for example in medical imaging [ZLZ13]. Mathematically, the i-th response or observation is given as

  $$y_i = \langle \boldsymbol{\mathcal{A}}_i, \boldsymbol{\mathcal{X}}_\star \rangle = \sum_{i_1, i_2, i_3} \boldsymbol{\mathcal{A}}_i(i_1, i_2, i_3) \boldsymbol{\mathcal{X}}_\star(i_1, i_2, i_3), \qquad i = 1, 2, \ldots, m,$$

  where $\boldsymbol{\mathcal{A}}_i$ is the i-th covariate or measurement tensor. The goal is then to recover the low-rank tensor $\boldsymbol{\mathcal{X}}_\star$ from the responses $\boldsymbol{y} = \{y_i\}_{i=1}^m$.

1.2 A gradient descent approach?

Recent years have seen remarkable successes in developing a plethora of provably efficient algorithms for low-rank *matrix* estimation (i.e. the special case of order-2 tensors) via both convex and nonconvex optimization. However, unique challenges arise when dealing with tensors, since tensors have more sophisticated algebraic structures [Hac12]. For instance, while nuclear norm minimization achieves near-optimal statistical guarantees for low-rank matrix estimation [CT10] with a polynomial run time, computing the nuclear norm of a tensor turns out to be NP-hard [FL18]. Therefore, there have been a number of efforts to develop polynomial-time algorithms for tensor recovery, including but not limited to the sum-of-squares hierarchy [BM16, PS17], nuclear norm minimization with unfolding [GRY11, MHWG14], regularized gradient descent [HWZ20], to name a few; see Section 1.4 for further discussions.

In view of the low-rank Tucker decomposition, a natural approach is to seek to recover the factor quadruple $\boldsymbol{F}_\star \coloneqq (\boldsymbol{U}_\star, \boldsymbol{V}_\star, \boldsymbol{W}_\star, \boldsymbol{\mathcal{S}}_\star)$ directly by optimizing the unconstrained least-squares loss function:

$$\min_{\boldsymbol{F}} \quad \mathcal{L}(\boldsymbol{F}) \coloneqq \frac{1}{2} \left\| \mathcal{A}\left((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}} \right) - \boldsymbol{y} \right\|_2^2, \tag{1}$$

where $\boldsymbol{F} \coloneqq (\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}, \boldsymbol{\mathcal{S}})$ consists of $\boldsymbol{U} \in \mathbb{R}^{n_1 \times r_1}$, $\boldsymbol{V} \in \mathbb{R}^{n_2 \times r_2}$, $\boldsymbol{W} \in \mathbb{R}^{n_3 \times r_3}$, and $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$. Since the factors have a much lower complexity than the tensor itself due to the low-rank structure, it is expected

[2] Other popular notation for Tucker decomposition in the literature includes $[\![\boldsymbol{\mathcal{S}}_\star; \boldsymbol{U}_\star, \boldsymbol{V}_\star, \boldsymbol{W}_\star]\!]$ and $\boldsymbol{\mathcal{S}}_\star \times_1 \boldsymbol{U}_\star \times_2 \boldsymbol{V}_\star \times_3 \boldsymbol{W}_\star$. In this work, we adopt the same notation $(\boldsymbol{U}_\star, \boldsymbol{V}_\star, \boldsymbol{W}_\star) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_\star$ as in [XY19] for convenience of our theoretical developments.

that manipulating the factors results in more scalable algorithms in terms of both computation and storage. This optimization problem is however, highly nonconvex, since the factors are not uniquely determined.[3] Nonetheless, one might be tempted to solve the problem (1) via gradient descent (GD), which updates the factors according to

$$\boldsymbol{F}_{t+1} = \boldsymbol{F}_t - \eta \nabla \mathcal{L}(\boldsymbol{F}_t), \qquad t = 0, 1, \dots, \tag{2}$$

where $\boldsymbol{F}_t$ is the estimate at the t-th iteration, $\eta > 0$ is the step size or learning rate, and $\nabla \mathcal{L}(\boldsymbol{F})$ is the gradient of $\mathcal{L}(\boldsymbol{F})$ at $\boldsymbol{F}$. Despite a flurry of activities for understanding factored gradient descent in the matrix setting [CLC19], this line of algorithmic thinkings has been severely under-explored for the tensor setting, especially when it comes to provable guarantees for both sample and computational complexities.

The closest existing theory that one comes across is [HWZ20] for tensor regression, which adds regularization terms to promote the orthogonality of the factors $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}$:

$$\mathcal{L}_{\mathsf{reg}}(\boldsymbol{F}) \coloneqq \mathcal{L}(\boldsymbol{F}) + \frac{\alpha}{4}\left(\|\boldsymbol{U}^\top \boldsymbol{U} - \beta \boldsymbol{I}_{r_1}\|_{\mathsf{F}}^2 + \|\boldsymbol{V}^\top \boldsymbol{V} - \beta \boldsymbol{I}_{r_2}\|_{\mathsf{F}}^2 + \|\boldsymbol{W}^\top \boldsymbol{W} - \beta \boldsymbol{I}_{r_3}\|_{\mathsf{F}}^2\right), \tag{3}$$

and perform GD on the regularized loss. Here, $\alpha > 0$ and $\beta > 0$ are two parameters to be specified. While encouraging, theoretical guarantees of this regularized GD algorithm [HWZ20] still fall short of achieving computational efficiency. In truth, its convergence speed is rather slow: it takes an order of $\kappa^2 \log(1/\varepsilon)$ iterations to attain an ε-accurate estimate of the ground truth tensor, where κ is a sort of condition number of $\boldsymbol{\mathcal{X}}_\star$ to be defined momentarily. Therefore, the computational efficacy of the regularized GD algorithm is severely hampered even when $\boldsymbol{\mathcal{X}}_\star$ is moderately ill-conditioned, a situation frequently encountered in practice. In addition, the regularization term introduces additional parameters that may be difficult to tune optimally in practice.

Turning to tensor completion, the situation is even worse: to the best of our knowledge, there is *no* provably linearly-convergent algorithm that accommodates low-rank tensor completion under the Tucker decomposition. The question is thus:

> *Can we develop a factored gradient-based algorithm that converges fast even for highly ill-conditioned tensors with near-optimal sample complexities for tensor completion and tensor regression?*

In this paper, we provide an affirmative answer to the above question.

1.3 A new algorithm: scaled gradient descent

We propose a novel algorithm—dubbed scaled gradient descent (ScaledGD)—to solve the tensor recovery problem. More specifically, at the core it performs the following iterative updates[4] to minimize the loss function (1):

$$\begin{aligned}
\boldsymbol{U}_{t+1} &= \boldsymbol{U}_t - \eta \nabla_{\boldsymbol{U}} \mathcal{L}(\boldsymbol{F}_t)\big(\breve{\boldsymbol{U}}_t^\top \breve{\boldsymbol{U}}_t\big)^{-1}, \\
\boldsymbol{V}_{t+1} &= \boldsymbol{V}_t - \eta \nabla_{\boldsymbol{V}} \mathcal{L}(\boldsymbol{F}_t)\big(\breve{\boldsymbol{V}}_t^\top \breve{\boldsymbol{V}}_t\big)^{-1}, \\
\boldsymbol{W}_{t+1} &= \boldsymbol{W}_t - \eta \nabla_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{F}_t)\big(\breve{\boldsymbol{W}}_t^\top \breve{\boldsymbol{W}}_t\big)^{-1}, \\
\boldsymbol{\mathcal{S}}_{t+1} &= \boldsymbol{\mathcal{S}}_t - \eta \left((\boldsymbol{U}_t^\top \boldsymbol{U}_t)^{-1}, (\boldsymbol{V}_t^\top \boldsymbol{V}_t)^{-1}, (\boldsymbol{W}_t^\top \boldsymbol{W}_t)^{-1}\right) \boldsymbol{\cdot} \nabla_{\boldsymbol{\mathcal{S}}} \mathcal{L}(\boldsymbol{F}_t),
\end{aligned} \tag{4}$$

where $\nabla_{\boldsymbol{U}} \mathcal{L}(\boldsymbol{F})$, $\nabla_{\boldsymbol{V}} \mathcal{L}(\boldsymbol{F})$, $\nabla_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{F})$, and $\nabla_{\boldsymbol{\mathcal{S}}} \mathcal{L}(\boldsymbol{F})$ are the partial derivatives of $\mathcal{L}(\boldsymbol{F})$ with respect to the corresponding variables, and

$$\breve{\boldsymbol{U}}_t \coloneqq (\boldsymbol{V}_t \otimes \boldsymbol{W}_t)\mathcal{M}_1(\boldsymbol{\mathcal{S}}_t)^\top, \qquad \breve{\boldsymbol{V}}_t \coloneqq (\boldsymbol{U}_t \otimes \boldsymbol{W}_t)\mathcal{M}_2(\boldsymbol{\mathcal{S}}_t)^\top, \qquad \breve{\boldsymbol{W}}_t \coloneqq (\boldsymbol{U}_t \otimes \boldsymbol{V}_t)\mathcal{M}_3(\boldsymbol{\mathcal{S}}_t)^\top. \tag{5}$$

Here, $\mathcal{M}_k(\boldsymbol{\mathcal{S}})$ is the matricization of the tensor $\boldsymbol{\mathcal{S}}$ along the k-th mode ($k = 1, 2, 3$), and $\otimes$ denotes the Kronecker product. Inspired by its counterpart in the matrix setting [TMC20], the ScaledGD algorithm (4) exploits the structures of Tucker decomposition and possesses many desirable properties:

[3] For any invertible matrices $\boldsymbol{Q}_k \in \mathbb{R}^{r_k \times r_k}$, $k = 1, 2, 3$, one has $(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}} = (\boldsymbol{U}\boldsymbol{Q}_1, \boldsymbol{V}\boldsymbol{Q}_2, \boldsymbol{W}\boldsymbol{Q}_3) \boldsymbol{\cdot} ((\boldsymbol{Q}_1^{-1}, \boldsymbol{Q}_2^{-1}, \boldsymbol{Q}_3^{-1}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}})$.

[4] The matrix inverses in ScaledGD always exist under the assumptions of our theory.

Algorithms	Sample complexity	Iteration complexity	Parameter space
Unfolding + nuclear norm min. [HMGW15]	$n^2 r \log^2 n$	polynomial	tensor
Tensor nuclear norm min. [YZ16]	$n^{3/2} r^{1/2} \log^{3/2} n$	NP-hard	tensor
Grassmannian GD [XY19]	$n^{3/2} r^{7/2} \kappa^4 \log^{7/2} n$	N/A	factor
ScaledGD (this paper)	$n^{3/2} r^2 \kappa (\sqrt{r} \vee \kappa^2) \log^3 n$	$\log \frac{1}{\varepsilon}$	factor

Table 1: Comparisons of ScaledGD with existing algorithms for tensor completion when the tensor is incoherent and low-rank under the Tucker decomposition. Here, we say that the output $\boldsymbol{\mathcal{X}}$ of an algorithm reaches ε-accuracy, if it satisfies $\|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \le \varepsilon \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$. Here, κ and $\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$ are the condition number and the minimum singular value of $\boldsymbol{\mathcal{X}}_\star$ (defined in Section 2.1). For simplicity, we let $n = \max_{k=1,2,3} n_k$ and $r = \max_{k=1,2,3} r_k$, and assume $r \vee \kappa \ll n^\delta$ for some small constant δ to keep only terms with dominating orders of n.

Algorithms	Sample complexity	Iteration complexity	Parameter space
Unfolding + nuclear norm min. [MHWG14]	$n^2 r$	polynomial	tensor
Projected GD [CRY19]	$n^2 r$	$\kappa^2 \log \frac{1}{\varepsilon}$	tensor
Regularized GD [HWZ20]	$n^{3/2} r \kappa^4$	$\kappa^2 \log \frac{1}{\varepsilon}$	factor
Riemannian Gauss-Newton [LZ21] (concurrent)[5]	$n^{3/2} r^{3/2} \kappa^4$	$\log \log \frac{1}{\varepsilon}$	tensor
ScaledGD (this paper)	$n^{3/2} r \kappa^2$	$\log \frac{1}{\varepsilon}$	factor

Table 2: Comparisons of ScaledGD with existing algorithms for tensor regression when the tensor is low-rank under the Tucker decomposition. Here, we say that the output $\boldsymbol{\mathcal{X}}$ of an algorithm reaches ε-accuracy, if it satisfies $\|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \le \varepsilon \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$. Here, κ and $\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$ are the condition number and minimum singular value of $\boldsymbol{\mathcal{X}}_\star$ (defined in Section 2.1). For simplicity, we let $n = \max_{k=1,2,3} n_k$, and $r = \max_{k=1,2,3} r_k$, and assume $r \vee \kappa \ll n^\delta$ for some small constant δ to keep only terms with dominating orders of n.

- *Low per-iteration cost:* as a preconditioned GD or quasi-Newton algorithm, ScaledGD updates the factors along the descent direction of a scaled gradient, where the preconditioners can be viewed as the inverse of the diagonal blocks of the Hessian for the population loss (i.e. tensor factorization). As the sizes of the preconditioners are proportional to the size of the multilinear rank, the matrix inverses are cheap to compute with a minimal overhead and the overall per-iteration cost is still low and linear in the time it takes to read the input data.

- *Equivariance to parameterization:* one crucial property of ScaledGD is that if we reparameterize the factors by some invertible transforms (i.e. replacing $(\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t, \boldsymbol{\mathcal{S}}_t)$ by $(\boldsymbol{U}_t \boldsymbol{Q}_1, \boldsymbol{V}_t \boldsymbol{Q}_2, \boldsymbol{W}_t \boldsymbol{Q}_3, (\boldsymbol{Q}_1^{-1}, \boldsymbol{Q}_2^{-1}, \boldsymbol{Q}_3^{-1}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_t)$ for some invertible matrices $\{\boldsymbol{Q}_k\}_{k=1}^3$), the entire trajectory will go through the same reparameterization, leading to an *invariant* sequence of low-rank tensor updates $\boldsymbol{\mathcal{X}}_t = (\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_t$ regardless of the parameterization being adopted.

- *Implicit balancing:* ScaledGD optimizes the natural loss function (1) in an *unconstrained* manner without requiring additional regularizations or orthogonalizations used in prior literature [HWZ20, FG20, KM16], and the factors stay balanced in an automatic manner—a feature sometimes referred to as implicit regularization [MLC21].

[5] [LZ21, Theorem 3] states the sample complexity $n^{3/2}\sqrt{r}\kappa^2 \|\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}^2 / \sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star)$, where $\|\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}^2 / \sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star)$ has an order of $r\kappa^2$.

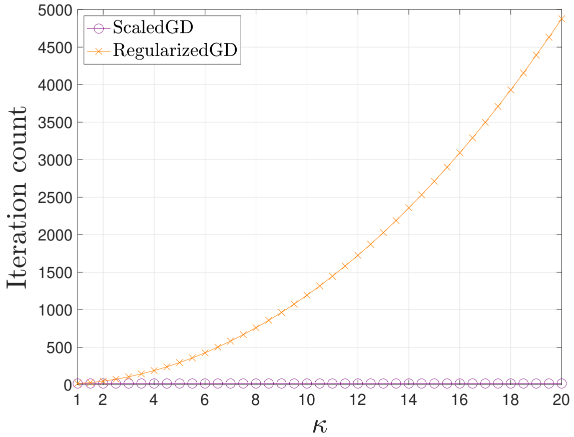


Figure 1: The iteration complexities of ScaledGD (this paper) and regularized GD to achieve $\|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{X}}_\star\|_\mathsf{F} \le 10^{-3}\|\boldsymbol{\mathcal{X}}_\star\|_\mathsf{F}$ with respect to different condition numbers for low-rank tensor completion with $n_1 = n_2 = n_3 = 100$, $r_1 = r_2 = r_3 = 5$, and the probability of observation $p = 0.1$.

Theoretical guarantees. We investigate the theoretical properties of ScaledGD for both tensor completion and tensor regression, which are notably more challenging than the matrix counterpart. It is demonstrated that ScaledGD—when initialized properly using appropriate spectral methods —achieves linear convergence at a rate *independent* of the condition number of the ground truth tensor with near-optimal sample complexities. In other words, ScaledGD needs no more than $O(\log(1/\varepsilon))$ iterations to reach ε-accuracy; together with its low computational and memory costs by operating in the factor space, this makes ScaledGD a highly scalable method for a wide range of low-rank tensor estimation tasks. More specifically, we have the following guarantees (assume $n = \max_{k=1,2,3} n_k$ and $r = \max_{k=1,2,3} r_k$):

- *Tensor completion.* Under the Bernoulli sampling model, ScaledGD (with an additional scaled projection step) succeeds with high probability as long as the sample complexity is above the order of $n^{3/2}r^2\kappa(\sqrt{r} \vee \kappa^2)\log^3 n$, where $a \vee b := \max\{a, b\}$. Connected to some well-reckoned conjecture on computational barriers, it is widely believed that no polynomial-time algorithm will be successful if the sample complexity is less than the order of $n^{3/2}$ for tensor completion [BM16], which suggests the near-optimality of the sample complexity of ScaledGD. Compare with existing approaches (cf. Table 1), ScaledGD provides the first computationally efficient algorithm with a near-linear run time at the near-optimal sample complexity.

- *Tensor regression.* Under the Gaussian design, ScaledGD succeeds with high probability as long as the sample complexity is above the order of $n^{3/2}r\kappa^2$. Our analysis of local convergence is more general, based on the tensor restricted isometry property (TRIP) [RSS17], and is therefore applicable to various measurement ensembles that satisfy TRIP. Compared with existing approaches (cf. Table 2), ScaledGD achieves competitive performance guarantees in terms of sample and iteration complexities with a low per-iteration cost in the factor space.

Figure 1 illustrates the number of iterations needed to achieve a relative error $\|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{X}}_\star\|_\mathsf{F} \le 10^{-3}\|\boldsymbol{\mathcal{X}}_\star\|_\mathsf{F}$ for ScaledGD and regularized GD [HWZ20] under different condition numbers for tensor completion under the Bernoulli sampling model (see Section 4 for experimental settings). Clearly, the iteration complexity of regularized GD [HWZ20] deteriorates at a super linear rate with respect to the condition number κ, while ScaledGD enjoys an iteration complexity that is independent of κ as predicted by our theory. Indeed, with a seemingly small modification, ScaledGD takes merely 17 iterations to achieve the desired accuracy over the entire range of κ, while regularized GD takes thousands of iterations even with a moderate condition number!

1.4 Additional related works

Low-rank tensor estimation with Tucker decomposition. [FG20] analyzed the landscape of Tucker decomposition for tensor factorization, and showed benign landscape properties with suitable regularizations. [GRY11, MHWG14] developed convex relaxation algorithms based on minimizing the nuclear norms of unfolded tensors for tensor regression, and similar approaches were developed in [HMGW15] for robust tensor completion. However, unfolding-based approaches typically result in sub-optimal sample complexities since they do not fully exploit the tensor structure. [YZ16] studied directly minimizing the nuclear norm of the tensor, which regrettably is not computationally tractable. [XY19] proposed a Grassmannian gradient descent algorithm over the factors other than the core tensor for exact tensor completion, whose iteration complexity is not characterized. The statistical rates of tensor completion, together with a spectral method, were investigated in [XYZ17, ZX18], and uncertainty quantifications were recently dealt with in [XZZ20]. Besides the entrywise i.i.d. observation models for tensor completion, [Zha19, KS13] considered tailored or adaptive observation patterns to improve the sample complexity. In addition, for low-rank tensor regression, [RYC19] proposed a general convex optimization approach based on decomposable regularizers, and [RSS17] developed an iterative hard thresholding algorithm. [CRY19] proposed projected gradient descent algorithms with respect to the tensors, which have larger computation and memory footprints than the factored gradient descent approaches taken in this paper. [ARB20] proposed a tensor regression model where the tensor is simultaneously low-rank and sparse in the Tucker decomposition. A concurrent work [LZ21] proposed a Riemannian Gauss-Newton algorithm, and obtained an impressive quadratic convergence rate for tensor regression (see Table 2). Compared with ScaledGD, this algorithm runs in the tensor space, and the update rule is more sophisticated with higher per-iteration cost by solving a least-squares problem and performing a truncated HOSVD every iteration.

Last but not least, many scalable algorithms for low-rank tensor estimation have been proposed in the literature of numerical optimization [XY13, GQ14], where preconditioning has long been recognized as a key idea to accelerate convergence [KM16, KSV14]. In particular, if we constrain $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}$ to be orthonormal, i.e. on the Grassmanian manifold, the preconditioners used in ScaledGD degenerate to the ones investigated in [KM16], which was a Riemannian manifold gradient algorithm under a scaled metric. On the other hand, ScaledGD does not assume orthonormality of the factors, therefore is conceptually simpler to understand and avoids complicated manifold operations (e.g. geodesics, retraction). Furthermore, none of the prior algorithmic developments [KM16, KSV14] are endowed with the type of global performance guarantees with linear convergence rate as developed herein.

Provable low-rank tensor estimation with other decompositions. Complementary to ours, there have also been a growing number of algorithms proposed for estimating a low-rank tensor adopting the CP decomposition. Examples include sum-of-squares hierarchy [BM16, PS17], gradient descent [CLPC19, CPC20, HZC20], alternating minimization [JO14, LM20], spectral methods [MS18, CCFM20, CLC+21], atomic norm minimization [LPST15, GPY19], to name a few. [GM20] studied the optimization landscape of overcomplete CP tensor decomposition. Beyond the CP decomposition, [ZA16] developed exact tensor completion algorithms under the so-called tensor-SVD [ZEA+14], and [LAAW19, LFLY18] studied low-tubal-rank tensor recovery. We will not elaborate further since these algorithms are not directly comparable to ours due to the difference in models.

Nonconvex optimization for statistical estimation. Our work contributes to the recent strand of works that develop provable nonconvex methods for statistical estimation, including but not limited to low-rank matrix estimation [SL16, CW15, MWCC20, CCD+21, MLC21, PKCS17, CLL20], phase retrieval [CLS15, WGE18, CC17, ZZLC17, ZCL16, CCFM19], quadratic sampling [LMCC19], dictionary learning [SQW17a, SQW17b, BJS18], neural network training [BGW20, FCL20, HV19], and blind deconvolution [LLSW19, MWCC20, SC19]; the readers are referred to the overviews [CLC19, CC18, ZQW20] for further references. The proposed ScaledGD algorithm is inspired by its counterpart in the matrix setting, which was first proposed in [TMC20] with theoretical guarantees; see [TMC21, CMPC20] for further developments that enable robust recovery in corrupted and mixture models.

1.5 A primer on tensor algebra and notation

We end this section with a primer on some useful tensor algebra; for a more detailed exposition, see [KB09, SDLF+17]. 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 $\boldsymbol{M}$, we use $\sigma_i(\boldsymbol{M})$ to denote its i-th largest singular value, and $\|\boldsymbol{M}\|$, $\|\boldsymbol{M}\|_{\mathsf{F}}$, $\|\boldsymbol{M}\|_{1,\infty}$, $\|\boldsymbol{M}\|_{2,\infty}$, and $\|\boldsymbol{M}\|_{\infty}$ stand for the spectral norm (i.e. the largest singular value), the Frobenius norm, the $\ell_{1,\infty}$ norm (i.e. the largest ℓ_1 norm of the rows), the $\ell_{2,\infty}$ norm (i.e. the largest ℓ_2 norm of the rows), and the entrywise ℓ_∞ norm (the largest magnitude of all entries) of a matrix $\boldsymbol{M}$.

To begin with, we define the unfolding (i.e. flattening) operations of tensors and matrices.

- The mode-1 matricization $\mathcal{M}_1(\boldsymbol{\mathcal{X}}) \in \mathbb{R}^{n_1 \times (n_2 n_3)}$ of a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is given by $[\mathcal{M}_1(\boldsymbol{\mathcal{X}})]\big(i_1, (i_2 - 1)n_3 + i_3\big) = \boldsymbol{\mathcal{X}}(i_1, i_2, i_3)$, for $1 \leq i_k \leq n_k$, $k = 1, 2, 3$; $\mathcal{M}_2(\boldsymbol{\mathcal{X}})$ and $\mathcal{M}_3(\boldsymbol{\mathcal{X}})$ can be defined in a similar manner.
- The vectorization $\mathrm{vec}(\boldsymbol{\mathcal{X}}) \in \mathbb{R}^{n_1 n_2 n_3}$ of a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is given by $[\mathrm{vec}(\boldsymbol{\mathcal{X}})]\big((i_1 - 1)n_2 n_3 + (i_2 - 1)n_3 + i_3\big) = \boldsymbol{\mathcal{X}}(i_1, i_2, i_3)$ for $1 \leq i_k \leq n_k$, $k = 1, 2, 3$.
- The row-major vectorization $\mathrm{vec}(\boldsymbol{M}) \in \mathbb{R}^{n_1 n_2}$ of a matrix $\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2}$ is given by $[\mathrm{vec}(\boldsymbol{M})]\big((i_1 - 1)n_2 + i_2\big) = \boldsymbol{M}(i_1, i_2)$ for $1 \leq i_k \leq n_k$, $k = 1, 2$.

The vectorization of a tensor is related to the Kronecker product as

$$\mathrm{vec}((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \cdot \boldsymbol{\mathcal{S}}) = \mathrm{vec}\left(\boldsymbol{U}\mathcal{M}_1(\boldsymbol{\mathcal{S}})(\boldsymbol{V} \otimes \boldsymbol{W})^\top\right) = (\boldsymbol{U} \otimes \boldsymbol{V} \otimes \boldsymbol{W})\,\mathrm{vec}(\boldsymbol{\mathcal{S}}). \tag{6a}$$

Tensor norms. The inner product between two tensors is defined as

$$\langle \boldsymbol{\mathcal{X}}_1, \boldsymbol{\mathcal{X}}_2 \rangle = \sum_{i_1, i_2, i_3} \boldsymbol{\mathcal{X}}_1(i_1, i_2, i_3)\boldsymbol{\mathcal{X}}_2(i_1, i_2, i_3).$$

A useful relation is that

$$\langle \boldsymbol{\mathcal{X}}_1, \boldsymbol{\mathcal{X}}_2 \rangle = \langle \mathcal{M}_k(\boldsymbol{\mathcal{X}}_1), \mathcal{M}_k(\boldsymbol{\mathcal{X}}_2) \rangle, \quad k = 1, 2, 3, \tag{6b}$$

which allows one to move between the tensor representation and the unfolded matrix representation. The Frobenius norm of a tensor is defined as $\|\boldsymbol{\mathcal{X}}\|_{\mathsf{F}} = \sqrt{\langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}} \rangle}$. In addition, the following basic relations, which follow straightforwardly from analogous matrix relations after applying matricizations, will be proven useful:

$$(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \cdot \big((\boldsymbol{Q}_1, \boldsymbol{Q}_2, \boldsymbol{Q}_3) \cdot \boldsymbol{\mathcal{S}}\big) = (\boldsymbol{U}\boldsymbol{Q}_1, \boldsymbol{V}\boldsymbol{Q}_2, \boldsymbol{W}\boldsymbol{Q}_3) \cdot \boldsymbol{\mathcal{S}}, \tag{6c}$$

$$\langle (\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \cdot \boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{X}} \rangle = \left\langle \boldsymbol{\mathcal{S}}, (\boldsymbol{U}^\top, \boldsymbol{V}^\top, \boldsymbol{W}^\top) \cdot \boldsymbol{\mathcal{X}} \right\rangle, \tag{6d}$$

$$\|(\boldsymbol{Q}_1, \boldsymbol{Q}_2, \boldsymbol{Q}_3) \cdot \boldsymbol{\mathcal{S}}\|_{\mathsf{F}} \leq \|\boldsymbol{Q}_1\|\|\boldsymbol{Q}_2\|\|\boldsymbol{Q}_3\|\|\boldsymbol{\mathcal{S}}\|_{\mathsf{F}}, \tag{6e}$$

where $\boldsymbol{Q}_k \in \mathbb{R}^{r_k \times r_k}$, $k = 1, 2, 3$.

The ℓ_∞ norm of $\boldsymbol{\mathcal{X}}$ is defined as $\|\boldsymbol{\mathcal{X}}\|_\infty = \max_{i_1, i_2, i_3} |\boldsymbol{\mathcal{X}}(i_1, i_2, i_3)|$. With slight abuse of terminology, denote

$$\sigma_{\max}(\boldsymbol{\mathcal{X}}) = \max_{k=1,2,3} \sigma_{\max}(\mathcal{M}_k(\boldsymbol{\mathcal{X}})), \qquad \text{and} \qquad \sigma_{\min}(\boldsymbol{\mathcal{X}}) = \min_{k=1,2,3} \sigma_{\max}(\mathcal{M}_k(\boldsymbol{\mathcal{X}}))$$

as the maximum and minimum singular values of $\boldsymbol{\mathcal{X}}$, where $\sigma_{\max}(\mathcal{M}_k(\boldsymbol{\mathcal{X}}))$ and $\sigma_{\min}(\mathcal{M}_k(\boldsymbol{\mathcal{X}}))$ are the largest and the smallest nonzero singular values of $\mathcal{M}_k(\boldsymbol{\mathcal{X}})$, respectively. In addition, define the spectral norm of a tensor $\boldsymbol{\mathcal{X}}$ as

$$\|\boldsymbol{\mathcal{X}}\| = \sup_{\boldsymbol{u}_k \in \mathbb{R}^{n_k} :\, \|\boldsymbol{u}_k\|_2 \leq 1} |\langle \boldsymbol{\mathcal{X}}, (\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3) \cdot 1 \rangle|\,.$$

Note that $\|\boldsymbol{\mathcal{X}}\| \neq \sigma_{\max}(\boldsymbol{\mathcal{X}})$ in general. For a tensor $\boldsymbol{\mathcal{X}}$ of multilinear rank at most $\boldsymbol{r} = (r_1, r_2, r_3)$, its spectral norm is related to the Frobenius norm as [JYZ17, LNSU18]

$$\|\boldsymbol{\mathcal{X}}\|_{\mathsf{F}} \leq \sqrt{\frac{r_1 r_2 r_3}{r}}\|\boldsymbol{\mathcal{X}}\|, \qquad \text{where } r = \max_k r_k. \tag{7}$$

Higher-order SVD. For a general tensor $\boldsymbol{\mathcal{X}}$, define $\mathcal{H}_{\boldsymbol{r}}(\boldsymbol{\mathcal{X}})$ as the top-$\boldsymbol{r}$ higher-order SVD (HOSVD) of $\boldsymbol{\mathcal{X}}$ with $\boldsymbol{r} = (r_1, r_2, r_3)$, given by

$$\mathcal{H}_{\boldsymbol{r}}(\boldsymbol{\mathcal{X}}) = (\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}, \tag{8a}$$

where $\boldsymbol{U}$ is the top-r_1 left singular vectors of $\mathcal{M}_1(\boldsymbol{\mathcal{X}})$, $\boldsymbol{V}$ is the top-r_2 left singular vectors of $\mathcal{M}_2(\boldsymbol{\mathcal{X}})$, $\boldsymbol{W}$ is the top-r_3 left singular vectors of $\mathcal{M}_3(\boldsymbol{\mathcal{X}})$, and $\boldsymbol{\mathcal{S}} = (\boldsymbol{U}^\top, \boldsymbol{V}^\top, \boldsymbol{W}^\top) \boldsymbol{\cdot} \boldsymbol{\mathcal{X}}$ is the core tensor. Equivalently, we denote

$$(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}, \boldsymbol{\mathcal{S}}) = \mathrm{HOSVD}_{\boldsymbol{r}}(\boldsymbol{\mathcal{X}}) \tag{8b}$$

as the output to the HOSVD procedure described above with the multilinear rank $\boldsymbol{r}$. In contrast to the matrix case, HOSVD is not guaranteed to yield the optimal rank-$\boldsymbol{r}$ approximation of $\boldsymbol{\mathcal{X}}$ (which is NP-hard [HL13] to find). Nevertheless, it yields a quasi-optimal approximation [Hac12] in the sense that

$$\|\boldsymbol{\mathcal{X}} - \mathcal{H}_{\boldsymbol{r}}(\boldsymbol{\mathcal{X}})\|_{\mathsf{F}} \le \sqrt{3} \inf_{\widetilde{\boldsymbol{\mathcal{X}}}:\, \mathrm{rank}(\mathcal{M}_k(\widetilde{\boldsymbol{\mathcal{X}}})) \le r_k} \|\boldsymbol{\mathcal{X}} - \widetilde{\boldsymbol{\mathcal{X}}}\|_{\mathsf{F}}. \tag{9}$$

There are many other variants of or alternatives to HOSVD in the literature, e.g. successive HOSVD [Hac12], alternating least squares (ALS) [Hac12], higher-order orthogonal iteration (HOOI) [ZX18], etc. These methods compute truncated singular value decompositions in successive or alternating manners, to either reduce the computational costs or pursue a better (but still quasi-optimal) approximation. We will not delve into the details of these variants; interested readers can consult [Hac12].

Additional notation. The set of invertible matrices in $\mathbb{R}^{r\times r}$ is denoted by $\mathrm{GL}(r)$. Let $\boldsymbol{M}(i,:)$ and $\boldsymbol{M}(:,j)$ denote the i-th row and j-th column of $\boldsymbol{M}$, respectively. Let $a \vee b = \max\{a,b\}$ and $a \wedge b = \min\{a,b\}$. Throughout, $f(n) \lesssim g(n)$ or $f(n) = O(g(n))$ means $|f(n)|/|g(n)| \le C$ for some constant $C > 0$ when n is sufficiently large; $f(n) \gtrsim g(n)$ means $|f(n)|/|g(n)| \ge C$ for some constant $C > 0$ when n is sufficiently large. Additionally, we use $f(n) \gg g(n)$ (resp. $f(n) \ll g(n)$) to indicate that there exists some very large (resp. small) universal constant $c > 0$ such that $|f(n)| \ge c\,|g(n)|$ (resp. $|f(n)| \le c\,|g(n)|$). We use $C, C_1, C_2, c_1, c_2, \ldots$ to represent positive constants, whose values may differ from line to line. Last but not least, we use the terminology "with overwhelming probability" to denote the event happens with probability at least $1 - c_1 n^{-c_2}$.

2 Main Results

2.1 Models and assumptions

We assume the ground truth tensor $\boldsymbol{\mathcal{X}}_\star = [\boldsymbol{\mathcal{X}}_\star(i_1,i_2,i_3)] \in \mathbb{R}^{n_1\times n_2\times n_3}$ admits the following Tucker decomposition

$$\boldsymbol{\mathcal{X}}_\star(i_1,i_2,i_3) = \sum_{j_1=1}^{r_1}\sum_{j_2=1}^{r_2}\sum_{j_3=1}^{r_3} \boldsymbol{U}_\star(i_1,j_1)\boldsymbol{V}_\star(i_2,j_2)\boldsymbol{W}_\star(i_3,j_3)\boldsymbol{\mathcal{S}}_\star(j_1,j_2,j_3), \quad 1 \le i_k \le n_k, \tag{10}$$

or more compactly,

$$\boldsymbol{\mathcal{X}}_\star = (\boldsymbol{U}_\star, \boldsymbol{V}_\star, \boldsymbol{W}_\star) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_\star, \tag{11}$$

where $\boldsymbol{\mathcal{S}}_\star = [\boldsymbol{\mathcal{S}}_\star(j_1,j_2,j_3)] \in \mathbb{R}^{r_1\times r_2\times r_3}$ is the core tensor of multilinear rank $\boldsymbol{r} = (r_1,r_2,r_3)$, and $\boldsymbol{U}_\star = [\boldsymbol{U}_\star(i_1,j_1)] \in \mathbb{R}^{n_1\times r_1}$, $\boldsymbol{V}_\star = [\boldsymbol{V}_\star(i_2,j_2)] \in \mathbb{R}^{n_2\times r_2}$, $\boldsymbol{W}_\star = [\boldsymbol{W}_\star(i_3,j_3)] \in \mathbb{R}^{n_3\times r_3}$ are the factor matrices of each mode. Let $\mathcal{M}_k(\boldsymbol{\mathcal{X}}_\star)$ be the mode-k matricization of $\boldsymbol{\mathcal{X}}_\star$, we have

$$\mathcal{M}_1(\boldsymbol{\mathcal{X}}_\star) = \boldsymbol{U}_\star \mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{V}_\star \otimes \boldsymbol{W}_\star)^\top, \tag{12a}$$

$$\mathcal{M}_2(\boldsymbol{\mathcal{X}}_\star) = \boldsymbol{V}_\star \mathcal{M}_2(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{U}_\star \otimes \boldsymbol{W}_\star)^\top, \tag{12b}$$

$$\mathcal{M}_3(\boldsymbol{\mathcal{X}}_\star) = \boldsymbol{W}_\star \mathcal{M}_3(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{U}_\star \otimes \boldsymbol{V}_\star)^\top. \tag{12c}$$

It is straightforward to see that the Tucker decomposition is not uniquely specified: for any invertible matrices $\boldsymbol{Q}_k \in \mathbb{R}^{r_k \times r_k}$, $k = 1, 2, 3$, one has

$$(\boldsymbol{U}_\star, \boldsymbol{V}_\star, \boldsymbol{W}_\star) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_\star = (\boldsymbol{U}_\star \boldsymbol{Q}_1, \boldsymbol{V}_\star \boldsymbol{Q}_2, \boldsymbol{W}_\star \boldsymbol{Q}_3) \boldsymbol{\cdot} ((\boldsymbol{Q}_1^{-1}, \boldsymbol{Q}_2^{-1}, \boldsymbol{Q}_3^{-1}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_\star).$$

We shall fix the ground truth factor such that $\boldsymbol{U}_\star$, $\boldsymbol{V}_\star$ and $\boldsymbol{W}_\star$ are orthonormal matrices consisting of left singular vectors in each mode. Furthermore, the core tensor $\boldsymbol{\mathcal{S}}_\star$ is related to the singular values in each mode as

$$\mathcal{M}_k(\boldsymbol{\mathcal{S}}_\star)\mathcal{M}_k(\boldsymbol{\mathcal{S}}_\star)^\top = \boldsymbol{\Sigma}_{\star,k}^2, \qquad k = 1, 2, 3, \tag{13}$$

where $\boldsymbol{\Sigma}_{\star,k} \coloneqq \mathrm{diag}[\sigma_1(\mathcal{M}_k(\boldsymbol{\mathcal{X}}_\star)), \ldots, \sigma_{r_k}(\mathcal{M}_k(\boldsymbol{\mathcal{X}}_\star))]$ is a diagonal matrix where the diagonal elements are composed of the nonzero singular values of $\mathcal{M}_k(\boldsymbol{\mathcal{X}}_\star)$ and $r_k = \mathrm{rank}(\mathcal{M}_k(\boldsymbol{\mathcal{X}}_\star))$ for $k = 1, 2, 3$.

Key parameters. Of particular interest is a sort of condition number of $\boldsymbol{\mathcal{X}}_\star$, which plays an important role in governing the computational efficiency of first-order algorithms.

Definition 1 (Condition number)**.** The condition number of $\boldsymbol{\mathcal{X}}_\star$ is defined as

$$\kappa \coloneqq \frac{\sigma_{\max}(\boldsymbol{\mathcal{X}}_\star)}{\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)} = \frac{\max_{k=1,2,3} \sigma_{\max}(\mathcal{M}_k(\boldsymbol{\mathcal{X}}_\star))}{\min_{k=1,2,3} \sigma_{\min}(\mathcal{M}_k(\boldsymbol{\mathcal{X}}_\star))}. \tag{14}$$

Another parameter is the incoherence parameter, which plays an important role in governing the well-posedness of low-rank tensor completion.

Definition 2 (Incoherence)**.** The incoherence parameter of $\boldsymbol{\mathcal{X}}_\star$ is defined as

$$\mu \coloneqq \max\left\{\frac{n_1}{r_1}\|\boldsymbol{U}_\star\|_{2,\infty}^2, \frac{n_2}{r_2}\|\boldsymbol{V}_\star\|_{2,\infty}^2, \frac{n_3}{r_3}\|\boldsymbol{W}_\star\|_{2,\infty}^2\right\}. \tag{15}$$

Roughly speaking, a small incoherence parameter ensures that the energy of the tensor is evenly distributed across its entries, so that a small random subset of its elements still reveals substantial information about the latent structure of the entire tensor.

2.2 ScaledGD for tensor completion

Assume that we have observed a subset of entries in $\boldsymbol{\mathcal{X}}_\star$, given as $\boldsymbol{\mathcal{Y}} = \mathcal{P}_\Omega(\boldsymbol{\mathcal{X}}_\star)$, where $\mathcal{P}_\Omega : \mathbb{R}^{n_1 \times n_2 \times n_3} \mapsto \mathbb{R}^{n_1 \times n_2 \times n_3}$ is a projection such that

$$[\mathcal{P}_\Omega(\boldsymbol{\mathcal{X}}_\star)](i_1, i_2, i_3) = \begin{cases} \boldsymbol{\mathcal{X}}_\star(i_1, i_2, i_3), & \text{if } (i_1, i_2, i_3) \in \Omega, \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

Here, Ω is generated according to the Bernoulli observation model in the sense that

$$(i_1, i_2, i_3) \in \Omega \quad \text{independently with probability } p \in (0, 1]. \tag{17}$$

The goal of tensor completion is to recover the tensor $\boldsymbol{\mathcal{X}}_\star$ from its partial observation $\mathcal{P}_\Omega(\boldsymbol{\mathcal{X}}_\star)$. Similar to the tensor regression case, this can be achieved by minimizing the loss function

$$\min_{\boldsymbol{F}=(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W},\boldsymbol{\mathcal{S}})} \mathcal{L}(\boldsymbol{F}) \coloneqq \frac{1}{2p}\left\|\mathcal{P}_\Omega\big((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}\big) - \boldsymbol{\mathcal{Y}}\right\|_{\mathsf{F}}^2. \tag{18}$$

Preparation: a scaled projection operator. To guarantee faithful recovery from partial observations, the underlying low-rank tensor $\boldsymbol{\mathcal{X}}_\star$ needs to be incoherent (cf. Definition 2) to avoid ill-posedness. One typical strategy, frequently employed in the matrix setting, to ensure the incoherence condition is to trim the rows of the factors [CW15] after the gradient update. For ScaledGD, this needs to be done in a careful manner to preserve the equivariance with respect to invertible transforms. Motivated by [TMC20], we introduce the scaled projection as follows,

$$(\boldsymbol{U}, \boldsymbol{W}, \boldsymbol{W}, \boldsymbol{\mathcal{S}}) = \mathcal{P}_B(\boldsymbol{U}_+, \boldsymbol{V}_+, \boldsymbol{W}_+, \boldsymbol{\mathcal{S}}_+), \tag{19a}$$

where $B > 0$ is the projection radius, and

$$\begin{aligned}
\boldsymbol{U}(i_1,:) &= \left(1 \wedge \frac{B}{\sqrt{n_1}\|\boldsymbol{U}_+(i_1,:)\breve{\boldsymbol{U}}_+^\top\|_2}\right)\boldsymbol{U}_+(i_1,:), \qquad 1 \le i_1 \le n_1;\\
\boldsymbol{V}(i_2,:) &= \left(1 \wedge \frac{B}{\sqrt{n_2}\|\boldsymbol{V}_+(i_2,:)\breve{\boldsymbol{V}}_+^\top\|_2}\right)\boldsymbol{V}_+(i_2,:), \qquad 1 \le i_2 \le n_2;\\
\boldsymbol{W}(i_3,:) &= \left(1 \wedge \frac{B}{\sqrt{n_3}\|\boldsymbol{W}_+(i_3,:)\breve{\boldsymbol{W}}_+^\top\|_2}\right)\boldsymbol{W}_+(i_3,:), \qquad 1 \le i_3 \le n_3;\\
\boldsymbol{\mathcal{S}} &= \boldsymbol{\mathcal{S}}_+.
\end{aligned} \tag{19b}$$

Here, we recall $\breve{\boldsymbol{U}}_+$, $\breve{\boldsymbol{V}}_+$, $\breve{\boldsymbol{W}}_+$ are analogously defined in (5) using $(\boldsymbol{U}_+, \boldsymbol{V}_+, \boldsymbol{W}_+, \boldsymbol{\mathcal{S}}_+)$. As can be seen, each row of $\boldsymbol{U}_+$ (resp. $\boldsymbol{V}_+$ and $\boldsymbol{W}_+$) is scaled by a scalar based on the row ℓ_2 norms of $\boldsymbol{U}_+\breve{\boldsymbol{U}}_+^\top$ (resp. $\boldsymbol{V}_+\breve{\boldsymbol{V}}_+^\top$ and $\boldsymbol{W}_+\breve{\boldsymbol{W}}_+^\top$), which is the mode-1 (resp. mode-2 and mode-3) matricization of the tensor $(\boldsymbol{U}_+, \boldsymbol{V}_+, \boldsymbol{W}_+) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_+$. It is a straightforward observation that the projection can be computed efficiently.

Algorithm description. With the scaled projection $\mathcal{P}_B(\cdot)$ defined in hand, we are in a position to describe the details of the proposed ScaledGD algorithm, summarized in Algorithm 1. It consists of two stages: spectral initialization followed by iterative refinements using the scaled projected gradient updates in (20). It is worth emphasizing that all the factors are updated simultaneously, which can be achieved in a parallel manner to accelerate computation run time.

For the spectral initialization, we take advantage of the subspace estimators proposed in [CLC+21] for highly unbalanced data matrices. Specifically, we estimate the subspace spanned by $\boldsymbol{U}_\star$ by that spanned by the top-r_1 left singular vectors $\boldsymbol{U}_+$ of the diagonally-deleted Gram matrix of $p^{-1}\mathcal{M}_1(\boldsymbol{\mathcal{Y}})$, denoted as $\mathcal{P}_{\mathsf{off\text{-}diag}}(p^{-2}\mathcal{M}_1(\boldsymbol{\mathcal{Y}})\mathcal{M}_1(\boldsymbol{\mathcal{Y}})^\top)$, where $\mathcal{P}_{\mathsf{off\text{-}diag}}(\boldsymbol{M})$ sets the diagonal entries of the matrix $\boldsymbol{M}$ as zeros; the other two factors $\boldsymbol{V}_+$ and $\boldsymbol{W}_+$ are estimated similarly. The core tensor is then estimated as

$$\boldsymbol{\mathcal{S}}_+ = p^{-1}(\boldsymbol{U}_+^\top, \boldsymbol{V}_+^\top, \boldsymbol{W}_+^\top) \boldsymbol{\cdot} \boldsymbol{\mathcal{Y}},$$

which is consistent with its estimation in the HOSVD procedure. To ensure the initialization is incoherent, we pass it through the scaled projection operator to obtain the final initial estimate:

$$(\boldsymbol{U}_0, \boldsymbol{V}_0, \boldsymbol{W}_0, \boldsymbol{\mathcal{S}}_0) = \mathcal{P}_B\big(\boldsymbol{U}_+, \boldsymbol{V}_+, \boldsymbol{W}_+, \boldsymbol{\mathcal{S}}_+\big).$$

Theoretical guarantees. The following theorem establishes the performance guarantee of ScaledGD for tensor completion, as soon as the sample size is sufficiently large.

Theorem 1 (ScaledGD for tensor completion)**.** *Suppose that $\boldsymbol{\mathcal{X}}_\star$ is μ-incoherent, $\mu r^2\kappa^2 \ll n$, and that p satisfies*

$$pn_1n_2n_3 \gtrsim \epsilon_0^{-2}\mu^{3/2}r^2\kappa(\sqrt{r} \vee \kappa^2)\sqrt{n_1n_2n_3}\log^3 n + \epsilon_0^{-4}\mu^3r^4\kappa^6 n\log^5 n,$$

for some small constant $\epsilon_0 > 0$. Set the projection radius as $B = C_B\sqrt{\mu r}\sigma_{\max}(\boldsymbol{\mathcal{X}}_\star)$ for some constant $C_B \ge (1+\epsilon_0)^3$. If the step size obeys $0 < \eta \le 2/5$, then with probability at least $1 - c_1 n^{-c_2}$ for universal constants $c_1, c_2 > 0$, for all $t \ge 0$, the iterates of Algorithm 1 satisfy

$$\|(\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_t - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \le 3\epsilon_0(1 - 0.6\eta)^t\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star).$$

Algorithm 1 ScaledGD for low-rank tensor completion

```
**Input parameters:** step size $\eta > 0$, multilinear rank $\boldsymbol{r} = (r_1, r_2, r_3)$, probability of observation $p$.
**Spectral initialization:** Let $\boldsymbol{U}_+$ be the top-$r_1$ left singular vectors of $\mathcal{P}_{\mathsf{off\text{-}diag}}(p^{-2}\mathcal{M}_1(\boldsymbol{\mathcal{Y}})\mathcal{M}_1(\boldsymbol{\mathcal{Y}})^\top)$, and similarly for $\boldsymbol{V}_+, \boldsymbol{W}_+$, and $\boldsymbol{\mathcal{S}}_+ = p^{-1}(\boldsymbol{U}_+^\top, \boldsymbol{V}_+^\top, \boldsymbol{W}_+^\top) \boldsymbol{\cdot} \boldsymbol{\mathcal{Y}}$. Set $(\boldsymbol{U}_0, \boldsymbol{V}_0, \boldsymbol{W}_0, \boldsymbol{\mathcal{S}}_0) = \mathcal{P}_B\big(\boldsymbol{U}_+, \boldsymbol{V}_+, \boldsymbol{W}_+, \boldsymbol{\mathcal{S}}_+\big)$.
**Scaled projected gradient updates:** for $t = 0, 1, 2, \dots, T-1$ do
    $$\begin{aligned} \boldsymbol{U}_{t+} &= \boldsymbol{U}_t - \frac{\eta}{p}\mathcal{M}_1\left(\mathcal{P}_\Omega\big((\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_t\big) - \boldsymbol{\mathcal{Y}}\right) \breve{\boldsymbol{U}}_t(\breve{\boldsymbol{U}}_t^\top \breve{\boldsymbol{U}}_t)^{-1}, \\ \boldsymbol{V}_{t+} &= \boldsymbol{V}_t - \frac{\eta}{p}\mathcal{M}_2\left(\mathcal{P}_\Omega\big((\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_t\big) - \boldsymbol{\mathcal{Y}}\right) \breve{\boldsymbol{V}}_t(\breve{\boldsymbol{V}}_t^\top \breve{\boldsymbol{V}}_t)^{-1}, \\ \boldsymbol{W}_{t+} &= \boldsymbol{W}_t - \frac{\eta}{p}\mathcal{M}_3\left(\mathcal{P}_\Omega\big((\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_t\big) - \boldsymbol{\mathcal{Y}}\right) \breve{\boldsymbol{W}}_t(\breve{\boldsymbol{W}}_t^\top \breve{\boldsymbol{W}}_t)^{-1}, \\ \boldsymbol{\mathcal{S}}_{t+} &= \boldsymbol{\mathcal{S}}_t - \frac{\eta}{p}\left((\boldsymbol{U}_t^\top \boldsymbol{U}_t)^{-1}\boldsymbol{U}_t^\top, (\boldsymbol{V}_t^\top \boldsymbol{V}_t)^{-1}\boldsymbol{V}_t^\top, (\boldsymbol{W}_t^\top \boldsymbol{W}_t)^{-1}\boldsymbol{W}_t^\top\right) \boldsymbol{\cdot} \left(\mathcal{P}_\Omega\big((\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_t\big) - \boldsymbol{\mathcal{Y}}\right), \end{aligned} \tag{20}$$
    where $\breve{\boldsymbol{U}}_t$, $\breve{\boldsymbol{V}}_t$, and $\breve{\boldsymbol{W}}_t$ are defined in (5). Set $(\boldsymbol{U}_{t+1}, \boldsymbol{V}_{t+1}, \boldsymbol{W}_{t+1}, \boldsymbol{\mathcal{S}}_{t+1}) = \mathcal{P}_B(\boldsymbol{U}_{t+}, \boldsymbol{V}_{t+}, \boldsymbol{W}_{t+}, \boldsymbol{\mathcal{S}}_{t+})$.
```

Theorem 1 ensures that ScaledGD finds an ε-accurate estimate, i.e. $\|(\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_t - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \le \varepsilon\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$, in at most $O(\log(1/\varepsilon))$ iterations, which is independent of the condition number of $\boldsymbol{\mathcal{X}}_\star$, as long as the sample complexity is large enough. Assuming that $\mu = O(1)$ and $r \vee \kappa \ll n^\delta$ for some small constant δ to keep only terms with dominating orders of n, the sample complexity simplifies to

$$pn_1n_2n_3 \gtrsim n^{3/2}r^2\kappa(\sqrt{r} \vee \kappa^2)\log^3 n,$$

which is near-optimal in view of the conjecture that no polynomial-time algorithm will be successful if the sample complexity is less than the order of $n^{3/2}$ for tensor completion [BM16]. Compared with existing algorithms collected in Table 1, ScaledGD is the *first* algorithm that simultaneously achieves a near-optimal sample complexity and a near-linear run time complexity in a provable manner. In particular, while [YZ16, XY19] achieve a sample complexity comparable to ours, the tensor nuclear norm minimization algorithm in [YZ16] is NP-hard to compute, and the Grassmannian GD algorithm in [XY19] does not offer an explicit iteration complexity, except that each iteration can be computed in a polynomial time.

2.3 ScaledGD for tensor regression

Now we move on to another tensor recovery problem—tensor regression with Gaussian design. Assume that we have access to a set of observations given as

$$y_i = \langle \boldsymbol{\mathcal{A}}_i, \boldsymbol{\mathcal{X}}_\star \rangle, \qquad i = 1, \dots, m, \qquad \text{or concisely,} \qquad \boldsymbol{y} = \mathcal{A}(\boldsymbol{\mathcal{X}}_\star), \tag{21}$$

where $\boldsymbol{\mathcal{A}}_i \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is the i-th measurement tensor composed of i.i.d. Gaussian entries drawn from $\mathcal{N}(0, 1/m)$, and $\mathcal{A}(\boldsymbol{\mathcal{X}}) = \{\langle \boldsymbol{\mathcal{A}}_i, \boldsymbol{\mathcal{X}} \rangle\}_{i=1}^m$ is a linear map from $\mathbb{R}^{n_1 \times n_2 \times n_3}$ to $\mathbb{R}^m$, whose adjoint operator is given by $\mathcal{A}^*(\boldsymbol{y}) = \sum_{i=1}^m y_i \boldsymbol{\mathcal{A}}_i$. The goal of tensor regression is to recover $\boldsymbol{\mathcal{X}}_\star$ from $\boldsymbol{y}$, by leveraging the low-rank structure of $\boldsymbol{\mathcal{X}}_\star$. This can be achieved by minimizing the following loss function

$$\min_{\boldsymbol{F} = (\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}, \boldsymbol{\mathcal{S}})} \mathcal{L}(\boldsymbol{F}) \coloneqq \frac{1}{2}\left\|\mathcal{A}((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}) - \boldsymbol{y}\right\|_2^2. \tag{22}$$

The proposed ScaledGD algorithm to minimize (22) is described in Algorithm 2, where the algorithm is initialized by applying HOSVD to $\mathcal{A}^*(\boldsymbol{y})$, followed by scaled gradient updates given in (23).

Theoretical guarantees. Encouragingly, we can guarantee that ScaledGD provably recovers the ground truth tensor as long as the sample size is sufficiently large, which is given in the following theorem.

Theorem 2 (ScaledGD for tensor regression)**.** *For tensor regression with Gaussian design, suppose that m satisfies*

$$m \gtrsim \epsilon_0^{-2}(\sqrt{n_1n_2n_3}r\kappa^2 + nr^2\kappa^3),$$

```
Algorithm 2 ScaledGD for low-rank tensor regression
  Input parameters: step size $\eta > 0$, multilinear rank $\boldsymbol{r} = (r_1, r_2, r_3)$.
  Spectral initialization: Let $(\boldsymbol{U}_0, \boldsymbol{V}_0, \boldsymbol{W}_0, \boldsymbol{\mathcal{S}}_0) = \mathrm{HOSVD}_{\boldsymbol{r}}(\mathcal{A}^*(\boldsymbol{y}))$ defined in (8b).
  Scaled gradient updates: for $t = 0, 1, 2, \dots, T-1$
      $$\begin{aligned} \boldsymbol{U}_{t+1} &= \boldsymbol{U}_t - \eta \mathcal{M}_1\left(\mathcal{A}^*(\mathcal{A}((\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_t) - \boldsymbol{y})\right) \breve{\boldsymbol{U}}_t^\top \left(\breve{\boldsymbol{U}}_t^\top \breve{\boldsymbol{U}}_t\right)^{-1}, \\ \boldsymbol{V}_{t+1} &= \boldsymbol{V}_t - \eta \mathcal{M}_2\left(\mathcal{A}^*(\mathcal{A}((\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_t) - \boldsymbol{y})\right) \breve{\boldsymbol{V}}_t^\top \left(\breve{\boldsymbol{V}}_t^\top \breve{\boldsymbol{V}}_t\right)^{-1}, \\ \boldsymbol{W}_{t+1} &= \boldsymbol{W}_t - \eta \mathcal{M}_3\left(\mathcal{A}^*(\mathcal{A}((\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_t) - \boldsymbol{y})\right) \breve{\boldsymbol{W}}_t^\top \left(\breve{\boldsymbol{W}}_t^\top \breve{\boldsymbol{W}}_t\right)^{-1}, \\ \boldsymbol{\mathcal{S}}_{t+1} &= \boldsymbol{\mathcal{S}}_t - \eta \left((\boldsymbol{U}_t^\top \boldsymbol{U}_t)^{-1} \boldsymbol{U}_t^\top, (\boldsymbol{V}_t^\top \boldsymbol{V}_t)^{-1} \boldsymbol{V}_t^\top, (\boldsymbol{W}_t^\top \boldsymbol{W}_t)^{-1} \boldsymbol{W}_t^\top\right) \boldsymbol{\cdot} \mathcal{A}^*(\mathcal{A}((\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_t) - \boldsymbol{y}), \end{aligned} \tag{23}$$
  where $\breve{\boldsymbol{U}}_t$, $\breve{\boldsymbol{V}}_t$, and $\breve{\boldsymbol{W}}_t$ are defined in (5).
```

for some small constant $\epsilon_0 > 0$. If the step size obeys $0 < \eta \le 2/5$, then with probability at least $1 - c_1 n^{-c_2}$ for universal constants $c_1, c_2 > 0$, for all $t \ge 0$, the iterates of Algorithm 2 satisfy

$$\|(\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_t - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \le 3\epsilon_0 (1 - 0.6\eta)^t \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star).$$

Theorem 2 ensures that ScaledGD finds an ε-accurate estimate, i.e. $\|(\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_t - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \le \varepsilon \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$, in at most $O(\log(1/\varepsilon))$ iterations, which is independent of the condition number of $\boldsymbol{\mathcal{X}}_\star$, as long as the sample complexity satisfies

$$m \gtrsim n^{3/2} r \kappa^2,$$

where again we keep only terms with dominating orders of n. Compared with the regularized GD algorithm [HWZ20], ScaledGD achieves a low computation complexity with robustness to ill-conditioning, improving its iteration complexity by a factor of κ^2, and does not require any explicit regularization.

3 Analysis

In this section, we provide some intuitions and sketch the proof of our main theorems. Before continuing, we highlight an important property of ScaledGD: if starting from an equivalent estimate

$$\widetilde{\boldsymbol{U}}_t = \boldsymbol{U}_t \boldsymbol{Q}_1, \quad \widetilde{\boldsymbol{V}}_t = \boldsymbol{V}_t \boldsymbol{Q}_2, \quad \widetilde{\boldsymbol{W}}_t = \boldsymbol{W}_t \boldsymbol{Q}_3, \quad \widetilde{\boldsymbol{\mathcal{S}}}_t = (\boldsymbol{Q}_1^{-1}, \boldsymbol{Q}_2^{-1}, \boldsymbol{Q}_3^{-1}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_t$$

for some invertible matrices $\boldsymbol{Q}_k \in \mathrm{GL}(r_k)$ (i.e. replacing $\boldsymbol{U}_t$ by $\boldsymbol{U}_t \boldsymbol{Q}_1$, and so on), by plugging the above estimate in (4) it is easy to check that the next iterate of ScaledGD is covariant with respect to invertible transforms, meaning

$$\widetilde{\boldsymbol{U}}_{t+1} = \boldsymbol{U}_{t+1} \boldsymbol{Q}_1, \quad \widetilde{\boldsymbol{V}}_{t+1} = \boldsymbol{V}_{t+1} \boldsymbol{Q}_2, \quad \widetilde{\boldsymbol{W}}_{t+1} = \boldsymbol{W}_{t+1} \boldsymbol{Q}_3, \quad \widetilde{\boldsymbol{\mathcal{S}}}_{t+1} = (\boldsymbol{Q}_1^{-1}, \boldsymbol{Q}_2^{-1}, \boldsymbol{Q}_3^{-1}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_{t+1}.$$

In other words, ScaledGD produces an invariant sequence of low-rank tensor estimates

$$\boldsymbol{\mathcal{X}}_t = (\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_t = (\widetilde{\boldsymbol{U}}_t, \widetilde{\boldsymbol{V}}_t, \widetilde{\boldsymbol{W}}_t) \boldsymbol{\cdot} \widetilde{\boldsymbol{\mathcal{S}}}_t$$

regardless of the representation of the tensor factors with respect to the underlying symmetry group. This is one of the key reasons behind the insensitivity of ScaledGD to ill-conditioning and factor unbalancedness.

A key scaled distance metric. To track the progress of ScaledGD throughout the entire trajectory, one needs a distance metric that properly takes account of the factor ambiguity due to invertible transforms, as well as the effect of scaling. To that end, we define the scaled distance between factor quadruples $\boldsymbol{F} = (\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}, \boldsymbol{\mathcal{S}})$ and $\boldsymbol{F}_\star = (\boldsymbol{U}_\star, \boldsymbol{V}_\star, \boldsymbol{W}_\star, \boldsymbol{\mathcal{S}}_\star)$ as

$$\mathrm{dist}^2(\boldsymbol{F}, \boldsymbol{F}_\star) \coloneqq \inf_{\boldsymbol{Q}_k \in \mathrm{GL}(r_k)} \|(\boldsymbol{U}\boldsymbol{Q}_1 - \boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 + \|(\boldsymbol{V}\boldsymbol{Q}_2 - \boldsymbol{V}_\star)\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}^2 + \|(\boldsymbol{W}\boldsymbol{Q}_3 - \boldsymbol{W}_\star)\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}^2$$

$$+\left\|(\boldsymbol{Q}_1^{-1}, \boldsymbol{Q}_2^{-1}, \boldsymbol{Q}_3^{-1}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{S}}_\star\right\|_{\mathsf{F}}^2. \tag{24}$$

The distance is closely related to the ℓ_2 distances between the corresponding tensors. In fact, it can be shown that as long as $\boldsymbol{F}$ and $\boldsymbol{F}_\star$ are not too far apart, i.e. $\operatorname{dist}(\boldsymbol{F}, \boldsymbol{F}_\star) \le 0.2\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$, it holds that $\operatorname{dist}(\boldsymbol{F}, \boldsymbol{F}_\star) \asymp \|(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}$ in the sense that (see Appendix A.1 for proofs):

$$\tfrac{1}{3}\|(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \le \operatorname{dist}(\boldsymbol{F}, \boldsymbol{F}_\star) \le (\sqrt{2}+1)^{3/2}\|(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}.$$

3.1 A warm-up case: ScaledGD for tensor factorization

To shed light on the design insights as well as the proof techniques, we now introduce the ScaledGD algorithm for the tensor factorization problem, which aims to minimize the following loss function:

$$\mathcal{L}(\boldsymbol{F}) \coloneqq \frac{1}{2}\|(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}^2 = \frac{1}{2}\|\mathcal{M}_k\left((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_\star\right)\|_{\mathsf{F}}^2, \quad k = 1,2,3, \tag{25}$$

where the last equality follows from (6b). Recalling the update rule (4), ScaledGD proceeds as

$$\begin{aligned}
\boldsymbol{U}_{t+1} &= \boldsymbol{U}_t - \eta\mathcal{M}_1\left(\boldsymbol{\mathcal{X}}_t - \boldsymbol{\mathcal{X}}_\star\right)\breve{\boldsymbol{U}}_t\left(\breve{\boldsymbol{U}}_t^\top\breve{\boldsymbol{U}}_t\right)^{-1}, \\
\boldsymbol{V}_{t+1} &= \boldsymbol{V}_t - \eta\mathcal{M}_2\left(\boldsymbol{\mathcal{X}}_t - \boldsymbol{\mathcal{X}}_\star\right)\breve{\boldsymbol{V}}_t\left(\breve{\boldsymbol{V}}_t^\top\breve{\boldsymbol{V}}_t\right)^{-1}, \\
\boldsymbol{W}_{t+1} &= \boldsymbol{W}_t - \eta\mathcal{M}_3\left(\boldsymbol{\mathcal{X}}_t - \boldsymbol{\mathcal{X}}_\star\right)\breve{\boldsymbol{W}}_t\left(\breve{\boldsymbol{W}}_t^\top\breve{\boldsymbol{W}}_t\right)^{-1}, \\
\boldsymbol{\mathcal{S}}_{t+1} &= \boldsymbol{\mathcal{S}}_t - \eta\left((\boldsymbol{U}_t^\top\boldsymbol{U}_t)^{-1}\boldsymbol{U}_t^\top, (\boldsymbol{V}_t^\top\boldsymbol{V}_t)^{-1}\boldsymbol{V}_t^\top, (\boldsymbol{W}_t^\top\boldsymbol{W}_t)^{-1}\boldsymbol{W}_t^\top\right) \boldsymbol{\cdot} \left(\boldsymbol{\mathcal{X}}_t - \boldsymbol{\mathcal{X}}_\star\right),
\end{aligned} \tag{26}$$

where $\boldsymbol{\mathcal{X}}_t = (\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_t$, with $\breve{\boldsymbol{U}}_t$, $\breve{\boldsymbol{V}}_t$, and $\breve{\boldsymbol{W}}_t$ defined in (5).

ScaledGD as a quasi-Newton algorithm. One way to think of ScaledGD is through the lens of quasi-Newton methods, by equivalently rewriting the ScaledGD update (26) as

$$\operatorname{vec}(\boldsymbol{F}_{t+1}) = \operatorname{vec}(\boldsymbol{F}_t) - \eta\boldsymbol{H}_t^{-1}\nabla_{\operatorname{vec}(\boldsymbol{F})}\mathcal{L}(\boldsymbol{F}_t), \tag{27}$$

where $\boldsymbol{H}_t \coloneqq \operatorname{diag}\left[\nabla^2_{\operatorname{vec}(\boldsymbol{U}),\operatorname{vec}(\boldsymbol{U})}\mathcal{L}(\boldsymbol{F}_t),\ \nabla^2_{\operatorname{vec}(\boldsymbol{V}),\operatorname{vec}(\boldsymbol{V})}\mathcal{L}(\boldsymbol{F}_t),\ \nabla^2_{\operatorname{vec}(\boldsymbol{W}),\operatorname{vec}(\boldsymbol{W})}\mathcal{L}(\boldsymbol{F}_t),\ \nabla^2_{\operatorname{vec}(\boldsymbol{\mathcal{S}}),\operatorname{vec}(\boldsymbol{\mathcal{S}})}\mathcal{L}(\boldsymbol{F}_t)\right]$. To see this, it is straightforward to check that the diagonal blocks of the Hessian of the loss function (25) are given precisely as

$$\begin{aligned}
\nabla^2_{\operatorname{vec}(\boldsymbol{U}),\operatorname{vec}(\boldsymbol{U})}\mathcal{L}(\boldsymbol{F}_t) &= (\breve{\boldsymbol{U}}_t^\top\breve{\boldsymbol{U}}_t) \otimes \boldsymbol{I}_{n_1}, \\
\nabla^2_{\operatorname{vec}(\boldsymbol{V}),\operatorname{vec}(\boldsymbol{V})}\mathcal{L}(\boldsymbol{F}_t) &= (\breve{\boldsymbol{V}}_t^\top\breve{\boldsymbol{V}}_t) \otimes \boldsymbol{I}_{n_2}, \\
\nabla^2_{\operatorname{vec}(\boldsymbol{W}),\operatorname{vec}(\boldsymbol{W})}\mathcal{L}(\boldsymbol{F}_t) &= (\breve{\boldsymbol{W}}_t^\top\breve{\boldsymbol{W}}_t) \otimes \boldsymbol{I}_{n_3}, \\
\nabla^2_{\operatorname{vec}(\boldsymbol{\mathcal{S}}),\operatorname{vec}(\boldsymbol{\mathcal{S}})}\mathcal{L}(\boldsymbol{F}_t) &= (\boldsymbol{U}_t^\top\boldsymbol{U}_t) \otimes (\boldsymbol{V}_t^\top\boldsymbol{V}_t) \otimes (\boldsymbol{W}_t^\top\boldsymbol{W}_t).
\end{aligned} \tag{28}$$

Therefore, by vectorization of (26), ScaledGD can be regarded as a quasi-Newton method where the preconditioner is designed as the inverse of the diagonal approximation of the Hessian.

Guarantees for tensor factorization. Fortunately, ScaledGD admits a κ-independent convergence rate for tensor factorization, as long as the initialization is not too far from the ground truth. This is summarized in Theorem 3, whose proof can be found in Appendix B.

Theorem 3. *For tensor factorization (25), suppose that the initialization satisfies* $\operatorname{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \le \epsilon_0\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$ *for some small constant* $\epsilon_0 > 0$*, then for all* $t \ge 0$*, the iterates of ScaledGD in (26) satisfy*

$$\operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \le (1-0.7\eta)^t\epsilon_0\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star), \quad \textit{and} \quad \|(\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_t - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \le 3\epsilon_0(1-0.7\eta)^t\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star),$$

as long as the step size satisfies $0 < \eta \le 2/5$.

Intuition of the proof. Let us provide some intuitions to facilitate understanding by examining a toy case, where all factors become scalars, and the loss function with respect to the factor $\boldsymbol{f} = [u, v, w, s]^\top$ becomes

$$\mathcal{L}(\boldsymbol{f}) = \frac{1}{2}(uvws - u_\star v_\star w_\star s_\star)^2 = \frac{1}{2}(uvws - s_\star)^2,$$

where $u_\star = v_\star = w_\star = 1$, and the ground truth is $\boldsymbol{f}_\star = [1,\ 1,\ 1,\ s_\star]^\top$. The gradient and the diagonal entries of the Hessian are given respectively as

$$\begin{aligned} \nabla\mathcal{L}(\boldsymbol{f}) &= (uvws - s_\star)[vws,\ uws,\ uvs,\ uvw]^\top, \\ \operatorname{diag}(\nabla^2\mathcal{L}(\boldsymbol{f})) &= \operatorname{diag}[(vws)^2,\ (uws)^2,\ (uvs)^2,\ (uvw)^2]. \end{aligned}$$

Moreover, the Hessian matrix at the ground truth is given by

$$\nabla^2\mathcal{L}(\boldsymbol{f}_\star) = [s_\star,\ s_\star,\ s_\star,\ 1]^\top[s_\star,\ s_\star,\ s_\star,\ 1].$$

With these in mind, the ScaledGD update rule in (26) and the scaled distance in (24) reduce respectively to

$$\begin{aligned} \boldsymbol{f}_{t+1} &= \boldsymbol{f}_t - \eta \operatorname{diag}^{-1}(\nabla^2\mathcal{L}(\boldsymbol{f}_t))\nabla\mathcal{L}(\boldsymbol{f}_t), \\ \operatorname{dist}(\boldsymbol{f}, \boldsymbol{f}_\star) &= \inf_{\boldsymbol{Q}=\operatorname{diag}(q_1,q_2,q_3,(q_1q_2q_3)^{-1})} \left\|\operatorname{diag}^{1/2}(\nabla^2\mathcal{L}(\boldsymbol{f}_\star))(\boldsymbol{Q}\boldsymbol{f} - \boldsymbol{f}_\star)\right\|_2. \end{aligned}$$

Consequently, we can bound the distance between $\boldsymbol{f}_{t+1}$ and $\boldsymbol{f}_\star$ as

$$\begin{aligned} \operatorname{dist}(\boldsymbol{f}_{t+1}, \boldsymbol{f}_\star) &\overset{\text{(i)}}{\leq} \left\|\operatorname{diag}^{1/2}(\nabla^2\mathcal{L}(\boldsymbol{f}_\star))\left(\boldsymbol{Q}_t(\boldsymbol{f}_t - \eta \operatorname{diag}^{-1}(\nabla^2\mathcal{L}(\boldsymbol{f}_t))\nabla\mathcal{L}(\boldsymbol{f}_t)) - \boldsymbol{f}_\star\right)\right\|_2 \\ &\overset{\text{(ii)}}{=} \left\|\operatorname{diag}^{1/2}(\nabla^2\mathcal{L}(\boldsymbol{f}_\star))\left(\boldsymbol{Q}_t\boldsymbol{f}_t - \eta \operatorname{diag}^{-1}(\nabla^2\mathcal{L}(\boldsymbol{Q}_t\boldsymbol{f}_t))\nabla\mathcal{L}(\boldsymbol{Q}_t\boldsymbol{f}_t) - \boldsymbol{f}_\star\right)\right\|_2 \\ &\overset{\text{(iii)}}{\approx} \left\|\left(\boldsymbol{I} - \eta \operatorname{diag}^{-1/2}(\nabla^2\mathcal{L}(\boldsymbol{f}_\star))\nabla^2\mathcal{L}(\boldsymbol{f}_\star)\operatorname{diag}^{-1/2}(\nabla^2\mathcal{L}(\boldsymbol{f}_\star))\right)\operatorname{diag}^{1/2}(\nabla^2\mathcal{L}(\boldsymbol{f}_\star))(\boldsymbol{Q}_t\boldsymbol{f}_t - \boldsymbol{f}_\star)\right\|_2 \\ &\overset{\text{(iv)}}{=} \left\|(\boldsymbol{I} - \eta\boldsymbol{1}\boldsymbol{1}^\top)\operatorname{diag}^{1/2}(\nabla^2\mathcal{L}(\boldsymbol{f}_\star))(\boldsymbol{Q}_t\boldsymbol{f}_t - \boldsymbol{f}_\star)\right\|_2 \end{aligned}$$

where (i) follows from replacing $\boldsymbol{Q}$ by the optimal alignment matrix $\boldsymbol{Q}_t$ between $\boldsymbol{f}_t$ and $\boldsymbol{f}_\star$, (ii) follows from the scaling invariance of the iterates, and (iii) holds approximately as long as $\boldsymbol{Q}_t\boldsymbol{f}_t$ is sufficiently close to $\boldsymbol{f}_\star$, which is made precise in the formal proof. The last line (iv) follows from that the scaled Hessian matrix obeys

$$\operatorname{diag}^{-1/2}(\nabla^2\mathcal{L}(\boldsymbol{f}_\star))\nabla^2\mathcal{L}(\boldsymbol{f}_\star)\operatorname{diag}^{-1/2}(\nabla^2\mathcal{L}(\boldsymbol{f}_\star)) = \boldsymbol{1}\boldsymbol{1}^\top.$$

By the optimality condition for $\boldsymbol{Q}_t$ (see Lemma 7), it follows that $\operatorname{diag}^{1/2}(\nabla^2\mathcal{L}(\boldsymbol{f}_\star))(\boldsymbol{Q}_t\boldsymbol{f}_t - \boldsymbol{f}_\star)$ is approximately parallel to $\boldsymbol{1}$. Thus, $\operatorname{dist}(\boldsymbol{f}_{t+1}, \boldsymbol{f}_\star)$ contracts at a constant rate as long as the step size η is set as a small constant obeying $0 < \eta \leq 2/5$.

3.2 Proof outline for tensor completion (Theorem 1)

Armed with the insights from the tensor factorization case, we now provide a proof outline of our main theorems on tensor completion and tensor regression, both of which can be viewed as perturbations of tensor factorization with incomplete measurements, combined with properly designed initialization schemes. We start with the guarantee for the spectral initialization for tensor completion.

Lemma 1 (Initialization for tensor completion)**.** *Suppose that $\boldsymbol{\mathcal{X}}_\star$ is μ-incoherent, $\mu r^2\kappa^2 \ll n$, and that p satisfies*

$$pn_1n_2n_3 \gtrsim \epsilon_0^{-2}\mu^{3/2}r^2\kappa(\sqrt{r} \vee \kappa)\sqrt{n_1n_2n_3}\log^3 n + \epsilon_0^{-4}\mu^2r^4\kappa^4 n\log^5 n$$

for some small constant $\epsilon_0 > 0$. Then with overwhelming probability, the spectral initialization before projection $\boldsymbol{F}_+ = (\boldsymbol{U}_+, \boldsymbol{V}_+, \boldsymbol{W}_+, \boldsymbol{S}_+)$ for low-rank tensor completion in Algorithm 1 satisfies

$$\operatorname{dist}(\boldsymbol{F}_+, \boldsymbol{F}_\star) \leq \epsilon_0\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star).$$

Under a suitable sample-size condition, Lemma 1 guarantees that $\mathrm{dist}(\boldsymbol{F}_{+},\boldsymbol{F}_{\star}) \le \epsilon_0\sigma_{\min}(\boldsymbol{\mathcal{X}}_{\star})$ for some small constant ϵ_0. To proceed, we need to know what would happen for the spectral estimate $\boldsymbol{F}_0 = \mathcal{P}_B\big(\boldsymbol{F}_{+}\big)$ after projection. In fact, the scaled projection is non-expansive w.r.t. the scaled distance. More importantly, the output is guaranteed to be incoherent. Both properties are stated in the following lemma.

Lemma 2 (Properties of scaled projection)**.** *Suppose that $\boldsymbol{\mathcal{X}}_{\star}$ is μ-incoherent, and* $\mathrm{dist}(\boldsymbol{F}_{+},\boldsymbol{F}_{\star}) \le \epsilon\sigma_{\min}(\boldsymbol{\mathcal{X}}_{\star})$ *for some $\epsilon < 1$. Set $B = C_B\sqrt{\mu r}\sigma_{\max}(\boldsymbol{\mathcal{X}}_{\star})$ for some constant $C_B \ge (1+\epsilon)^3$, then $\boldsymbol{F} = (\boldsymbol{U},\boldsymbol{V},\boldsymbol{W},\boldsymbol{\mathcal{S}}) \coloneqq \mathcal{P}_B(\boldsymbol{F}_{+})$ satisfies the non-expansiveness property*

$$\mathrm{dist}(\boldsymbol{F},\boldsymbol{F}_{\star}) \le \mathrm{dist}(\boldsymbol{F}_{+},\boldsymbol{F}_{\star}) \le \epsilon\sigma_{\min}(\boldsymbol{\mathcal{X}}_{\star}),$$

and the incoherence condition

$$\sqrt{n_1}\|\boldsymbol{U}\breve{\boldsymbol{U}}^\top\|_{2,\infty} \vee \sqrt{n_2}\|\boldsymbol{V}\breve{\boldsymbol{V}}^\top\|_{2,\infty} \vee \sqrt{n_3}\|\boldsymbol{W}\breve{\boldsymbol{W}}^\top\|_{2,\infty} \le C_B\sqrt{\mu r}\sigma_{\max}(\boldsymbol{\mathcal{X}}_{\star}). \tag{29}$$

Now we are ready to state the following lemma that ensures the linear contraction of the iterative refinements given by the ScaledGD updates.

Lemma 3 (Local refinements for tensor completion)**.** *Suppose that $\boldsymbol{\mathcal{X}}_{\star}$ is μ-incoherent, and p satisfies*

$$pn_1n_2n_3 \gtrsim \mu^{3/2}r^2\kappa^3\sqrt{n_1n_2n_3}\log^3 n + \mu^3 r^4\kappa^6 n\log^5 n.$$

Under an event $\mathcal{E}$ which happens with overwhelming probability (i.e. at least $1-c_1n^{-c_2}$), if the t-th iterate satisfies $\mathrm{dist}(\boldsymbol{F}_t,\boldsymbol{F}_{\star}) \le \epsilon\sigma_{\min}(\boldsymbol{\mathcal{X}}_{\star})$ *for some small constant ϵ, then* $\|(\boldsymbol{U}_t,\boldsymbol{V}_t,\boldsymbol{W}_t)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_t - \boldsymbol{\mathcal{X}}_{\star}\|_{\mathsf{F}} \le 3\,\mathrm{dist}(\boldsymbol{F}_t,\boldsymbol{F}_{\star})$. *In addition, if the t-th iterate satisfies the incoherence condition*

$$\sqrt{n_1}\|\boldsymbol{U}_t\breve{\boldsymbol{U}}_t^\top\|_{2,\infty} \vee \sqrt{n_2}\|\boldsymbol{V}_t\breve{\boldsymbol{V}}_t^\top\|_{2,\infty} \vee \sqrt{n_3}\|\boldsymbol{W}_t\breve{\boldsymbol{W}}_t^\top\|_{2,\infty} \le B,$$

with $B = C_B\sqrt{\mu r}\sigma_{\max}(\boldsymbol{\mathcal{X}}_{\star})$ for some constant $C_B \ge (1+\epsilon)^3$, then the $(t+1)$-th iterate of Algorithm 1 satisfies

$$\mathrm{dist}(\boldsymbol{F}_{t+1},\boldsymbol{F}_{\star}) \le (1-0.6\eta)\,\mathrm{dist}(\boldsymbol{F}_t,\boldsymbol{F}_{\star}),$$

and the incoherence condition

$$\sqrt{n_1}\|\boldsymbol{U}_{t+1}\breve{\boldsymbol{U}}_{t+1}^\top\|_{2,\infty} \vee \sqrt{n_2}\|\boldsymbol{V}_{t+1}\breve{\boldsymbol{V}}_{t+1}^\top\|_{2,\infty} \vee \sqrt{n_3}\|\boldsymbol{W}_{t+1}\breve{\boldsymbol{W}}_{t+1}^\top\|_{2,\infty} \le B.$$

By combining Lemma 1 and Lemma 2, we can ensure that the spectral initialization $\boldsymbol{F}_0 = \mathcal{P}_B(\boldsymbol{F}_{+})$ satisfies the conditions required in Lemma 3, which further enables us to repetitively apply Lemma 3 to finish the proof of Theorem 1. The proofs of the above three lemmas are provided in Appendix C.

3.3 Proof outline for tensor regression (Theorem 2)

Now we turn to the proof outline for tensor regression (cf. Theorem 2). To begin with, we show that the local linear convergence of ScaledGD can be guaranteed more generally, as long as the measurement operator $\mathcal{A}(\cdot)$ satisfies the so-called tensor restricted isometry property (TRIP), which is formally defined as follows.

Definition 3 (TRIP [RSS17])**.** The linear map $\mathcal{A}:\mathbb{R}^{n_1\times n_2\times n_3}\mapsto\mathbb{R}^m$ is said to obey the rank-$\boldsymbol{r}$ TRIP with $\delta_{\boldsymbol{r}}\in(0,1)$, if for all tensor $\boldsymbol{\mathcal{X}}\in\mathbb{R}^{n_1\times n_2\times n_3}$ of multilinear rank at most $\boldsymbol{r}=(r_1,r_2,r_3)$, one has

$$(1-\delta_{\boldsymbol{r}})\|\boldsymbol{\mathcal{X}}\|_{\mathsf{F}}^2 \le \|\mathcal{A}(\boldsymbol{\mathcal{X}})\|_{\mathsf{F}}^2 \le (1+\delta_{\boldsymbol{r}})\|\boldsymbol{\mathcal{X}}\|_{\mathsf{F}}^2.$$

If $\mathcal{A}(\cdot)$ satisfies rank-$2\boldsymbol{r}$ TRIP with $\delta_{2\boldsymbol{r}}\in(0,1)$, then for any two tensors $\boldsymbol{\mathcal{X}}_1,\boldsymbol{\mathcal{X}}_2\in\mathbb{R}^{n_1\times n_2\times n_3}$ of multilinear rank at most $\boldsymbol{r}=(r_1,r_2,r_3)$, we have

$$(1-\delta_{2\boldsymbol{r}})\|\boldsymbol{\mathcal{X}}_1-\boldsymbol{\mathcal{X}}_2\|_{\mathsf{F}}^2 \le \|\mathcal{A}(\boldsymbol{\mathcal{X}}_1-\boldsymbol{\mathcal{X}}_2)\|_{\mathsf{F}}^2 \le (1+\delta_{2\boldsymbol{r}})\|\boldsymbol{\mathcal{X}}_1-\boldsymbol{\mathcal{X}}_2\|_{\mathsf{F}}^2.$$

In other words, the distance between any pair of rank-$\boldsymbol{r}$ tensors $\boldsymbol{\mathcal{X}}_1$ and $\boldsymbol{\mathcal{X}}_2$ is approximately preserved after the linear map $\mathcal{A}(\cdot)$. The TRIP has been investigated extensively, where [RSS17, Theorem 1] stated that if $\boldsymbol{\mathcal{A}}_i$'s are composed of i.i.d. sub-Gaussian entries, TRIP holds with high probability provided that $m \gtrsim \delta_{\boldsymbol{r}}^{-2}(nr+r^3)$. TRIP also holds for more structured measurement ensembles such as the random Fourier mapping [RSS17]. With the TRIP of $\mathcal{A}(\cdot)$ in hand, we have the following theorem regarding the local linear convergence of ScaledGD as long as the iterates are close to the ground truth.

Lemma 4 (Local refinements for tensor regression)**.** *Suppose that* $\mathcal{A}(\cdot)$ *obeys the* $2\boldsymbol{r}$*-TRIP with a small constant* $\delta_{2\boldsymbol{r}} \lesssim 1$*. If the* t*-th iterate satisfies* $\operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \le \epsilon\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$ *for some small constant* ϵ*, then* $\|(\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_t - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \le 3 \operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$*. In addition, if the step size obeys* $0 < \eta < 2/5$*, then the* $(t+1)$*-th iterate of Algorithm 2 satisfies*

$$\operatorname{dist}(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) \le (1 - 0.6\eta) \operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star).$$

Therefore, ScaledGD converges linearly as long as the sample size $m \gtrsim nr + r^3$ under the Gaussian design, when initialized properly. Unfortunately, obtaining a desired initialization turns out to be a major roadblock and requires a substantially higher sample size, which has been studied extensively for tensor regression, say, in [LZ21, HWZ20, ZLRY20]. Under the Gaussian design, we have the following guarantee for the spectral initialization scheme that invokes HOSVD in Algorithm 2.

Lemma 5 (Initialization for tensor regression)**.** *Suppose that* $\{\boldsymbol{\mathcal{A}}_i\}_{i=1}^m$ *are composed of i.i.d.* $\mathcal{N}(0, 1/m)$ *entries, and that* m *satisfies*

$$m \gtrsim \epsilon_0^{-2}(\sqrt{n_1 n_2 n_3} r \kappa^2 + n r^2 \kappa^3)$$

for some small constant $\epsilon_0 > 0$*. Then with overwhelming probability, the spectral initialization for low-rank tensor regression in Algorithm 2 satisfies*

$$\operatorname{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \le \epsilon_0 \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star).$$

Combining Lemma 4 and Lemma 5 finishes the proof of Theorem 2. Their proofs can be found in Appendix D.

4 Numerical Experiments

We illustrate the numerical performance of ScaledGD for tensor completion to corroborate our findings, especially its computational advantage over the regularized GD algorithm [HWZ20] that is closest to our design. We remark that similar results can be obtained for tensor regression, but the expensive sensing operator limits the appeal of the experiments and therefore is omitted. Since the scaled projection does not visibly impact the performance, we implement ScaledGD without performing the projection. For simplicity, we set $n_1 = n_2 = n_3 = n$, and $r_1 = r_2 = r_3 = r$. Each entry of the tensor is observed i.i.d. with probability $p \in (0, 1]$.

Phase transition of ScaledGD. We construct the ground truth tensor $\boldsymbol{\mathcal{X}}_\star = (\boldsymbol{U}_\star, \boldsymbol{V}_\star, \boldsymbol{W}_\star) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_\star$ by generating $\boldsymbol{U}_\star$, $\boldsymbol{V}_\star$ and $\boldsymbol{W}_\star$ as random orthonormal matrices, and the core tensor $\boldsymbol{\mathcal{S}}_\star$ composed of i.i.d. standard Gaussian entries, i.e. $\boldsymbol{\mathcal{S}}_\star(j_1, j_2, j_3) \sim \mathcal{N}(0, 1)$ for $1 \le j_k \le r$, $k = 1, 2, 3$. For each set of parameters, we run 100 random tests and count the success rate, where the recovery is regarded as successful if the recovered tensor has a relative error $\|\boldsymbol{\mathcal{X}}_T - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} / \|\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \le 10^{-3}$. Figure 2 illustrates the success rate with respect to the (scaled) sample size for different tensor sizes n, which implies that the recovery is successful when the sample size is moderately large.

Comparison with regularized GD. We next compare the performance of ScaledGD with the regularized GD algorithm proposed in [HWZ20], which aims to minimize the loss function in (3). Since [HWZ20] adopts a different scaling of the factors from ours, we rescale it first so that both ScaledGD and regularized GD start from the same spectral initialization. Following the choice of parameters in [HWZ20], we obtain the equivalent update rule of regularized GD as (see Appendix A.3 for additional discussions)

$$\begin{aligned}
\boldsymbol{U}_{t+1} &= \boldsymbol{U}_t - \eta \left[\sigma_{\max}^{-2}(\boldsymbol{\mathcal{X}}_\star) \nabla_{\boldsymbol{U}} \mathcal{L}(\boldsymbol{F}_t) + \boldsymbol{U}_t(\boldsymbol{U}_t^\top \boldsymbol{U}_t - \boldsymbol{I}_{r_1})\right], \\
\boldsymbol{V}_{t+1} &= \boldsymbol{V}_t - \eta \left[\sigma_{\max}^{-2}(\boldsymbol{\mathcal{X}}_\star) \nabla_{\boldsymbol{V}} \mathcal{L}(\boldsymbol{F}_t) + \boldsymbol{V}_t(\boldsymbol{V}_t^\top \boldsymbol{V}_t - \boldsymbol{I}_{r_2})\right], \\
\boldsymbol{W}_{t+1} &= \boldsymbol{W}_t - \eta \left[\sigma_{\max}^{-2}(\boldsymbol{\mathcal{X}}_\star) \nabla_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{F}_t) + \boldsymbol{W}_t(\boldsymbol{W}_t^\top \boldsymbol{W}_t - \boldsymbol{I}_{r_3})\right], \\
\boldsymbol{\mathcal{S}}_{t+1} &= \boldsymbol{\mathcal{S}}_t - \eta \nabla_{\boldsymbol{\mathcal{S}}} \mathcal{L}(\boldsymbol{F}_t).
\end{aligned} \tag{30}$$

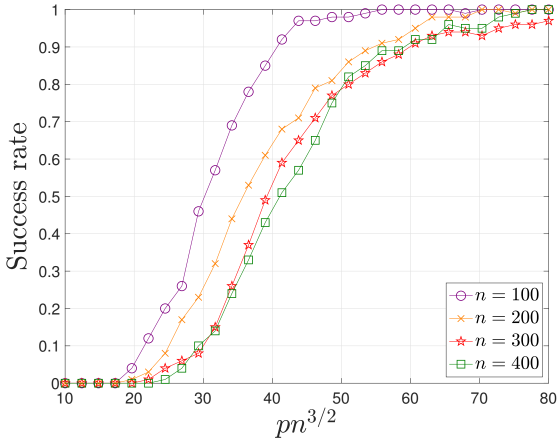


Figure 2: The success rate of ScaledGD with respect to the scaled sample size for tensor completion with $r = 5$, when the core tensor is composed of i.i.d. standard Gaussian entries, for various tensor size n.

Throughout the experiments, we used the ground truth value $\sigma_{\max}(\boldsymbol{\mathcal{X}}_\star)$ in running (30), while in practice, this parameter needs to estimated; to put it differently, the step size of regularized GD is not *scale-invariant*, whereas the step size of ScaledGD is.

To ensure the ground truth tensor $\boldsymbol{\mathcal{X}}_\star = (\boldsymbol{U}_\star, \boldsymbol{V}_\star, \boldsymbol{W}_\star) \cdot \boldsymbol{\mathcal{S}}_\star$ has a prescribed condition number κ, we generate the core tensor $\boldsymbol{\mathcal{S}}_\star \in \mathbb{R}^{r\times r\times r}$ according to $\boldsymbol{\mathcal{S}}_\star(j_1, j_2, j_3) = \sigma_{j_1}/\sqrt{r}$ if $j_1 + j_2 + j_3 \equiv 0 \pmod r$ and 0 otherwise, where $\{\sigma_{j_1}\}_{1\leq j_1\leq r}$ take values that are spaced equally from 1 to κ^{-1}. It then follows that $\sigma_{\max}(\boldsymbol{\mathcal{X}}_\star) = 1$, $\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star) = \kappa^{-1}$, and the condition number of $\boldsymbol{\mathcal{X}}_\star$ is exactly κ. Figure 3 illustrates the convergence speed of ScaledGD and regularized GD under different step sizes, where we plot the relative error after at most 80 iterations (the algorithm is terminated if the relative error exceeds 10^2 following an excessive step size). It can be seen that ScaledGD outperforms regularized GD quite significantly even when the step size of regularized GD is optimized for its performance. Hence, we will fix $\eta = 0.3$ for the rest of the comparisons for both ScaledGD and regularized GD without hurting the conclusions.

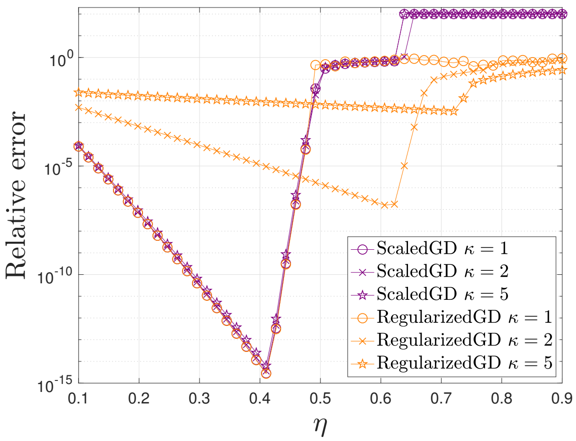
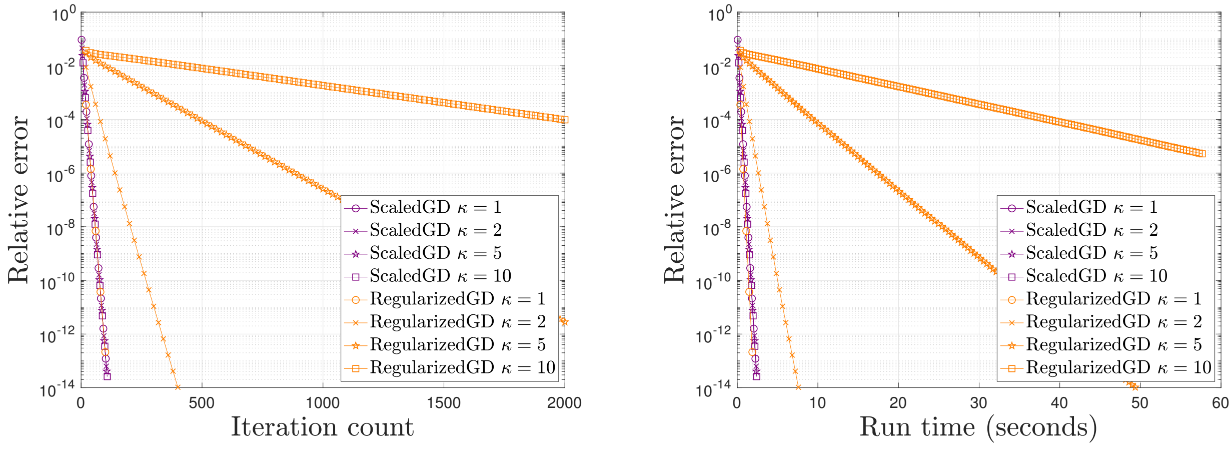


Figure 3: The relative errors of ScaledGD and regularized GD after 80 iterations with respect to different step sizes η from 0.1 to 0.9 for tensor completion with $n = 100$, $r = 5$, $p = 0.1$.

Figure 4 compares the relative errors of ScaledGD and regularized GD for tensor completion with respect to the iteration count and run time (in seconds) under different condition numbers $\kappa = 1, 2, 5, 10$. This experiment verifies that ScaledGD converges rapidly at a rate independent of the condition number, and



(a)

(b)

Figure 4: The relative errors of ScaledGD and regularized GD with respect to (a) the iteration count and (b) run time (in seconds) under different condition numbers $\kappa = 1, 2, 5, 10$ for tensor completion with $n = 100$, $r = 5$, and $p = 0.1$.

matches the fastest rate of regularized GD with perfect conditioning $\kappa = 1$. In contrast, the convergence rate of regularized GD deteriorates quickly with the increase of κ even at a moderate level. The advantage of ScaledGD carries over to the run time as well, since the scaled gradient only adds a negligible overhead to the gradient computation.

We next examine the performance of ScaledGD and regularized GD when randomly initialized. Here, we initialize $\boldsymbol{U}_0, \boldsymbol{V}_0, \boldsymbol{W}_0$ composed of i.i.d. Gaussian entries sampled from $\mathcal{N}(0, 1/n)$, and $\boldsymbol{\mathcal{S}}_0$ composed of i.i.d. Gaussian entries sampled from $\mathcal{N}(0, \|\boldsymbol{\mathcal{Y}}\|_{\mathsf{F}}^2/(pn^3))$. Figure 5 plots the relative errors of ScaledGD and regularized GD under different condition numbers $\kappa = 1, 2, 5$, using the same random initialization. Surprisingly, while regularized GD stuck in a flat region before entering the phase of linear convergence, ScaledGD seems to be quite insensitive to the choice of initialization, and converges almost in the same fashion as the case with spectral initialization.

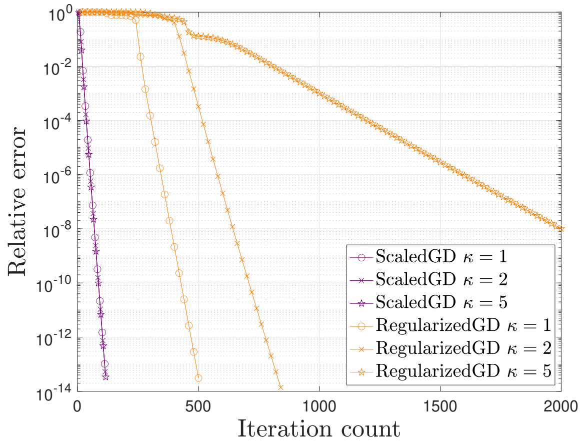


Figure 5: The relative errors of random-initialized ScaledGD and regularized GD with respect to the iteration count under different condition numbers $\kappa = 1, 2, 5$ for tensor completion with $n = 100$, $r = 5$, $p = 0.1$.

Finally, we examine the performance of ScaledGD when the observations are corrupted by additive noise, where we assume the noisy observations are given by $\boldsymbol{\mathcal{Y}} = \mathcal{P}_\Omega(\boldsymbol{\mathcal{X}}_\star + \boldsymbol{\mathcal{W}})$, with $\boldsymbol{\mathcal{W}}(i_1, i_2, i_3) \sim \mathcal{N}(0, \sigma_w^2)$

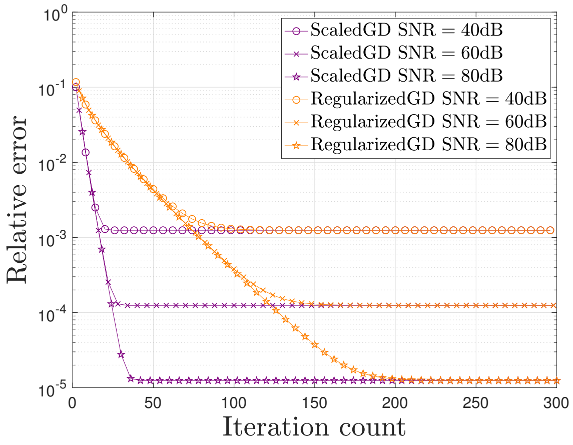


Figure 6: The relative errors of ScaledGD and regularized GD with respect to the iteration count under signal-to-noise ratios SNR = 40, 60, 80dB for tensor completion with $n = 100$, $r = 5$, and $p = 0.1$.

composed of i.i.d. Gaussian entries. Denote the signal-to-noise ratio as $\mathsf{SNR} := 10\log_{10}\frac{\|\boldsymbol{\mathcal{X}}_\star\|_\mathsf{F}^2}{n^3\sigma_w^2}$ in dB. Figure 6 demonstrates the robustness of ScaledGD, by plotting the relative errors with respect to the iteration count under SNR = 40, 60, 80dB. Here, the ground truth tensor $\boldsymbol{\mathcal{X}}_\star$ is constructed in the same manner as Figure 2, where its condition number is approximately $\kappa \approx 2.5$. It can been seen that ScaledGD reaches the same statistical error as regularized GD, but at a much faster rate. In addition, the convergence speeds are not impacted by the noise levels.

5 Discussions

This paper develops a scaled gradient descent algorithm over the factor space for low-rank tensor estimation (i.e. completion and regression) with provable sample and computational guarantees, leading to a highly scalable approach especially when the ground truth tensor is ill-conditioned and high-dimensional. There are several future directions that are worth exploring, which we briefly discuss below.

- *Preconditioning for other tensor decompositions.* The use of preconditioning will likely also accelerate vanilla gradient descent for low-rank tensor estimation using other decomposition models, such as CP decomposition [CLPC19], which is worth investigating.
- *Entrywise error control for tensor completion.* In this paper, we focused on controlling the ℓ_2 error of the reconstructed tensor in tensor completion, whereas another strong form of statistical guarantees deals with the ℓ_∞ error, as done in [MWCC20] for matrix completion and in [CLPC19] for tensor completion with CP decomposition. It is hence of interest to develop similar strong entrywise error guarantees of ScaledGD for tensor completion with Tucker decomposition.
- *Stable and robust low-rank tensor estimation.* In practice, the observations are corrupted by noise and even outliers [LCZL20], therefore, it is necessary to examine the stability and robustness of ScaledGD in more depths, such as by pinning down the statistical error rates and extending the scaled subgradient algorithm in [TMC21] to the tensor case.
- *Random initialization?* As evident from the numerical experiment in Figure 5, ScaledGD works remarkably well even from a random initialization, which requires us to go beyond the local geometry and pursue a further understanding of the global landscape of the optimization geometry.

Acknowledgements

The work of T. Tong and Y. Chi is supported in part by Office of Naval Research under N00014-18-1-2142 and N00014-19-1-2404, by Army Research Office under W911NF-18-1-0303, by Air Force Research Laboratory under FA8750-20-2-0504, and by National Science Foundation under CAREER ECCS-1818571, CCF-1806154 and CCF-1901199. 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.

[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.

[BGW20] S. Buchanan, D. Gilboa, and J. Wright. Deep networks and the multiple manifold problem. *arXiv preprint arXiv:2008.11245*, 2020.

[BJS18] Y. Bai, Q. Jiang, and J. Sun. Subgradient descent learns orthogonal dictionaries. *arXiv preprint arXiv:1810.10702*, 2018.

[BM16] B. Barak and A. Moitra. Noisy tensor completion via the sum-of-squares hierarchy. In *Conference on Learning Theory*, pages 417–445. PMLR, 2016.

[CC17] Y. Chen and E. Candès. Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Communications on Pure and Applied Mathematics*, 70(5):822–883, 2017.

[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.

[CCD$^+$21] V. Charisopoulos, Y. Chen, D. Davis, M. Díaz, L. Ding, and D. Drusvyatskiy. Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. *Foundations of Computational Mathematics*, pages 1–89, 2021.

[CCFM19] Y. Chen, Y. Chi, J. Fan, and C. Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176(1-2):5–37, 2019.

[CCFM20] Y. Chen, Y. Chi, J. Fan, and C. Ma. Spectral methods for data science: A statistical perspective. *arXiv preprint arXiv:2012.08496*, 2020.

[CL19] J. Chen and X. Li. Model-free nonconvex matrix completion: Local minima analysis and applications in memory-efficient kernel PCA. *Journal of Machine Learning Research*, 20(142):1–39, 2019.

[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.

[CLC$^+$21] C. Cai, G. Li, Y. Chi, H. V. Poor, and Y. Chen. Subspace estimation from unbalanced and incomplete data matrices: $\ell_{2,\infty}$ statistical guarantees. *The Annals of Statistics*, 49(2):944–967, 2021.

[CLL20] J. Chen, D. Liu, and X. Li. Nonconvex rectangular matrix completion via gradient descent without $\ell_{2,\infty}$ regularization. *IEEE Transactions on Information Theory*, 66(9):5806–5841, 2020.

[CLPC19] C. Cai, G. Li, H. V. Poor, and Y. Chen. Nonconvex low-rank tensor completion from noisy data. In *Advances in Neural Information Processing Systems*, pages 1863–1874, 2019.

[CLS15] E. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *Information Theory, IEEE Transactions on*, 61(4):1985–2007, 2015.

[CMPC20] Y. Chen, C. Ma, H. V. Poor, and Y. Chen. Learning mixtures of low-rank models. *arXiv preprint arXiv:2009.11282*, 2020.

[CPC20] C. Cai, H. V. Poor, and Y. Chen. Uncertainty quantification for nonconvex tensor completion: Confidence intervals, heteroscedasticity and optimality. In *International Conference on Machine Learning*, pages 1271–1282. PMLR, 2020.

[CRY19] H. Chen, G. Raskutti, and M. Yuan. Non-convex projected gradient descent for generalized low-rank tensor regression. *The Journal of Machine Learning Research*, 20(1):172–208, 2019.

[CT10] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[CW15] Y. Chen and M. J. Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

[DFL17] R. Dian, L. Fang, and S. Li. Hyperspectral image super-resolution via non-local sparse tensor factorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5344–5353, 2017.

[FCL20] H. Fu, Y. Chi, and Y. Liang. Guaranteed recovery of one-hidden-layer neural networks via cross entropy. *IEEE transactions on signal processing*, 68:3225–3235, 2020.

[FG20] A. Frandsen and R. Ge. Optimization landscape of Tucker decomposition. *Mathematical Programming*, pages 1–26, 2020.

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

[GM20] R. Ge and T. Ma. On the optimization landscape of tensor decompositions. *Mathematical Programming*, pages 1–47, 2020.

[GPY19] N. Ghadermarzy, Y. Plan, and Ö. Yilmaz. Near-optimal sample complexity for convex tensor completion. *Information and Inference: A Journal of the IMA*, 8(3):577–619, 2019.

[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.

[GRY11] S. Gandy, B. Recht, and I. Yamada. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 27(2):025010, 2011.

[Hac12] W. Hackbusch. *Tensor spaces and numerical tensor calculus*, volume 42. Springer, 2012.

[HL13] C. J. Hillar and L.-H. Lim. Most tensor problems are NP-hard. *Journal of the ACM (JACM)*, 60(6):1–39, 2013.

[HMGW15] B. Huang, C. Mu, D. Goldfarb, and J. Wright. Provable models for robust low-rank tensor completion. *Pacific Journal of Optimization*, 11(2):339–364, 2015.

[HV19] P. Hand and V. Voroninski. Global guarantees for enforcing deep generative priors by empirical risk. *IEEE Transactions on Information Theory*, 66(1):401–418, 2019.

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

[HZC20] B. Hao, A. Zhang, and G. Cheng. Sparse and low-rank tensor estimation via cubic sketchings. *IEEE Transactions on Information Theory*, 66(9):5927–5964, 2020.

[JO14] P. Jain and S. Oh. Provable tensor factorization with missing data. *Advances in Neural Information Processing Systems*, 2:1431–1439, 2014.

[JYZ17] B. Jiang, F. Yang, and S. Zhang. Tensor and its Tucker core: the invariance relationships. *Numerical Linear Algebra with Applications*, 24(3):e2086, 2017.

[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.

[KM16] H. Kasai and B. Mishra. Low-rank tensor completion: a Riemannian manifold preconditioning approach. In *International Conference on Machine Learning*, pages 1012–1021, 2016.

[KS13] A. Krishnamurthy and A. Singh. Low-rank matrix and tensor completion via adaptive sampling. *Advances in Neural Information Processing Systems*, 26:836–844, 2013.

[KSV14] D. Kressner, M. Steinlechner, and B. Vandereycken. Low-rank tensor completion by Riemannian optimization. *BIT Numerical Mathematics*, 54(2):447–468, 2014.

[LAAW19] X.-Y. Liu, S. Aeron, V. Aggarwal, and X. Wang. Low-tubal-rank tensor completion using alternating minimization. *IEEE Transactions on Information Theory*, 66(3):1714–1737, 2019.

[LCZL20] Y. Li, Y. Chi, H. Zhang, and Y. Liang. Non-convex low-rank matrix recovery with arbitrary outliers via median-truncated gradient descent. *Information and Inference: A Journal of the IMA*, 9(2):289–325, 2020.

[LFLY18] C. Lu, J. Feng, Z. Lin, and S. Yan. Exact low tubal rank tensor recovery from Gaussian measurements. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 2504–2510, 2018.

[LLSW19] X. Li, S. Ling, T. Strohmer, and K. Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *Applied and computational harmonic analysis*, 47(3):893–934, 2019.

[LM20] A. Liu and A. Moitra. Tensor completion made practical. *Advances in Neural Information Processing Systems*, 33, 2020.

[LMCC19] Y. Li, C. Ma, Y. Chen, and Y. Chi. Nonconvex matrix factorization from rank-one measurements. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1496–1505, 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.

[LNSU18] Z. Li, Y. Nakatsukasa, T. Soma, and A. Uschmajew. On orthogonal tensors and best rank-one approximation ratio. *SIAM Journal on Matrix Analysis and Applications*, 39(1):400–425, 2018.

[LPST15] Q. Li, A. Prater, L. Shen, and G. Tang. Overcomplete tensor decomposition via convex optimization. In *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 53–56. 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.

[MHWG14] C. Mu, B. Huang, J. Wright, and D. Goldfarb. Square deal: Lower bounds and improved relaxations for tensor recovery. In *International conference on machine learning*, pages 73–81. PMLR, 2014.

[MLC21] C. Ma, Y. Li, and Y. Chi. Beyond Procrustes: Balancing-free gradient descent for asymmetric low-rank matrix sensing. *IEEE Transactions on Signal Processing*, 69:867–877, 2021.

[MS18] A. Montanari and N. Sun. Spectral algorithms for tensor completion. *Communications on Pure and Applied Mathematics*, 71(11):2381–2425, 2018.

[MWCC20] C. Ma, K. Wang, Y. Chi, and Y. Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Foundations of Computational Mathematics*, 20:451–632, 2020.

[Paa00] P. Paatero. Construction and analysis of degenerate PARAFAC models. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 14(3):285–299, 2000.

[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.

[PKCS17] D. Park, A. Kyrillidis, C. Carmanis, and S. Sanghavi. Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach. In *Artificial Intelligence and Statistics*, pages 65–74, 2017.

[PS17] A. Potechin and D. Steurer. Exact tensor completion with sum-of-squares. In *Conference on Learning Theory*, pages 1619–1673. PMLR, 2017.

[RSS17] H. Rauhut, R. Schneider, and Ž. Stojanac. Low rank tensor recovery via iterative hard thresholding. *Linear Algebra and its Applications*, 523:220–262, 2017.

[RYC19] G. Raskutti, M. Yuan, and H. Chen. Convex regularization for high-dimensional multiresponse tensor regression. *The Annals of Statistics*, 47(3):1554–1584, 2019.

[SC19] L. Shi and Y. Chi. Manifold gradient descent solves multi-channel sparse blind deconvolution provably and efficiently. *arXiv preprint arXiv:1911.11167*, 2019.

[SDLF$^+$17] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 65(13):3551–3582, 2017.

[SL16] R. Sun and Z.-Q. Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

[SQW17a] J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere I: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2017.

[SQW17b] J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere II: Recovery by Riemannian trust-region method. *IEEE Transactions on Information Theory*, 63(2):885–914, 2017.

[TMC20] T. Tong, C. Ma, and Y. Chi. Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *arXiv preprint arXiv:2005.08898*, 2020.

[TMC21] 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*, 2021.

[Tuc66] L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

[WGE18] G. Wang, G. B. Giannakis, and Y. C. Eldar. Solving systems of random quadratic equations via truncated amplitude flow. *IEEE Transactions on Information Theory*, 64(2):773–794, 2018.

[XCH+10] L. Xiong, X. Chen, T.-K. Huang, J. Schneider, and J. G. Carbonell. Temporal collaborative filtering with bayesian probabilistic tensor factorization. In *Proceedings of the 2010 SIAM international conference on data mining*, pages 211–222. SIAM, 2010.

[XY13] Y. Xu and W. Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on imaging sciences*, 6(3):1758–1789, 2013.

[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.

[XYZ17] D. Xia, M. Yuan, and C.-H. Zhang. Statistically optimal and computationally efficient low rank tensor completion from noisy entries. *arXiv preprint arXiv:1711.04934*, 2017.

[XZZ20] D. Xia, A. R. Zhang, and Y. Zhou. Inference for low-rank tensors–no need to debias. *arXiv preprint arXiv:2012.14844*, 2020.

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

[ZA16] Z. Zhang and S. Aeron. Exact tensor completion using t-SVD. *IEEE Transactions on Signal Processing*, 65(6):1511–1526, 2016.

[ZCL16] H. Zhang, Y. Chi, and Y. Liang. Provable non-convex phase retrieval with outliers: Median truncated Wirtinger flow. In *International conference on machine learning*, pages 1022–1031, 2016.

[ZEA+14] Z. Zhang, G. Ely, S. Aeron, N. Hao, and M. Kilmer. Novel methods for multilinear data completion and de-noising based on tensor-SVD. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3842–3849, 2014.

[Zha19] A. Zhang. Cross: Efficient low-rank tensor completion. *Annals of Statistics*, 47(2):936–964, 2019.

[ZLRY20] A. Zhang, Y. Luo, G. Raskutti, and M. Yuan. ISLET: Fast and optimal low-rank tensor regression via importance sketching. *SIAM Journal on Mathematics of Data Science*, 2(2):444–479, 2020.

[ZLZ13] H. Zhou, L. Li, and H. Zhu. Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502):540–552, 2013.

[ZQW20] Y. Zhang, Q. Qu, and J. Wright. From symmetry to geometry: Tractable nonconvex problems. *arXiv preprint arXiv:2007.06753*, 2020.

[ZX18] A. Zhang and D. Xia. Tensor SVD: Statistical and computational limits. *IEEE Transactions on Information Theory*, 64(11):7311–7338, 2018.

[ZZLC17] H. Zhang, Y. Zhou, Y. Liang, and Y. Chi. A nonconvex approach for phase retrieval: Reshaped Wirtinger flow and incremental algorithms. *Journal of Machine Learning Research*, 18(141):1–35, 2017.

A Preliminaries

This section gathers several technical lemmas that will be used later in the proof. More specifically, Section A.1 is devoted to understanding the scaled distance defined in the equation (24), and in Section A.2, we derive several useful perturbation bounds related to the tensor factors and the tensor itself. All the proofs are collected in the end of each subsection.

A.1 Understanding the scaled distance

To begin, recall the scaled distance between $\boldsymbol{F} = (\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}, \boldsymbol{\mathcal{S}})$ and $\boldsymbol{F}_\star = (\boldsymbol{U}_\star, \boldsymbol{V}_\star, \boldsymbol{W}_\star, \boldsymbol{\mathcal{S}}_\star)$:

$$\begin{aligned}\mathrm{dist}^2(\boldsymbol{F}, \boldsymbol{F}_\star) \coloneqq \inf_{\boldsymbol{Q}_k \in \mathrm{GL}(r_k)} \; & \|(\boldsymbol{U}\boldsymbol{Q}_1 - \boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 + \|(\boldsymbol{V}\boldsymbol{Q}_2 - \boldsymbol{V}_\star)\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}^2 + \|(\boldsymbol{W}\boldsymbol{Q}_3 - \boldsymbol{W}_\star)\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}^2 \\ & + \left\|(\boldsymbol{Q}_1^{-1}, \boldsymbol{Q}_2^{-1}, \boldsymbol{Q}_3^{-1}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{S}}_\star\right\|_{\mathsf{F}}^2, \end{aligned} \tag{31}$$

where we call the matrices $\{\boldsymbol{Q}_k\}_{k=1,2,3}$ (if exist) that attain the infimum the optimal alignment matrices between $\boldsymbol{F}$ and $\boldsymbol{F}_\star$; in particular, $\boldsymbol{F}$ and $\boldsymbol{F}_\star$ are said to be aligned if the optimal alignment matrices are identity matrices.

In what follows, we provide several useful lemmas whose proof can be found at the end of this subsection. We start with a lemma that ensures the attainability of the infimum in the definition (31) as long as $\mathrm{dist}(\boldsymbol{F}, \boldsymbol{F}_\star)$ is sufficiently small.

Lemma 6. *Fix any factor quadruple $\boldsymbol{F} = (\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}, \boldsymbol{\mathcal{S}})$. Suppose that $\mathrm{dist}(\boldsymbol{F}, \boldsymbol{F}_\star) < \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$, then the infimum of (31) is attained at some $\boldsymbol{Q}_k \in \mathrm{GL}(r_k)$, i.e., the alignment matrices between $\boldsymbol{F}$ and $\boldsymbol{F}_\star$ exist.*

With the existence of the optimal alignment matrices in place, the following lemma delineates the optimality conditions they need to satisfy.

Lemma 7. *The optimal alignment matrices $\{\boldsymbol{Q}_k\}_{k=1,2,3}$ between $\boldsymbol{F}$ and $\boldsymbol{F}_\star$, if exist, must satisfy*

$$\begin{aligned}
(\boldsymbol{U}\boldsymbol{Q}_1)^\top(\boldsymbol{U}\boldsymbol{Q}_1 - \boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1}^2 &= \mathcal{M}_1\left((\boldsymbol{Q}_1^{-1}, \boldsymbol{Q}_2^{-1}, \boldsymbol{Q}_3^{-1}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{S}}_\star\right) \mathcal{M}_1\left((\boldsymbol{Q}_1^{-1}, \boldsymbol{Q}_2^{-1}, \boldsymbol{Q}_3^{-1}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}\right)^\top, \\
(\boldsymbol{V}\boldsymbol{Q}_2)^\top(\boldsymbol{V}\boldsymbol{Q}_2 - \boldsymbol{V}_\star)\boldsymbol{\Sigma}_{\star,2}^2 &= \mathcal{M}_2\left((\boldsymbol{Q}_1^{-1}, \boldsymbol{Q}_2^{-1}, \boldsymbol{Q}_3^{-1}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{S}}_\star\right) \mathcal{M}_2\left((\boldsymbol{Q}_1^{-1}, \boldsymbol{Q}_2^{-1}, \boldsymbol{Q}_3^{-1}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}\right)^\top, \\
(\boldsymbol{W}\boldsymbol{Q}_3)^\top(\boldsymbol{W}\boldsymbol{Q}_3 - \boldsymbol{W}_\star)\boldsymbol{\Sigma}_{\star,3}^2 &= \mathcal{M}_3\left((\boldsymbol{Q}_1^{-1}, \boldsymbol{Q}_2^{-1}, \boldsymbol{Q}_3^{-1}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{S}}_\star\right) \mathcal{M}_3\left((\boldsymbol{Q}_1^{-1}, \boldsymbol{Q}_2^{-1}, \boldsymbol{Q}_3^{-1}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}\right)^\top.
\end{aligned}$$

The next lemma relates the scaled distance between the factors to the Euclidean distance between the tensors.

Lemma 8. *For any factor quadruple $\boldsymbol{F} = (\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}, \boldsymbol{\mathcal{S}})$, the scaled distance (31) satisfies*

$$\mathrm{dist}(\boldsymbol{F}, \boldsymbol{F}_\star) \le (\sqrt{2}+1)^{3/2} \left\|(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_\star\right\|_{\mathsf{F}}.$$

A.1.1 Proof of Lemma 6

This proof mimics that of [TMC20, Lemma 9]. The high level idea is to translate the optimization problem (31) into an equivalent continuous optimization problem over a *compact* set. Then an application of the Weierstrass extreme value theorem ensures the existence of the minimizer.

Under the condition $\mathrm{dist}(\boldsymbol{F}, \boldsymbol{F}_\star) < \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$, one knows that there exist matrices $\bar{\boldsymbol{Q}}_k \in \mathrm{GL}(r_k)$ such that

$$\begin{aligned}\Big(& \left\|(\boldsymbol{U}\bar{\boldsymbol{Q}}_1 - \boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{V}\bar{\boldsymbol{Q}}_2 - \boldsymbol{V}_\star)\boldsymbol{\Sigma}_{\star,2}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{W}\bar{\boldsymbol{Q}}_3 - \boldsymbol{W}_\star)\boldsymbol{\Sigma}_{\star,3}\right\|_{\mathsf{F}}^2 \\ & + \left\|(\bar{\boldsymbol{Q}}_1^{-1}, \bar{\boldsymbol{Q}}_2^{-1}, \bar{\boldsymbol{Q}}_3^{-1}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{S}}_\star\right\|_{\mathsf{F}}^2\Big)^{1/2} \le \epsilon\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star),\end{aligned}$$

for some ϵ obeying $0 < \epsilon < 1$. The above relation further implies that

$$\begin{aligned}\left\|\boldsymbol{U}\bar{\boldsymbol{Q}}_1 - \boldsymbol{U}_\star\right\| \vee \left\|\boldsymbol{V}\bar{\boldsymbol{Q}}_2 - \boldsymbol{V}_\star\right\| \vee \left\|\boldsymbol{W}\bar{\boldsymbol{Q}}_3 - \boldsymbol{W}_\star\right\| \le \epsilon, \qquad \text{and} \\ \sigma_{\max}\left((\bar{\boldsymbol{Q}}_1^{-1}, \bar{\boldsymbol{Q}}_2^{-1}, \bar{\boldsymbol{Q}}_3^{-1}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{S}}_\star\right) \le \epsilon\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star).\end{aligned}$$

Invoke Weyl's inequality, and use the fact that $\boldsymbol{U}_\star, \boldsymbol{V}_\star, \boldsymbol{W}_\star$ are orthonormal to obtain

$$\sigma_{\min}(\boldsymbol{U}\bar{\boldsymbol{Q}}_1) \wedge \sigma_{\min}(\boldsymbol{V}\bar{\boldsymbol{Q}}_2) \wedge \sigma_{\min}(\boldsymbol{W}\bar{\boldsymbol{Q}}_3) \ge 1 - \epsilon, \text{ and } \sigma_{\min}\left((\bar{\boldsymbol{Q}}_1^{-1}, \bar{\boldsymbol{Q}}_2^{-1}, \bar{\boldsymbol{Q}}_3^{-1}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}\right) \ge (1-\epsilon)\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star). \tag{32}$$

In addition, it is straightforward to see that the minimization problem on the right-hand side of (31) is equivalent to

$$\inf_{\boldsymbol{H}_k \in \mathrm{GL}(r_k)} \left\|(\boldsymbol{U}\bar{\boldsymbol{Q}}_1\boldsymbol{H}_1 - \boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{V}\bar{\boldsymbol{Q}}_2\boldsymbol{H}_2 - \boldsymbol{V}_\star)\boldsymbol{\Sigma}_{\star,2}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{W}\bar{\boldsymbol{Q}}_3\boldsymbol{H}_3 - \boldsymbol{W}_\star)\boldsymbol{\Sigma}_{\star,3}\right\|_{\mathsf{F}}^2 + \left\|\left(\boldsymbol{H}_1^{-1}\bar{\boldsymbol{Q}}_1^{-1}, \boldsymbol{H}_2^{-1}\bar{\boldsymbol{Q}}_2^{-1}, \boldsymbol{H}_3^{-1}\bar{\boldsymbol{Q}}_3^{-1}\right)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{S}}_\star\right\|_{\mathsf{F}}^2. \quad (33)$$

Therefore, it suffices to establish the infimum is attainable for the above problem instead. By the optimality of $\bar{\boldsymbol{Q}}_k\boldsymbol{H}_k$ over $\bar{\boldsymbol{Q}}_k$, to yield a smaller distance than $\bar{\boldsymbol{Q}}_k$, $\boldsymbol{H}_k$ must obey

$$\Big(\left\|(\boldsymbol{U}\bar{\boldsymbol{Q}}_1\boldsymbol{H}_1 - \boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{V}\bar{\boldsymbol{Q}}_2\boldsymbol{H}_2 - \boldsymbol{V}_\star)\boldsymbol{\Sigma}_{\star,2}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{W}\bar{\boldsymbol{Q}}_3\boldsymbol{H}_3 - \boldsymbol{W}_\star)\boldsymbol{\Sigma}_{\star,3}\right\|_{\mathsf{F}}^2 + \left\|\left(\boldsymbol{H}_1^{-1}\bar{\boldsymbol{Q}}_1^{-1}, \boldsymbol{H}_2^{-1}\bar{\boldsymbol{Q}}_2^{-1}, \boldsymbol{H}_3^{-1}\bar{\boldsymbol{Q}}_3^{-1}\right)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{S}}_\star\right\|_{\mathsf{F}}^2\Big)^{1/2} \le \epsilon\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star).$$

Follow similar reasonings as earlier and invoke Weyl's inequality again to obtain

$$\begin{aligned} \sigma_{\max}(\boldsymbol{U}\bar{\boldsymbol{Q}}_1\boldsymbol{H}_1) \vee \sigma_{\max}(\boldsymbol{V}\bar{\boldsymbol{Q}}_2\boldsymbol{H}_2) \vee \sigma_{\max}(\boldsymbol{W}\bar{\boldsymbol{Q}}_3\boldsymbol{H}_3) &\le 1+\epsilon, \qquad \text{and} \\ \sigma_{\max}\left(\left(\boldsymbol{H}_1^{-1}\bar{\boldsymbol{Q}}_1^{-1}, \boldsymbol{H}_2^{-1}\bar{\boldsymbol{Q}}_2^{-1}, \boldsymbol{H}_3^{-1}\bar{\boldsymbol{Q}}_3^{-1}\right)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}\right) &\le (1+\epsilon)\sigma_{\max}(\boldsymbol{\mathcal{X}}_\star). \end{aligned}$$

Use the relation $\sigma_{\min}(\boldsymbol{A})\sigma_{\max}(\boldsymbol{B}) \le \sigma_{\max}(\boldsymbol{A}\boldsymbol{B})$ to further obtain

$$\begin{aligned} \sigma_{\min}(\boldsymbol{U}\bar{\boldsymbol{Q}}_1)\sigma_{\max}(\boldsymbol{H}_1) \vee \sigma_{\min}(\boldsymbol{V}\bar{\boldsymbol{Q}}_2)\sigma_{\max}(\boldsymbol{H}_2) \vee \sigma_{\min}(\boldsymbol{W}\bar{\boldsymbol{Q}}_3)\sigma_{\max}(\boldsymbol{H}_3) &\le 1+\epsilon, \\ \sigma_{\max}(\boldsymbol{H}_1^{-1})\sigma_{\max}(\boldsymbol{H}_2^{-1})\sigma_{\max}(\boldsymbol{H}_3^{-1})\sigma_{\min}\left((\bar{\boldsymbol{Q}}_1^{-1}, \bar{\boldsymbol{Q}}_2^{-1}, \bar{\boldsymbol{Q}}_3^{-1})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}\right) &\le (1+\epsilon)\sigma_{\max}(\boldsymbol{\mathcal{X}}_\star), \end{aligned}$$

which, combined with (32), leads to

$$\sigma_{\max}(\boldsymbol{H}_k) \le \frac{1+\epsilon}{1-\epsilon}, \quad k = 1,2,3.$$

$$\sigma_{\max}(\boldsymbol{H}_1^{-1})\sigma_{\max}(\boldsymbol{H}_2^{-1})\sigma_{\max}(\boldsymbol{H}_3^{-1}) \le \frac{1+\epsilon}{1-\epsilon}\kappa \implies \sigma_{\min}(\boldsymbol{H}_1)\sigma_{\min}(\boldsymbol{H}_2)\sigma_{\min}(\boldsymbol{H}_3) \ge \frac{1-\epsilon}{1+\epsilon}\kappa^{-1}.$$

As a result, the minimization problem (33) is equivalent to the constrained problem:

$$\begin{aligned} \min_{\boldsymbol{H}_k \in \mathrm{GL}(r_k)} \quad & \left\|(\boldsymbol{U}\bar{\boldsymbol{Q}}_1\boldsymbol{H}_1 - \boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{V}\bar{\boldsymbol{Q}}_2\boldsymbol{H}_2 - \boldsymbol{V}_\star)\boldsymbol{\Sigma}_{\star,2}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{W}\bar{\boldsymbol{Q}}_3\boldsymbol{H}_3 - \boldsymbol{W}_\star)\boldsymbol{\Sigma}_{\star,3}\right\|_{\mathsf{F}}^2 \\ & + \left\|\left(\boldsymbol{H}_1^{-1}\bar{\boldsymbol{Q}}_1^{-1}, \boldsymbol{H}_2^{-1}\bar{\boldsymbol{Q}}_2^{-1}, \boldsymbol{H}_3^{-1}\bar{\boldsymbol{Q}}_3^{-1}\right)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{S}}_\star\right\|_{\mathsf{F}}^2 \\ \text{s.t.} \quad & \sigma_{\max}(\boldsymbol{H}_k) \le \frac{1+\epsilon}{1-\epsilon}, \quad \sigma_{\min}(\boldsymbol{H}_1)\sigma_{\min}(\boldsymbol{H}_2)\sigma_{\min}(\boldsymbol{H}_3) \ge \frac{1-\epsilon}{1+\epsilon}\kappa^{-1}, \quad k = 1,2,3. \end{aligned}$$

Since this is a continuous optimization problem over a compact set, applying the Weierstrass extreme value theorem finishes the proof.

A.1.2 Proof of Lemma 7

Set the gradient of the expression on the right-hand side of (31) with respect to $\boldsymbol{Q}_1$ as zero to see

$$\boldsymbol{U}^\top(\boldsymbol{U}\boldsymbol{Q}_1 - \boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1}^2 - \boldsymbol{Q}_1^{-\top}\mathcal{M}_1\left((\boldsymbol{Q}_1^{-1}, \boldsymbol{Q}_2^{-1}, \boldsymbol{Q}_3^{-1})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{S}}_\star\right)\mathcal{M}_1\left((\boldsymbol{Q}_1^{-1}, \boldsymbol{Q}_2^{-1}, \boldsymbol{Q}_3^{-1})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}\right)^\top = \boldsymbol{0}.$$

We conclude the proof by similarly setting the gradient with respect to $\boldsymbol{Q}_2$ or $\boldsymbol{Q}_3$ to zero.

A.1.3 Proof of Lemma 8

We first state a lemma from [TMC20, Lemma 11], which will be used repeatedly for matricization over different modes.

Lemma 9 ([TMC20]). *Suppose that $\boldsymbol{X}_\star \in \mathbb{R}^{n_1 \times n_2}$ has the compact rank-r SVD $\boldsymbol{X}_\star = \boldsymbol{U}_\star\boldsymbol{\Sigma}_\star\boldsymbol{V}_\star^\top$. For any $\boldsymbol{L} \in \mathbb{R}^{n_1 \times r}$ and $\boldsymbol{R} \in \mathbb{R}^{n_2 \times r}$, one has*

$$\inf_{\boldsymbol{Q} \in \mathrm{GL}(r)} \left\|\boldsymbol{L}\boldsymbol{Q}\boldsymbol{\Sigma}_\star^{1/2} - \boldsymbol{U}_\star\boldsymbol{\Sigma}_\star\right\|_{\mathsf{F}}^2 + \left\|\boldsymbol{R}\boldsymbol{Q}^{-\top}\boldsymbol{\Sigma}_\star^{1/2} - \boldsymbol{V}_\star\boldsymbol{\Sigma}_\star\right\|_{\mathsf{F}}^2 \le (\sqrt{2}+1)\|\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star\|_{\mathsf{F}}^2.$$

We begin by applying the mode-1 matricization (see (12)), and invoking Lemma 9 with $\boldsymbol{L} \coloneqq \boldsymbol{U}$, $\boldsymbol{R} \coloneqq (\boldsymbol{V} \otimes \boldsymbol{W})\mathcal{M}_1(\boldsymbol{\mathcal{S}})^\top$, $\boldsymbol{X}_\star \coloneqq \boldsymbol{U}_\star\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{V}_\star \otimes \boldsymbol{W}_\star)^\top$, $\boldsymbol{\Sigma}_\star \coloneqq \boldsymbol{\Sigma}_{\star,1}$ to arrive at

$$\begin{aligned}
&\|(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}^2 = \left\|\boldsymbol{U}\mathcal{M}_1(\boldsymbol{\mathcal{S}})(\boldsymbol{V}\otimes\boldsymbol{W})^\top - \boldsymbol{U}_\star\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{V}_\star\otimes\boldsymbol{W}_\star)^\top\right\|_{\mathsf{F}}^2 \\
&\quad\ge (\sqrt{2}-1)\inf_{\boldsymbol{Q}\in\mathrm{GL}(r_1)} \left\|\boldsymbol{U}\boldsymbol{Q}\boldsymbol{\Sigma}_{\star,1}^{1/2} - \boldsymbol{U}_\star\boldsymbol{\Sigma}_{\star,1}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{V}\otimes\boldsymbol{W})\mathcal{M}_1(\boldsymbol{\mathcal{S}})^\top\boldsymbol{Q}^{-\top}\boldsymbol{\Sigma}_{\star,1}^{1/2} - (\boldsymbol{V}_\star\otimes\boldsymbol{W}_\star)\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)^\top\right\|_{\mathsf{F}}^2 \\
&\quad= (\sqrt{2}-1)\inf_{\boldsymbol{Q}_1\in\mathrm{GL}(r_1)} \|(\boldsymbol{U}\boldsymbol{Q}_1 - \boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{V}\otimes\boldsymbol{W})\mathcal{M}_1(\boldsymbol{\mathcal{S}})^\top\boldsymbol{Q}_1^{-\top} - (\boldsymbol{V}_\star\otimes\boldsymbol{W}_\star)\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)^\top\right\|_{\mathsf{F}}^2 \\
&\quad= (\sqrt{2}-1)\inf_{\boldsymbol{Q}_1\in\mathrm{GL}(r_1)} \|(\boldsymbol{U}\boldsymbol{Q}_1 - \boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{Q}_1^{-1},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}} - (\boldsymbol{I}_{r_1},\boldsymbol{V}_\star,\boldsymbol{W}_\star)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star\right\|_{\mathsf{F}}^2,
\end{aligned}$$

where we have applied a change-of-variable as $\boldsymbol{Q}_1 = \boldsymbol{Q}\boldsymbol{\Sigma}_{\star,1}^{-1/2}$ in the third line, and converted back to the tensor space in the last line. Continue in a similar manner, by applying the mode-2 matricization to the second term (see (12)), and invoke Lemma 9 with $\boldsymbol{L} \coloneqq \boldsymbol{V}$, $\boldsymbol{R} \coloneqq (\boldsymbol{Q}_1^{-1}\otimes\boldsymbol{W})\mathcal{M}_2(\boldsymbol{\mathcal{S}})^\top$, $\boldsymbol{X}_\star \coloneqq \boldsymbol{V}_\star\mathcal{M}_2(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{I}_{r_1}\otimes\boldsymbol{W}_\star)^\top$, $\boldsymbol{\Sigma}_\star \coloneqq \boldsymbol{\Sigma}_{\star,2}$ to arrive at

$$\begin{aligned}
&\left\|(\boldsymbol{Q}_1^{-1},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}} - (\boldsymbol{I}_{r_1},\boldsymbol{V}_\star,\boldsymbol{W}_\star)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star\right\|_{\mathsf{F}}^2 = \left\|\boldsymbol{V}\mathcal{M}_2(\boldsymbol{\mathcal{S}})(\boldsymbol{Q}_1^{-1}\otimes\boldsymbol{W})^\top - \boldsymbol{V}_\star\mathcal{M}_2(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{I}_{r_1}\otimes\boldsymbol{W}_\star)^\top\right\|_{\mathsf{F}}^2 \\
&\quad\ge (\sqrt{2}-1)\inf_{\boldsymbol{Q}\in\mathrm{GL}(r_2)} \left\|\boldsymbol{V}\boldsymbol{Q}\boldsymbol{\Sigma}_{\star,2}^{1/2} - \boldsymbol{V}_\star\boldsymbol{\Sigma}_{\star,2}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{Q}_1^{-1}\otimes\boldsymbol{W})\mathcal{M}_2(\boldsymbol{\mathcal{S}})^\top\boldsymbol{Q}^{-\top}\boldsymbol{\Sigma}_{\star,2}^{1/2} - (\boldsymbol{I}_{r_1}\otimes\boldsymbol{W}_\star)\mathcal{M}_2(\boldsymbol{\mathcal{S}}_\star)^\top\right\|_{\mathsf{F}}^2 \\
&\quad= (\sqrt{2}-1)\inf_{\boldsymbol{Q}_2\in\mathrm{GL}(r_2)} \|(\boldsymbol{V}\boldsymbol{Q}_2 - \boldsymbol{V}_\star)\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}^2 + \|(\boldsymbol{Q}_1^{-1},\boldsymbol{Q}_2^{-1},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}} - (\boldsymbol{I}_{r_1},\boldsymbol{I}_{r_2},\boldsymbol{W}_\star)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star\|_{\mathsf{F}}^2.
\end{aligned}$$

where we have applied a change-of-variable as $\boldsymbol{Q}_2 = \boldsymbol{Q}\boldsymbol{\Sigma}_{\star,2}^{-1/2}$ as well as tensorization in the last line. Repeating the same argument by applying the mode-3 matricization to the second term, we obtain

$$\begin{aligned}
\left\|(\boldsymbol{Q}_1^{-1},\boldsymbol{Q}_2^{-1},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}} - (\boldsymbol{I}_{r_1},\boldsymbol{I}_{r_2},\boldsymbol{W}_\star)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star\right\|_{\mathsf{F}}^2 &= \left\|\boldsymbol{W}\mathcal{M}_3(\boldsymbol{\mathcal{S}})(\boldsymbol{Q}_1^{-1}\otimes\boldsymbol{Q}_2^{-1})^\top - \boldsymbol{W}_\star\mathcal{M}_3(\boldsymbol{\mathcal{S}}_\star)\right\|_{\mathsf{F}}^2 \\
&\ge (\sqrt{2}-1)\inf_{\boldsymbol{Q}_3\in\mathrm{GL}(r_3)} \|(\boldsymbol{W}\boldsymbol{Q}_3 - \boldsymbol{W}_\star)\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{Q}_1^{-1},\boldsymbol{Q}_2^{-1},\boldsymbol{Q}_3^{-1})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{S}}_\star\right\|_{\mathsf{F}}^2.
\end{aligned}$$

Finally, combine these results to conclude

$$\begin{aligned}
\|(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}^2 &\ge \inf_{\boldsymbol{Q}_k\in\mathrm{GL}(r_k)} (\sqrt{2}-1)\,\|(\boldsymbol{U}\boldsymbol{Q}_1 - \boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 + (\sqrt{2}-1)^2\,\|(\boldsymbol{V}\boldsymbol{Q}_2 - \boldsymbol{V}_\star)\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}^2 \\
&\qquad + (\sqrt{2}-1)^3\,\|(\boldsymbol{W}\boldsymbol{Q}_3 - \boldsymbol{W}_\star)\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}^2 + (\sqrt{2}-1)^3\left\|(\boldsymbol{Q}_1^{-1},\boldsymbol{Q}_2^{-1},\boldsymbol{Q}_3^{-1})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{S}}_\star\right\|_{\mathsf{F}}^2 \\
&\ge (\sqrt{2}-1)^3\,\mathrm{dist}^2(\boldsymbol{F},\boldsymbol{F}_\star),
\end{aligned}$$

where the last relation uses the definition of $\mathrm{dist}^2(\boldsymbol{F},\boldsymbol{F}_\star)$.

A.2 Several perturbation bounds

We now collect several perturbation bounds that will be used repeatedly in the proof. Without loss of generality, assume that $\boldsymbol{F} = (\boldsymbol{U},\boldsymbol{V},\boldsymbol{W},\boldsymbol{\mathcal{S}})$ and $\boldsymbol{F}_\star = (\boldsymbol{U}_\star,\boldsymbol{V}_\star,\boldsymbol{W}_\star,\boldsymbol{\mathcal{S}}_\star)$ are aligned, and introduce the following notation that will be used repeatedly:

$$\begin{array}{llll}
\boldsymbol{\Delta}_U \coloneqq \boldsymbol{U} - \boldsymbol{U}_\star, & \boldsymbol{\Delta}_V \coloneqq \boldsymbol{V} - \boldsymbol{V}_\star, & \boldsymbol{\Delta}_W \coloneqq \boldsymbol{W} - \boldsymbol{W}_\star, & \boldsymbol{\Delta}_{\mathcal{S}} \coloneqq \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{S}}_\star, \\
\breve{\boldsymbol{U}} \coloneqq (\boldsymbol{V}\otimes\boldsymbol{W})\mathcal{M}_1(\boldsymbol{\mathcal{S}})^\top, & & \breve{\boldsymbol{V}} \coloneqq (\boldsymbol{U}\otimes\boldsymbol{W})\mathcal{M}_2(\boldsymbol{\mathcal{S}})^\top, & \breve{\boldsymbol{W}} \coloneqq (\boldsymbol{U}\otimes\boldsymbol{V})\mathcal{M}_3(\boldsymbol{\mathcal{S}})^\top, \\
\breve{\boldsymbol{U}}_\star \coloneqq (\boldsymbol{V}_\star\otimes\boldsymbol{W}_\star)\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)^\top, & & \breve{\boldsymbol{V}}_\star \coloneqq (\boldsymbol{U}_\star\otimes\boldsymbol{W}_\star)\mathcal{M}_2(\boldsymbol{\mathcal{S}}_\star)^\top, & \breve{\boldsymbol{W}}_\star \coloneqq (\boldsymbol{U}_\star\otimes\boldsymbol{V}_\star)\mathcal{M}_3(\boldsymbol{\mathcal{S}}_\star)^\top, \\
\boldsymbol{\mathcal{T}}_U \coloneqq (\boldsymbol{U}_\star^\top\boldsymbol{\Delta}_U,\boldsymbol{I}_{r_2},\boldsymbol{I}_{r_3})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star, & & \boldsymbol{\mathcal{T}}_V \coloneqq (\boldsymbol{I}_{r_1},\boldsymbol{V}_\star^\top\boldsymbol{\Delta}_V,\boldsymbol{I}_{r_3})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star, & \boldsymbol{\mathcal{T}}_W \coloneqq (\boldsymbol{I}_{r_1},\boldsymbol{I}_{r_2},\boldsymbol{W}_\star^\top\boldsymbol{\Delta}_W)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star, \\
\boldsymbol{D}_U \coloneqq (\boldsymbol{U}^\top\boldsymbol{U})^{-1/2}\boldsymbol{U}^\top\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}, & & \boldsymbol{D}_V \coloneqq (\boldsymbol{V}^\top\boldsymbol{V})^{-1/2}\boldsymbol{V}^\top\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}, & \boldsymbol{D}_W \coloneqq (\boldsymbol{W}^\top\boldsymbol{W})^{-1/2}\boldsymbol{W}^\top\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}.
\end{array} \tag{34}$$

With these notation, we can rephrase the consequences of Lemma 7 as:

$$\boldsymbol{U}^\top\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}^2 = \mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})\mathcal{M}_1(\boldsymbol{\mathcal{S}})^\top, \tag{35a}$$

$$\boldsymbol{V}^\top \boldsymbol{\Delta}_V \boldsymbol{\Sigma}_{\star,2}^2 = \mathcal{M}_2(\boldsymbol{\Delta}_{\mathcal{S}})\mathcal{M}_2(\boldsymbol{\mathcal{S}})^\top, \tag{35b}$$
$$\boldsymbol{W}^\top \boldsymbol{\Delta}_W \boldsymbol{\Sigma}_{\star,3}^2 = \mathcal{M}_3(\boldsymbol{\Delta}_{\mathcal{S}})\mathcal{M}_3(\boldsymbol{\mathcal{S}})^\top. \tag{35c}$$

Now we are ready to state the lemma on perturbation bounds.

Lemma 10. *Suppose* $\boldsymbol{F} = (\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}, \boldsymbol{\mathcal{S}})$ *and* $\boldsymbol{F}_\star = (\boldsymbol{U}_\star, \boldsymbol{V}_\star, \boldsymbol{W}_\star, \boldsymbol{\mathcal{S}}_\star)$ *are aligned and satisfy* $\operatorname{dist}(\boldsymbol{F}, \boldsymbol{F}_\star) \le \epsilon\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$ *for some* $\epsilon < 1$. *Then the following bounds hold regarding the spectral norm:*

$$\|\boldsymbol{\Delta}_U\| \vee \|\boldsymbol{\Delta}_V\| \vee \|\boldsymbol{\Delta}_W\| \vee \|\mathcal{M}_k(\boldsymbol{\Delta}_{\mathcal{S}})^\top \boldsymbol{\Sigma}_{\star,k}^{-1}\| \le \epsilon, \qquad k = 1,2,3; \tag{36a}$$
$$\|\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\| \le (1-\epsilon)^{-1}; \tag{36b}$$
$$\left\|\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1} - \boldsymbol{U}_\star\right\| \le \frac{\sqrt{2}\epsilon}{1-\epsilon}; \tag{36c}$$
$$\left\|(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\right\| \le (1-\epsilon)^{-2}; \tag{36d}$$
$$\left\|(\breve{\boldsymbol{U}} - \breve{\boldsymbol{U}}_\star)\boldsymbol{\Sigma}_{\star,1}^{-1}\right\| \le 3\epsilon + 3\epsilon^2 + \epsilon^3; \tag{36e}$$
$$\left\|\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}\right\| \le (1-\epsilon)^{-3}; \tag{36f}$$
$$\left\|\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1} - \breve{\boldsymbol{U}}_\star\boldsymbol{\Sigma}_{\star,1}^{-1}\right\| \le \frac{\sqrt{2}(3\epsilon + 3\epsilon^2 + \epsilon^3)}{(1-\epsilon)^3}; \tag{36g}$$
$$\left\|\boldsymbol{\Sigma}_{\star,1}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}\right\| \le (1-\epsilon)^{-6}; \tag{36h}$$
$$\left\|\boldsymbol{\Sigma}_{\star,1}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\mathcal{M}_1(\boldsymbol{\mathcal{S}})\right\| \le (1-\epsilon)^{-5}. \tag{36i}$$

By symmetry, a corresponding set of bounds holds for $\boldsymbol{V}, \breve{\boldsymbol{V}}$ *and* $\boldsymbol{W}, \breve{\boldsymbol{W}}$.

In addition, the following bounds hold regarding the Frobenius norm:

$$\|(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \le (1 + \frac{3}{2}\epsilon + \epsilon^2 + \frac{\epsilon^3}{4})\left(\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\right); \tag{37a}$$
$$\|(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_\star - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \le (1 + \epsilon + \frac{\epsilon^2}{3})\left(\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}\right); \tag{37b}$$
$$\left\|\breve{\boldsymbol{U}} - \breve{\boldsymbol{U}}_\star\right\|_{\mathsf{F}} \le (1 + \epsilon + \frac{\epsilon^2}{3})\left(\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\right). \tag{37c}$$

As a straightforward consequence of (37a), *the following important relation holds when* $\epsilon \le 0.2$:

$$\|(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \le 2(1 + \frac{3}{2}\epsilon + \epsilon^2 + \frac{\epsilon^3}{4})\operatorname{dist}(\boldsymbol{F}, \boldsymbol{F}_\star) \le 3\operatorname{dist}(\boldsymbol{F}, \boldsymbol{F}_\star). \tag{38}$$

Hence, the scaled distance serves as a metric to gauge the quality of the tensor recovery.

A.2.1 Proof of Lemma 10

Proof of spectral norm perturbation bounds. To begin, recalling the notation in (34), (36a) follows directly from the definition

$$\operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) = \sqrt{\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}^2} \le \epsilon\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$$

together with the relation $\|\boldsymbol{A}\boldsymbol{B}\|_{\mathsf{F}} \ge \|\boldsymbol{A}\|_{\mathsf{F}}\sigma_{\min}(\boldsymbol{B})$.

For (36b), Weyl's inequality tells $\sigma_{\min}(\boldsymbol{U}) \ge \sigma_{\min}(\boldsymbol{U}_\star) - \|\boldsymbol{\Delta}_U\| \ge 1 - \epsilon$, and use that

$$\left\|\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\right\| = \frac{1}{\sigma_{\min}(\boldsymbol{U})} \le \frac{1}{1-\epsilon}.$$

For (36c), decompose

$$\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1} - \boldsymbol{U}_\star = -\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\boldsymbol{\Delta}_U^\top\boldsymbol{U}_\star + \left(\boldsymbol{I}_{n_1} - \boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\boldsymbol{U}^\top\right)\boldsymbol{\Delta}_U,$$

and use that the two terms are orthogonal to obtain

$$\begin{aligned}\left\|\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}-\boldsymbol{U}_\star\right\|^2 &= \left\|\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\boldsymbol{\Delta}_U^\top\boldsymbol{U}_\star\right\|^2 + \left\|\left(\boldsymbol{I}_{n_1}-\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\boldsymbol{U}^\top\right)\boldsymbol{\Delta}_U\right\|^2\\ &\le \|\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\|^2\|\boldsymbol{\Delta}_U\|^2+\|\boldsymbol{\Delta}_U\|^2\\ &\le \left((1-\epsilon)^{-2}+1\right)\epsilon^2.\end{aligned}$$

It follows from $\epsilon < 1$ that

$$\left\|\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}-\boldsymbol{U}_\star\right\| \le \frac{\sqrt{2}\epsilon}{1-\epsilon}.$$

For (36d), recognizing that

$$(\boldsymbol{U}^\top\boldsymbol{U})^{-1} = (\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1})^\top\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1} \qquad\Longrightarrow\qquad \|(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\| = \|\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\|^2 \le \frac{1}{(1-\epsilon)^2},$$

where the last inequality follows from (36b).

For (36e), we first expand the expression as

$$\begin{aligned}\breve{\boldsymbol{U}}-\breve{\boldsymbol{U}}_\star &= (\boldsymbol{V}\otimes\boldsymbol{W})\mathcal{M}_1(\boldsymbol{\mathcal{S}})^\top-(\boldsymbol{V}_\star\otimes\boldsymbol{W}_\star)\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)^\top\\ &= (\boldsymbol{V}\otimes\boldsymbol{W}-\boldsymbol{V}_\star\otimes\boldsymbol{W}_\star)\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)^\top+(\boldsymbol{V}\otimes\boldsymbol{W})\mathcal{M}_1(\boldsymbol{\mathcal{S}})^\top-(\boldsymbol{V}\otimes\boldsymbol{W})\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)^\top\\ &= (\boldsymbol{\Delta}_V\otimes\boldsymbol{W}+\boldsymbol{V}_\star\otimes\boldsymbol{\Delta}_W)\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)^\top+(\boldsymbol{V}\otimes\boldsymbol{W})\mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})^\top.\end{aligned}\tag{39}$$

Apply the triangle inequality to obtain

$$\begin{aligned}\|(\breve{\boldsymbol{U}}-\breve{\boldsymbol{U}}_\star)\boldsymbol{\Sigma}_{\star,1}^{-1}\| &\le \left\|(\boldsymbol{\Delta}_V\otimes\boldsymbol{W}+\boldsymbol{V}_\star\otimes\boldsymbol{\Delta}_W)\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)^\top\boldsymbol{\Sigma}_{\star,1}^{-1}\right\|+\left\|(\boldsymbol{V}\otimes\boldsymbol{W})\mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})^\top\boldsymbol{\Sigma}_{\star,1}^{-1}\right\|\\ &\le (\|\boldsymbol{\Delta}_V\|\|\boldsymbol{W}\|+\|\boldsymbol{V}_\star\|\|\boldsymbol{\Delta}_W\|)\,\|\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)^\top\boldsymbol{\Sigma}_{\star,1}^{-1}\|+\|\boldsymbol{V}\|\|\boldsymbol{W}\|\|\mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})^\top\boldsymbol{\Sigma}_{\star,1}^{-1}\|\\ &\le \epsilon(1+\epsilon)+\epsilon+(1+\epsilon)^2\epsilon = 3\epsilon+3\epsilon^2+\epsilon^3,\end{aligned}$$

where we have used (36a) and the fact $\|\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)^\top\boldsymbol{\Sigma}_{\star,1}^{-1}\| = 1$ (see (13)) in the last line.

(36f) follows from combining

$$\begin{aligned}\sigma_{\min}\left(\breve{\boldsymbol{U}}\boldsymbol{\Sigma}_{\star,1}^{-1}\right) &\ge \sigma_{\min}(\boldsymbol{V})\sigma_{\min}(\boldsymbol{W})\sigma_{\min}\left(\mathcal{M}_1(\boldsymbol{\mathcal{S}})\boldsymbol{\Sigma}_{\star,1}^{-1}\right) \ge (1-\epsilon)^3,\\ \text{and}\quad \left\|\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}\right\| &= \frac{1}{\sigma_{\min}\left(\breve{\boldsymbol{U}}\boldsymbol{\Sigma}_{\star,1}^{-1}\right)} \le \frac{1}{(1-\epsilon)^3}.\end{aligned}$$

With regard to (36g), repeat the same proof as (36c), decompose

$$\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}-\breve{\boldsymbol{U}}_\star\boldsymbol{\Sigma}_{\star,1}^{-1} = -\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}(\breve{\boldsymbol{U}}-\breve{\boldsymbol{U}}_\star)^\top\breve{\boldsymbol{U}}_\star\boldsymbol{\Sigma}_{\star,1}^{-1}+\left(\boldsymbol{I}_{n_2n_3}-\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\breve{\boldsymbol{U}}^\top\right)(\breve{\boldsymbol{U}}-\breve{\boldsymbol{U}}_\star)\boldsymbol{\Sigma}_{\star,1}^{-1},$$

and use that the two terms are orthogonal to obtain

$$\begin{aligned}\left\|\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}-\breve{\boldsymbol{U}}_\star\boldsymbol{\Sigma}_{\star,1}^{-1}\right\|^2 &= \left\|\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}(\breve{\boldsymbol{U}}-\breve{\boldsymbol{U}}_\star)^\top\breve{\boldsymbol{U}}_\star\boldsymbol{\Sigma}_{\star,1}^{-1}\right\|^2+\left\|\left(\boldsymbol{I}_{n_2n_3}-\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\breve{\boldsymbol{U}}^\top\right)(\breve{\boldsymbol{U}}-\breve{\boldsymbol{U}}_\star)\boldsymbol{\Sigma}_{\star,1}^{-1}\right\|^2\\ &\le \|\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}\|^2\|(\breve{\boldsymbol{U}}-\breve{\boldsymbol{U}}_\star)\boldsymbol{\Sigma}_{\star,1}^{-1}\|^2+\|(\breve{\boldsymbol{U}}-\breve{\boldsymbol{U}}_\star)\boldsymbol{\Sigma}_{\star,1}^{-1}\|^2\\ &\le \left((1-\epsilon)^{-6}+1\right)(3\epsilon+3\epsilon^2+\epsilon^3)^2.\end{aligned}$$

It follows from $\epsilon < 1$ that

$$\left\|\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}-\breve{\boldsymbol{U}}_\star\boldsymbol{\Sigma}_{\star,1}^{-1}\right\| \le \frac{\sqrt{2}(3\epsilon+3\epsilon^2+\epsilon^3)}{(1-\epsilon)^3}.$$

The relation (36h) follows from the relation below and (36f):

$$\|\boldsymbol{\Sigma}_{\star,1}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}\| = \|\boldsymbol{\Sigma}_{\star,1}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}\| = \|\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}\|^2.$$

With regard to (36i), we have

$$\begin{aligned}\left\|\boldsymbol{\Sigma}_{\star,1}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\mathcal{M}_1(\boldsymbol{\mathcal{S}})\right\| &= \left\|\boldsymbol{\Sigma}_{\star,1}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\breve{\boldsymbol{U}}^\top\left(\boldsymbol{V}(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\otimes\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\right)\right\| \\ &\le \left\|\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}\right\|\left\|\boldsymbol{V}(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\right\|\left\|\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\right\| \\ &\le (1-\epsilon)^{-5},\end{aligned}$$

where the first line follows from

$$\breve{\boldsymbol{U}}^\top = \mathcal{M}_1(\boldsymbol{S})(\boldsymbol{V}\otimes\boldsymbol{W})^\top \qquad \Longrightarrow \qquad \mathcal{M}_1(\boldsymbol{\mathcal{S}}) = \breve{\boldsymbol{U}}^\top\left(\boldsymbol{V}(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\otimes\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\right), \tag{40}$$

and the last inequality uses (36c) and (36f).

Proof of Frobenius norm perturbation bounds. We proceed to prove the perturbation bounds regarding the Frobenius norm. For (37a), we begin with the following decomposition

$$\begin{aligned}(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\cdot\boldsymbol{\mathcal{S}}-\boldsymbol{\mathcal{X}}_\star &= (\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\cdot\boldsymbol{\mathcal{S}}-(\boldsymbol{U}_\star,\boldsymbol{V}_\star,\boldsymbol{W}_\star)\cdot\boldsymbol{\mathcal{S}}_\star \\ &= (\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\cdot\boldsymbol{\Delta}_{\mathcal{S}}+(\boldsymbol{\Delta}_U,\boldsymbol{V},\boldsymbol{W})\cdot\boldsymbol{\mathcal{S}}_\star+(\boldsymbol{U}_\star,\boldsymbol{\Delta}_V,\boldsymbol{W})\cdot\boldsymbol{\mathcal{S}}_\star+(\boldsymbol{U}_\star,\boldsymbol{V}_\star,\boldsymbol{\Delta}_W)\cdot\boldsymbol{\mathcal{S}}_\star.\end{aligned} \tag{41}$$

Apply the triangle inequality, together with the invariance of the Frobenius norm to matricization, to obtain

$$\begin{aligned}\|(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\cdot\boldsymbol{\mathcal{S}}-\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} &\le \|(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\cdot\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}+\left\|\boldsymbol{\Delta}_U\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{V}\otimes\boldsymbol{W})^\top\right\|_{\mathsf{F}} \\ &\quad+\left\|\boldsymbol{\Delta}_V\mathcal{M}_2(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{U}_\star\otimes\boldsymbol{W})^\top\right\|_{\mathsf{F}}+\left\|\boldsymbol{\Delta}_W\mathcal{M}_3(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{U}_\star\otimes\boldsymbol{V}_\star)^\top\right\|_{\mathsf{F}} \\ &\le \|\boldsymbol{U}\|\|\boldsymbol{V}\|\|\boldsymbol{W}\|\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_U\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)\|_{\mathsf{F}}\|\boldsymbol{V}\otimes\boldsymbol{W}\| \\ &\quad+\|\boldsymbol{\Delta}_V\mathcal{M}_2(\boldsymbol{\mathcal{S}}_\star)\|_{\mathsf{F}}\|\boldsymbol{U}_\star\otimes\boldsymbol{W}\|+\|\boldsymbol{\Delta}_W\mathcal{M}_3(\boldsymbol{\mathcal{S}}_\star)\|_{\mathsf{F}}\|\boldsymbol{U}_\star\otimes\boldsymbol{V}_\star\| \\ &\le (1+\epsilon)^3\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}+(1+\epsilon)^2\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}+(1+\epsilon)\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}},\end{aligned}$$

where the second inequality follows from (6e), and the last inequality follows from (13) and (36a). By symmetry, one can permute the occurrence of $\boldsymbol{\Delta}_U,\boldsymbol{\Delta}_V,\boldsymbol{\Delta}_W,\boldsymbol{\Delta}_{\mathcal{S}}$ in the decomposition (41). For example, invoking another viable decomposition of $(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\cdot\boldsymbol{\mathcal{S}}-\boldsymbol{\mathcal{X}}_\star$ as

$$(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\cdot\boldsymbol{\mathcal{S}}-\boldsymbol{\mathcal{X}}_\star = (\boldsymbol{U},\boldsymbol{\Delta}_V,\boldsymbol{W})\cdot\boldsymbol{\mathcal{S}}+(\boldsymbol{U},\boldsymbol{V}_\star,\boldsymbol{\Delta}_W)\cdot\boldsymbol{\mathcal{S}}+(\boldsymbol{U},\boldsymbol{V}_\star,\boldsymbol{W}_\star)\cdot\boldsymbol{\Delta}_{\mathcal{S}}+(\boldsymbol{\Delta}_U,\boldsymbol{V}_\star,\boldsymbol{W}_\star)\cdot\boldsymbol{\mathcal{S}}_\star$$

leads to the perturbation bound

$$\|(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\cdot\boldsymbol{\mathcal{S}}-\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \le (1+\epsilon)^3\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}+(1+\epsilon)^2\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}+(1+\epsilon)\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}.$$

To complete the proof of (37a), we take an average of all viable bounds from $4! = 24$ permutations to balance their coefficients as

$$\frac{1}{4}\left((1+\epsilon)^3+(1+\epsilon)^2+(1+\epsilon)+1\right) = 1+\frac{3}{2}\epsilon+\epsilon^2+\frac{1}{4}\epsilon^3,$$

thus we obtain

$$\|(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\cdot\boldsymbol{\mathcal{S}}-\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \le (1+\frac{3}{2}\epsilon+\epsilon^2+\frac{1}{4}\epsilon^3)\big(\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\big).$$

The relation (37b) can be proved in a similar fashion; for the sake of brevity, we omit its proof.

Turning to (37c), apply the triangle inequality to (39) to obtain

$$\|\breve{\boldsymbol{U}}-\breve{\boldsymbol{U}}_\star\|_{\mathsf{F}} \le \left\|(\boldsymbol{\Delta}_V\otimes\boldsymbol{W})\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)^\top\right\|_{\mathsf{F}}+\left\|(\boldsymbol{V}_\star\otimes\boldsymbol{\Delta}_W)\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)^\top\right\|_{\mathsf{F}}+\|(\boldsymbol{V}\otimes\boldsymbol{W})\mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})\|_{\mathsf{F}}\,. \tag{42}$$

To bound the first term, change the mode of matricization (see (12)) to arrive at

$$\begin{aligned}\left\|(\boldsymbol{\Delta}_V\otimes\boldsymbol{W})\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)^\top\right\|_{\mathsf{F}} &= \|(\boldsymbol{I}_{r_1},\boldsymbol{\Delta}_V,\boldsymbol{W})\cdot\boldsymbol{\mathcal{S}}_\star\|_{\mathsf{F}} = \left\|\boldsymbol{\Delta}_V\mathcal{M}_2(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{I}_{r_1}\otimes\boldsymbol{W})^\top\right\|_{\mathsf{F}} \\ &\le \|\boldsymbol{\Delta}_V\mathcal{M}_2(\boldsymbol{\mathcal{S}}_\star)\|_{\mathsf{F}}\|\boldsymbol{W}\| \le (1+\epsilon)\|\boldsymbol{\Delta}_V\mathcal{M}_2(\boldsymbol{\mathcal{S}}_\star)\|_{\mathsf{F}},\end{aligned}$$

where the last inequality uses (36a). Similarly, the last two terms in (42) can be bounded as

$$\left\|(\boldsymbol{V}_\star \otimes \boldsymbol{\Delta}_W)\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)^\top\right\|_{\mathsf{F}} \le \|\boldsymbol{\Delta}_W \mathcal{M}_3(\boldsymbol{\mathcal{S}}_\star)\|_{\mathsf{F}}, \quad \text{and} \quad \|(\boldsymbol{V} \otimes \boldsymbol{W})\mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})\|_{\mathsf{F}} \le (1+\epsilon)^2 \|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}.$$

Plugging the above bounds back to (42), we have

$$\|\breve{\boldsymbol{U}} - \breve{\boldsymbol{U}}_\star\|_{\mathsf{F}} \le (1+\epsilon)\|\boldsymbol{\Delta}_V \mathcal{M}_2(\boldsymbol{\mathcal{S}}_\star)\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_W \mathcal{M}_3(\boldsymbol{\mathcal{S}}_\star)\|_{\mathsf{F}} + (1+\epsilon)^2 \|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}.$$

Using a similar symmetrization trick as earlier, by permuting the occurrences of $\boldsymbol{\Delta}_V, \boldsymbol{\Delta}_W, \boldsymbol{\Delta}_{\mathcal{S}}$ in the decomposition (39), we arrive at the final advertised bound (37c).

A.3 A useful reformulation of regularized GD

Here, we discuss the original regularized GD algorithm in [HWZ20] and detail how it is mapped to (30) when using the same initialization as ScaledGD. To be exact, the algorithm in [HWZ20] uses an initialization

$$\widetilde{\boldsymbol{F}}_0 = (\widetilde{\boldsymbol{U}}_0, \widetilde{\boldsymbol{V}}_0, \widetilde{\boldsymbol{W}}_0, \widetilde{\boldsymbol{\mathcal{S}}}_0) = (\sigma_{\max}^{1/4}\boldsymbol{U}_0,\, \sigma_{\max}^{1/4}\boldsymbol{V}_0,\, \sigma_{\max}^{1/4}\boldsymbol{W}_0,\, \sigma_{\max}^{-3/4}\boldsymbol{\mathcal{S}}_0),$$

where $(\boldsymbol{U}_0, \boldsymbol{V}_0, \boldsymbol{W}_0, \boldsymbol{\mathcal{S}}_0)$ is the initialization of ScaledGD, and $\sigma_{\max} = \sigma_{\max}(\boldsymbol{\mathcal{X}}_\star)$ for brevity. According to [HWZ20], we set $\alpha = \sigma_{\max}$, $\beta = \sigma_{\max}^{1/2}$ in (3) and obtain the update rule as

$$\begin{aligned}
\widetilde{\boldsymbol{U}}_{t+1} &= \widetilde{\boldsymbol{U}}_t - \widetilde{\eta}\left[\nabla_{\widetilde{\boldsymbol{U}}}\mathcal{L}(\widetilde{\boldsymbol{F}}_t) + \sigma_{\max}\widetilde{\boldsymbol{U}}_t(\widetilde{\boldsymbol{U}}_t^\top \widetilde{\boldsymbol{U}}_t - \sigma_{\max}^{1/2}\boldsymbol{I}_{r_1})\right],\\
\widetilde{\boldsymbol{V}}_{t+1} &= \widetilde{\boldsymbol{V}}_t - \widetilde{\eta}\left[\nabla_{\widetilde{\boldsymbol{V}}}\mathcal{L}(\widetilde{\boldsymbol{F}}_t) + \sigma_{\max}\widetilde{\boldsymbol{V}}_t(\widetilde{\boldsymbol{V}}_t^\top \widetilde{\boldsymbol{V}}_t - \sigma_{\max}^{1/2}\boldsymbol{I}_{r_2})\right],\\
\widetilde{\boldsymbol{W}}_{t+1} &= \widetilde{\boldsymbol{W}}_t - \widetilde{\eta}\left[\nabla_{\widetilde{\boldsymbol{W}}}\mathcal{L}(\widetilde{\boldsymbol{F}}_t) + \sigma_{\max}\widetilde{\boldsymbol{W}}_t(\widetilde{\boldsymbol{W}}_t^\top \widetilde{\boldsymbol{W}}_t - \sigma_{\max}^{1/2}\boldsymbol{I}_{r_3})\right],\\
\widetilde{\boldsymbol{\mathcal{S}}}_{t+1} &= \widetilde{\boldsymbol{\mathcal{S}}}_t - \widetilde{\eta}\nabla_{\widetilde{\boldsymbol{\mathcal{S}}}}\mathcal{L}(\widetilde{\boldsymbol{F}}_t).
\end{aligned}$$

By rescaling the factors as (so that $\boldsymbol{F}_0$ becomes the same initialization as ScaledGD)

$$\boldsymbol{F}_t = (\sigma_{\max}^{-1/4}\widetilde{\boldsymbol{U}}_t, \sigma_{\max}^{-1/4}\widetilde{\boldsymbol{V}}_t, \sigma_{\max}^{-1/4}\widetilde{\boldsymbol{W}}_t, \sigma_{\max}^{3/4}\widetilde{\boldsymbol{\mathcal{S}}}_t),$$

we can equivalently rewrite the update rule as

$$\begin{aligned}
\boldsymbol{U}_{t+1} &= \boldsymbol{U}_t - \widetilde{\eta}\sigma_{\max}^{3/2}\left[\sigma_{\max}^{-2}\nabla_{\boldsymbol{U}}\mathcal{L}(\boldsymbol{F}_t) + \boldsymbol{U}_t(\boldsymbol{U}_t^\top \boldsymbol{U}_t - \boldsymbol{I}_{r_1})\right],\\
\boldsymbol{V}_{t+1} &= \boldsymbol{V}_t - \widetilde{\eta}\sigma_{\max}^{3/2}\left[\sigma_{\max}^{-2}\nabla_{\boldsymbol{V}}\mathcal{L}(\boldsymbol{F}_t) + \boldsymbol{V}_t(\boldsymbol{V}_t^\top \boldsymbol{V}_t - \boldsymbol{I}_{r_2})\right],\\
\boldsymbol{W}_{t+1} &= \boldsymbol{W}_t - \widetilde{\eta}\sigma_{\max}^{3/2}\left[\sigma_{\max}^{-2}\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{F}_t) + \boldsymbol{W}_t(\boldsymbol{W}_t^\top \boldsymbol{W}_t - \boldsymbol{I}_{r_3})\right],\\
\boldsymbol{\mathcal{S}}_{t+1} &= \boldsymbol{\mathcal{S}}_t - \widetilde{\eta}\sigma_{\max}^{3/2}\nabla_{\boldsymbol{\mathcal{S}}}\mathcal{L}(\boldsymbol{F}_t).
\end{aligned}$$

Hence, by scaling the step size as $\eta = \widetilde{\eta}\sigma_{\max}^{3/2}$, we obtain the update rule (30).

B Proof for Tensor Factorization (Theorem 3)

We prove Theorem 3 via induction. Suppose that for some $t \ge 0$, one has $\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \le \epsilon\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$ for some sufficiently small ϵ whose size will be specified later in the proof. Our goal is to bound the scaled distance from the ground truth to the next iterate, i.e. $\mathrm{dist}(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star)$.

Since $\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \le \epsilon\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$, Lemma 6 ensures that the optimal alignment matrices $\{\boldsymbol{Q}_{t,k}\}_{k=1,2,3}$ between $\boldsymbol{F}_t$ and $\boldsymbol{F}_\star$ exist. Therefore, in view of the definition of $\mathrm{dist}(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star)$, one has

$$\begin{aligned}
\mathrm{dist}^2(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) \le{}& \|(\boldsymbol{U}_{t+1}\boldsymbol{Q}_{t,1} - \boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 + \|(\boldsymbol{V}_{t+1}\boldsymbol{Q}_{t,2} - \boldsymbol{V}_\star)\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}^2 + \|(\boldsymbol{W}_{t+1}\boldsymbol{Q}_{t,3} - \boldsymbol{W}_\star)\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}^2\\
&+ \left\|(\boldsymbol{Q}_{t,1}^{-1}, \boldsymbol{Q}_{t,2}^{-1}, \boldsymbol{Q}_{t,3}^{-1}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_{t+1} - \boldsymbol{\mathcal{S}}_\star\right\|_{\mathsf{F}}^2. \qquad (43)
\end{aligned}$$

To avoid notational clutter, we denote $\boldsymbol{F} \coloneqq (\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}, \boldsymbol{\mathcal{S}})$ with

$$\boldsymbol{U} \coloneqq \boldsymbol{U}_t\boldsymbol{Q}_{t,1}, \qquad \boldsymbol{V} \coloneqq \boldsymbol{V}_t\boldsymbol{Q}_{t,2}, \qquad \boldsymbol{W} \coloneqq \boldsymbol{W}_t\boldsymbol{Q}_{t,3}, \qquad \boldsymbol{\mathcal{S}} \coloneqq (\boldsymbol{Q}_{t,1}^{-1}, \boldsymbol{Q}_{t,2}^{-1}, \boldsymbol{Q}_{t,3}^{-1}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_t, \qquad (44)$$

and adopt the set of notation defined in (34) for the rest of the proof. Clearly, $\boldsymbol{F}$ is aligned with $\boldsymbol{F}_\star$.

We aim to establish the following bounds for the four terms in (43) as long as $\eta < 1$:

$$\begin{aligned}
\|(\boldsymbol{U}_{t+1}\boldsymbol{Q}_{t,1} - \boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 \le (1-\eta)^2\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 & \\
- 2\eta(1-\eta)\left\langle \boldsymbol{\mathcal{T}}_U, \boldsymbol{\mathcal{T}}_U + \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W\right\rangle + \eta^2 \left\|\boldsymbol{\mathcal{T}}_U + \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W\right\|_{\mathsf{F}}^2 & \\
+ 2\eta(1-\eta)C_1\epsilon \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star) + \eta^2 C_2\epsilon \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star); & \qquad (45\text{a})
\end{aligned}$$

$$\begin{aligned}
\|(\boldsymbol{V}_{t+1}\boldsymbol{Q}_{t,2} - \boldsymbol{V}_\star)\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}^2 \le (1-\eta)^2\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}^2 & \\
- 2\eta(1-\eta)\left\langle \boldsymbol{\mathcal{T}}_V, \boldsymbol{\mathcal{T}}_U + \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W\right\rangle + \eta^2 \left\|\boldsymbol{\mathcal{T}}_U + \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W\right\|_{\mathsf{F}}^2 & \\
+ 2\eta(1-\eta)C_1\epsilon \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star) + \eta^2 C_2\epsilon \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star); & \qquad (45\text{b})
\end{aligned}$$

$$\begin{aligned}
\|(\boldsymbol{W}_{t+1}\boldsymbol{Q}_{t,3} - \boldsymbol{W}_\star)\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}^2 \le (1-\eta)^2\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}^2 & \\
- 2\eta(1-\eta)\left\langle \boldsymbol{\mathcal{T}}_W, \boldsymbol{\mathcal{T}}_U + \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W\right\rangle + \eta^2 \left\|\boldsymbol{\mathcal{T}}_U + \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W\right\|_{\mathsf{F}}^2 & \\
+ 2\eta(1-\eta)C_1\epsilon \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star) + \eta^2 C_2\epsilon \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star); & \qquad (45\text{c})
\end{aligned}$$

$$\begin{aligned}
\left\|(\boldsymbol{Q}_{t,1}^{-1}, \boldsymbol{Q}_{t,2}^{-1}, \boldsymbol{Q}_{t,3}^{-1}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_{t+1} - \boldsymbol{\mathcal{S}}_\star\right\|_{\mathsf{F}}^2 \le (1-\eta)^2\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}^2 - \eta(2-5\eta)\left(\|\boldsymbol{D}_U\|_{\mathsf{F}}^2 + \|\boldsymbol{D}_V\|_{\mathsf{F}}^2 + \|\boldsymbol{D}_W\|_{\mathsf{F}}^2\right) & \\
+ 2\eta(1-\eta)C_1\epsilon \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star) + \eta^2 C_2\epsilon \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star), & \qquad (45\text{d})
\end{aligned}$$

where $C_1, C_2 > 1$ are two universal constants. Suppose for the moment that the four bounds (45) hold. We can then combine them all to deduce

$$\begin{aligned}
\operatorname{dist}^2(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) \le (1-\eta)^2\Big(\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}^2\Big) & \\
- \eta(2-5\eta)\left\|\boldsymbol{\mathcal{T}}_U + \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W\right\|_{\mathsf{F}}^2 - \eta(2-5\eta)\Big(\|\boldsymbol{D}_U\|_{\mathsf{F}}^2 + \|\boldsymbol{D}_V\|_{\mathsf{F}}^2 + \|\boldsymbol{D}_W\|_{\mathsf{F}}^2\Big) & \\
+ 2\eta(1-\eta)C\epsilon \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star) + \eta^2 C\epsilon \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star). & \qquad (46)
\end{aligned}$$

Here $C \coloneqq 4(C_1 \vee C_2)$. As long as $\eta \le 2/5$ and $\epsilon \le 0.2/C$, one has

$$\operatorname{dist}^2(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) \le \big((1-\eta)^2 + 2\eta(1-\eta)C\epsilon + \eta^2 C\epsilon\big) \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star) \le (1-0.7\eta)^2 \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star),$$

and therefore we arrive at the conclusion that $\operatorname{dist}(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) \le (1-0.7\eta)\operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$. In addition, the relation (38) in Lemma 10 guarantees that $\|(\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_t - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \le 3\operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$.

It then boils down to demonstrating the four bounds (45). Due to the symmetry among $\boldsymbol{U}, \boldsymbol{V}$ and $\boldsymbol{W}$, we will focus on proving the bounds (45a) and (45d), omitting the proofs for the other two.

Proof of the bound (45a). Utilize the ScaledGD update rule (26) to write

$$\begin{aligned}
(\boldsymbol{U}_{t+1}\boldsymbol{Q}_{t,1} - \boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1} &= \Big(\boldsymbol{U} - \eta\mathcal{M}_1\left((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_\star\right)\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1} - \boldsymbol{U}_\star\Big)\boldsymbol{\Sigma}_{\star,1} \\
&= (1-\eta)\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1} - \eta\boldsymbol{U}_\star(\breve{\boldsymbol{U}} - \breve{\boldsymbol{U}}_\star)^\top\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}, \qquad (47)
\end{aligned}$$

where we use the decomposition of the mode-1 matricization

$$\begin{aligned}
\mathcal{M}_1\left((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_\star\right) &= \boldsymbol{U}\mathcal{M}_1(\boldsymbol{\mathcal{S}})(\boldsymbol{V}\otimes\boldsymbol{W})^\top - \boldsymbol{U}_\star\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{V}_\star\otimes\boldsymbol{W}_\star)^\top \\
&= \boldsymbol{\Delta}_U\mathcal{M}_1(\boldsymbol{\mathcal{S}})(\boldsymbol{V}\otimes\boldsymbol{W})^\top + \boldsymbol{U}_\star\left(\mathcal{M}_1(\boldsymbol{\mathcal{S}})(\boldsymbol{V}\otimes\boldsymbol{W})^\top - \mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{V}_\star\otimes\boldsymbol{W}_\star)^\top\right) \\
&= \boldsymbol{\Delta}_U\breve{\boldsymbol{U}}^\top + \boldsymbol{U}_\star(\breve{\boldsymbol{U}} - \breve{\boldsymbol{U}}_\star)^\top.
\end{aligned}$$

Take the squared norm of both sides of the identity (47) to obtain

$$\begin{aligned}
\|(\boldsymbol{U}_{t+1}\boldsymbol{Q}_{t,1} - \boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 = (1-\eta)^2\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 - 2\eta(1-\eta)\underbrace{\left\langle\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}, \boldsymbol{U}_\star(\breve{\boldsymbol{U}} - \breve{\boldsymbol{U}}_\star)^\top\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}\right\rangle}_{=:\mathfrak{U}_1} & \\
+ \eta^2\underbrace{\left\|\boldsymbol{U}_\star(\breve{\boldsymbol{U}} - \breve{\boldsymbol{U}}_\star)^\top\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}\right\|_{\mathsf{F}}^2}_{=:\mathfrak{U}_2}. &
\end{aligned}$$

The following two claims bound the two terms $\mathfrak{U}_1$ and $\mathfrak{U}_2$, whose proofs can be found in Appendix B.1 and Appendix B.2, respectively.

Claim 1. $\mathfrak{U}_1 \geq \langle \boldsymbol{\mathcal{T}}_U, \boldsymbol{\mathcal{T}}_U + \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W \rangle - C_1 \epsilon \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star)$.

Claim 2. $\mathfrak{U}_2 \leq \|\boldsymbol{\mathcal{T}}_U + \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W\|_{\mathsf{F}}^2 + C_2 \epsilon \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star)$.

We can combine the above two claims to obtain that

$$\begin{aligned}\|(\boldsymbol{U}_{t+1}\boldsymbol{Q}_{t,1} - \boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 &\leq (1-\eta)^2 \|\boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 - 2\eta(1-\eta)\,\langle \boldsymbol{\mathcal{T}}_U, \boldsymbol{\mathcal{T}}_U + \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W \rangle \\ &\quad + \eta^2 \,\|\boldsymbol{\mathcal{T}}_U + \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W\|_{\mathsf{F}}^2 + 2\eta(1-\eta) C_1 \epsilon \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star) + \eta^2 C_2 \epsilon \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star),\end{aligned}$$

as long as $\eta < 1$. This proves the bound (45a).

Proof of the bound (45d). Again, we use the ScaledGD update rule (26) and the decomposition $\boldsymbol{\mathcal{S}} = \boldsymbol{\Delta}_{\mathcal{S}} + \boldsymbol{\mathcal{S}}_\star$ to obtain

$$\begin{aligned}&(\boldsymbol{Q}_{t,1}^{-1}, \boldsymbol{Q}_{t,2}^{-1}, \boldsymbol{Q}_{t,3}^{-1}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_{t+1} - \boldsymbol{\mathcal{S}}_\star \\ &\quad = \boldsymbol{\mathcal{S}} - \eta \left((\boldsymbol{U}^\top \boldsymbol{U})^{-1} \boldsymbol{U}^\top, (\boldsymbol{V}^\top \boldsymbol{V})^{-1} \boldsymbol{V}^\top, (\boldsymbol{W}^\top \boldsymbol{W})^{-1} \boldsymbol{W}^\top \right) \boldsymbol{\cdot} \left((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_\star \right) - \boldsymbol{\mathcal{S}}_\star \\ &\quad = (1-\eta) \boldsymbol{\Delta}_{\mathcal{S}} - \eta \left((\boldsymbol{U}^\top \boldsymbol{U})^{-1} \boldsymbol{U}^\top, (\boldsymbol{V}^\top \boldsymbol{V})^{-1} \boldsymbol{V}^\top, (\boldsymbol{W}^\top \boldsymbol{W})^{-1} \boldsymbol{W}^\top \right) \boldsymbol{\cdot} \left((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_\star - \boldsymbol{\mathcal{X}}_\star \right),\end{aligned} \tag{48}$$

where we used (6c) in the last line. Expand the squared norm of both sides to reach

$$\begin{aligned}&\left\|(\boldsymbol{Q}_{t,1}^{-1}, \boldsymbol{Q}_{t,2}^{-1}, \boldsymbol{Q}_{t,3}^{-1}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_{t+1} - \boldsymbol{\mathcal{S}}_\star\right\|_{\mathsf{F}}^2 = (1-\eta)^2 \|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}^2 \\ &\quad - 2\eta(1-\eta) \underbrace{\left\langle \boldsymbol{\Delta}_{\mathcal{S}}, \left((\boldsymbol{U}^\top \boldsymbol{U})^{-1} \boldsymbol{U}^\top, (\boldsymbol{V}^\top \boldsymbol{V})^{-1} \boldsymbol{V}^\top, (\boldsymbol{W}^\top \boldsymbol{W})^{-1} \boldsymbol{W}^\top \right) \boldsymbol{\cdot} \left((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_\star - \boldsymbol{\mathcal{X}}_\star \right) \right\rangle}_{=:\mathfrak{S}_1} \\ &\quad + \eta^2 \underbrace{\left\| \left((\boldsymbol{U}^\top \boldsymbol{U})^{-1} \boldsymbol{U}^\top, (\boldsymbol{V}^\top \boldsymbol{V})^{-1} \boldsymbol{V}^\top, (\boldsymbol{W}^\top \boldsymbol{W})^{-1} \boldsymbol{W}^\top \right) \boldsymbol{\cdot} \left((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_\star - \boldsymbol{\mathcal{X}}_\star \right) \right\|_{\mathsf{F}}^2}_{=:\mathfrak{S}_2}.\end{aligned}$$

We collect the bounds of the two relevant terms $\mathfrak{S}_1$ and $\mathfrak{S}_2$ in the following two claims, whose proofs can be found in Appendix B.3 and Appendix B.4, respectively.

Claim 3. $\mathfrak{S}_1 \geq \|\boldsymbol{D}_U\|_{\mathsf{F}}^2 + \|\boldsymbol{D}_V\|_{\mathsf{F}}^2 + \|\boldsymbol{D}_W\|_{\mathsf{F}}^2 - C_1 \epsilon \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star)$.

Claim 4. $\mathfrak{S}_2 \leq 3\left(\|\boldsymbol{D}_U\|_{\mathsf{F}}^2 + \|\boldsymbol{D}_V\|_{\mathsf{F}}^2 + \|\boldsymbol{D}_W\|_{\mathsf{F}}^2\right) + C_2 \epsilon \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star)$.

Take the bounds on $\mathfrak{S}_1$ and $\mathfrak{S}_2$ collectively to reach

$$\begin{aligned}\left\|(\boldsymbol{Q}_{t,1}^{-1}, \boldsymbol{Q}_{t,2}^{-1}, \boldsymbol{Q}_{t,3}^{-1}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_{t+1} - \boldsymbol{\mathcal{S}}_\star\right\|_{\mathsf{F}}^2 &\leq (1-\eta)^2 \|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}^2 - \eta(2-5\eta)\left(\|\boldsymbol{D}_U\|_{\mathsf{F}}^2 + \|\boldsymbol{D}_V\|_{\mathsf{F}}^2 + \|\boldsymbol{D}_W\|_{\mathsf{F}}^2\right) \\ &\quad + 2\eta(1-\eta) C_1 \epsilon \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star) + \eta^2 C_2 \epsilon \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star)\end{aligned} \tag{49}$$

as long as $\eta < 1$. This recovers the bound (45d).

B.1 Proof of Claim 1

Use the relation (39) to decompose $\mathfrak{U}_1$ as

$$\begin{aligned}\mathfrak{U}_1 &= \left\langle \boldsymbol{U}_\star^\top \boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}, (\breve{\boldsymbol{U}} - \breve{\boldsymbol{U}}_\star)^\top \breve{\boldsymbol{U}} (\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1} \boldsymbol{\Sigma}_{\star,1} \right\rangle \\ &= \underbrace{\left\langle \boldsymbol{U}_\star^\top \boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}, \mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{\Delta}_V \otimes \boldsymbol{W} + \boldsymbol{V}_\star \otimes \boldsymbol{\Delta}_W)^\top \breve{\boldsymbol{U}} (\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1} \boldsymbol{\Sigma}_{\star,1} \right\rangle}_{=:\mathfrak{U}_{1,1}} \\ &\quad + \underbrace{\left\langle \boldsymbol{U}_\star^\top \boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}, \mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})(\boldsymbol{V} \otimes \boldsymbol{W})^\top \breve{\boldsymbol{U}} (\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1} \boldsymbol{\Sigma}_{\star,1} \right\rangle}_{=:\mathfrak{U}_{1,2}}.\end{aligned}$$

In what follows, we bound $\mathfrak{U}_{1,1}$ and $\mathfrak{U}_{1,2}$ separately.

Step 1: tackling the term $\mathfrak{U}_{1,1}$. We can further decompose $\mathfrak{U}_{1,1}$ into the following four terms

$$
\begin{aligned}
\mathfrak{U}_{1,1} = &\underbrace{\left\langle \boldsymbol{U}_\star^\top \boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}, \mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{\Delta}_V \otimes \boldsymbol{W}_\star + \boldsymbol{V}_\star \otimes \boldsymbol{\Delta}_W)^\top \breve{\boldsymbol{U}}_\star \boldsymbol{\Sigma}_{\star,1}^{-1} \right\rangle}_{=:\mathfrak{U}_{1,1}^{\mathrm{m}}} \\
&+ \underbrace{\left\langle \boldsymbol{U}_\star^\top \boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}, \mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{\Delta}_V \otimes \boldsymbol{W}_\star)^\top \left(\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1} \boldsymbol{\Sigma}_{\star,1} - \breve{\boldsymbol{U}}_\star \boldsymbol{\Sigma}_{\star,1}^{-1} \right) \right\rangle}_{=:\mathfrak{U}_{1,1}^{\mathrm{p},1}} \\
&+ \underbrace{\left\langle \boldsymbol{U}_\star^\top \boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}, \mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{V}_\star \otimes \boldsymbol{\Delta}_W)^\top \left(\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1} \boldsymbol{\Sigma}_{\star,1} - \breve{\boldsymbol{U}}_\star \boldsymbol{\Sigma}_{\star,1}^{-1} \right) \right\rangle}_{=:\mathfrak{U}_{1,1}^{\mathrm{p},2}} \\
&+ \underbrace{\left\langle \boldsymbol{U}_\star^\top \boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}, \mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{\Delta}_V \otimes \boldsymbol{\Delta}_W)^\top \breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1} \boldsymbol{\Sigma}_{\star,1} \right\rangle}_{=:\mathfrak{U}_{1,1}^{\mathrm{p},3}},
\end{aligned}
$$

where $\mathfrak{U}_{1,1}^{\mathrm{m}}$ denotes the main term and the remaining ones are perturbation terms.

Utilizing the definition of $\breve{\boldsymbol{U}}_\star$ in (34) and the relation (12), the main term $\mathfrak{U}_{1,1}^{\mathrm{m}}$ can be rewritten as an inner product in the tensor space:

$$
\begin{aligned}
\mathfrak{U}_{1,1}^{\mathrm{m}} &= \left\langle \boldsymbol{U}_\star^\top \boldsymbol{\Delta}_U \mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star), \mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{\Delta}_V^\top \boldsymbol{V}_\star \otimes \boldsymbol{I}_{r_3} + \boldsymbol{I}_{r_2} \otimes \boldsymbol{\Delta}_W^\top \boldsymbol{W}_\star) \right\rangle \\
&= \left\langle \boldsymbol{\mathcal{T}}_U, \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W \right\rangle.
\end{aligned}
$$

To control the other three perturbation terms, Lemma 10 turns out to be extremely useful. For instance, the perturbation term $\mathfrak{U}_{1,1}^{\mathrm{p},1}$ is bounded by

$$
\begin{aligned}
|\mathfrak{U}_{1,1}^{\mathrm{p},1}| &\le \left\| \boldsymbol{U}_\star^\top \boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1} \right\|_{\mathsf{F}} \left\| \mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{\Delta}_V \otimes \boldsymbol{W}_\star)^\top \right\|_{\mathsf{F}} \left\| \breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1} \boldsymbol{\Sigma}_{\star,1} - \breve{\boldsymbol{U}}_\star \boldsymbol{\Sigma}_{\star,1}^{-1} \right\| \\
&\le \frac{\sqrt{2}(3\epsilon + 3\epsilon^2 + \epsilon^3)}{(1-\epsilon)^3} \|\boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}} \|\boldsymbol{\Delta}_V \boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}.
\end{aligned}
$$

Here in the last inequality, we used the upper bound (36g) and changed the matricization mode to obtain

$$
\left\| \mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{\Delta}_V \otimes \boldsymbol{W}_\star)^\top \right\|_{\mathsf{F}} = \|(\boldsymbol{I}_{r_1}, \boldsymbol{\Delta}_V, \boldsymbol{W}_\star) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_\star\|_{\mathsf{F}} = \left\| \boldsymbol{\Delta}_V \mathcal{M}_2(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{I}_{r_1} \otimes \boldsymbol{W}_\star)^\top \right\|_{\mathsf{F}} \le \|\boldsymbol{\Delta}_V \boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}.
$$

Similarly, the remaining two perturbation terms $\mathfrak{U}_{1,1}^{\mathrm{p},2}$ and $\mathfrak{U}_{1,1}^{\mathrm{p},3}$ obey

$$
\begin{aligned}
|\mathfrak{U}_{1,1}^{\mathrm{p},2}| &\le \frac{\sqrt{2}(3\epsilon + 3\epsilon^2 + \epsilon^3)}{(1-\epsilon)^3} \|\boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}} \|\boldsymbol{\Delta}_W \boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}, \\
|\mathfrak{U}_{1,1}^{\mathrm{p},3}| &\le \frac{\epsilon}{(1-\epsilon)^3} \|\boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}} \|\boldsymbol{\Delta}_V \boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}.
\end{aligned}
$$

Step 2: tackling the term $\mathfrak{U}_{1,2}$. Now we move on to the term $\mathfrak{U}_{1,2}$, which can be decomposed as

$$
\begin{aligned}
\mathfrak{U}_{1,2} = &\left\langle \boldsymbol{U}_\star^\top \boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}, \mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}}) \mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)^\top \boldsymbol{\Sigma}_{\star,1}^{-1} \right\rangle \\
&+ \underbrace{\left\langle \boldsymbol{U}_\star^\top \boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}, \mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})(\boldsymbol{V}_\star \otimes \boldsymbol{W}_\star)^\top \left(\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1} \boldsymbol{\Sigma}_{\star,1} - \breve{\boldsymbol{U}}_\star \boldsymbol{\Sigma}_{\star,1}^{-1} \right) \right\rangle}_{=:\mathfrak{U}_{1,2}^{\mathrm{p},1}} \\
&+ \underbrace{\left\langle \boldsymbol{U}_\star^\top \boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}, \mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})(\boldsymbol{V} \otimes \boldsymbol{W} - \boldsymbol{V}_\star \otimes \boldsymbol{W}_\star)^\top \breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1} \boldsymbol{\Sigma}_{\star,1} \right\rangle}_{=:\mathfrak{U}_{1,2}^{\mathrm{p},2}} \\
= &\left\langle \boldsymbol{U}_\star^\top \boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}, \mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}}) \mathcal{M}_1(\boldsymbol{\mathcal{S}})^\top \boldsymbol{\Sigma}_{\star,1}^{-1} \right\rangle \underbrace{- \left\langle \boldsymbol{U}_\star^\top \boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}, \mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}}) \mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})^\top \boldsymbol{\Sigma}_{\star,1}^{-1} \right\rangle}_{=:\mathfrak{U}_{1,2}^{\mathrm{p},3}} + \mathfrak{U}_{1,2}^{\mathrm{p},1} + \mathfrak{U}_{1,2}^{\mathrm{p},2}
\end{aligned}
$$

$$
\begin{aligned}
&= \big\langle \boldsymbol{U}_\star^\top \boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}, \boldsymbol{U}^\top \boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1} \big\rangle + \mathfrak{L}_{1,2}^{\mathrm{p},1} + \mathfrak{L}_{1,2}^{\mathrm{p},2} + \mathfrak{L}_{1,2}^{\mathrm{p},3} \\
&= \|\boldsymbol{\mathcal{T}}_U\|_{\mathsf{F}}^2 + \mathfrak{L}_{1,2}^{\mathrm{p},1} + \mathfrak{L}_{1,2}^{\mathrm{p},2} + \mathfrak{L}_{1,2}^{\mathrm{p},3} + \underbrace{\big\langle \boldsymbol{U}_\star^\top \boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}, \boldsymbol{\Delta}_U^\top \boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1} \big\rangle}_{=:\mathfrak{L}_{1,2}^{\mathrm{p},4}},
\end{aligned}
$$

where in the penultimate identity we have applied the identity (35) to replace $\mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})\mathcal{M}_1(\boldsymbol{\mathcal{S}})^\top$. Again, by Lemma 10, the perturbation term $\mathfrak{L}_{1,2}^{\mathrm{p},1}$ is bounded by

$$
\begin{aligned}
|\mathfrak{L}_{1,2}^{\mathrm{p},1}| &\le \big\|\boldsymbol{U}_\star^\top \boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}\big\|_{\mathsf{F}} \big\|\mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})(\boldsymbol{V}_\star \otimes \boldsymbol{W}_\star)^\top\big\|_{\mathsf{F}} \big\|\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1} - \breve{\boldsymbol{U}}_\star \boldsymbol{\Sigma}_{\star,1}^{-1}\big\| \\
&\le \frac{\sqrt{2}(3\epsilon + 3\epsilon^2 + \epsilon^3)}{(1-\epsilon)^3} \|\boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}} \|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}.
\end{aligned}
$$

In addition, the term $\mathfrak{L}_{1,2}^{\mathrm{p},2}$ is bounded by

$$
\begin{aligned}
|\mathfrak{L}_{1,2}^{\mathrm{p},2}| &\le \big\|\boldsymbol{U}_\star^\top \boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}\big\|_{\mathsf{F}} \|\mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})\|_{\mathsf{F}} \|\boldsymbol{V} \otimes \boldsymbol{W} - \boldsymbol{V}_\star \otimes \boldsymbol{W}_\star\| \big\|\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}\big\| \\
&\le \frac{2\epsilon + \epsilon^2}{(1-\epsilon)^3} \|\boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}} \|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}},
\end{aligned}
$$

where we have used

$$
\begin{aligned}
\|\boldsymbol{V} \otimes \boldsymbol{W} - \boldsymbol{V}_\star \otimes \boldsymbol{W}_\star\| &\le \|\boldsymbol{V}_\star \otimes \boldsymbol{\Delta}_W\| + \|\boldsymbol{\Delta}_V \otimes \boldsymbol{W}_\star\| + \|\boldsymbol{\Delta}_V \otimes \boldsymbol{\Delta}_W\| \\
&\le \|\boldsymbol{\Delta}_W\| + \|\boldsymbol{\Delta}_V\| + \|\boldsymbol{\Delta}_V\|\|\boldsymbol{\Delta}_W\| \le 2\epsilon + \epsilon^2.
\end{aligned}
$$

Following similar arguments (i.e. repeatedly using Lemma 10), we can bound $\mathfrak{L}_{1,2}^{\mathrm{p},3}$ and $\mathfrak{L}_{1,2}^{\mathrm{p},4}$ as

$$
\begin{aligned}
|\mathfrak{L}_{1,2}^{\mathrm{p},3}| &\le \big\|\boldsymbol{U}_\star^\top \boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}\big\|_{\mathsf{F}} \|\mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})\|_{\mathsf{F}} \big\|\mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})^\top \boldsymbol{\Sigma}_{\star,1}^{-1}\big\| \le \epsilon \|\boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}} \|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}; \\
|\mathfrak{L}_{1,2}^{\mathrm{p},4}| &\le \big\|\boldsymbol{U}_\star^\top \boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}\big\|_{\mathsf{F}} \|\boldsymbol{\Delta}_U\| \|\boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}} \le \epsilon \|\boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2.
\end{aligned}
$$

Step 3: putting the bound together. Combine these results on $\mathfrak{L}_{1,1}$ and $\mathfrak{L}_{1,2}$ to see

$$
\mathfrak{L}_1 = \langle \boldsymbol{\mathcal{T}}_U, \boldsymbol{\mathcal{T}}_U + \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W \rangle + \mathfrak{L}_1^{\mathrm{p}},
$$

where the perturbation term $\mathfrak{L}_1^{\mathrm{p}} := \sum_{i=1}^3 \mathfrak{L}_{1,1}^{\mathrm{p},i} + \sum_{i=1}^4 \mathfrak{L}_{1,2}^{\mathrm{p},i}$ obeys

$$
\begin{aligned}
|\mathfrak{L}_1^{\mathrm{p}}| \le \epsilon \|\boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}} \Big(&\|\boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}} + \frac{1 + \sqrt{2}(3 + 3\epsilon + \epsilon^2)}{(1-\epsilon)^3} \|\boldsymbol{\Delta}_V \boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + \frac{\sqrt{2}(3 + 3\epsilon + \epsilon^2)}{(1-\epsilon)^3} \|\boldsymbol{\Delta}_W \boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}} \\
&+ (1 + \frac{2 + \epsilon + \sqrt{2}(3 + 3\epsilon + \epsilon^2)}{(1-\epsilon)^3}) \|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}} \Big). \qquad (50)
\end{aligned}
$$

Using the Cauchy–Schwarz inequality, we can further simplify it as $|\mathfrak{L}_1^{p}| \le C_1 \epsilon \, \mathrm{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star)$ for some universal constant $C_1 > 1$.

B.2 Proof of Claim 2

Note that

$$
\begin{aligned}
\mathfrak{L}_2 &= \big\|(\breve{\boldsymbol{U}} - \breve{\boldsymbol{U}}_\star)^\top \breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1} \boldsymbol{\Sigma}_{\star,1}\big\|_{\mathsf{F}}^2 \\
&\le \big\|(\breve{\boldsymbol{U}} - \breve{\boldsymbol{U}}_\star)^\top \breve{\boldsymbol{U}} \boldsymbol{\Sigma}_{\star,1}^{-1}\big\|_{\mathsf{F}}^2 \big\|\boldsymbol{\Sigma}_{\star,1}(\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1} \boldsymbol{\Sigma}_{\star,1}\big\|^2 \\
&\le \big\|(\breve{\boldsymbol{U}} - \breve{\boldsymbol{U}}_\star)^\top \breve{\boldsymbol{U}} \boldsymbol{\Sigma}_{\star,1}^{-1}\big\|_{\mathsf{F}}^2 (1-\epsilon)^{-12}, \qquad (51)
\end{aligned}
$$

where the last relation arises from the bound (36h) in Lemma 10. We can then use the decomposition (39) to obtain

$$
\big\|(\breve{\boldsymbol{U}} - \breve{\boldsymbol{U}}_\star)^\top \breve{\boldsymbol{U}} \boldsymbol{\Sigma}_{\star,1}^{-1}\big\|_{\mathsf{F}} = \Big\|\Big(\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{\Delta}_V \otimes \boldsymbol{W} + \boldsymbol{V}_\star \otimes \boldsymbol{\Delta}_W)^\top + \mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})(\boldsymbol{V} \otimes \boldsymbol{W})^\top\Big)(\boldsymbol{V} \otimes \boldsymbol{W})\mathcal{M}_1(\boldsymbol{\mathcal{S}})^\top \boldsymbol{\Sigma}_{\star,1}^{-1}\Big\|_{\mathsf{F}}
$$

$$
\begin{aligned}
&\leq \underbrace{\left\|\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)\left(\boldsymbol{\Delta}_V^\top \boldsymbol{V}_\star \otimes \boldsymbol{I}_{r_3} + \boldsymbol{I}_{r_2} \otimes \boldsymbol{\Delta}_W^\top \boldsymbol{W}_\star\right)\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)^\top \boldsymbol{\Sigma}_{\star,1}^{-1} + \mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})\mathcal{M}_1(\boldsymbol{\mathcal{S}})^\top \boldsymbol{\Sigma}_{\star,1}^{-1}\right\|_{\mathsf{F}}}_{=:\mathfrak{U}_2^{\mathrm{m}}} \\
&\quad + \underbrace{\left\|\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)\left(\boldsymbol{\Delta}_V^\top \boldsymbol{V} \otimes \boldsymbol{W}^\top \boldsymbol{W} - \boldsymbol{\Delta}_V^\top \boldsymbol{V}_\star \otimes \boldsymbol{I}_{r_3}\right)\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)^\top \boldsymbol{\Sigma}_{\star,1}^{-1}\right\|_{\mathsf{F}}}_{=:\mathfrak{U}_2^{\mathrm{p},1}} \\
&\quad + \underbrace{\left\|\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)\left(\boldsymbol{V}_\star^\top \boldsymbol{V} \otimes \boldsymbol{\Delta}_W^\top \boldsymbol{W} - \boldsymbol{I}_{r_2} \otimes \boldsymbol{\Delta}_W^\top \boldsymbol{W}_\star\right)\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)^\top \boldsymbol{\Sigma}_{\star,1}^{-1}\right\|_{\mathsf{F}}}_{=:\mathfrak{U}_2^{\mathrm{p},2}} \\
&\quad + \underbrace{\left\|\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)\left(\boldsymbol{\Delta}_V^\top \boldsymbol{V} \otimes \boldsymbol{W}^\top \boldsymbol{W} + \boldsymbol{V}_\star^\top \boldsymbol{V} \otimes \boldsymbol{\Delta}_W^\top \boldsymbol{W}\right)\mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})^\top \boldsymbol{\Sigma}_{\star,1}^{-1}\right\|_{\mathsf{F}}}_{=:\mathfrak{U}_2^{\mathrm{p},3}} \\
&\quad + \underbrace{\left\|\mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})\left(\boldsymbol{V}^\top \boldsymbol{V} \otimes \boldsymbol{W}^\top \boldsymbol{W} - \boldsymbol{I}_{r_2} \otimes \boldsymbol{I}_{r_3}\right)\mathcal{M}_1(\boldsymbol{\mathcal{S}})^\top \boldsymbol{\Sigma}_{\star,1}^{-1}\right\|_{\mathsf{F}}}_{=:\mathfrak{U}_2^{\mathrm{p},4}}.
\end{aligned}
$$

Here, $\mathfrak{U}_2^{\mathrm{m}}$ is the main term while the remaining four are perturbation terms. Use the relation (35) again to replace $\mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})\mathcal{M}_1(\boldsymbol{\mathcal{S}})^\top$ in the main term $\mathfrak{U}_2^{\mathrm{m}}$ and see

$$
\begin{aligned}
\mathfrak{U}_2^{\mathrm{m}} &= \left\|\Big(\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{\Delta}_V^\top \boldsymbol{V}_\star \otimes \boldsymbol{I}_{r_3} + \boldsymbol{I}_{r_2} \otimes \boldsymbol{\Delta}_W^\top \boldsymbol{W}_\star) + \boldsymbol{U}_\star^\top \boldsymbol{\Delta}_U \mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)\Big)\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)^\top \boldsymbol{\Sigma}_{\star,1}^{-1}\right\|_{\mathsf{F}} \\
&\leq \left\|\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{\Delta}_V^\top \boldsymbol{V}_\star \otimes \boldsymbol{I}_{r_3} + \boldsymbol{I}_{r_2} \otimes \boldsymbol{\Delta}_W^\top \boldsymbol{W}_\star) + \boldsymbol{U}_\star^\top \boldsymbol{\Delta}_U \mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)\right\|_{\mathsf{F}} \left\|\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)^\top \boldsymbol{\Sigma}_{\star,1}^{-1}\right\| \\
&= \|\boldsymbol{\mathcal{T}}_U + \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W\|_{\mathsf{F}},
\end{aligned}
$$

where the last equality uses $\|\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)^\top \boldsymbol{\Sigma}_{\star,1}^{-1}\| = 1$. The perturbation terms are bounded by

$$
\begin{aligned}
\mathfrak{U}_2^{\mathrm{p},1} &\leq ((1+\epsilon)^3 - 1)\|\boldsymbol{\Delta}_V \boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}; \\
\mathfrak{U}_2^{\mathrm{p},2} &\leq ((1+\epsilon)^2 - 1)\|\boldsymbol{\Delta}_W \boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}; \\
\mathfrak{U}_2^{\mathrm{p},3} &\leq \epsilon(1+\epsilon)^3\|\boldsymbol{\Delta}_V \boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + \epsilon(1+\epsilon)^2\|\boldsymbol{\Delta}_W \boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}; \\
\mathfrak{U}_2^{\mathrm{p},4} &\leq ((1+\epsilon)^4 - 1)(1+\epsilon)\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}.
\end{aligned}
$$

They follow from similar calculations as those in bounding $\mathfrak{U}_1$ with the aid of Lemma 10; hence we omit the details for brevity. Combine these results to see

$$
\left\|(\breve{\boldsymbol{U}} - \breve{\boldsymbol{U}}_\star)^\top \breve{\boldsymbol{U}} \boldsymbol{\Sigma}_{\star,1}^{-1}\right\|_{\mathsf{F}} \leq \|\boldsymbol{\mathcal{T}}_U + \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W\|_{\mathsf{F}} + \mathfrak{U}_2^{\mathrm{p}},
$$

with $\mathfrak{U}_2^{\mathrm{p}} := \sum_{i=1}^4 \mathfrak{U}_2^{\mathrm{p},i}$ obeying

$$
\begin{aligned}
\mathfrak{U}_2^{\mathrm{p}} &\leq ((1+\epsilon)^4 - 1)\|\boldsymbol{\Delta}_V \boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + ((1+\epsilon)^3 - 1)\|\boldsymbol{\Delta}_W \boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}} + ((1+\epsilon)^4 - 1)(1+\epsilon)\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}} \\
&\lesssim \epsilon\big(\|\boldsymbol{\Delta}_V \boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_W \boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\big) \lesssim \epsilon \operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star).
\end{aligned}
$$

Next take the square to obtain

$$
\left\|(\breve{\boldsymbol{U}} - \breve{\boldsymbol{U}}_\star)^\top \breve{\boldsymbol{U}} \boldsymbol{\Sigma}_{\star,1}^{-1}\right\|_{\mathsf{F}}^2 \leq \|\boldsymbol{\mathcal{T}}_U + \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W\|_{\mathsf{F}}^2 + 2\mathfrak{U}_2^{\mathrm{p}}\, \|\boldsymbol{\mathcal{T}}_U + \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W\|_{\mathsf{F}} + (\mathfrak{U}_2^{\mathrm{p}})^2.
$$

Finally plug this back into (51) to conclude

$$
\begin{aligned}
\mathfrak{U}_2 &\leq (1-\epsilon)^{-12}\, \|\boldsymbol{\mathcal{T}}_U + \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W\|_{\mathsf{F}}^2 + 2(1-\epsilon)^{-12}\mathfrak{U}_2^{\mathrm{p}}\, \|\boldsymbol{\mathcal{T}}_U + \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W\|_{\mathsf{F}} + (1-\epsilon)^{-12}(\mathfrak{U}_2^{\mathrm{p}})^2 \\
&\leq \|\boldsymbol{\mathcal{T}}_U + \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W\|_{\mathsf{F}}^2 + \big((1-\epsilon)^{-12} - 1\big)\left(\|\boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_V \boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_W \boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}\right)^2 \\
&\qquad + 2(1-\epsilon)^{-12}\mathfrak{U}_2^{\mathrm{p}}\left(\|\boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_V \boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_W \boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}\right) + (1-\epsilon)^{-12}(\mathfrak{U}_2^{\mathrm{p}})^2 \\
&\leq \|\boldsymbol{\mathcal{T}}_U + \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W\|_{\mathsf{F}}^2 + C_2\epsilon \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star),
\end{aligned}
$$

for some universal constant $C_2 > 1$. Here in the second inequality, we use the fact that $\|\boldsymbol{\mathcal{T}}_U\|_{\mathsf{F}} \leq \|\boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}$, $\|\boldsymbol{\mathcal{T}}_V\|_{\mathsf{F}} \leq \|\boldsymbol{\Delta}_V \boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}$, and $\|\boldsymbol{\mathcal{T}}_W\|_{\mathsf{F}} \leq \|\boldsymbol{\Delta}_W \boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}$. This finishes the proof of the claim.

B.3 Proof of Claim 3

Use the decomposition

$$(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\cdot\boldsymbol{\mathcal{S}}_\star-\boldsymbol{\mathcal{X}}_\star=(\boldsymbol{\Delta}_U,\boldsymbol{V},\boldsymbol{W})\cdot\boldsymbol{\mathcal{S}}_\star+(\boldsymbol{U}_\star,\boldsymbol{\Delta}_V,\boldsymbol{W})\cdot\boldsymbol{\mathcal{S}}_\star+(\boldsymbol{U}_\star,\boldsymbol{V}_\star,\boldsymbol{\Delta}_W)\cdot\boldsymbol{\mathcal{S}}_\star \tag{52}$$

to rewrite $\mathfrak{S}_1$ as

$$\begin{aligned}\mathfrak{S}_1&=\underbrace{\left\langle\boldsymbol{\Delta}_{\mathcal{S}},((\boldsymbol{U}^\top\boldsymbol{U})^{-1}\boldsymbol{U}^\top\boldsymbol{\Delta}_U,\boldsymbol{I}_{r_2},\boldsymbol{I}_{r_3})\cdot\boldsymbol{\mathcal{S}}_\star\right\rangle}_{=:\mathfrak{S}_{1,1}}+\underbrace{\left\langle\boldsymbol{\Delta}_{\mathcal{S}},((\boldsymbol{U}^\top\boldsymbol{U})^{-1}\boldsymbol{U}^\top\boldsymbol{U}_\star,(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\boldsymbol{V}^\top\boldsymbol{\Delta}_V,\boldsymbol{I}_{r_3})\cdot\boldsymbol{\mathcal{S}}_\star\right\rangle}_{=:\mathfrak{S}_{1,2}}\\&\quad+\underbrace{\left\langle\boldsymbol{\Delta}_{\mathcal{S}},((\boldsymbol{U}^\top\boldsymbol{U})^{-1}\boldsymbol{U}^\top\boldsymbol{U}_\star,(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\boldsymbol{V}^\top\boldsymbol{V}_\star,(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{\Delta}_W)\cdot\boldsymbol{\mathcal{S}}_\star\right\rangle}_{=:\mathfrak{S}_{1,3}}.\end{aligned}$$

Step 1: tackling the term $\mathfrak{S}_{1,1}$. Translating the inner product from the tensor space to the matrix space via the mode-1 matricization yields

$$\begin{aligned}\mathfrak{S}_{1,1}&=\left\langle\mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}}),(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\boldsymbol{U}^\top\boldsymbol{\Delta}_U\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)\right\rangle\\&=\underbrace{\left\langle\mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}}),(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\boldsymbol{U}^\top\boldsymbol{\Delta}_U\mathcal{M}_1(\boldsymbol{\mathcal{S}})\right\rangle}_{=:\mathfrak{S}^{\mathrm{m}}_{1,1}}-\underbrace{\left\langle\mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}}),(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\boldsymbol{U}^\top\boldsymbol{\Delta}_U\mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})\right\rangle}_{=:\mathfrak{S}^{\mathrm{p}}_{1,1}}.\end{aligned}$$

Again, the identity (35) is helpful in characterizing the main term $\mathfrak{S}^{\mathrm{m}}_{1,1}$:

$$\mathfrak{S}^{\mathrm{m}}_{1,1}=\left\langle\boldsymbol{U}^\top\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}^2,(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\boldsymbol{U}^\top\boldsymbol{\Delta}_U\right\rangle=\left\|(\boldsymbol{U}^\top\boldsymbol{U})^{-1/2}\boldsymbol{U}^\top\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\right\|_{\mathsf{F}}^2.$$

The perturbation term $\mathfrak{S}^{\mathrm{p}}_{1,1}$ is bounded by

$$|\mathfrak{S}^{\mathrm{p}}_{1,1}|\le\|\mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})\|_{\mathsf{F}}\left\|\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\right\|\|\boldsymbol{\Delta}_U\|\|\mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})\|_{\mathsf{F}}\le\epsilon(1-\epsilon)^{-1}\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}^2,$$

which follows directly from Lemma 10.

Step 2: tackling the term $\mathfrak{S}_{1,2}$. Following the same recipe as above, we can apply the mode-2 matricization to $\mathfrak{S}_{1,2}$ to see

$$\begin{aligned}\mathfrak{S}_{1,2}&=\left\langle\mathcal{M}_2(\boldsymbol{\Delta}_{\mathcal{S}}),(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\boldsymbol{V}^\top\boldsymbol{\Delta}_V\mathcal{M}_2(\boldsymbol{\mathcal{S}}_\star)\left(\boldsymbol{U}_\star^\top\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\otimes\boldsymbol{I}_{r_3}\right)\right\rangle\\&=\underbrace{\left\langle\mathcal{M}_2(\boldsymbol{\Delta}_{\mathcal{S}}),(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\boldsymbol{V}^\top\boldsymbol{\Delta}_V\mathcal{M}_2(\boldsymbol{\mathcal{S}})\right\rangle}_{=:\mathfrak{S}^{\mathrm{m}}_{1,2}}-\underbrace{\left\langle\mathcal{M}_2(\boldsymbol{\Delta}_{\mathcal{S}}),(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\boldsymbol{V}^\top\boldsymbol{\Delta}_V\mathcal{M}_2(\boldsymbol{\Delta}_{\mathcal{S}})\right\rangle}_{=:\mathfrak{S}^{\mathrm{p},1}_{1,2}}\\&\quad+\underbrace{\left\langle\mathcal{M}_2(\boldsymbol{\Delta}_{\mathcal{S}}),(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\boldsymbol{V}^\top\boldsymbol{\Delta}_V\mathcal{M}_2(\boldsymbol{\mathcal{S}}_\star)\left((\boldsymbol{U}_\star^\top\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}-\boldsymbol{I}_{r_1})\otimes\boldsymbol{I}_{r_3}\right)\right\rangle}_{=:\mathfrak{S}^{\mathrm{p},2}_{1,2}}.\end{aligned}$$

In view of the relation (35), we can rewrite the main term $\mathfrak{S}^{\mathrm{m}}_{1,2}$ as

$$\mathfrak{S}^{\mathrm{m}}_{1,2}=\left\|(\boldsymbol{V}^\top\boldsymbol{V})^{-1/2}\boldsymbol{V}^\top\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\right\|_{\mathsf{F}}^2.$$

In addition, for the perturbation terms, Lemma 10 allows us to obtain

$$|\mathfrak{S}^{\mathrm{p},1}_{1,2}|\le\|\mathcal{M}_2(\boldsymbol{\Delta}_{\mathcal{S}})\|_{\mathsf{F}}\left\|\boldsymbol{V}(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\right\|\|\boldsymbol{\Delta}_V\|\|\mathcal{M}_2(\boldsymbol{\Delta}_{\mathcal{S}})\|_{\mathsf{F}}\le\epsilon(1-\epsilon)^{-1}\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}^2.$$

Moreover, we can write $\boldsymbol{U}_\star^\top\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}-\boldsymbol{I}_{r_1}=\boldsymbol{U}_\star^\top(\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}-\boldsymbol{U}_\star)$, and bound $\mathfrak{S}^{\mathrm{p},2}_{1,2}$ as

$$\begin{aligned}|\mathfrak{S}^{\mathrm{p},2}_{1,2}|&\le\|\mathcal{M}_2(\boldsymbol{\Delta}_{\mathcal{S}})\|_{\mathsf{F}}\|\boldsymbol{V}(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\|\|\boldsymbol{\Delta}_V\mathcal{M}_2(\boldsymbol{\mathcal{S}}_\star)\|_{\mathsf{F}}\|\boldsymbol{U}_\star\|\|\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}-\boldsymbol{U}_\star\|\\&\le\sqrt{2}\epsilon(1-\epsilon)^{-2}\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}.\end{aligned}$$

Step 3: tackling the term $\mathfrak{S}_{1,3}$. Similar to before, we rewrite $\mathfrak{S}_{1,3}$ by applying the mode-3 matricization as

$$\begin{aligned}\mathfrak{S}_{1,3} &= \left\langle \mathcal{M}_3(\boldsymbol{\Delta}_{\mathcal{S}}), (\boldsymbol{W}^\top \boldsymbol{W})^{-1}\boldsymbol{W}^\top \boldsymbol{\Delta}_W \mathcal{M}_3(\boldsymbol{\mathcal{S}}_\star)\left(\boldsymbol{U}_\star^\top \boldsymbol{U}(\boldsymbol{U}^\top \boldsymbol{U})^{-1} \otimes \boldsymbol{V}_\star^\top \boldsymbol{V}(\boldsymbol{V}^\top \boldsymbol{V})^{-1}\right)\right\rangle \\ &= \underbrace{\left\langle \mathcal{M}_3(\boldsymbol{\Delta}_{\mathcal{S}}), (\boldsymbol{W}^\top \boldsymbol{W})^{-1}\boldsymbol{W}^\top \boldsymbol{\Delta}_W \mathcal{M}_3(\boldsymbol{\mathcal{S}})\right\rangle}_{=:\mathfrak{S}_{1,3}^{\mathrm{m}}} - \underbrace{\left\langle \mathcal{M}_3(\boldsymbol{\Delta}_{\mathcal{S}}), (\boldsymbol{W}^\top \boldsymbol{W})^{-1}\boldsymbol{W}^\top \boldsymbol{\Delta}_W \mathcal{M}_3(\boldsymbol{\Delta}_{\mathcal{S}})\right\rangle}_{=:\mathfrak{S}_{1,3}^{\mathrm{p},1}} \\ &\quad + \underbrace{\left\langle \mathcal{M}_3(\boldsymbol{\Delta}_{\mathcal{S}}), (\boldsymbol{W}^\top \boldsymbol{W})^{-1}\boldsymbol{W}^\top \boldsymbol{\Delta}_W \mathcal{M}_3(\boldsymbol{\mathcal{S}}_\star)\left(\boldsymbol{U}_\star^\top \boldsymbol{U}(\boldsymbol{U}^\top \boldsymbol{U})^{-1} \otimes \boldsymbol{V}_\star^\top \boldsymbol{V}(\boldsymbol{V}^\top \boldsymbol{V})^{-1} - \boldsymbol{I}_{r_1} \otimes \boldsymbol{I}_{r_2}\right)\right\rangle}_{=:\mathfrak{S}_{1,3}^{\mathrm{p},2}}.\end{aligned}$$

The main term obeys (thanks again to the identity (35))

$$\mathfrak{S}_{1,3}^{\mathrm{m}} = \left\|(\boldsymbol{W}^\top \boldsymbol{W})^{-1/2}\boldsymbol{W}^\top \boldsymbol{\Delta}_W \boldsymbol{\Sigma}_{\star,3}\right\|_{\mathsf{F}}^2.$$

As the same time, the perturbation term $\mathfrak{S}_{1,3}^{\mathrm{p},1}$ can be bounded by

$$|\mathfrak{S}_{1,3}^{\mathrm{p},1}| \le \|\mathcal{M}_3(\boldsymbol{\Delta}_{\mathcal{S}})\|_{\mathsf{F}} \left\|\boldsymbol{W}(\boldsymbol{W}^\top \boldsymbol{W})^{-1}\right\| \|\boldsymbol{\Delta}_W\| \|\mathcal{M}_3(\boldsymbol{\Delta}_{\mathcal{S}})\|_{\mathsf{F}} \le \epsilon(1-\epsilon)^{-1}\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}^2.$$

Similarly, we have

$$\begin{aligned}|\mathfrak{S}_{1,3}^{\mathrm{p},2}| &\le \|\mathcal{M}_3(\boldsymbol{\Delta}_{\mathcal{S}})\|_{\mathsf{F}} \|\boldsymbol{W}(\boldsymbol{W}^\top \boldsymbol{W})^{-1}\| \|\boldsymbol{\Delta}_W \mathcal{M}_3(\boldsymbol{\mathcal{S}}_\star)\|_{\mathsf{F}} \left\|\boldsymbol{U}_\star^\top \boldsymbol{U}(\boldsymbol{U}^\top \boldsymbol{U})^{-1} \otimes \boldsymbol{V}_\star^\top \boldsymbol{V}(\boldsymbol{V}^\top \boldsymbol{V})^{-1} - \boldsymbol{I}_{r_1} \otimes \boldsymbol{I}_{r_2}\right\| \\ &\le \sqrt{2}\epsilon((1-\epsilon)^{-2} + (1-\epsilon)^{-3})\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}} \|\boldsymbol{\Delta}_W \boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}},\end{aligned}$$

where we use the decomposition

$$\begin{aligned}&\boldsymbol{U}_\star^\top \boldsymbol{U}(\boldsymbol{U}^\top \boldsymbol{U})^{-1} \otimes \boldsymbol{V}_\star^\top \boldsymbol{V}(\boldsymbol{V}^\top \boldsymbol{V})^{-1} - \boldsymbol{I}_{r_1} \otimes \boldsymbol{I}_{r_2} \\ &\qquad = \boldsymbol{U}_\star^\top(\boldsymbol{U}(\boldsymbol{U}^\top \boldsymbol{U})^{-1} - \boldsymbol{U}_\star) \otimes \boldsymbol{V}_\star^\top \boldsymbol{V}(\boldsymbol{V}^\top \boldsymbol{V})^{-1} + \boldsymbol{I}_{r_1} \otimes \boldsymbol{V}_\star^\top(\boldsymbol{V}(\boldsymbol{V}^\top \boldsymbol{V})^{-1} - \boldsymbol{V}_\star)\end{aligned}$$

and its immediate consequence

$$\begin{aligned}&\left\|\boldsymbol{U}_\star^\top \boldsymbol{U}(\boldsymbol{U}^\top \boldsymbol{U})^{-1} \otimes \boldsymbol{V}_\star^\top \boldsymbol{V}(\boldsymbol{V}^\top \boldsymbol{V})^{-1} - \boldsymbol{I}_{r_1} \otimes \boldsymbol{I}_{r_2}\right\| \\ &\qquad \le \left\|\boldsymbol{U}(\boldsymbol{U}^\top \boldsymbol{U})^{-1} - \boldsymbol{U}_\star\right\| \left\|\boldsymbol{V}(\boldsymbol{V}^\top \boldsymbol{V})^{-1}\right\| + \left\|\boldsymbol{V}(\boldsymbol{V}^\top \boldsymbol{V})^{-1} - \boldsymbol{V}_\star\right\| \le \sqrt{2}\epsilon(1-\epsilon)^{-2} + \sqrt{2}\epsilon(1-\epsilon)^{-1}.\end{aligned}$$

Step 4: putting all pieces together. Denote $\mathfrak{S}_1^{\mathrm{p}}$ as the sum of the perturbation terms in $\mathfrak{S}_1$. Combine results of $\mathfrak{S}_{1,1}, \mathfrak{S}_{1,2}, \mathfrak{S}_{1,3}$ to see

$$\mathfrak{S}_1 = \left\|(\boldsymbol{U}^\top \boldsymbol{U})^{-1/2}\boldsymbol{U}^\top \boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{V}^\top \boldsymbol{V})^{-1/2}\boldsymbol{V}^\top \boldsymbol{\Delta}_V \boldsymbol{\Sigma}_{\star,2}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{W}^\top \boldsymbol{W})^{-1/2}\boldsymbol{W}^\top \boldsymbol{\Delta}_W \boldsymbol{\Sigma}_{\star,3}\right\|_{\mathsf{F}}^2 + \mathfrak{S}_{1,p},$$

where the perturbation term $\mathfrak{S}_1^{\mathrm{p}}$ obeys

$$|\mathfrak{S}_1^{\mathrm{p}}| \le \epsilon\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}} \left(\sqrt{2}(1-\epsilon)^{-2}\|\boldsymbol{\Delta}_V \boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + \sqrt{2}((1-\epsilon)^{-2} + (1-\epsilon)^{-3})\|\boldsymbol{\Delta}_W \boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}} + 3(1-\epsilon)^{-1}\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\right).$$

It is straightforward to check that $|\mathfrak{S}_{1,p}| \le C_1\epsilon\,\mathrm{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star)$ for some absolute constant $C_1 > 1$.

B.4 Proof of Claim 4

Reuse the decomposition (52) and the elementary inequality $(a+b+c)^2 \le 3(a^2+b^2+c^2)$ to obtain

$$\begin{aligned}\mathfrak{S}_2 \le 3\underbrace{\left\|((\boldsymbol{U}^\top \boldsymbol{U})^{-1}\boldsymbol{U}^\top \boldsymbol{\Delta}_U, \boldsymbol{I}_{r_2}, \boldsymbol{I}_{r_3}) \cdot \boldsymbol{\mathcal{S}}_\star\right\|_{\mathsf{F}}^2}_{=:\mathfrak{S}_{2,1}} &+ 3\underbrace{\left\|((\boldsymbol{U}^\top \boldsymbol{U})^{-1}\boldsymbol{U}^\top \boldsymbol{U}_\star, (\boldsymbol{V}^\top \boldsymbol{V})^{-1}\boldsymbol{V}^\top \boldsymbol{\Delta}_V, \boldsymbol{I}_{r_3}) \cdot \boldsymbol{\mathcal{S}}_\star\right\|_{\mathsf{F}}^2}_{=:\mathfrak{S}_{2,2}} \\ + 3\underbrace{\left\|((\boldsymbol{U}^\top \boldsymbol{U})^{-1}\boldsymbol{U}^\top \boldsymbol{U}_\star, (\boldsymbol{V}^\top \boldsymbol{V})^{-1}\boldsymbol{V}^\top \boldsymbol{V}_\star, (\boldsymbol{W}^\top \boldsymbol{W})^{-1}\boldsymbol{W}^\top \boldsymbol{\Delta}_W) \cdot \boldsymbol{\mathcal{S}}_\star\right\|_{\mathsf{F}}^2}_{=:\mathfrak{S}_{2,3}}&.\end{aligned}$$

Apply the mode-1 matricization and Lemma 10 to $\mathfrak{S}_{2,1}$ to see

$$\begin{aligned}\mathfrak{S}_{2,1} &= \left\|(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\boldsymbol{U}^\top\boldsymbol{\Delta}_U\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)\right\|_{\mathsf{F}}^2 \\ &\le \|(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\|\left\|(\boldsymbol{U}^\top\boldsymbol{U})^{-1/2}\boldsymbol{U}^\top\boldsymbol{\Delta}_U\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)\right\|_{\mathsf{F}}^2 \\ &\le (1-\epsilon)^{-2}\left\|(\boldsymbol{U}^\top\boldsymbol{U})^{-1/2}\boldsymbol{U}^\top\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\right\|_{\mathsf{F}}^2.\end{aligned}$$

Similarly, apply the mode-2 (resp. mode-3) matricization to $\mathfrak{S}_{2,2}$ (resp. $\mathfrak{S}_{2,3}$) to see

$$\begin{aligned}\mathfrak{S}_{2,2} &= \left\|(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\boldsymbol{V}^\top\boldsymbol{\Delta}_V\mathcal{M}_2(\boldsymbol{\mathcal{S}}_\star)\left(\boldsymbol{U}_\star^\top\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\otimes\boldsymbol{I}_{r_3}\right)\right\|_{\mathsf{F}}^2 \\ &\le \|(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\|\left\|(\boldsymbol{V}^\top\boldsymbol{V})^{-1/2}\boldsymbol{V}^\top\boldsymbol{\Delta}_V\mathcal{M}_2(\boldsymbol{\mathcal{S}}_\star)\right\|_{\mathsf{F}}^2\|\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\|^2 \\ &\le (1-\epsilon)^{-4}\left\|(\boldsymbol{V}^\top\boldsymbol{V})^{-1/2}\boldsymbol{V}^\top\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\right\|_{\mathsf{F}}^2,\end{aligned}$$

and

$$\begin{aligned}\mathfrak{S}_{2,3} &= \left\|(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{\Delta}_W\mathcal{M}_3(\boldsymbol{\mathcal{S}}_\star)\left(\boldsymbol{U}_\star^\top\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\otimes\boldsymbol{V}_\star^\top\boldsymbol{V}(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\right)\right\|_{\mathsf{F}}^2 \\ &\le \|(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\|\left\|(\boldsymbol{W}^\top\boldsymbol{W})^{-1/2}\boldsymbol{W}^\top\boldsymbol{\Delta}_W\mathcal{M}_3(\boldsymbol{\mathcal{S}}_\star)\right\|_{\mathsf{F}}^2\|\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\|^2\|\boldsymbol{V}(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\|^2 \\ &\le (1-\epsilon)^{-6}\left\|(\boldsymbol{W}^\top\boldsymbol{W})^{-1/2}\boldsymbol{W}^\top\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\right\|_{\mathsf{F}}^2.\end{aligned}$$

Combine the bounds on $\mathfrak{S}_{2,1}, \mathfrak{S}_{2,2}, \mathfrak{S}_{2,3}$ to write $\mathfrak{S}_2$ as

$$\begin{aligned}\mathfrak{S}_2 \le{}& 3(1-\epsilon)^{-2}\left\|(\boldsymbol{U}^\top\boldsymbol{U})^{-1/2}\boldsymbol{U}^\top\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\right\|_{\mathsf{F}}^2 + 3(1-\epsilon)^{-4}\left\|(\boldsymbol{V}^\top\boldsymbol{V})^{-1/2}\boldsymbol{V}^\top\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\right\|_{\mathsf{F}}^2 \\ &+ 3(1-\epsilon)^{-6}\left\|(\boldsymbol{W}^\top\boldsymbol{W})^{-1/2}\boldsymbol{W}^\top\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\right\|_{\mathsf{F}}^2.\end{aligned}$$

By symmetry, one can permute $\boldsymbol{\Delta}_U, \boldsymbol{\Delta}_V, \boldsymbol{\Delta}_W$, and take the average to balance their coefficients and reach the conclusion that

$$\mathfrak{S}_2 \le 3\left(\left\|(\boldsymbol{U}^\top\boldsymbol{U})^{-1/2}\boldsymbol{U}^\top\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{V}^\top\boldsymbol{V})^{-1/2}\boldsymbol{V}^\top\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{W}^\top\boldsymbol{W})^{-1/2}\boldsymbol{W}^\top\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\right\|_{\mathsf{F}}^2\right) + \mathfrak{S}_2^{\mathrm{p}},$$

where the perturbation term $\mathfrak{S}_2^{\mathrm{p}}$ obeys

$$\mathfrak{S}_2^{\mathrm{p}} \le \left((1-\epsilon)^{-2} + (1-\epsilon)^{-4} + (1-\epsilon)^{-6} - 3\right)\left(\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}^2\right).$$

A bit simplification yields $\mathfrak{S}_2^{\mathrm{p}} \le C_2\epsilon\,\mathrm{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star)$.

C Proof for Tensor Completion

This section is devoted to the proofs of claims related to tensor completion. To begin with, we state a perturbation bound that will be repeatedly used throughout this section.

Lemma 11. *Suppose that $\boldsymbol{\mathcal{X}}_\star$ is μ-incoherent, and $\boldsymbol{F} = (\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}, \boldsymbol{\mathcal{S}})$ satisfies $\mathrm{dist}(\boldsymbol{F}, \boldsymbol{F}_\star) \le \epsilon\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$ for $\epsilon < 1$ and the incoherence condition (29). Then one has the following perturbation bounds regarding the $\ell_{2,\infty}$ norm:*

$$\sqrt{n_1}\|\boldsymbol{U}\mathcal{M}_1(\boldsymbol{\mathcal{S}})\|_{2,\infty} \le (1-\epsilon)^{-2}C_B\sqrt{\mu r}\sigma_{\max}(\boldsymbol{\mathcal{X}}_\star); \tag{53a}$$

$$\sqrt{n_1}\|\boldsymbol{U}\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)\|_{2,\infty} = \sqrt{n_1}\|\boldsymbol{U}\boldsymbol{\Sigma}_{\star,1}\|_{2,\infty} \le (1-\epsilon)^{-3}C_B\sqrt{\mu r}\sigma_{\max}(\boldsymbol{\mathcal{X}}_\star); \tag{53b}$$

$$\sqrt{n_1}\|\boldsymbol{U}\|_{2,\infty} \le (1-\epsilon)^{-3}C_B\kappa\sqrt{\mu r}. \tag{53c}$$

By symmetry, a corresponding set of bounds hold for $\boldsymbol{V}, \breve{\boldsymbol{V}}$ and $\boldsymbol{W}, \breve{\boldsymbol{W}}$.

Proof. For (53a), we have

$$\|\boldsymbol{U}\mathcal{M}_1(\boldsymbol{\mathcal{S}})\|_{2,\infty} = \left\|\boldsymbol{U}\breve{\boldsymbol{U}}^\top\left(\boldsymbol{V}(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\otimes\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\right)\right\|_{2,\infty}$$

$$
\begin{aligned}
&\leq \|\boldsymbol{U}\breve{\boldsymbol{U}}^\top\|_{2,\infty} \left\|\boldsymbol{V}(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\right\| \left\|\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\right\| \\
&\leq \|\boldsymbol{U}\breve{\boldsymbol{U}}^\top\|_{2,\infty}(1-\epsilon)^{-2},
\end{aligned}
$$

where the first line uses (40), the second line follows from $\|\boldsymbol{A}\boldsymbol{B}\|_{2,\infty} \leq \|\boldsymbol{A}\|_{2,\infty}\|\boldsymbol{B}\|$, and the last inequality uses (36c). This combined with condition (29) leads to the declared bound.

Similarly for (53b), we have

$$
\begin{aligned}
\|\boldsymbol{U}\boldsymbol{\Sigma}_{\star,1}\|_{2,\infty} &= \left\|\boldsymbol{U}\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}\right\|_{2,\infty} \\
&\leq \|\boldsymbol{U}\breve{\boldsymbol{U}}^\top\|_{2,\infty}\left\|\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}\right\| \\
&\leq \|\boldsymbol{U}\breve{\boldsymbol{U}}^\top\|_{2,\infty}(1-\epsilon)^{-3},
\end{aligned}
$$

where the last line follows from (36f).

Finally, observe that

$$
\|\boldsymbol{U}\boldsymbol{\Sigma}_{\star,1}\|_{2,\infty} \geq \|\boldsymbol{U}\|_{2,\infty}\sigma_{\min}(\boldsymbol{\Sigma}_{\star,1}) \geq \|\boldsymbol{U}\|_{2,\infty}\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star).
$$

Combining the above inequality with (53b), we reach the bound (53c). □

C.1 Proof of Lemma 2

A crucial operation, which aims to preserve the desirable incoherence property with respect to the scaled distance, is the scaled projection $\boldsymbol{F} = \mathcal{P}_B(\boldsymbol{F}_+)$ defined in (19). For the purpose of understanding, it is instructive to view $\boldsymbol{F}$ as the solution to the following optimization problems:

$$
\begin{aligned}
\boldsymbol{U} &= \underset{\boldsymbol{U}}{\operatorname{argmin}} \ \left\|(\boldsymbol{U}-\boldsymbol{U}_+)\breve{\boldsymbol{U}}_+^\top\right\|_{\mathsf{F}}^2 \qquad \text{s.t.} \quad \sqrt{n_1}\|\boldsymbol{U}\breve{\boldsymbol{U}}_+^\top\|_{2,\infty} \leq B, \\
\boldsymbol{V} &= \underset{\boldsymbol{V}}{\operatorname{argmin}} \ \left\|(\boldsymbol{V}-\boldsymbol{V}_+)\breve{\boldsymbol{V}}_+^\top\right\|_{\mathsf{F}}^2 \qquad \text{s.t.} \quad \sqrt{n_2}\|\boldsymbol{V}\breve{\boldsymbol{V}}_+^\top\|_{2,\infty} \leq B, \\
\boldsymbol{W} &= \underset{\boldsymbol{W}}{\operatorname{argmin}} \ \left\|(\boldsymbol{W}-\boldsymbol{W}_+)\breve{\boldsymbol{W}}_+^\top\right\|_{\mathsf{F}}^2 \qquad \text{s.t.} \quad \sqrt{n_3}\|\boldsymbol{W}\breve{\boldsymbol{W}}_+^\top\|_{2,\infty} \leq B.
\end{aligned}
\tag{54}
$$

The remaining proof follows similar arguments as [TMC20]. To begin, we collect a useful claim as follows.

Claim 5 ([TMC20, Claim 5])**.** *For vectors $\boldsymbol{u}, \boldsymbol{u}_\star \in \mathbb{R}^n$ and $\lambda \geq \|\boldsymbol{u}_\star\|_2/\|\boldsymbol{u}\|_2$, it holds that*

$$
\|(1\wedge\lambda)\boldsymbol{u}-\boldsymbol{u}_\star\|_2 \leq \|\boldsymbol{u}-\boldsymbol{u}_\star\|_2.
$$

Proof of the non-expansive property. We begin with proving the non-expansive property. Denote the optimal alignment matrices between $\boldsymbol{F}_+$ and $\boldsymbol{F}_\star$ as $\{\boldsymbol{Q}_{+,k}\}_{k=1,2,3}$, whose existence is guaranteed by Lemma 6. Assume for now (which shall be established at the end of the proof) that for any $1 \leq i_1 \leq n_1$, we have

$$
\frac{B}{\sqrt{n_1}\left\|\boldsymbol{U}_+(i_1,:)\breve{\boldsymbol{U}}_+^\top\right\|_2} \geq \frac{\left\|\boldsymbol{U}_\star(i_1,:)\boldsymbol{\Sigma}_{\star,1}\right\|_2}{\left\|\boldsymbol{U}_+(i_1,:)\boldsymbol{Q}_{+,1}\boldsymbol{\Sigma}_{\star,1}\right\|_2}.
\tag{55}
$$

This taken together with Claim 5 immediately implies

$$
\begin{aligned}
&\left\|\boldsymbol{U}(i_1,:)\boldsymbol{Q}_{+,1}\boldsymbol{\Sigma}_{\star,1}-\boldsymbol{U}_\star(i_1,:)\boldsymbol{\Sigma}_{\star,1}\right\|_2 \leq \left\|\boldsymbol{U}_+(i_1,:)\boldsymbol{Q}_{+,1}\boldsymbol{\Sigma}_{\star,1}-\boldsymbol{U}_\star(i_1,:)\boldsymbol{\Sigma}_{\star,1}\right\|_2, \qquad 1 \leq i_1 \leq n_1, \\
\Longrightarrow \qquad &\left\|(\boldsymbol{U}\boldsymbol{Q}_{+,1}-\boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1}\right\|_{\mathsf{F}} \leq \left\|(\boldsymbol{U}_+\boldsymbol{Q}_{+,1}-\boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1}\right\|_{\mathsf{F}}.
\end{aligned}
$$

Repeating similar arguments for the other two factors, we obtain

$$
\left\|(\boldsymbol{V}\boldsymbol{Q}_{+,2}-\boldsymbol{V}_\star)\boldsymbol{\Sigma}_{\star,2}\right\|_{\mathsf{F}} \leq \left\|(\boldsymbol{V}_+\boldsymbol{Q}_{+,2}-\boldsymbol{V}_\star)\boldsymbol{\Sigma}_{\star,2}\right\|_{\mathsf{F}}, \quad \left\|(\boldsymbol{W}\boldsymbol{Q}_{+,3}-\boldsymbol{W}_\star)\boldsymbol{\Sigma}_{\star,3}\right\|_{\mathsf{F}} \leq \left\|(\boldsymbol{W}_+\boldsymbol{Q}_{+,3}-\boldsymbol{W}_\star)\boldsymbol{\Sigma}_{\star,3}\right\|_{\mathsf{F}}.
$$

Combining the above bounds, we have

$$
\begin{aligned}
\operatorname{dist}^2(\boldsymbol{F},\boldsymbol{F}_\star) &\leq \|(\boldsymbol{U}\boldsymbol{Q}_{+,1}-\boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 + \|(\boldsymbol{V}\boldsymbol{Q}_{+,2}-\boldsymbol{V}_\star)\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}^2 \\
&\quad + \|(\boldsymbol{W}\boldsymbol{Q}_{+,3}-\boldsymbol{W}_\star)\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{Q}_{+,1}^{-1},\boldsymbol{Q}_{+,2}^{-1},\boldsymbol{Q}_{+,3}^{-1})\cdot\boldsymbol{\mathcal{S}}-\boldsymbol{\mathcal{S}}_\star\right\|_{\mathsf{F}}^2 = \operatorname{dist}^2(\boldsymbol{F}_+,\boldsymbol{F}_\star).
\end{aligned}
$$

Proof of the incoherence condition. Turning to the incoherence condition, it follows that for any $1 \le i_1 \le n_1$,

$$\begin{aligned}
\left\|\boldsymbol{U}(i_1,:)\breve{\boldsymbol{U}}^\top\right\|_2^2 &= \sum_{i_2=1}^{n_2}\sum_{i_3=1}^{n_3}\left\langle \boldsymbol{U}(i_1,:)\mathcal{M}_1(\boldsymbol{\mathcal{S}}), \boldsymbol{V}(i_2,:)\otimes\boldsymbol{W}(i_3,:)\right\rangle^2 \\
&\overset{\text{(i)}}{=} \sum_{i_2=1}^{n_2}\sum_{i_3=1}^{n_3}\left\langle \boldsymbol{U}(i_1,:)\mathcal{M}_1(\boldsymbol{\mathcal{S}}), \boldsymbol{V}_+(i_2,:)\otimes\boldsymbol{W}_+(i_3,:)\right\rangle^2 \left(1\wedge\frac{B}{\sqrt{n_2}\|\boldsymbol{V}_+(i_2,:)\breve{\boldsymbol{V}}_+^\top\|_2}\right)^2\left(1\wedge\frac{B}{\sqrt{n_3}\|\boldsymbol{W}_+(i_3,:)\breve{\boldsymbol{W}}_+^\top\|_2}\right)^2 \\
&\overset{\text{(ii)}}{\le} \sum_{i_2=1}^{n_2}\sum_{i_3=1}^{n_3}\left\langle \boldsymbol{U}(i_1,:)\mathcal{M}_1(\boldsymbol{\mathcal{S}}), \boldsymbol{V}_+(i_2,:)\otimes\boldsymbol{W}_+(i_3,:)\right\rangle^2 \\
&\overset{\text{(iii)}}{=} \sum_{i_2=1}^{n_2}\sum_{i_3=1}^{n_3}\left(1\wedge\frac{B}{\sqrt{n_1}\|\boldsymbol{U}_+(i_1,:)\breve{\boldsymbol{U}}_+^\top\|_2}\right)^2\left\langle \boldsymbol{U}_+(i_1,:)\mathcal{M}_1(\boldsymbol{\mathcal{S}}_+), \boldsymbol{V}_+(i_2,:)\otimes\boldsymbol{W}_+(i_3,:)\right\rangle^2 \\
&= \left(1\wedge\frac{B}{\sqrt{n_1}\|\boldsymbol{U}_+(i_1,:)\breve{\boldsymbol{U}}_+^\top\|_2}\right)^2\left\|\boldsymbol{U}_+(i_1,:)\breve{\boldsymbol{U}}_+^\top\right\|_2^2 \overset{\text{(iv)}}{\le} \frac{B^2}{n_1}.
\end{aligned}$$

Here, (i) and (iii) follow from the definition of the scaled projection (19), (ii) and (iv) follow from the basic relations $a\wedge b\le a$ and $a\wedge b\le b$. By symmetry, one has

$$\sqrt{n_1}\|\boldsymbol{U}\breve{\boldsymbol{U}}^\top\|_{2,\infty}\vee\sqrt{n_2}\|\boldsymbol{V}\breve{\boldsymbol{V}}^\top\|_{2,\infty}\vee\sqrt{n_3}\|\boldsymbol{W}\breve{\boldsymbol{W}}^\top\|_{2,\infty}\le B.$$

The proof is then finished once we prove inequality (55).

Proof of the relation (55). Under the condition $\operatorname{dist}(\boldsymbol{F}_+,\boldsymbol{F}_\star)\le\epsilon\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$, invoke (36a) in Lemma 10 on the factor quadruple $\left(\boldsymbol{U}_+\boldsymbol{Q}_{+,1},\boldsymbol{V}_+\boldsymbol{Q}_{+,2},\boldsymbol{W}_+\boldsymbol{Q}_{+,3},(\boldsymbol{Q}_{+,1}^{-1},\boldsymbol{Q}_{+,2}^{-1},\boldsymbol{Q}_{+,3}^{-1})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_+\right)$ to see

$$\|\boldsymbol{V}_+\boldsymbol{Q}_{+,2}\|\vee\|\boldsymbol{W}_+\boldsymbol{Q}_{+,3}\|\vee\left\|\mathcal{M}_1\left((\boldsymbol{Q}_{+,1}^{-1},\boldsymbol{Q}_{+,2}^{-1},\boldsymbol{Q}_{+,3}^{-1})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_+\right)^\top\boldsymbol{\Sigma}_{\star,1}^{-1}\right\|\le 1+\epsilon,$$

which further implies that

$$\left\|\breve{\boldsymbol{U}}_+\boldsymbol{Q}_{+,1}^{-\top}\boldsymbol{\Sigma}_{\star,1}^{-1}\right\|\le\|\boldsymbol{V}_+\boldsymbol{Q}_{+,2}\|\|\boldsymbol{W}_+\boldsymbol{Q}_{+,3}\|\left\|\mathcal{M}_1\left((\boldsymbol{Q}_{+,1}^{-1},\boldsymbol{Q}_{+,2}^{-1},\boldsymbol{Q}_{+,3}^{-1})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_+\right)^\top\boldsymbol{\Sigma}_{\star,1}^{-1}\right\|\le(1+\epsilon)^3. \tag{56}$$

For any $1\le i_1\le n_1$, one has

$$\begin{aligned}
\left\|\boldsymbol{U}_+(i_1,:)\breve{\boldsymbol{U}}_+^\top\right\|_2 &\le \left\|\boldsymbol{U}_+(i_1,:)\boldsymbol{Q}_{+,1}\boldsymbol{\Sigma}_{\star,1}\right\|_2\left\|\breve{\boldsymbol{U}}_+\boldsymbol{Q}_{+,1}^{-\top}\boldsymbol{\Sigma}_{\star,1}^{-1}\right\| \\
&\le \left\|\boldsymbol{U}_+(i_1,:)\boldsymbol{Q}_{+,1}\boldsymbol{\Sigma}_{\star,1}\right\|_2(1+\epsilon)^3,
\end{aligned}$$

where the second line follows from the bound (56). In addition, the incoherence assumption of $\boldsymbol{\mathcal{X}}_\star$ (15) implies that

$$\sqrt{n_1}\left\|\boldsymbol{U}_\star(i_1,:)\boldsymbol{\Sigma}_{\star,1}\right\|_2\le\sqrt{n_1}\left\|\boldsymbol{U}_\star(i_1,:)\right\|_2\left\|\boldsymbol{\Sigma}_{\star,1}\right\|\le\sqrt{\mu r}\sigma_{\max}(\boldsymbol{\mathcal{X}}_\star)\le B(1+\epsilon)^{-3},$$

where the last inequality follows from the choice of B. Take the above two relations collectively to reach the advertised bound (55).

C.2 Concentration bounds

We gather several useful concentration bounds regarding the partial observation operator $\mathcal{P}_\Omega(\cdot)$.

Lemma 12. *Suppose that $\boldsymbol{\mathcal{X}}_\star$ is μ-incoherent, and Ω satisfies the Bernoulli observation model in (17) with $pn_1n_2n_3\gtrsim\mu^2r^2n\log n$. With overwhelming probability, one has*

$$\left|\left\langle(p^{-1}\mathcal{P}_\Omega-\mathcal{I})(\boldsymbol{\mathcal{X}}_A),\boldsymbol{\mathcal{X}}_B\right\rangle\right|\le C_T\sqrt{\frac{\mu^2r^2n\log n}{pn_1n_2n_3}}\|\boldsymbol{\mathcal{X}}_A\|_{\mathsf{F}}\|\boldsymbol{\mathcal{X}}_B\|_{\mathsf{F}}$$

simultaneously for all tensors $\boldsymbol{\mathcal{X}}_A, \boldsymbol{\mathcal{X}}_B \in \mathbb{R}^{n_1\times n_2\times n_3}$ in the form of

$$\begin{aligned}\boldsymbol{\mathcal{X}}_A &= (\boldsymbol{U}_A, \boldsymbol{V}_\star, \boldsymbol{W}_\star)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_{A,1} + (\boldsymbol{U}_\star, \boldsymbol{V}_A, \boldsymbol{W}_\star)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_{A,2} + (\boldsymbol{U}_\star, \boldsymbol{V}_\star, \boldsymbol{W}_A)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_{A,3},\\ \boldsymbol{\mathcal{X}}_B &= (\boldsymbol{U}_B, \boldsymbol{V}_\star, \boldsymbol{W}_\star)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_{B,1} + (\boldsymbol{U}_\star, \boldsymbol{V}_B, \boldsymbol{W}_\star)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_{B,2} + (\boldsymbol{U}_\star, \boldsymbol{V}_\star, \boldsymbol{W}_B)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_{B,3},\end{aligned}$$

where $\boldsymbol{U}_A, \boldsymbol{U}_B \in \mathbb{R}^{n_1\times r_1}$, $\boldsymbol{V}_A, \boldsymbol{V}_B \in \mathbb{R}^{n_2\times r_2}$, $\boldsymbol{W}_A, \boldsymbol{W}_B \in \mathbb{R}^{n_3\times r_3}$, and $\boldsymbol{\mathcal{S}}_{A,k}, \boldsymbol{\mathcal{S}}_{B,k} \in \mathbb{R}^{r_1\times r_2\times r_3}$ are arbitrary factors, and $C_T > 0$ is some universal constant.

Lemma 13 ([CLPC19, Lemma D.2])**.** *Suppose that Ω satisfies the Bernoulli observation model in (17), then for any fixed $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1\times n_2\times n_3}$, with overwhelming probability, one has*

$$\left\|(p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{\mathcal{X}})\right\| \le C_Y\left(p^{-1}\log^3 n\|\boldsymbol{\mathcal{X}}\|_\infty + \sqrt{p^{-1}\log^5 n}\max_{k=1,2,3}\|\mathcal{M}_k(\boldsymbol{\mathcal{X}})^\top\|_{2,\infty}\right),$$

where $C_Y > 0$ is some universal constant.

Lemma 14. *Suppose that Ω satisfies the Bernoulli observation model in (17). Then with overwhelming probability, one has*

$$\left|\left\langle(p^{-1}\mathcal{P}_\Omega - \mathcal{I})((\boldsymbol{U}_A, \boldsymbol{V}_A, \boldsymbol{W}_A)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_A), (\boldsymbol{U}_B, \boldsymbol{V}_B, \boldsymbol{W}_B)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_B\right\rangle\right| \le C_Y\left(p^{-1}\log^3 n + \sqrt{p^{-1}n\log^5 n}\right)\mathfrak{N},$$

simultaneously for all tensors $(\boldsymbol{U}_A, \boldsymbol{V}_A, \boldsymbol{W}_A)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_A$ and $(\boldsymbol{U}_B, \boldsymbol{V}_B, \boldsymbol{W}_B)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_B$, where the quantity $\mathfrak{N}$ obeys

$$\begin{aligned}\mathfrak{N} \le& \big(\|\boldsymbol{U}_A\mathcal{M}_1(\boldsymbol{\mathcal{S}}_A)\|_{2,\infty}\|\boldsymbol{U}_B\mathcal{M}_1(\boldsymbol{\mathcal{S}}_B)\|_{\mathsf{F}} \wedge \|\boldsymbol{U}_A\mathcal{M}_1(\boldsymbol{\mathcal{S}}_A)\|_{\mathsf{F}}\|\boldsymbol{U}_B\mathcal{M}_1(\boldsymbol{\mathcal{S}}_B)\|_{2,\infty}\big)\\ &\big(\|\boldsymbol{V}_A\|_{2,\infty}\|\boldsymbol{V}_B\|_{\mathsf{F}} \wedge \|\boldsymbol{V}_A\|_{\mathsf{F}}\|\boldsymbol{V}_B\|_{2,\infty}\big)\big(\|\boldsymbol{W}_A\|_{2,\infty}\|\boldsymbol{W}_B\|_{\mathsf{F}} \wedge \|\boldsymbol{W}_A\|_{\mathsf{F}}\|\boldsymbol{W}_B\|_{2,\infty}\big).\end{aligned}$$

By symmetry, the above bound continues to hold if we permute the occurrences of $\boldsymbol{U}$, $\boldsymbol{V}$, and $\boldsymbol{W}$.

C.2.1 Proof of Lemma 12

This lemma is essentially [YZ16, Lemma 5] under the Bernoulli observation model. Here, we provide a simpler proof based on the matrix Bernstein inequality. Let $\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3}$ be the tensor with only the (i_1, i_2, i_3)-th entry as 1 and all the other entries as 0, and let $\delta_{i_1,i_2,i_3} \sim \mathrm{Bernoulli}(p)$ be an i.i.d. Bernoulli random variable for $1 \le i_k \le n_k$, $k = 1, 2, 3$. Define an operator $\mathcal{P}_T : \mathbb{R}^{n_1\times n_2\times n_3} \mapsto \mathbb{R}^{n_1\times n_2\times n_3}$ as

$$\mathcal{P}_T(\boldsymbol{\mathcal{X}}) = (\boldsymbol{I}_{n_1}, \boldsymbol{V}_\star\boldsymbol{V}_\star^\top, \boldsymbol{W}_\star\boldsymbol{W}_\star^\top)\boldsymbol{\cdot}\boldsymbol{\mathcal{X}} + (\boldsymbol{U}_\star\boldsymbol{U}_\star^\top, \boldsymbol{I}_{n_2} - \boldsymbol{V}_\star\boldsymbol{V}_\star^\top, \boldsymbol{W}_\star\boldsymbol{W}_\star^\top)\boldsymbol{\cdot}\boldsymbol{\mathcal{X}} + (\boldsymbol{U}_\star\boldsymbol{U}_\star^\top, \boldsymbol{V}_\star\boldsymbol{V}_\star^\top, \boldsymbol{I}_{n_3} - \boldsymbol{W}_\star\boldsymbol{W}_\star^\top)\boldsymbol{\cdot}\boldsymbol{\mathcal{X}}.$$

It is straightforward to verify that $\mathcal{P}_T(\cdot)$ defines a projection, and that

$$\begin{aligned}\boldsymbol{\mathcal{X}}_A &= (\boldsymbol{U}_A, \boldsymbol{V}_\star, \boldsymbol{W}_\star)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_{A,1} + (\boldsymbol{U}_\star, \boldsymbol{V}_A, \boldsymbol{W}_\star)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_{A,2} + (\boldsymbol{U}_\star, \boldsymbol{V}_\star, \boldsymbol{W}_A)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_{A,3}\\ &= \mathcal{P}_T((\boldsymbol{U}_A, \boldsymbol{V}_\star, \boldsymbol{W}_\star)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_{A,1}) + \mathcal{P}_T((\boldsymbol{U}_\star, \boldsymbol{V}_A, \boldsymbol{W}_\star)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_{A,2}) + \mathcal{P}_T((\boldsymbol{U}_\star, \boldsymbol{V}_\star, \boldsymbol{W}_A)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_{A,3})\\ &= \mathcal{P}_T(\boldsymbol{\mathcal{X}}_A) = \sum_{i_1,i_2,i_3}\langle\mathcal{P}_T(\boldsymbol{\mathcal{X}}_A), \boldsymbol{\mathcal{E}}_{i_1,i_2,i_3}\rangle\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3} = \sum_{i_1,i_2,i_3}\langle\boldsymbol{\mathcal{X}}_A, \mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3})\rangle\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3}.\end{aligned}$$

A similar expression holds for $\boldsymbol{\mathcal{X}}_B$. Hence, we have

$$\begin{aligned}\left|\left\langle(p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{\mathcal{X}}_A), \boldsymbol{\mathcal{X}}_B\right\rangle\right| &= \left|\sum_{i_1,i_2,i_3}\left(p^{-1}\delta_{i_1,i_2,i_3} - 1\right)\langle\boldsymbol{\mathcal{X}}_A, \mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3})\rangle\langle\boldsymbol{\mathcal{X}}_B, \mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3})\rangle\right|\\ &= \left|\left\langle\mathrm{vec}(\boldsymbol{\mathcal{X}}_A), \sum_{i_1,i_2,i_3}(p^{-1}\delta_{i_1,i_2,i_3} - 1)\,\mathrm{vec}\left(\mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3})\right)\mathrm{vec}\left(\mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3})\right)^\top\mathrm{vec}(\boldsymbol{\mathcal{X}}_B)\right\rangle\right|\\ &\le \|\boldsymbol{\mathcal{X}}_A\|_{\mathsf{F}}\|\boldsymbol{\mathcal{X}}_B\|_{\mathsf{F}}\left\|\sum_{i_1,i_2,i_3}(p^{-1}\delta_{i_1,i_2,i_3} - 1)\,\mathrm{vec}\left(\mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3})\right)\mathrm{vec}\left(\mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3})\right)^\top\right\|.\end{aligned}$$

Therefore it suffices to bound the last term in the above inequality, which we resort to the matrix Bernstein inequality: with overwhelming probability, one has

$$\begin{aligned}\left\| \sum_{i_1,i_2,i_3} (p^{-1}\delta_{i_1,i_2,i_3} - 1) \operatorname{vec}\left(\mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3})\right) \operatorname{vec}\left(\mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3})\right)^\top \right\| &\lesssim \left(\frac{\mu^2 r^2 n \log n}{p n_1 n_2 n_3} + \sqrt{\frac{\mu^2 r^2 n \log n}{p n_1 n_2 n_3}} \right) \qquad (57)\\ &\lesssim \sqrt{\frac{\mu^2 r^2 n \log n}{p n_1 n_2 n_3}},\end{aligned}$$

where the second line holds as long as $pn_1n_2n_3 \gtrsim \mu^2 r^2 n \log n$. Plugging the above bound (which will be proved at the end of the proof) in the previous one, we immediately arrive at the desired result:

$$\left|\left\langle (p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{\mathcal{X}}_A), \boldsymbol{\mathcal{X}}_B \right\rangle\right| \lesssim \sqrt{\frac{\mu^2 r^2 n \log n}{p n_1 n_2 n_3}} \|\boldsymbol{\mathcal{X}}_A\|_{\mathsf{F}} \|\boldsymbol{\mathcal{X}}_B\|_{\mathsf{F}}.$$

Proof of (57). By standard matrix Bernstein inequality, we have

$$\left\| \sum_{i_1,i_2,i_3} (p^{-1}\delta_{i_1,i_2,i_3} - 1) \operatorname{vec}\left(\mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3})\right) \operatorname{vec}\left(\mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3})\right)^\top \right\| \lesssim B \log n + \sigma \sqrt{\log n},$$

where

$$\begin{aligned} B &= \max_{i_1,i_2,i_3} \left\| (p^{-1}\delta_{i_1,i_2,i_3} - 1) \operatorname{vec}\left(\mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3})\right) \operatorname{vec}\left(\mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3})\right)^\top \right\|, \\ \sigma^2 &= \left\| \sum_{i_1,i_2,i_3} \mathbb{E}(p^{-1}\delta_{i_1,i_2,i_3} - 1)^2 \operatorname{vec}\left(\mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3})\right) \operatorname{vec}\left(\mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3})\right)^\top \operatorname{vec}\left(\mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3})\right) \operatorname{vec}\left(\mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3})\right)^\top \right\|. \end{aligned}$$

- Here, B obeys

  $$B = \max_{i_1,i_2,i_3} \left\| (p^{-1}\delta_{i_1,i_2,i_3} - 1) \operatorname{vec}\left(\mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3})\right) \operatorname{vec}\left(\mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3})\right)^\top \right\| \le p^{-1} \max_{i_1,i_2,i_3} \|\mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3})\|_{\mathsf{F}}^2,$$

  where the last inequality uses $|(p^{-1}\delta_{i_1,i_2,i_3} - 1)| \le p^{-1}$. To proceed, first notice that the three terms in $\mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3})$ are mutually orthogonal, which allows

  $$\begin{aligned}\|\mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3})\|_{\mathsf{F}}^2 &= \left\| (\boldsymbol{I}_{n_1}, \boldsymbol{V}_\star\boldsymbol{V}_\star^\top, \boldsymbol{W}_\star\boldsymbol{W}_\star^\top) \boldsymbol{\cdot} \boldsymbol{\mathcal{E}}_{i_1,i_2,i_3} \right\|_{\mathsf{F}}^2 + \left\| (\boldsymbol{U}_\star\boldsymbol{U}_\star^\top, \boldsymbol{I}_{n_2} - \boldsymbol{V}_\star\boldsymbol{V}_\star^\top, \boldsymbol{W}_\star\boldsymbol{W}_\star^\top) \boldsymbol{\cdot} \boldsymbol{\mathcal{E}}_{i_1,i_2,i_3} \right\|_{\mathsf{F}}^2 \\ &\quad + \left\| (\boldsymbol{U}_\star\boldsymbol{U}_\star^\top, \boldsymbol{V}_\star\boldsymbol{V}_\star^\top, \boldsymbol{I}_{n_3} - \boldsymbol{W}_\star\boldsymbol{W}_\star^\top) \boldsymbol{\cdot} \boldsymbol{\mathcal{E}}_{i_1,i_2,i_3} \right\|_{\mathsf{F}}^2.\end{aligned}$$

  Since $\boldsymbol{U}_\star, \boldsymbol{V}_\star, \boldsymbol{W}_\star$ have orthonormal columns, it is straightforward to see

  $$\begin{aligned} \left\| (\boldsymbol{I}_{n_1}, \boldsymbol{V}_\star\boldsymbol{V}_\star^\top, \boldsymbol{W}_\star\boldsymbol{W}_\star^\top) \boldsymbol{\cdot} \boldsymbol{\mathcal{E}}_{i_1,i_2,i_3} \right\|_{\mathsf{F}}^2 &= \|\boldsymbol{I}_{n_1}(i_1,:)\|_2^2 \left\|\boldsymbol{V}_\star(i_2,:)\boldsymbol{V}_\star^\top\right\|_2^2 \left\|\boldsymbol{W}_\star(i_3,:)\boldsymbol{W}_\star^\top\right\|_2^2 \\ &\le \|\boldsymbol{V}_\star\|_{2,\infty}^2 \|\boldsymbol{W}_\star\|_{2,\infty}^2; \\ \left\| (\boldsymbol{U}_\star\boldsymbol{U}_\star^\top, \boldsymbol{I}_{n_2} - \boldsymbol{V}_\star\boldsymbol{V}_\star^\top, \boldsymbol{W}_\star\boldsymbol{W}_\star^\top) \boldsymbol{\cdot} \boldsymbol{\mathcal{E}}_{i_1,i_2,i_3} \right\|_{\mathsf{F}}^2 &= \left\|\boldsymbol{U}_\star(i_1,:)\boldsymbol{U}_\star^\top\right\|_2^2 \left\| \left[\boldsymbol{I}_{n_2} - \boldsymbol{V}_\star\boldsymbol{V}_\star^\top\right](i_2,:)\right\|_2^2 \left\|\boldsymbol{W}_\star(i_3,:)\boldsymbol{W}_\star^\top\right\|_2^2 \\ &\le \|\boldsymbol{U}_\star\|_{2,\infty}^2 \|\boldsymbol{W}_\star\|_{2,\infty}^2; \\ \left\| (\boldsymbol{U}_\star\boldsymbol{U}_\star^\top, \boldsymbol{V}_\star\boldsymbol{V}_\star^\top, \boldsymbol{I}_{n_3} - \boldsymbol{W}_\star\boldsymbol{W}_\star^\top) \boldsymbol{\cdot} \boldsymbol{\mathcal{E}}_{i_1,i_2,i_3} \right\|_{\mathsf{F}}^2 &= \left\|\boldsymbol{U}_\star(i_1,:)\boldsymbol{U}_\star^\top\right\|_2^2 \left\|\boldsymbol{V}_\star(i_2,:)\boldsymbol{V}_\star^\top\right\|_2^2 \left\| \left[\boldsymbol{I}_{r_3} - \boldsymbol{W}_\star\boldsymbol{W}_\star^\top\right](i_3,:)\right\|_2^2 \\ &\le \|\boldsymbol{U}_\star\|_{2,\infty}^2 \|\boldsymbol{V}_\star\|_{2,\infty}^2. \end{aligned}$$

  Finally use the definition of incoherence (cf. Definition 2) to conclude

  $$B \le p^{-1}\left(\|\boldsymbol{V}_\star\|_{2,\infty}^2 \|\boldsymbol{W}_\star\|_{2,\infty}^2 + \|\boldsymbol{U}_\star\|_{2,\infty}^2 \|\boldsymbol{W}_\star\|_{2,\infty}^2 + \|\boldsymbol{U}_\star\|_{2,\infty}^2 \|\boldsymbol{V}_\star\|_{2,\infty}^2 \right) \le \frac{3\mu^2 r^2 n}{p n_1 n_2 n_3}.$$

- In addition, σ^2 obeys

$$\sigma^2 \leq p^{-1} \max_{i_1,i_2,i_3} \|\mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3})\|_{\mathsf{F}}^2 \left\| \sum_{i_1,i_2,i_3} \mathrm{vec}\left(\mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3})\right) \mathrm{vec}\left(\mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3})\right)^\top \right\| \leq \frac{3\mu^2 r^2 n}{p n_1 n_2 n_3},$$

where we have used the variational representation to conclude

$$\begin{aligned} \left\| \sum_{i_1,i_2,i_3} \mathrm{vec}\left(\mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3})\right) \mathrm{vec}\left(\mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3})\right)^\top \right\| &= \sup_{\widetilde{\boldsymbol{\mathcal{X}}}:\|\widetilde{\boldsymbol{\mathcal{X}}}\|_{\mathsf{F}}\leq 1} \sum_{i_1,i_2,i_3} \langle \widetilde{\boldsymbol{\mathcal{X}}}, \mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3}) \rangle^2 \\ &= \sup_{\widetilde{\boldsymbol{\mathcal{X}}}:\|\widetilde{\boldsymbol{\mathcal{X}}}\|_{\mathsf{F}}\leq 1} \|\mathcal{P}_T(\widetilde{\boldsymbol{\mathcal{X}}})\|_{\mathsf{F}}^2 \leq 1. \end{aligned}$$

Plugging the expressions of B and σ leads to the advertised bound (57).

C.2.2 Proof of Lemma 14

This lemma generalizes [CL19, Lemma 8] to the tensor setting, which is a powerful tool in the analysis of matrix completion [CLL20, TMC20]. We begin by decomposing $(\boldsymbol{U}_A, \boldsymbol{V}_A, \boldsymbol{W}_A) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_A$ into a sum of $r_2 r_3$ rank-**1** tensors:

$$(\boldsymbol{U}_A, \boldsymbol{V}_A, \boldsymbol{W}_A) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_A = \sum_{a_2=1}^{r_2} \sum_{a_3=1}^{r_3} (\boldsymbol{u}_{a_2,a_3}, \boldsymbol{v}_{a_2}, \boldsymbol{w}_{a_3}) \boldsymbol{\cdot} 1,$$

where we denote the column vectors $\boldsymbol{u}_{a_2,a_3} \coloneqq [\boldsymbol{U}_A \mathcal{M}_1(\boldsymbol{\mathcal{S}}_A)](:, (r_3 - 1)a_2 + a_3)$, $\boldsymbol{v}_{a_2} \coloneqq \boldsymbol{V}_A(:, a_2)$, and $\boldsymbol{w}_{a_3} \coloneqq \boldsymbol{W}_A(:, a_3)$ for notational convenience. Similarly, we can decompose $(\boldsymbol{U}_B, \boldsymbol{V}_B, \boldsymbol{W}_B) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_B$ as

$$(\boldsymbol{U}_B, \boldsymbol{V}_B, \boldsymbol{W}_B) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_B = \sum_{b_2=1}^{r_2} \sum_{b_3=1}^{r_3} (\boldsymbol{u}_{b_2,b_3}, \boldsymbol{v}_{b_2}, \boldsymbol{w}_{b_3}) \boldsymbol{\cdot} 1,$$

with $\boldsymbol{u}_{b_2,b_3}$, $\boldsymbol{v}_{b_2}$ and $\boldsymbol{w}_{b_3}$ defined analogously. We further denote $\boldsymbol{\mathcal{J}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ as the tensor with all-one entries, i.e. $\boldsymbol{\mathcal{J}}(i_1, i_2, i_3) = 1$ for all $1 \leq i_k \leq n_k$, $k = 1, 2, 3$. With these preparation in hand, by the triangle inequality we have

$$\begin{aligned} &\left|\left\langle (p^{-1}\mathcal{P}_\Omega - \mathcal{I})((\boldsymbol{U}_A, \boldsymbol{V}_A, \boldsymbol{W}_A) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_A), (\boldsymbol{U}_B, \boldsymbol{V}_B, \boldsymbol{W}_B) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_B \right\rangle\right| \\ &\quad \leq \sum_{a_2,b_2=1}^{r_2} \sum_{a_3,b_3=1}^{r_3} \left|\left\langle (p^{-1}\mathcal{P}_\Omega - \mathcal{I})((\boldsymbol{u}_{a_2,a_3}, \boldsymbol{v}_{a_2}, \boldsymbol{w}_{a_3}) \boldsymbol{\cdot} 1), (\boldsymbol{u}_{b_2,b_3}, \boldsymbol{v}_{b_2}, \boldsymbol{w}_{b_3}) \boldsymbol{\cdot} 1 \right\rangle\right| \\ &\quad = \sum_{a_2,b_2=1}^{r_2} \sum_{a_3,b_3=1}^{r_3} \left|\left\langle (p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{\mathcal{J}}), (\boldsymbol{u}_{a_2,a_3} \odot \boldsymbol{u}_{b_2,b_3}, \boldsymbol{v}_{a_2} \odot \boldsymbol{v}_{b_2}, \boldsymbol{w}_{a_3} \odot \boldsymbol{w}_{b_3}) \boldsymbol{\cdot} 1 \right\rangle\right| \\ &\quad \leq \sum_{a_2,b_2=1}^{r_2} \sum_{a_3,b_3=1}^{r_3} \|(p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{\mathcal{J}})\| \|\boldsymbol{u}_{a_2,a_3} \odot \boldsymbol{u}_{b_2,b_3}\|_2 \|\boldsymbol{v}_{a_2} \odot \boldsymbol{v}_{b_2}\|_2 \|\boldsymbol{w}_{a_3} \odot \boldsymbol{w}_{b_3}\|_2 \\ &\quad = \|(p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{\mathcal{J}})\| \mathfrak{N}, \end{aligned}$$

where $\odot$ denotes the Hadamard (entrywise) product, and

$$\mathfrak{N} \coloneqq \sum_{a_2,b_2=1}^{r_2} \sum_{a_3,b_3=1}^{r_3} \|\boldsymbol{u}_{a_2,a_3} \odot \boldsymbol{u}_{b_2,b_3}\|_2 \|\boldsymbol{v}_{a_2} \odot \boldsymbol{v}_{b_2}\|_2 \|\boldsymbol{w}_{a_3} \odot \boldsymbol{w}_{b_3}\|_2.$$

Therefore, it boils down to controlling $\|(p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{\mathcal{J}})\|$ and $\mathfrak{N}$.

- Regarding $\|(p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{\mathcal{J}})\|$, Lemma 13 tells that, with overwhelming probability, it is bounded by
$$\|(p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{\mathcal{J}})\| \le C_Y\left(p^{-1}\log^3 n + \sqrt{p^{-1} n \log^5 n}\right),$$
where we use the fact $\|\boldsymbol{\mathcal{J}}\|_\infty = 1$ and $\max_{k=1,2,3}\|\mathcal{M}_k(\boldsymbol{\mathcal{J}})^\top\|_{2,\infty} \le \sqrt{n}$.
- Turning to $\mathfrak{N}$, applying the Cauchy-Schwarz inequality we have
$$\begin{aligned}
\mathfrak{N} &\le \sqrt{\sum_{a_2,b_2=1}^{r_2}\sum_{a_3,b_3=1}^{r_3}\|\boldsymbol{u}_{a_2,a_3}\odot\boldsymbol{u}_{b_2,b_3}\|_2^2}\sqrt{\sum_{a_2,b_2=1}^{r_2}\|\boldsymbol{v}_{a_2}\odot\boldsymbol{v}_{b_2}\|_2^2\sum_{a_3,b_3=1}^{r_3}\|\boldsymbol{w}_{a_3}\odot\boldsymbol{w}_{b_3}\|_2^2}\\
&= \sqrt{\sum_{i_1=1}^{n_1}\|\boldsymbol{U}_A(i_1,:)\mathcal{M}_1(\boldsymbol{\mathcal{S}}_A)\|_2^2\|\boldsymbol{U}_B(i_1,:)\mathcal{M}_1(\boldsymbol{\mathcal{S}}_B)\|_2^2}\\
&\qquad\sqrt{\sum_{i_2=1}^{n_2}\|\boldsymbol{V}_A(i_2,:)\|_2^2\|\boldsymbol{V}_B(i_2,:)\|_2^2}\sqrt{\sum_{i_3=1}^{n_3}\|\boldsymbol{W}_A(i_3,:)\|_2^2\|\boldsymbol{W}_B(i_3,:)\|_2^2}\\
&\le \big(\|\boldsymbol{U}_A\mathcal{M}_1(\boldsymbol{\mathcal{S}}_A)\|_{2,\infty}\|\boldsymbol{U}_B\mathcal{M}_1(\boldsymbol{\mathcal{S}}_B)\|_{\mathsf{F}} \wedge \|\boldsymbol{U}_A\mathcal{M}_1(\boldsymbol{\mathcal{S}}_A)\|_{\mathsf{F}}\|\boldsymbol{U}_B\mathcal{M}_1(\boldsymbol{\mathcal{S}}_B)\|_{2,\infty}\big)\\
&\qquad\big(\|\boldsymbol{V}_A\|_{2,\infty}\|\boldsymbol{V}_B\|_{\mathsf{F}} \wedge \|\boldsymbol{V}_A\|_{\mathsf{F}}\|\boldsymbol{V}_B\|_{2,\infty}\big)\big(\|\boldsymbol{W}_A\|_{2,\infty}\|\boldsymbol{W}_B\|_{\mathsf{F}} \wedge \|\boldsymbol{W}_A\|_{\mathsf{F}}\|\boldsymbol{W}_B\|_{2,\infty}\big).
\end{aligned}$$

The proof is complete by combining the above two bounds.

C.3 Proof of spectral initialization (Lemma 1)

In view of Lemma 8, we start by relating $\mathrm{dist}(\boldsymbol{F}_+,\boldsymbol{F}_\star)$ to $\|(\boldsymbol{U}_+,\boldsymbol{V}_+,\boldsymbol{W}_+)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_+ - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}$ as
$$\mathrm{dist}(\boldsymbol{F}_+,\boldsymbol{F}_\star) \le (\sqrt{2}+1)^{3/2}\,\|(\boldsymbol{U}_+,\boldsymbol{V}_+,\boldsymbol{W}_+)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_+ - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}.$$

With this bound in mind, it suffices to control $\|(\boldsymbol{U}_+,\boldsymbol{V}_+,\boldsymbol{W}_+)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_+ - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}$. To proceed, define $\boldsymbol{P}_U \coloneqq \boldsymbol{U}_+\boldsymbol{U}_+^\top$ as the projection matrix onto the column space of $\boldsymbol{U}_+$, $\boldsymbol{P}_U^\perp \coloneqq \boldsymbol{I}_{n_1} - \boldsymbol{P}_U$ as its orthogonal complement, and define $\boldsymbol{P}_V, \boldsymbol{P}_V^\perp, \boldsymbol{P}_W, \boldsymbol{P}_W^\perp$ analogously. We have the decomposition
$$\boldsymbol{\mathcal{X}}_\star = (\boldsymbol{P}_U,\boldsymbol{P}_V,\boldsymbol{P}_W)\boldsymbol{\cdot}\boldsymbol{\mathcal{X}}_\star + (\boldsymbol{P}_U^\perp,\boldsymbol{P}_V,\boldsymbol{P}_W)\boldsymbol{\cdot}\boldsymbol{\mathcal{X}}_\star + (\boldsymbol{I}_{n_1},\boldsymbol{P}_V^\perp,\boldsymbol{P}_W)\boldsymbol{\cdot}\boldsymbol{\mathcal{X}}_\star + (\boldsymbol{I}_{n_1},\boldsymbol{I}_{n_2},\boldsymbol{P}_W^\perp)\boldsymbol{\cdot}\boldsymbol{\mathcal{X}}_\star.$$

Expand the following squared norm and use that the four terms are mutually orthogonal to see
$$\begin{aligned}
&\|(\boldsymbol{U}_+,\boldsymbol{V}_+,\boldsymbol{W}_+)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_+ - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}^2\\
&\quad= \left\|(\boldsymbol{P}_U,\boldsymbol{P}_V,\boldsymbol{P}_W)\boldsymbol{\cdot}(p^{-1}\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}_\star) - (\boldsymbol{P}_U^\perp,\boldsymbol{P}_V,\boldsymbol{P}_W)\boldsymbol{\cdot}\boldsymbol{\mathcal{X}}_\star - (\boldsymbol{I}_{n_1},\boldsymbol{P}_V^\perp,\boldsymbol{P}_W)\boldsymbol{\cdot}\boldsymbol{\mathcal{X}}_\star - (\boldsymbol{I}_{n_1},\boldsymbol{I}_{n_2},\boldsymbol{P}_W^\perp)\boldsymbol{\cdot}\boldsymbol{\mathcal{X}}_\star\right\|_{\mathsf{F}}^2\\
&\quad= \left\|(\boldsymbol{P}_U,\boldsymbol{P}_V,\boldsymbol{P}_W)\boldsymbol{\cdot}(p^{-1}\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}_\star)\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{P}_U^\perp,\boldsymbol{P}_V,\boldsymbol{P}_W)\boldsymbol{\cdot}\boldsymbol{\mathcal{X}}_\star\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{I}_{n_1},\boldsymbol{P}_V^\perp,\boldsymbol{P}_W)\boldsymbol{\cdot}\boldsymbol{\mathcal{X}}_\star\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{I}_{n_1},\boldsymbol{I}_{n_2},\boldsymbol{P}_W^\perp)\boldsymbol{\cdot}\boldsymbol{\mathcal{X}}_\star\right\|_{\mathsf{F}}^2\\
&\quad\le \left\|(\boldsymbol{P}_U,\boldsymbol{P}_V,\boldsymbol{P}_W)\boldsymbol{\cdot}(p^{-1}\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}_\star)\right\|_{\mathsf{F}}^2 + \left\|\boldsymbol{P}_U^\perp\mathcal{M}_1(\boldsymbol{\mathcal{X}}_\star)\right\|_{\mathsf{F}}^2 + \left\|\boldsymbol{P}_V^\perp\mathcal{M}_2(\boldsymbol{\mathcal{X}}_\star)\right\|_{\mathsf{F}}^2 + \left\|\boldsymbol{P}_W^\perp\mathcal{M}_3(\boldsymbol{\mathcal{X}}_\star)\right\|_{\mathsf{F}}^2. \qquad (58)
\end{aligned}$$

We next control the terms in (58) one by one.

- For the first term in (58), since $(\boldsymbol{P}_U,\boldsymbol{P}_V,\boldsymbol{P}_W)\boldsymbol{\cdot}(p^{-1}\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}_\star)$ has a multilinear rank of at most $\boldsymbol{r}$, applying the relation (7) leads to
$$\left\|(\boldsymbol{P}_U,\boldsymbol{P}_V,\boldsymbol{P}_W)\boldsymbol{\cdot}(p^{-1}\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}_\star)\right\|_{\mathsf{F}} \le r\left\|(\boldsymbol{P}_U,\boldsymbol{P}_V,\boldsymbol{P}_W)\boldsymbol{\cdot}(p^{-1}\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}_\star)\right\| \le r\left\|(p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{\mathcal{X}}_\star)\right\|.$$
Therefore, it comes down to control $\left\|(p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{\mathcal{X}}_\star)\right\|$. Lemma 13 tells that, with overwhelming probability, one has
$$\left\|(p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{\mathcal{X}}_\star)\right\| \lesssim \left(p^{-1}\log^3 n\|\boldsymbol{\mathcal{X}}_\star\|_\infty + \sqrt{p^{-1}\log^5 n}\max_{k=1,2,3}\|\mathcal{M}_k(\boldsymbol{\mathcal{X}}_\star)^\top\|_{2,\infty}\right)$$

$$\lesssim \left(\frac{\mu^{3/2}r^{3/2}\log^3 n}{p\sqrt{n_1n_2n_3}} + \sqrt{\frac{\mu^2r^2n\log^5 n}{pn_1n_2n_3}}\right)\sigma_{\max}(\boldsymbol{X}_\star),$$

where the second line follows from the following relations in view of the incoherence property of $\boldsymbol{\mathcal{X}}_\star$ (cf. Definition 2):

$$\begin{aligned}
\|\boldsymbol{\mathcal{X}}_\star\|_\infty &\le \sigma_{\max}(\boldsymbol{\mathcal{X}}_\star)\|\boldsymbol{U}_\star\|_{2,\infty}\|\boldsymbol{V}_\star\|_{2,\infty}\|\boldsymbol{W}_\star\|_{2,\infty} \le \sigma_{\max}(\boldsymbol{\mathcal{X}}_\star)\sqrt{\frac{\mu^3r^3}{n_1n_2n_3}};\\
\|\mathcal{M}_1(\boldsymbol{\mathcal{X}}_\star)^\top\|_{2,\infty} &\le \|\boldsymbol{U}_\star\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)\|\|\boldsymbol{V}_\star\otimes\boldsymbol{W}_\star\|_{2,\infty} \le \sigma_{\max}(\boldsymbol{\mathcal{X}}_\star)\sqrt{\frac{\mu^2r^2}{n_2n_3}};\\
\|\mathcal{M}_2(\boldsymbol{\mathcal{X}}_\star)^\top\|_{2,\infty} &\le \|\boldsymbol{V}_\star\mathcal{M}_2(\boldsymbol{\mathcal{S}}_\star)\|\|\boldsymbol{U}_\star\otimes\boldsymbol{W}_\star\|_{2,\infty} \le \sigma_{\max}(\boldsymbol{\mathcal{X}}_\star)\sqrt{\frac{\mu^2r^2}{n_1n_3}};\\
\|\mathcal{M}_3(\boldsymbol{\mathcal{X}}_\star)^\top\|_{2,\infty} &\le \|\boldsymbol{W}_\star\mathcal{M}_3(\boldsymbol{\mathcal{S}}_\star)\|\|\boldsymbol{U}_\star\otimes\boldsymbol{V}_\star\|_{2,\infty} \le \sigma_{\max}(\boldsymbol{\mathcal{X}}_\star)\sqrt{\frac{\mu^2r^2}{n_1n_2}}.
\end{aligned}$$

In total, the first term in (58) is bounded by

$$\left\|(\boldsymbol{P}_U,\boldsymbol{P}_V,\boldsymbol{P}_W)\boldsymbol{\cdot}(p^{-1}\boldsymbol{\mathcal{Y}}-\boldsymbol{\mathcal{X}}_\star)\right\|_{\mathsf{F}}^2 \lesssim \left(\frac{\mu^{3/2}r^{5/2}\kappa\log^3 n}{p\sqrt{n_1n_2n_3}} + \sqrt{\frac{\mu^2r^4\kappa^2n\log^5 n}{pn_1n_2n_3}}\right)^2\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star).$$

- For the second term in (58), first bound it by the spectral norm as

$$\left\|\boldsymbol{P}_U^\perp\mathcal{M}_1(\boldsymbol{\mathcal{X}}_\star)\right\|_{\mathsf{F}}^2 \le r_1\left\|\boldsymbol{P}_U^\perp\boldsymbol{U}_\star\boldsymbol{\Sigma}_{\star,1}^2\boldsymbol{U}_\star^\top\boldsymbol{P}_U^\perp\right\|.$$

Denote $\boldsymbol{G}\coloneqq\mathcal{P}_{\mathsf{off\text{-}diag}}(p^{-2}\mathcal{M}_1(\boldsymbol{\mathcal{Y}})\mathcal{M}_1(\boldsymbol{\mathcal{Y}})^\top)$, and $\boldsymbol{G}_\star\coloneqq\boldsymbol{U}_\star\boldsymbol{\Sigma}_{\star,1}^2\boldsymbol{U}_\star^\top$. By the triangle inequality, we obtain

$$\begin{aligned}
\left\|\boldsymbol{P}_U^\perp\boldsymbol{G}_\star\boldsymbol{P}_U^\perp\right\| &\le \left\|\boldsymbol{P}_U^\perp\left(\boldsymbol{G}-\boldsymbol{G}_\star\right)\boldsymbol{P}_U^\perp\right\| + \left\|\boldsymbol{P}_U^\perp\boldsymbol{G}\boldsymbol{P}_U^\perp\right\| \le \|\boldsymbol{G}-\boldsymbol{G}_\star\| + \sigma_{r_1+1}(\boldsymbol{G})\\
&\le \|\boldsymbol{G}-\boldsymbol{G}_\star\| + \sigma_{r_1+1}(\boldsymbol{G}_\star) + \|\boldsymbol{G}-\boldsymbol{G}_\star\| = 2\|\boldsymbol{G}-\boldsymbol{G}_\star\|,
\end{aligned}$$

where the second line follows from Weyl's inequality and that $\boldsymbol{G}_\star$ has rank r_1. Then invoke [CLC+21, Lemma 1] to see

$$\|\boldsymbol{G}-\boldsymbol{G}_\star\| \lesssim \left(\frac{\mu r\log n}{p\sqrt{n_1n_2n_3}} + \sqrt{\frac{\mu rn\log n}{pn_1n_2n_3}} + \frac{\mu r}{n}\right)\sigma_{\max}^2(\boldsymbol{\mathcal{X}}_\star).$$

Combining the above relations, the second term in (58) is bounded by

$$\left\|\boldsymbol{P}_U^\perp\mathcal{M}_1(\boldsymbol{\mathcal{X}}_\star)\right\|_{\mathsf{F}}^2 \lesssim \left(\frac{\mu r^2\kappa^2\log n}{p\sqrt{n_1n_2n_3}} + \sqrt{\frac{\mu r^3\kappa^4n\log n}{pn_1n_2n_3}} + \frac{\mu r^2\kappa^2}{n}\right)\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star).$$

The third and fourth terms in (58) can be bounded similarly.

Under the conditions $\mu r^2\kappa^2 \ll n$ and

$$pn_1n_2n_3 \gtrsim \epsilon_0^{-2}\mu^{3/2}r^2\kappa(\sqrt{r}\vee\kappa)\sqrt{n_1n_2n_3}\log^3 n + \epsilon_0^{-4}\mu^2r^4\kappa^4n\log^5 n,$$

with overwhelming probability, we conclude that

$$\mathrm{dist}(\boldsymbol{F}_+,\boldsymbol{F}_\star) \le (\sqrt{2}+1)^{3/2}\left\|(\boldsymbol{U}_+,\boldsymbol{V}_+,\boldsymbol{W}_+)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_+ - \boldsymbol{\mathcal{X}}_\star\right\|_{\mathsf{F}} \le \epsilon_0\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star).$$

C.4 Proof of local convergence (Lemma 3)

Define the event $\mathcal{E}$ as the intersection of the events that Lemmas 12 and 14 hold, which happens with overwhelming probability. The rest of the proof is then performed under the event that $\mathcal{E}$ holds.

Given that $\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \le \epsilon \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$, the conclusion $\|(\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_t - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \le 3\, \mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$ follows from the relation (38) in Lemma 10. Again, we will reuse the notations in (44) and (34). By the definition of $\mathrm{dist}(\boldsymbol{F}_{t+}, \boldsymbol{F}_\star)$, where $\boldsymbol{F}_{t+}$ is the update before projection, one has

$$\begin{aligned} \mathrm{dist}^2(\boldsymbol{F}_{t+}, \boldsymbol{F}_\star) \le \|(\boldsymbol{U}_{t+}\boldsymbol{Q}_{t,1} - \boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 + \|(\boldsymbol{V}_{t+}\boldsymbol{Q}_{t,2} - \boldsymbol{V}_\star)\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}^2 + \|(\boldsymbol{W}_{t+}\boldsymbol{Q}_{t,3} - \boldsymbol{W}_\star)\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}^2 \\ + \left\|(\boldsymbol{Q}_{t,1}^{-1}, \boldsymbol{Q}_{t,2}^{-1}, \boldsymbol{Q}_{t,3}^{-1}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_{t+} - \boldsymbol{\mathcal{S}}_\star\right\|_{\mathsf{F}}^2. \end{aligned} \tag{59}$$

In the sequel, we shall bound each square on the right hand side of equation (59) separately. After a long journey of computation, the final result is

$$\begin{aligned} \mathrm{dist}^2(\boldsymbol{F}_{t+}, \boldsymbol{F}_\star) \le (1-\eta)^2 \left(\|\boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_V \boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_W \boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}^2\right) \\ - \eta(2-5\eta)\, \|\boldsymbol{\mathcal{T}}_U + \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W\|_{\mathsf{F}}^2 - \eta(2-5\eta) \left(\|\boldsymbol{D}_U\|_{\mathsf{F}}^2 + \|\boldsymbol{D}_V\|_{\mathsf{F}}^2 + \|\boldsymbol{D}_W\|_{\mathsf{F}}^2\right) \\ + 2\eta(1-\eta) C(\epsilon + \delta + \delta^2)\, \mathrm{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star) + \eta^2 C(\epsilon + \delta + \delta^2)\, \mathrm{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star), \end{aligned} \tag{60}$$

where $C > 1$ is some universal constant, and δ is defined as

$$\delta \coloneqq C_T \sqrt{\frac{\mu^2 r^2 n \log n}{p n_1 n_2 n_3}} + C_Y \left(p^{-1} \log^3 n + \sqrt{p^{-1} n \log^5 n}\right) \sqrt{\frac{\mu^3 r^4}{n_1 n_2 n_3}} C_B^3 \kappa^3. \tag{61}$$

Under the condition

$$p n_1 n_2 n_3 \gtrsim \mu^{3/2} r^2 \kappa^3 \sqrt{n_1 n_2 n_3} \log^3 n + \mu^3 r^4 \kappa^6 n \log^5 n,$$

δ is a sufficiently small constant. As long as $\eta \le 2/5$ and ϵ is small, one has $\mathrm{dist}(\boldsymbol{F}_{t+}, \boldsymbol{F}_\star) \le (1 - 0.6\eta)\, \mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$. Finally Lemma 2 implies $\mathrm{dist}(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) \le \mathrm{dist}^2(\boldsymbol{F}_{t+}, \boldsymbol{F}_\star) \le (1 - 0.6\eta)\, \mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$ and the incoherence condition.

It then boils down to expanding and bounding the four terms in (59). As before, we omit the control of the terms pertaining to $\boldsymbol{V}$ and $\boldsymbol{W}$.

C.4.1 Bounding the term related to $\boldsymbol{U}$

The first term in (59) is related to

$$\begin{aligned} (\boldsymbol{U}_{t+}\boldsymbol{Q}_{t,1} - \boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1} &= \Big(\boldsymbol{U} - \eta \mathcal{M}_1\left(p^{-1}\mathcal{P}_\Omega((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_\star)\right) \breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1} - \boldsymbol{U}_\star\Big)\boldsymbol{\Sigma}_{\star,1} \\ &= (1-\eta)\boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1} - \eta \boldsymbol{U}_\star (\breve{\boldsymbol{U}} - \breve{\boldsymbol{U}}_\star)^\top \breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1} \boldsymbol{\Sigma}_{\star,1} \\ &\quad - \eta \mathcal{M}_1\left((p^{-1}\mathcal{P}_\Omega - \mathcal{I})((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_\star)\right) \breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1} \boldsymbol{\Sigma}_{\star,1}. \end{aligned}$$

Take the squared norm of both sides to reach

$$\begin{aligned} \|(\boldsymbol{U}_{t+}\boldsymbol{Q}_{t,1} - \boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 &= \underbrace{\left\|(1-\eta)\boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1} - \eta \boldsymbol{U}_\star (\breve{\boldsymbol{U}} - \breve{\boldsymbol{U}}_\star)^\top \breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1} \boldsymbol{\Sigma}_{\star,1}\right\|_{\mathsf{F}}^2}_{=:\mathfrak{P}_U^{\mathsf{m}}} \\ &- 2\eta(1-\eta) \underbrace{\left\langle \boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}, \mathcal{M}_1\left((p^{-1}\mathcal{P}_\Omega - \mathcal{I})((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_\star)\right) \breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1} \boldsymbol{\Sigma}_{\star,1} \right\rangle}_{=:\mathfrak{P}_U^{\mathsf{p},1}} \\ &+ 2\eta^2 \underbrace{\left\langle \boldsymbol{U}_\star (\breve{\boldsymbol{U}} - \breve{\boldsymbol{U}}_\star)^\top \breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1} \boldsymbol{\Sigma}_{\star,1}, \mathcal{M}_1\left((p^{-1}\mathcal{P}_\Omega - \mathcal{I})((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_\star)\right) \breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1} \boldsymbol{\Sigma}_{\star,1} \right\rangle}_{=:\mathfrak{P}_U^{\mathsf{p},2}} \end{aligned}$$

$$+\,\eta^2 \underbrace{\left\|\mathcal{M}_1\left((p^{-1}\mathcal{P}_\Omega-\mathcal{I})((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}-\boldsymbol{\mathcal{X}}_\star)\right)\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}\right\|_{\mathsf{F}}^2}_{=:\mathfrak{P}_U^{\mathrm{p},3}}.$$

As before, the main term $\mathfrak{P}_U^{\mathrm{m}}$ has been handled in the tensor factorization problem in Section B; see (47) and the bound (45a). Hence we shall focus on the perturbation terms.

Step 1: bounding the term $\mathfrak{P}_U^{\mathrm{p},1}$. First, rewrite $\mathfrak{P}_U^{\mathrm{p},1}$ as the inner product in the tensor space:

$$\mathfrak{P}_U^{\mathrm{p},1}=\left\langle\left(\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}^2(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1},\boldsymbol{V},\boldsymbol{W}\right)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}},(p^{-1}\mathcal{P}_\Omega-\mathcal{I})((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}-\boldsymbol{\mathcal{X}}_\star)\right\rangle.$$

Apply the decomposition

$$\begin{aligned}(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}-\boldsymbol{\mathcal{X}}_\star&=(\boldsymbol{U},\boldsymbol{\Delta}_V,\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}+(\boldsymbol{U},\boldsymbol{V}_\star,\boldsymbol{\Delta}_W)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}+(\boldsymbol{U},\boldsymbol{V}_\star,\boldsymbol{W}_\star)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}-(\boldsymbol{U}_\star,\boldsymbol{V}_\star,\boldsymbol{W}_\star)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star\\&=(\boldsymbol{U},\boldsymbol{\Delta}_V,\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}+(\boldsymbol{U},\boldsymbol{V}_\star,\boldsymbol{\Delta}_W)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}+(\boldsymbol{U},\boldsymbol{V}_\star,\boldsymbol{W}_\star)\boldsymbol{\cdot}\boldsymbol{\Delta}_{\mathcal{S}}+(\boldsymbol{\Delta}_U,\boldsymbol{V}_\star,\boldsymbol{W}_\star)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star\end{aligned}\tag{62}$$

to further expand $\mathfrak{P}_U^{\mathrm{p},1}$ as

$$\begin{aligned}\mathfrak{P}_U^{\mathrm{p},1}&=\underbrace{\left\langle\left(\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}^2(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1},\boldsymbol{V}_\star,\boldsymbol{W}_\star\right)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}},(p^{-1}\mathcal{P}_\Omega-\mathcal{I})\left((\boldsymbol{U},\boldsymbol{V}_\star,\boldsymbol{W}_\star)\boldsymbol{\cdot}\boldsymbol{\Delta}_{\mathcal{S}}+(\boldsymbol{\Delta}_U,\boldsymbol{V}_\star,\boldsymbol{W}_\star)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star\right)\right\rangle}_{=:\mathfrak{P}_U^{\mathrm{p},1,1}}\\&\quad+\underbrace{\left\langle\begin{matrix}\left(\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}^2(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1},\boldsymbol{\Delta}_V,\boldsymbol{W}\right)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}\\+\left(\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}^2(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1},\boldsymbol{V}_\star,\boldsymbol{\Delta}_W\right)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}\end{matrix},(p^{-1}\mathcal{P}_\Omega-\mathcal{I})\left((\boldsymbol{U},\boldsymbol{V}_\star,\boldsymbol{W}_\star)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}-(\boldsymbol{U}_\star,\boldsymbol{V}_\star,\boldsymbol{W}_\star)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star\right)\right\rangle}_{=:\mathfrak{P}_U^{\mathrm{p},1,2}}\\&\quad+\underbrace{\left\langle\left(\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}^2(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1},\boldsymbol{V},\boldsymbol{W}\right)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}},(p^{-1}\mathcal{P}_\Omega-\mathcal{I})\left((\boldsymbol{U},\boldsymbol{\Delta}_V,\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}+(\boldsymbol{U},\boldsymbol{V}_\star,\boldsymbol{\Delta}_W)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}\right)\right\rangle}_{=:\mathfrak{P}_U^{\mathrm{p},1,3}}.\end{aligned}$$

We shall bound each term in the sequel.

- For the first term $\mathfrak{P}_U^{\mathrm{p},1,1}$, we resort to Lemma 12, which leads to

  $$|\mathfrak{P}_U^{\mathrm{p},1,1}|\le C_T\sqrt{\frac{\mu^2r^2n\log n}{pn_1n_2n_3}}\left\|\left(\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}^2(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1},\boldsymbol{V}_\star,\boldsymbol{W}_\star\right)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}\right\|_{\mathsf{F}}\|(\boldsymbol{U},\boldsymbol{V}_\star,\boldsymbol{W}_\star)\boldsymbol{\cdot}\boldsymbol{\Delta}_{\mathcal{S}}+(\boldsymbol{\Delta}_U,\boldsymbol{V}_\star,\boldsymbol{W}_\star)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star\|_{\mathsf{F}}.$$

  Further use (36i) to bound that

  $$\begin{aligned}\left\|\left(\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}^2(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1},\boldsymbol{V}_\star,\boldsymbol{W}_\star\right)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}\right\|_{\mathsf{F}}&=\left\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}^2(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\mathcal{M}_1(\boldsymbol{\mathcal{S}})(\boldsymbol{V}_\star\otimes\boldsymbol{W}_\star)^\top\right\|_{\mathsf{F}}\\&\le\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}\left\|\boldsymbol{\Sigma}_{\star,1}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\mathcal{M}_1(\boldsymbol{\mathcal{S}})\right\|\\&\le\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}(1-\epsilon)^{-5},\end{aligned}$$

  and that

  $$\begin{aligned}\|(\boldsymbol{U},\boldsymbol{V}_\star,\boldsymbol{W}_\star)\boldsymbol{\cdot}\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}&\le\|\boldsymbol{U}\mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})\|_{\mathsf{F}}\le(1+\epsilon)\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}};\\\|(\boldsymbol{\Delta}_U,\boldsymbol{V}_\star,\boldsymbol{W}_\star)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star\|_{\mathsf{F}}&\le\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}.\end{aligned}$$

  Combine the preceding bounds to see

  $$|\mathfrak{P}_U^{\mathrm{p},1,1}|\le C_T\sqrt{\frac{\mu^2r^2n\log n}{pn_1n_2n_3}}\frac{\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}}{(1-\epsilon)^5}\left(\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}+(1+\epsilon)\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\right).$$

- For the second term $\mathfrak{P}_U^{\mathrm{p},1,2}$, our main hammer is Lemma 14, which implies

$$|\mathfrak{P}_U^{\mathrm{p},1,2}| \le C_Y \left(p^{-1}\log^3 n + \sqrt{p^{-1} n \log^5 n}\right) \left\|\boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}^2 (\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1} \mathcal{M}_1(\boldsymbol{\mathcal{S}})\right\|_{\mathsf{F}} \left(\|\boldsymbol{U}\mathcal{M}_1(\boldsymbol{\mathcal{S}})\|_{2,\infty} + \|\boldsymbol{U}_\star \mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)\|_{2,\infty}\right)$$
$$(\|\boldsymbol{\Delta}_V\|_{\mathsf{F}}\|\boldsymbol{W}\|_{\mathsf{F}} + \|\boldsymbol{V}_\star\|_{\mathsf{F}}\|\boldsymbol{\Delta}_W\|_{\mathsf{F}})\, \|\boldsymbol{V}_\star\|_{2,\infty}\|\boldsymbol{W}_\star\|_{2,\infty}.$$

  Use results in Lemma 11, together with the bounds

$$\|\boldsymbol{\Delta}_V\|_{\mathsf{F}} \le \frac{\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}}{\sigma_{\min}(\boldsymbol{\Sigma}_{\star,2})} \le \frac{\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}}{\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)}; \qquad \|\boldsymbol{\Delta}_W\|_{\mathsf{F}} \le \frac{\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}}{\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)};$$
$$\|\boldsymbol{W}\|_{\mathsf{F}} \le \sqrt{r_3}\|\boldsymbol{W}\| \le \sqrt{r_3}(1+\epsilon); \qquad \|\boldsymbol{V}_\star\|_{\mathsf{F}} = \sqrt{r_2};$$
$$\|\boldsymbol{U}_\star\mathcal{M}_1(\boldsymbol{S}_\star)\|_{2,\infty} \le \|\boldsymbol{U}_\star\|_{2,\infty}\|\mathcal{M}_1(\boldsymbol{S}_\star)\| \le \sqrt{\frac{\mu r}{n_1}}\sigma_{\max}(\boldsymbol{\mathcal{X}}_\star); \qquad \|\boldsymbol{V}_\star\|_{2,\infty} \le \sqrt{\frac{\mu r}{n_2}}; \qquad \|\boldsymbol{W}_\star\|_{2,\infty} \le \sqrt{\frac{\mu r}{n_3}},$$

  to arrive at the conclusion that

$$\begin{aligned}|\mathfrak{P}_U^{\mathrm{p},1,2}| &\le C_Y \left(p^{-1}\log^3 n + \sqrt{p^{-1} n \log^5 n}\right) \frac{\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}}{(1-\epsilon)^5}\left((1-\epsilon)^{-2}C_B + 1\right)\sqrt{\frac{\mu r}{n_1}}\sigma_{\max}(\boldsymbol{\mathcal{X}}_\star) \\ &\qquad \left(\frac{\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}}{\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)}\sqrt{r}(1+\epsilon) + \sqrt{r}\frac{\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}}{\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)}\right)\sqrt{\frac{\mu r}{n_2}}\sqrt{\frac{\mu r}{n_3}} \\ &= C_Y \left(p^{-1}\log^3 n + \sqrt{p^{-1} n \log^5 n}\right)\sqrt{\frac{\mu^3 r^4}{n_1 n_2 n_3}}\frac{(1-\epsilon)^{-2}C_B + 1}{(1-\epsilon)^5}\kappa \\ &\qquad \|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}\left((1+\epsilon)\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}\right).\end{aligned}$$

- Repeat similar arguments, we can obtain the bound on $\mathfrak{P}_U^{\mathrm{p},1,3}$:

$$\begin{aligned}|\mathfrak{P}_U^{\mathrm{p},1,3}| &\le C_Y \left(p^{-1}\log^3 n + \sqrt{p^{-1} n \log^5 n}\right) \left\|\boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}^2 (\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1} \mathcal{M}_1(\boldsymbol{\mathcal{S}})\right\|_{\mathsf{F}} \|\boldsymbol{U}\mathcal{M}_1(\boldsymbol{\mathcal{S}})\|_{2,\infty} \\ &\qquad \|\boldsymbol{V}\|_{2,\infty}\|\boldsymbol{W}\|_{2,\infty}(\|\boldsymbol{\Delta}_V\|_{\mathsf{F}}\|\boldsymbol{W}\|_{\mathsf{F}} + \|\boldsymbol{V}_\star\|_{\mathsf{F}}\|\boldsymbol{\Delta}_W\|_{\mathsf{F}}) \\ &\le C_Y \left(p^{-1}\log^3 n + \sqrt{p^{-1} n \log^5 n}\right)\frac{\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}}{(1-\epsilon)^5}\frac{C_B}{(1-\epsilon)^2}\sqrt{\frac{\mu r}{n_1}}\sigma_{\max}(\boldsymbol{\mathcal{X}}_\star) \\ &\qquad \frac{C_B\kappa}{(1-\epsilon)^3}\sqrt{\frac{\mu r}{n_2}}\frac{C_B\kappa}{(1-\epsilon)^3}\sqrt{\frac{\mu r}{n_3}}\left(\frac{\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}}{\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)}\sqrt{r}(1+\epsilon) + \sqrt{r}\frac{\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}}{\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)}\right) \\ &\le C_Y \left(p^{-1}\log^3 n + \sqrt{p^{-1} n \log^5 n}\right)\sqrt{\frac{\mu^3 r^4}{n_1 n_2 n_3}}\frac{C_B^3\kappa^3}{(1-\epsilon)^{13}} \\ &\qquad \|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}\left((1+\epsilon)\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}\right).\end{aligned}$$

In total, we have

$$|\mathfrak{P}_U^{\mathrm{p},1}| \le |\mathfrak{P}_U^{\mathrm{p},1,1}| + |\mathfrak{P}_U^{\mathrm{p},1,2}| + |\mathfrak{P}_U^{\mathrm{p},1,3}| \lesssim \delta\,\mathrm{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star),$$

where we recall the definition of δ in (61).

Step 2: bounding the term $\mathfrak{P}_U^{\mathrm{p},2}$. We begin by rewriting $\mathfrak{P}_U^{\mathrm{p},2}$ as

$$\mathfrak{P}_U^{\mathrm{p},2} = \left\langle \left(\boldsymbol{U}_\star(\breve{\boldsymbol{U}} - \breve{\boldsymbol{U}}_\star)^\top \breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}^2(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}, \boldsymbol{V}, \boldsymbol{W}\right)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}, (p^{-1}\mathcal{P}_\Omega - \mathcal{I})((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_\star)\right\rangle.$$

Compared to $\mathfrak{P}_U^{\mathrm{p},1}$, the only difference is that the leading term $\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}$ in the first argument of the inner product is replaced by $\boldsymbol{U}_\star(\breve{\boldsymbol{U}} - \breve{\boldsymbol{U}}_\star)^\top \breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}$. Note that

$$\left\|\boldsymbol{U}_\star(\breve{\boldsymbol{U}} - \breve{\boldsymbol{U}}_\star)^\top \breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}\right\|_{\mathsf{F}} \le \left\|\breve{\boldsymbol{U}} - \breve{\boldsymbol{U}}_\star\right\|_{\mathsf{F}}\left\|\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}\right\|_{\mathsf{F}}$$

$$\leq \frac{1+\epsilon+\frac{1}{3}\epsilon^2}{(1-\epsilon)^3}\left(\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\right).$$

Omitting the somewhat tedious details, we can go through the same argument as bounding $\mathfrak{P}_U^{\mathrm{p},1}$ and arrive at

$$\begin{aligned}
|\mathfrak{P}_U^{\mathrm{p},2}| &\leq C_T\sqrt{\frac{\mu^2r^2n\log n}{pn_1n_2n_3}}\frac{1+\epsilon+\frac{1}{3}\epsilon^2}{(1-\epsilon)^8}\left(\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\right)\left(\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}+(1+\epsilon)\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\right)\\
&\quad + C_Y\left(p^{-1}\log^3 n+\sqrt{p^{-1}n\log^5 n}\right)\sqrt{\frac{\mu^3r^4}{n_1n_2n_3}}\frac{(1+\epsilon+\frac{1}{3}\epsilon^2)((1-\epsilon)^{-2}C_B+1)}{(1-\epsilon)^8}\kappa\\
&\qquad \left(\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\right)\left((1+\epsilon)\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}\right)\\
&\quad + C_Y\left(p^{-1}\log^3 n+\sqrt{p^{-1}n\log^5 n}\right)\sqrt{\frac{\mu^3r^4}{n_1n_2n_3}}\frac{(1+\epsilon+\frac{1}{3}\epsilon^2)C_B^3\kappa^3}{(1-\epsilon)^{16}}\\
&\qquad \left(\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\right)\left((1+\epsilon)\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}\right)\\
&\lesssim \delta\,\mathrm{dist}^2(\boldsymbol{F}_t,\boldsymbol{F}_\star).
\end{aligned}$$

Step 3: bounding the term $\mathfrak{P}_U^{\mathrm{p},3}$. Use the variational representation of the Frobenius norm to write

$$\sqrt{\mathfrak{P}_U^{\mathrm{p},3}}=\left\langle\left(\widetilde{\boldsymbol{U}}\boldsymbol{\Sigma}_{\star,1}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1},\boldsymbol{V},\boldsymbol{W}\right)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}},(p^{-1}\mathcal{P}_\Omega-\mathcal{I})((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}-\boldsymbol{\mathcal{X}}_\star)\right\rangle$$

for some $\widetilde{\boldsymbol{U}}\in\mathbb{R}^{n_1\times r_1}$ obeying $\|\widetilde{\boldsymbol{U}}\|_{\mathsf{F}}=1$. Repeat the same argument as bounding $\mathfrak{P}_U^{\mathrm{p},1}$ with proper modifications to yield

$$\begin{aligned}
\sqrt{\mathfrak{P}_U^{\mathrm{p},3}} &\leq C_T\sqrt{\frac{\mu^2r^2n\log n}{pn_1n_2n_3}}(1-\epsilon)^{-5}\left(\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}+(1+\epsilon)\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\right)\\
&\quad + C_Y\left(p^{-1}\log^3 n+\sqrt{p^{-1}n\log^5 n}\right)\sqrt{\frac{\mu^3r^4}{n_1n_2n_3}}\frac{(1-\epsilon)^{-2}C_B+1}{(1-\epsilon)^5}\kappa\left((1+\epsilon)\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}\right)\\
&\quad + C_Y\left(p^{-1}\log^3 n+\sqrt{p^{-1}n\log^5 n}\right)\sqrt{\frac{\mu^3r^4}{n_1n_2n_3}}\frac{C_B^3\kappa^3}{(1-\epsilon)^{13}}\left((1+\epsilon)\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}\right)\\
&\lesssim \delta\,\mathrm{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star).
\end{aligned}$$

Then take the square of both sides to see

$$\mathfrak{P}_U^{\mathrm{p},3}\lesssim\delta^2\,\mathrm{dist}^2(\boldsymbol{F}_t,\boldsymbol{F}_\star).$$

C.4.2 Bounding the term related to $\boldsymbol{\mathcal{S}}$

The last term of (59) is related to

$$\begin{aligned}
&(\boldsymbol{Q}_{t,1}^{-1},\boldsymbol{Q}_{t,2}^{-1},\boldsymbol{Q}_{t,3}^{-1})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_{t+}-\boldsymbol{\mathcal{S}}_\star\\
&\quad=\boldsymbol{\mathcal{S}}-\eta\left((\boldsymbol{U}^\top\boldsymbol{U})^{-1}\boldsymbol{U}^\top,(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\boldsymbol{V}^\top,(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\right)\boldsymbol{\cdot}p^{-1}\mathcal{P}_\Omega\left((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}-\boldsymbol{\mathcal{X}}_\star\right)-\boldsymbol{\mathcal{S}}_\star\\
&\quad=(1-\eta)\boldsymbol{\Delta}_{\mathcal{S}}-\eta\left((\boldsymbol{U}^\top\boldsymbol{U})^{-1}\boldsymbol{U}^\top,(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\boldsymbol{V}^\top,(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\right)\boldsymbol{\cdot}\left((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star-\boldsymbol{\mathcal{X}}_\star\right)\\
&\qquad-\eta\left((\boldsymbol{U}^\top\boldsymbol{U})^{-1}\boldsymbol{U}^\top,(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\boldsymbol{V}^\top,(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\right)\boldsymbol{\cdot}(p^{-1}\mathcal{P}_\Omega-\mathcal{I})((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}-\boldsymbol{\mathcal{X}}_\star).
\end{aligned}$$

Expand its squared norm to obtain

$$\left\|(\boldsymbol{Q}_{t,1}^{-1},\boldsymbol{Q}_{t,2}^{-1},\boldsymbol{Q}_{t,3}^{-1})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_{t+}-\boldsymbol{\mathcal{S}}_\star\right\|_{\mathsf{F}}^2$$

$$
\begin{aligned}
&= \underbrace{\left\|(1-\eta)\boldsymbol{\Delta}_{\mathcal{S}} - \eta\left((\boldsymbol{U}^\top\boldsymbol{U})^{-1}\boldsymbol{U}^\top, (\boldsymbol{V}^\top\boldsymbol{V})^{-1}\boldsymbol{V}^\top, (\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\right)\boldsymbol{\cdot}((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star - \boldsymbol{\mathcal{X}}_\star)\right\|_{\mathsf{F}}^2}_{=:\mathfrak{P}^{\mathrm{m}}_{\mathcal{S}}} \\
&\quad - 2\eta(1-\eta)\underbrace{\left\langle\boldsymbol{\Delta}_{\mathcal{S}}, \left((\boldsymbol{U}^\top\boldsymbol{U})^{-1}\boldsymbol{U}^\top, (\boldsymbol{V}^\top\boldsymbol{V})^{-1}\boldsymbol{V}^\top, (\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\right)\boldsymbol{\cdot}(p^{-1}\mathcal{P}_\Omega - \mathcal{I})((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_\star)\right\rangle}_{=:\mathfrak{P}^{\mathrm{p},1}_{\mathcal{S}}} \\
&\quad + 2\eta^2 \underbrace{\left\langle \begin{array}{l}\left((\boldsymbol{U}^\top\boldsymbol{U})^{-1}\boldsymbol{U}^\top, (\boldsymbol{V}^\top\boldsymbol{V})^{-1}\boldsymbol{V}^\top, (\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\right)\boldsymbol{\cdot}((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star - \boldsymbol{\mathcal{X}}_\star)\,, \\ \qquad \left((\boldsymbol{U}^\top\boldsymbol{U})^{-1}\boldsymbol{U}^\top, (\boldsymbol{V}^\top\boldsymbol{V})^{-1}\boldsymbol{V}^\top, (\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\right)\boldsymbol{\cdot}(p^{-1}\mathcal{P}_\Omega - \mathcal{I})((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_\star\end{array}\right\rangle}_{=:\mathfrak{P}^{\mathrm{p},2}_{\mathcal{S}}} \\
&\quad + \eta^2 \underbrace{\left\|\left((\boldsymbol{U}^\top\boldsymbol{U})^{-1}\boldsymbol{U}^\top, (\boldsymbol{V}^\top\boldsymbol{V})^{-1}\boldsymbol{V}^\top, (\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\right)\boldsymbol{\cdot}(p^{-1}\mathcal{P}_\Omega - \mathcal{I})((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_\star)\right\|_{\mathsf{F}}^2}_{=:\mathfrak{P}^{\mathrm{p},3}_{\mathcal{S}}}.
\end{aligned}
$$

Recall that the main term $\mathfrak{P}^{\mathrm{m}}_{\mathcal{S}}$ has been controlled in Section B; see (48) and the bound (45d). We therefore concentrate on the remaining perturbation terms.

Step 1: bounding the term $\mathfrak{P}^{\mathrm{p},1}_{\mathcal{S}}$. Write $\mathfrak{P}^{\mathrm{p},1}_{\mathcal{S}}$ as

$$
\mathfrak{P}^{\mathrm{p},1}_{\mathcal{S}} = \left\langle\left(\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}, \boldsymbol{V}(\boldsymbol{V}^\top\boldsymbol{V})^{-1}, \boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\right)\boldsymbol{\cdot}\boldsymbol{\Delta}_{\mathcal{S}}, (p^{-1}\mathcal{P}_\Omega - \mathcal{I})((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_\star)\right\rangle.
$$

Use the decomposition (62) to further obtain

$$
\begin{aligned}
\mathfrak{P}^{\mathrm{p},1}_{\mathcal{S}} &= \underbrace{\left\langle\left(\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}, \boldsymbol{V}_\star(\boldsymbol{V}^\top\boldsymbol{V})^{-1}, \boldsymbol{W}_\star(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\right)\boldsymbol{\cdot}\boldsymbol{\Delta}_{\mathcal{S}}, (p^{-1}\mathcal{P}_\Omega - \mathcal{I})\left((\boldsymbol{U},\boldsymbol{V}_\star,\boldsymbol{W}_\star)\boldsymbol{\cdot}\boldsymbol{\Delta}_{\mathcal{S}} + (\boldsymbol{\Delta}_U,\boldsymbol{V}_\star,\boldsymbol{W}_\star)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star\right)\right\rangle}_{=:\mathfrak{P}^{\mathrm{p},1,1}_{\mathcal{S}}} \\
&\quad + \underbrace{\left\langle\begin{array}{l}\left(\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}, \boldsymbol{\Delta}_V(\boldsymbol{V}^\top\boldsymbol{V})^{-1}, \boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\right)\boldsymbol{\cdot}\boldsymbol{\Delta}_{\mathcal{S}} \\ + \left(\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}, \boldsymbol{V}_\star(\boldsymbol{V}^\top\boldsymbol{V})^{-1}, \boldsymbol{\Delta}_W(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\right)\boldsymbol{\cdot}\boldsymbol{\Delta}_{\mathcal{S}}\end{array}, (p^{-1}\mathcal{P}_\Omega - \mathcal{I})\left((\boldsymbol{U},\boldsymbol{V}_\star,\boldsymbol{W}_\star)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}} - (\boldsymbol{U}_\star,\boldsymbol{V}_\star,\boldsymbol{W}_\star)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star\right)\right\rangle}_{=:\mathfrak{P}^{\mathrm{p},1,2}_{\mathcal{S}}} \\
&\quad + \underbrace{\left\langle\left(\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}, \boldsymbol{V}(\boldsymbol{V}^\top\boldsymbol{V})^{-1}, \boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\right)\boldsymbol{\cdot}\boldsymbol{\Delta}_{\mathcal{S}}, (p^{-1}\mathcal{P}_\Omega - \mathcal{I})\left((\boldsymbol{U},\boldsymbol{\Delta}_V,\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}} + (\boldsymbol{U},\boldsymbol{V}_\star,\boldsymbol{\Delta}_W)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}\right)\right\rangle}_{=:\mathfrak{P}^{\mathrm{p},1,3}_{\mathcal{S}}}.
\end{aligned}
$$

We then bound each term in sequel.

- Regarding the first term $\mathfrak{P}^{\mathrm{p},1,1}_{\mathcal{S}}$, we can apply Lemma 12 to see

  $$
  \begin{aligned}
  |\mathfrak{P}^{\mathrm{p},1,1}_{\mathcal{S}}| \le C_T\sqrt{\frac{\mu^2r^2n\log n}{pn_1n_2n_3}}&\left\|\left(\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}, \boldsymbol{V}_\star(\boldsymbol{V}^\top\boldsymbol{V})^{-1}, \boldsymbol{W}_\star(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\right)\boldsymbol{\cdot}\boldsymbol{\Delta}_{\mathcal{S}}\right\|_{\mathsf{F}} \\
  &\left\|(\boldsymbol{U},\boldsymbol{V}_\star,\boldsymbol{W}_\star)\boldsymbol{\cdot}\boldsymbol{\Delta}_{\mathcal{S}} + (\boldsymbol{\Delta}_U,\boldsymbol{V}_\star,\boldsymbol{W}_\star)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star\right\|_{\mathsf{F}}.
  \end{aligned}
  $$

  In addition, notice that

  $$
  \begin{aligned}
  \left\|\left(\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}, \boldsymbol{V}_\star(\boldsymbol{V}^\top\boldsymbol{V})^{-1}, \boldsymbol{W}_\star(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\right)\boldsymbol{\cdot}\boldsymbol{\Delta}_{\mathcal{S}}\right\|_{\mathsf{F}} &\le \left\|\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\right\|\left\|(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\right\|\left\|(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\right\|\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}} \\
  &\le (1-\epsilon)^{-5}\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}},
  \end{aligned}
  $$

  which further implies

  $$
  |\mathfrak{P}^{\mathrm{p},1,1}_{\mathcal{S}}| \le C_T\sqrt{\frac{\mu^2r^2n\log n}{pn_1n_2n_3}}(1-\epsilon)^{-5}\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\left(\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}} + (1+\epsilon)\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\right).
  $$

- Now we turn to the second term $\mathfrak{P}^{\mathrm{p},1,2}_{\mathcal{S}}$, for which Lemma 14 yields

  $$
  |\mathfrak{P}^{\mathrm{p},1,2}_{\mathcal{S}}| \le C_Y\left(p^{-1}\log^3 n + \sqrt{p^{-1}n\log^5 n}\right)\left\|\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})\right\|_{\mathsf{F}}\left(\|\boldsymbol{U}\mathcal{M}_1(\boldsymbol{\mathcal{S}})\|_{2,\infty} + \|\boldsymbol{U}_\star\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)\|_{2,\infty}\right)
  $$

$$\left(\left\|\boldsymbol{\Delta}_V(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\right\|_{\mathsf{F}}\left\|\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\right\|_{\mathsf{F}}+\left\|\boldsymbol{V}_\star(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\right\|_{\mathsf{F}}\left\|\boldsymbol{\Delta}_W(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\right\|_{\mathsf{F}}\right)\|\boldsymbol{V}_\star\|_{2,\infty}\|\boldsymbol{W}_\star\|_{2,\infty}.$$

The results in Lemma 11 together with the bounds

$$\begin{aligned}
\left\|\boldsymbol{\Delta}_V(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\right\|_{\mathsf{F}} &\le \|\boldsymbol{\Delta}_V\|_{\mathsf{F}}\left\|(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\right\| \le (1-\epsilon)^{-2}\|\boldsymbol{\Delta}_V\|_{\mathsf{F}} \le \frac{\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}}{(1-\epsilon)^2\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)};\\
\left\|\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\right\|_{\mathsf{F}} &\le \sqrt{r_3}\left\|\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\right\| \le \sqrt{r_3}(1-\epsilon)^{-1};\\
\left\|\boldsymbol{V}_\star(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\right\|_{\mathsf{F}} &\le \|\boldsymbol{V}_\star\|_{\mathsf{F}}\left\|(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\right\| \le \sqrt{r_2}(1-\epsilon)^{-2};\\
\left\|\boldsymbol{\Delta}_W(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\right\|_{\mathsf{F}} &\le \|\boldsymbol{\Delta}_W\|_{\mathsf{F}}\left\|(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\right\| \le \|\boldsymbol{\Delta}_W\|_{\mathsf{F}}(1-\epsilon)^{-2} \le \frac{\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}}{(1-\epsilon)^2\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)},
\end{aligned}$$

allow us to continue the bound

$$|\mathfrak{P}_{\mathcal{S}}^{\mathrm{p},1,2}| \le C_Y\left(p^{-1}\log^3 n+\sqrt{p^{-1}n\log^5 n}\right)\sqrt{\frac{\mu^3 r^4}{n_1n_2n_3}}\frac{(1-\epsilon)^{-2}C_B+1}{(1-\epsilon)^5}\kappa\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\left((1-\epsilon)\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}\right).$$

- A similar strategy bounds $\mathfrak{P}_{\mathcal{S}}^{\mathrm{p},1,3}$ as

$$\begin{aligned}
|\mathfrak{P}_{\mathcal{S}}^{\mathrm{p},1,3}| \le C_Y\left(p^{-1}\log^3 n+\sqrt{p^{-1}n\log^5 n}\right)&\left\|\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})\right\|_{\mathsf{F}}\|\boldsymbol{U}\mathcal{M}_1(\boldsymbol{\mathcal{S}})\|_{2,\infty}\\
\left\|\boldsymbol{V}(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\right\|_{2,\infty}&\left\|\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\right\|_{2,\infty}\left(\|\boldsymbol{\Delta}_V\|_{\mathsf{F}}\|\boldsymbol{W}\|_{\mathsf{F}}+\|\boldsymbol{V}_\star\|_{\mathsf{F}}\|\boldsymbol{\Delta}_W\|_{\mathsf{F}}\right).
\end{aligned}$$

  Further combine (53c) and (36d) to see

$$\begin{aligned}
\left\|\boldsymbol{V}(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\right\|_{2,\infty} &\le \|\boldsymbol{V}\|_{2,\infty}\left\|(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\right\| \le (1-\epsilon)^{-5}C_B\sqrt{\frac{\mu r}{n_2}}\kappa;\\
\left\|\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\right\|_{2,\infty} &\le \|\boldsymbol{W}\|_{2,\infty}\left\|(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\right\| \le (1-\epsilon)^{-5}C_B\sqrt{\frac{\mu r}{n_3}}\kappa.
\end{aligned}$$

  These taken collectively with the results in Lemma 11 yield

$$|\mathfrak{P}_{\mathcal{S}}^{\mathrm{p},1,3}| \le C_Y\left(p^{-1}\log^3 n+\sqrt{p^{-1}n\log^5 n}\right)\sqrt{\frac{\mu^3 r^4}{n_1n_2n_3}}\frac{C_B^3\kappa^3}{(1-\epsilon)^{13}}\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\left((1+\epsilon)\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}\right).$$

  In the end, we conclude that

$$|\mathfrak{P}_{\mathcal{S}}^{\mathrm{p},1}| \le |\mathfrak{P}_{\mathcal{S}}^{\mathrm{p},1,1}|+|\mathfrak{P}_{\mathcal{S}}^{\mathrm{p},1,2}|+|\mathfrak{P}_{\mathcal{S}}^{\mathrm{p},1,3}| \lesssim \delta\,\mathrm{dist}^2(\boldsymbol{F}_t,\boldsymbol{F}_\star),$$

where we recall the definition of δ in (61).

Step 2: bounding the term $\mathfrak{P}_{\mathcal{S}}^{\mathrm{p},2}$. Write $\mathfrak{P}_{\mathcal{S}}^{\mathrm{p},2}$ as

$$\begin{aligned}
\mathfrak{P}_{\mathcal{S}}^{\mathrm{p},2} = \Big\langle &\left(\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-2}\boldsymbol{U}^\top,\boldsymbol{V}(\boldsymbol{V}^\top\boldsymbol{V})^{-2}\boldsymbol{V}^\top,\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-2}\boldsymbol{W}^\top\right)\boldsymbol{\cdot}\left((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star-\boldsymbol{\mathcal{X}}_\star\right),\\
&(p^{-1}\mathcal{P}_\Omega-\mathcal{I})\left((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}-\boldsymbol{\mathcal{X}}_\star\right)\Big\rangle.
\end{aligned}$$

Compared to $\mathfrak{P}_{\mathcal{S}}^{\mathrm{p},1}$, the only difference is that the quantity $\boldsymbol{\Delta}_{\mathcal{S}}$ in the first argument of the inner product is replaced by

$$\left((\boldsymbol{U}^\top\boldsymbol{U})^{-1}\boldsymbol{U}^\top,(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\boldsymbol{V}^\top,(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\right)\boldsymbol{\cdot}\left((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star-\boldsymbol{\mathcal{X}}_\star\right),$$

whose Frobenius norm can be bounded by

$$\left\|\left((\boldsymbol{U}^\top\boldsymbol{U})^{-1}\boldsymbol{U}^\top,(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\boldsymbol{V}^\top,(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\right)\boldsymbol{\cdot}\left((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star-\boldsymbol{\mathcal{X}}_\star\right)\right\|_{\mathsf{F}}$$

$$\begin{aligned}&\le \left\|\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\right\|_{\mathsf{F}}\left\|\boldsymbol{V}(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\right\|_{\mathsf{F}}\left\|\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\right\|_{\mathsf{F}}\|(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star-\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}\\ &\le \frac{1+\epsilon+\frac{1}{3}\epsilon^2}{(1-\epsilon)^3}\left(\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}\right).\end{aligned}$$

We can then repeat the same argument as bounding $\mathfrak{P}_{\mathcal{S}}^{\mathrm{p},1}$ to obtain

$$|\mathfrak{P}_{\mathcal{S}}^{\mathrm{p},2}|\lesssim\delta\,\mathrm{dist}^2(\boldsymbol{F}_t,\boldsymbol{F}_\star).$$

For the sake of space, we omit the details.

Step 3: bounding the term $\mathfrak{P}_{\mathcal{S}}^{\mathrm{p},3}$. Use the variational representation of the Frobenius norm to write

$$\sqrt{\mathfrak{P}_{\mathcal{S}}^{\mathrm{p},3}}=\left\langle\left(\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1},\boldsymbol{V}(\boldsymbol{V}^\top\boldsymbol{V})^{-1},\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\right)\boldsymbol{\cdot}\widetilde{\boldsymbol{\mathcal{S}}},(p^{-1}\mathcal{P}_\Omega-\mathcal{I})((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}-\boldsymbol{\mathcal{X}}_\star)\right\rangle$$

for some $\widetilde{\boldsymbol{\mathcal{S}}}\in\mathbb{R}^{n_1\times n_2\times n_3}$ obeying $\|\widetilde{\boldsymbol{\mathcal{S}}}\|_{\mathsf{F}}=1$. Repeating the same argument as bounding $\mathfrak{P}_{\mathcal{S}}^{\mathrm{p},1}$ with proper modifications to yield the bound

$$\mathfrak{P}_{\mathcal{S}}^{\mathrm{p},3}\lesssim\delta^2\,\mathrm{dist}^2(\boldsymbol{F}_t,\boldsymbol{F}_\star)$$

then complete the proof.

D Proof for Tensor Regression

Before embarking on the proof, we state a useful lemma regarding TRIP (cf. Definition 3).

Lemma 15 ([HWZ20, Lemma E.7]). *Suppose that $\mathcal{A}(\cdot)$ obeys the $2\boldsymbol{r}$-TRIP with a constant $\delta_{2\boldsymbol{r}}$. Then for all $\boldsymbol{\mathcal{X}}_1,\boldsymbol{\mathcal{X}}_2\in\mathbb{R}^{n_1\times n_2\times n_3}$ of multilinear rank at most $\boldsymbol{r}$, one has*

$$\left|\langle\mathcal{A}(\boldsymbol{\mathcal{X}}_1),\mathcal{A}(\boldsymbol{\mathcal{X}}_2)\rangle-\langle\boldsymbol{\mathcal{X}}_1,\boldsymbol{\mathcal{X}}_2\rangle\right|\le\delta_{2\boldsymbol{r}}\|\boldsymbol{\mathcal{X}}_1\|_{\mathsf{F}}\|\boldsymbol{\mathcal{X}}_2\|_{\mathsf{F}},$$

or equivalently,

$$\left|\langle(\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{\mathcal{X}}_1),\boldsymbol{\mathcal{X}}_2\rangle\right|\le\delta_{2\boldsymbol{r}}\|\boldsymbol{\mathcal{X}}_1\|_{\mathsf{F}}\|\boldsymbol{\mathcal{X}}_2\|_{\mathsf{F}}.$$

D.1 Proof of local convergence (Lemma 4)

Given that $\mathrm{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star)\le\epsilon\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$, the conclusion $\|(\boldsymbol{U}_t,\boldsymbol{V}_t,\boldsymbol{W}_t)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_t-\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}\le 3\,\mathrm{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star)$ directly follows from the relation (38) in Lemma 10. Hence we will focus on controlling $\mathrm{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star)$.

As in the proof of Theorem 3, we reuse the set of notation in (44) and (34), and the definition of $\mathrm{dist}(\boldsymbol{F}_{t+1},\boldsymbol{F}_\star)$ to obtain

$$\begin{aligned}\mathrm{dist}^2(\boldsymbol{F}_{t+1},\boldsymbol{F}_\star)\le{}&\|(\boldsymbol{U}_{t+1}\boldsymbol{Q}_{t,1}-\boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2+\|(\boldsymbol{V}_{t+1}\boldsymbol{Q}_{t,2}-\boldsymbol{V}_\star)\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}^2+\|(\boldsymbol{W}_{t+1}\boldsymbol{Q}_{t,3}-\boldsymbol{W}_\star)\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}^2\\&+\left\|(\boldsymbol{Q}_{t,1}^{-1},\boldsymbol{Q}_{t,2}^{-1},\boldsymbol{Q}_{t,3}^{-1})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_{t+1}-\boldsymbol{\mathcal{S}}_\star\right\|_{\mathsf{F}}^2.\end{aligned}\tag{63}$$

We shall bound each square in the right hand side of the bound (63) separately. The final result is

$$\begin{aligned}\mathrm{dist}^2(\boldsymbol{F}_{t+1},\boldsymbol{F}_\star)\le{}&(1-\eta)^2\left(\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2+\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}^2+\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}^2+\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}^2\right)\\&-\eta(2-5\eta)\,\|\boldsymbol{\mathcal{T}}_U+\boldsymbol{\mathcal{T}}_V+\boldsymbol{\mathcal{T}}_W\|_{\mathsf{F}}^2-\eta(2-5\eta)\left(\|\boldsymbol{D}_U\|_{\mathsf{F}}^2+\|\boldsymbol{D}_V\|_{\mathsf{F}}^2+\|\boldsymbol{D}_W\|_{\mathsf{F}}^2\right)\\&+2\eta(1-\eta)C(\epsilon+\delta_{2\boldsymbol{r}}+\delta_{2\boldsymbol{r}}^2)\,\mathrm{dist}^2(\boldsymbol{F}_t,\boldsymbol{F}_\star)+\eta^2C(\epsilon+\delta_{2\boldsymbol{r}}+\delta_{2\boldsymbol{r}}^2)\,\mathrm{dist}^2(\boldsymbol{F}_t,\boldsymbol{F}_\star),\end{aligned}\tag{64}$$

where $C>1$ is some universal constant. As long as $\eta\le 2/5$, and ϵ, $\delta_{2\boldsymbol{r}}$ are sufficiently small constants, one reaches the desired conclusion $\mathrm{dist}(\boldsymbol{F}_{t+1},\boldsymbol{F}_\star)\le(1-0.6\eta)\,\mathrm{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star)$.

In the following subsections, we provide bounds on the four terms in the right hand side of (63). In a nutshell, the bounds that are sought after are reminiscent of those established in (45), with additional perturbation terms introduced due to incomplete measurements, manifested via the TRIP parameter $\delta_{2\boldsymbol{r}}$. Once established, the claimed bound (64) easily follows. In light of the symmetry among $\boldsymbol{U},\boldsymbol{V}$, and $\boldsymbol{W}$, we omit the control of the terms pertaining to $\boldsymbol{V}$ and $\boldsymbol{W}$.

D.1.1 Bounding the term pertaining to $\boldsymbol{U}$

The first term in equation (63) is given by

$$\begin{aligned}(\boldsymbol{U}_{t+1}\boldsymbol{Q}_{t,1}-\boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1} &= \Big(\boldsymbol{U}-\eta\mathcal{M}_1\left(\mathcal{A}^*\mathcal{A}((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}-\boldsymbol{\mathcal{X}}_\star)\right)\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}-\boldsymbol{U}_\star\Big)\boldsymbol{\Sigma}_{\star,1}\\ &= (1-\eta)\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}-\eta\boldsymbol{U}_\star(\breve{\boldsymbol{U}}-\breve{\boldsymbol{U}}_\star)^\top\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}\\ &\qquad -\eta\mathcal{M}_1\left((\mathcal{A}^*\mathcal{A}-\mathcal{I})((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}-\boldsymbol{\mathcal{X}}_\star)\right)\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1},\end{aligned}$$

where we separate the population term from the perturbation term. Take the squared norm of both sides to see

$$\begin{aligned}\|(\boldsymbol{U}_{t+1}\boldsymbol{Q}_{t,1}-\boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 &= \underbrace{\left\|(1-\eta)\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}-\eta\boldsymbol{U}_\star(\breve{\boldsymbol{U}}-\breve{\boldsymbol{U}}_\star)^\top\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}\right\|_{\mathsf{F}}^2}_{=:\mathfrak{R}_U^{\mathsf{m}}}\\ &-2\eta(1-\eta)\underbrace{\left\langle\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1},\mathcal{M}_1\left((\mathcal{A}^*\mathcal{A}-\mathcal{I})((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}-\boldsymbol{\mathcal{X}}_\star)\right)\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}\right\rangle}_{=:\mathfrak{R}_U^{\mathsf{p},1}}\\ &+2\eta^2\underbrace{\left\langle\boldsymbol{U}_\star(\breve{\boldsymbol{U}}-\breve{\boldsymbol{U}}_\star)^\top\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1},\mathcal{M}_1\left((\mathcal{A}^*\mathcal{A}-\mathcal{I})((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}-\boldsymbol{\mathcal{X}}_\star)\right)\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}\right\rangle}_{=:\mathfrak{R}_U^{\mathsf{p},2}}\\ &+\eta^2\underbrace{\left\|\mathcal{M}_1\left((\mathcal{A}^*\mathcal{A}-\mathcal{I})((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}-\boldsymbol{\mathcal{X}}_\star)\right)\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}\right\|_{\mathsf{F}}^2}_{=:\mathfrak{R}_U^{\mathsf{p},3}}.\end{aligned}$$

The main term $\mathfrak{R}_U^{\mathsf{m}}$ has been handled in Section B; see (47) and the bound (45a). In the sequel, we shall bound the three perturbation terms.

Step 1: bounding $\mathfrak{R}_U^{\mathsf{p},1}$. Use the definition of $\breve{\boldsymbol{U}}$, we can translate the inner product in the matrix space to that in the tensor space

$$\begin{aligned}\mathfrak{R}_U^{\mathsf{p},1} &= \left\langle\Big(\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}^2(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1},\boldsymbol{V},\boldsymbol{W}\Big)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}},(\mathcal{A}^*\mathcal{A}-\mathcal{I})((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}-\boldsymbol{\mathcal{X}}_\star)\right\rangle\\ &= \left\langle\Big(\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}^2(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1},\boldsymbol{V},\boldsymbol{W}\Big)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}},(\mathcal{A}^*\mathcal{A}-\mathcal{I})((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\Delta}_{\mathcal{S}})\right\rangle\\ &\quad+\left\langle\Big(\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}^2(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1},\boldsymbol{V},\boldsymbol{W}\Big)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}},(\mathcal{A}^*\mathcal{A}-\mathcal{I})((\boldsymbol{\Delta}_U,\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star)\right\rangle\\ &\quad+\left\langle\Big(\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}^2(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1},\boldsymbol{V},\boldsymbol{W}\Big)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}},(\mathcal{A}^*\mathcal{A}-\mathcal{I})((\boldsymbol{U}_\star,\boldsymbol{\Delta}_V,\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star)\right\rangle\\ &\quad+\left\langle\Big(\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}^2(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1},\boldsymbol{V},\boldsymbol{W}\Big)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}},(\mathcal{A}^*\mathcal{A}-\mathcal{I})((\boldsymbol{U}_\star,\boldsymbol{V}_\star,\boldsymbol{\Delta}_W)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star)\right\rangle,\end{aligned}$$

where the second relation uses the decomposition (41). Apply Lemma 15 to each of the four terms to obtain

$$\begin{aligned}|\mathfrak{R}_U^{\mathsf{p},1}| &\le \delta_{2\boldsymbol{r}}\left\|\Big(\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}^2(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1},\boldsymbol{V},\boldsymbol{W}\Big)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}\right\|_{\mathsf{F}}\\ &\qquad(\|(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\Delta}_{\mathcal{S}})\|_{\mathsf{F}}+\|(\boldsymbol{\Delta}_U,\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star)\|_{\mathsf{F}}+\|(\boldsymbol{U}_\star,\boldsymbol{\Delta}_V,\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star)\|_{\mathsf{F}}+\|(\boldsymbol{U}_\star,\boldsymbol{V}_\star,\boldsymbol{\Delta}_W)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star)\|_{\mathsf{F}}).\end{aligned}$$

For the prefactor, we have

$$\begin{aligned}\left\|\Big(\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}^2(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1},\boldsymbol{V},\boldsymbol{W}\Big)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}\right\|_{\mathsf{F}} &= \left\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}^2(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\breve{\boldsymbol{U}}^\top\right\|_{\mathsf{F}}\\ &\le\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}\left\|\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}\right\|\\ &\le\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}(1-\epsilon)^{-3},\end{aligned}$$

where the last step arises from Lemma 10. In addition, the same argument as in (37a) yields

$$\|(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\Delta}_{\mathcal{S}})\|_{\mathsf{F}}+\|(\boldsymbol{\Delta}_U,\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star)\|_{\mathsf{F}}+\|(\boldsymbol{U}_\star,\boldsymbol{\Delta}_V,\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star)\|_{\mathsf{F}}+\|(\boldsymbol{U}_\star,\boldsymbol{V}_\star,\boldsymbol{\Delta}_W)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star)\|_{\mathsf{F}}$$

$$\le (1+\frac{3}{2}\epsilon+\epsilon^2+\frac{1}{4}\epsilon^3)\left(\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\right).$$

Take the previous two bounds collectively to arrive at

$$\begin{aligned}|\mathfrak{R}_{U,p1}| &\le \delta_{2\boldsymbol{r}}\frac{1+\frac{3}{2}\epsilon+\epsilon^2+\frac{1}{4}\epsilon^3}{(1-\epsilon)^3}\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}\left(\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\right)\\ &\lesssim \delta_{2\boldsymbol{r}}\operatorname{dist}^2(\boldsymbol{F}_t,\boldsymbol{F}_\star),\end{aligned}$$

with the proviso that ϵ is small enough.

Step 2: bounding $\mathfrak{R}_U^{\mathrm{p},2}$. Rewrite the inner product in the tensor space to see

$$\mathfrak{R}_U^{\mathrm{p},2}=\left\langle\left(\boldsymbol{U}_\star(\breve{\boldsymbol{U}}-\breve{\boldsymbol{U}}_\star)^\top\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}^2(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1},\boldsymbol{V},\boldsymbol{W}\right)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}},(\mathcal{A}^*\mathcal{A}-\mathcal{I})((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\cdot\boldsymbol{\mathcal{S}}-\boldsymbol{\mathcal{X}}_\star)\right\rangle.$$

Similar to the control of $\mathfrak{R}_U^{\mathrm{p},1}$, we have

$$\begin{aligned}|\mathfrak{R}_U^{\mathrm{p},2}| \le \delta_{2\boldsymbol{r}}&\left\|\boldsymbol{U}_\star(\breve{\boldsymbol{U}}-\breve{\boldsymbol{U}}_\star)^\top\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}^2(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\breve{\boldsymbol{U}}^\top\right\|_{\mathsf{F}}\\ &(1+\frac{3}{2}\epsilon+\epsilon^2+\frac{1}{4}\epsilon^3)\left(\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\right).\end{aligned}$$

For the prefactor, we can use (36f) and (37c) to obtain

$$\begin{aligned}\left\|\boldsymbol{U}_\star(\breve{\boldsymbol{U}}-\breve{\boldsymbol{U}}_\star)^\top\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}^2(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\breve{\boldsymbol{U}}^\top\right\|_{\mathsf{F}} &\le \|\breve{\boldsymbol{U}}-\breve{\boldsymbol{U}}_\star\|_{\mathsf{F}}\left\|\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}\right\|^2\\ &\le \frac{1+\epsilon+\frac{1}{3}\epsilon^2}{(1-\epsilon)^6}\left(\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\right),\end{aligned}$$

which further implies

$$\begin{aligned}|\mathfrak{R}_U^{\mathrm{p},2}| \le \delta_{2\boldsymbol{r}}&\frac{(1+\frac{3}{2}\epsilon+\epsilon^2+\frac{1}{4}\epsilon^3)(1+\epsilon+\frac{1}{3}\epsilon^2)}{(1-\epsilon)^6}\left(\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\right)\\ &\left(\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\right)\\ \lesssim \delta_{2\boldsymbol{r}}&\operatorname{dist}^2(\boldsymbol{F}_t,\boldsymbol{F}_\star),\end{aligned}$$

as long as ϵ is sufficiently small.

Step 3: bounding $\mathfrak{R}_U^{\mathrm{p},3}$. The last perturbation term needs special care. We first use the variational representation of the Frobenius norm to write

$$\sqrt{\mathfrak{R}_U^{\mathrm{p},3}}=\left\langle\left(\widetilde{\boldsymbol{U}}\boldsymbol{\Sigma}_{\star,1}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1},\boldsymbol{V},\boldsymbol{W}\right)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}},(\mathcal{A}^*\mathcal{A}-\mathcal{I})((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}-\boldsymbol{\mathcal{X}}_\star)\right\rangle$$

for some $\widetilde{\boldsymbol{U}}\in\mathbb{R}^{n_1\times r_1}$ obeying $\|\widetilde{\boldsymbol{U}}\|_{\mathsf{F}}=1$. Repeat the same argument as used in controlling $\mathfrak{R}_U^{\mathrm{p},1}$ to see

$$\begin{aligned}\sqrt{\mathfrak{R}_U^{\mathrm{p},3}} &\le \delta_{2\boldsymbol{r}}\left\|\widetilde{\boldsymbol{U}}\boldsymbol{\Sigma}_{\star,1}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\breve{\boldsymbol{U}}^\top\right\|_{\mathsf{F}}(1+\frac{3}{2}\epsilon+\epsilon^2+\frac{1}{4}\epsilon^3)\left(\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\right)\\ &\le \delta_{2\boldsymbol{r}}\frac{1+\frac{3}{2}\epsilon+\epsilon^2+\frac{1}{4}\epsilon^3}{(1-\epsilon)^3}\left(\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\right),\end{aligned}$$

where the last line uses the bound (36f) in Lemma 10. Then take the square on both sides to conclude

$$\begin{aligned}\mathfrak{R}_U^{\mathrm{p},3} &\le \delta_{2\boldsymbol{r}}^2\frac{(1+\frac{3}{2}\epsilon+\epsilon^2+\frac{1}{4}\epsilon^3)^2}{(1-\epsilon)^6}\left(\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\right)^2\\ &\lesssim \delta_{2\boldsymbol{r}}^2\operatorname{dist}^2(\boldsymbol{F}_t,\boldsymbol{F}_\star)\end{aligned}$$

as long as ϵ is sufficiently small.

D.1.2 Bounding the term pertaining to $\boldsymbol{\mathcal{S}}$

The last term of (63) can be rewritten as

$$\begin{aligned}
&(\boldsymbol{Q}_{t,1}^{-1}, \boldsymbol{Q}_{t,2}^{-1}, \boldsymbol{Q}_{t,3}^{-1}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_{t+1} - \boldsymbol{\mathcal{S}}_{\star} \\
&\quad = \boldsymbol{\mathcal{S}} - \eta \left((\boldsymbol{U}^{\top}\boldsymbol{U})^{-1}\boldsymbol{U}^{\top}, (\boldsymbol{V}^{\top}\boldsymbol{V})^{-1}\boldsymbol{V}^{\top}, (\boldsymbol{W}^{\top}\boldsymbol{W})^{-1}\boldsymbol{W}^{\top}\right) \boldsymbol{\cdot} \mathcal{A}^*\mathcal{A}\left((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_{\star}\right) - \boldsymbol{\mathcal{S}}_{\star} \\
&\quad = (1-\eta)\boldsymbol{\Delta}_{\mathcal{S}} - \eta\left((\boldsymbol{U}^{\top}\boldsymbol{U})^{-1}\boldsymbol{U}^{\top}, (\boldsymbol{V}^{\top}\boldsymbol{V})^{-1}\boldsymbol{V}^{\top}, (\boldsymbol{W}^{\top}\boldsymbol{W})^{-1}\boldsymbol{W}^{\top}\right) \boldsymbol{\cdot} \left((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_{\star} - \boldsymbol{\mathcal{X}}_{\star}\right) \\
&\qquad - \eta\left((\boldsymbol{U}^{\top}\boldsymbol{U})^{-1}\boldsymbol{U}^{\top}, (\boldsymbol{V}^{\top}\boldsymbol{V})^{-1}\boldsymbol{V}^{\top}, (\boldsymbol{W}^{\top}\boldsymbol{W})^{-1}\boldsymbol{W}^{\top}\right) \boldsymbol{\cdot} (\mathcal{A}^*\mathcal{A} - \mathcal{I})\left((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_{\star}\right),
\end{aligned}$$

which further gives

$$\begin{aligned}
&\left\|(\boldsymbol{Q}_{t,1}^{-1}, \boldsymbol{Q}_{t,2}^{-1}, \boldsymbol{Q}_{t,3}^{-1}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_{t+1} - \boldsymbol{\mathcal{S}}_{\star}\right\|_{\mathsf{F}}^2 \\
&= \underbrace{\left\|(1-\eta)\boldsymbol{\Delta}_{\mathcal{S}} - \eta\left((\boldsymbol{U}^{\top}\boldsymbol{U})^{-1}\boldsymbol{U}^{\top}, (\boldsymbol{V}^{\top}\boldsymbol{V})^{-1}\boldsymbol{V}^{\top}, (\boldsymbol{W}^{\top}\boldsymbol{W})^{-1}\boldsymbol{W}^{\top}\right) \boldsymbol{\cdot} \left((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_{\star} - \boldsymbol{\mathcal{X}}_{\star}\right)\right\|_{\mathsf{F}}^2}_{=:\mathfrak{R}_{\mathcal{S}}^{\mathrm{m}}} \\
&\quad - 2\eta(1-\eta)\underbrace{\left\langle \boldsymbol{\Delta}_{\mathcal{S}}, \left((\boldsymbol{U}^{\top}\boldsymbol{U})^{-1}\boldsymbol{U}^{\top}, (\boldsymbol{V}^{\top}\boldsymbol{V})^{-1}\boldsymbol{V}^{\top}, (\boldsymbol{W}^{\top}\boldsymbol{W})^{-1}\boldsymbol{W}^{\top}\right) \boldsymbol{\cdot} (\mathcal{A}^*\mathcal{A} - \mathcal{I})\left((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_{\star}\right)\right\rangle}_{=:\mathfrak{R}_{\mathcal{S}}^{\mathrm{p},1}} \\
&\quad + 2\eta^2 \underbrace{\begin{aligned}\Big\langle & \left((\boldsymbol{U}^{\top}\boldsymbol{U})^{-1}\boldsymbol{U}^{\top}, (\boldsymbol{V}^{\top}\boldsymbol{V})^{-1}\boldsymbol{V}^{\top}, (\boldsymbol{W}^{\top}\boldsymbol{W})^{-1}\boldsymbol{W}^{\top}\right) \boldsymbol{\cdot} \left((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}}_{\star} - \boldsymbol{\mathcal{X}}_{\star}\right), \\
& \left((\boldsymbol{U}^{\top}\boldsymbol{U})^{-1}\boldsymbol{U}^{\top}, (\boldsymbol{V}^{\top}\boldsymbol{V})^{-1}\boldsymbol{V}^{\top}, (\boldsymbol{W}^{\top}\boldsymbol{W})^{-1}\boldsymbol{W}^{\top}\right) \boldsymbol{\cdot} (\mathcal{A}^*\mathcal{A} - \mathcal{I})\left((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_{\star}\right)\Big\rangle\end{aligned}}_{=:\mathfrak{R}_{\mathcal{S}}^{\mathrm{p},2}} \\
&\quad + \eta^2 \underbrace{\left\|\left((\boldsymbol{U}^{\top}\boldsymbol{U})^{-1}\boldsymbol{U}^{\top}, (\boldsymbol{V}^{\top}\boldsymbol{V})^{-1}\boldsymbol{V}^{\top}, (\boldsymbol{W}^{\top}\boldsymbol{W})^{-1}\boldsymbol{W}^{\top}\right) \boldsymbol{\cdot} (\mathcal{A}^*\mathcal{A} - \mathcal{I})\left((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_{\star}\right)\right\|_{\mathsf{F}}^2}_{=:\mathfrak{R}_{\mathcal{S}}^{\mathrm{p},3}}.
\end{aligned}$$

Note that the main term $\mathfrak{R}_{\mathcal{S}}^{\mathrm{m}}$ has already been characterized in Section B (see (48) and the bound (45d)), and therefore we concentrate on the remaining perturbation terms.

Step 1: bounding $\mathfrak{R}_{\mathcal{S}}^{\mathrm{p},1}$. Use the property (6d) to write $\mathfrak{R}_{\mathcal{S}}^{\mathrm{p},1}$ as

$$\mathfrak{R}_{\mathcal{S}}^{\mathrm{p},1} = \left\langle \left(\boldsymbol{U}(\boldsymbol{U}^{\top}\boldsymbol{U})^{-1}, \boldsymbol{V}(\boldsymbol{V}^{\top}\boldsymbol{V})^{-1}, \boldsymbol{W}(\boldsymbol{W}^{\top}\boldsymbol{W})^{-1}\right) \boldsymbol{\cdot} \boldsymbol{\Delta}_{\mathcal{S}}, (\mathcal{A}^*\mathcal{A} - \mathcal{I})\left((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \boldsymbol{\cdot} \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_{\star}\right)\right\rangle.$$

We can use the decomposition (41) and Lemma 15 to derive

$$\begin{aligned}
|\mathfrak{R}_{\mathcal{S}}^{\mathrm{p},1}| \le \delta_{2\boldsymbol{r}} &\left\|\left(\boldsymbol{U}(\boldsymbol{U}^{\top}\boldsymbol{U})^{-1}, \boldsymbol{V}(\boldsymbol{V}^{\top}\boldsymbol{V})^{-1}, \boldsymbol{W}(\boldsymbol{W}^{\top}\boldsymbol{W})^{-1}\right) \boldsymbol{\cdot} \boldsymbol{\Delta}_{\mathcal{S}}\right\|_{\mathsf{F}} \\
&\qquad (1 + \frac{3}{2}\epsilon + \epsilon^2 + \frac{1}{4}\epsilon^3)\left(\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\right).
\end{aligned}$$

In addition, Lemma 10 tells us that

$$\begin{aligned}
&\left\|\left(\boldsymbol{U}(\boldsymbol{U}^{\top}\boldsymbol{U})^{-1}, \boldsymbol{V}(\boldsymbol{V}^{\top}\boldsymbol{V})^{-1}, \boldsymbol{W}(\boldsymbol{W}^{\top}\boldsymbol{W})^{-1}\right) \boldsymbol{\cdot} \boldsymbol{\Delta}_{\mathcal{S}}\right\|_{\mathsf{F}} \\
&\quad \le \left\|\boldsymbol{U}(\boldsymbol{U}^{\top}\boldsymbol{U})^{-1}\right\| \left\|\boldsymbol{V}(\boldsymbol{V}^{\top}\boldsymbol{V})^{-1}\right\| \left\|\boldsymbol{W}(\boldsymbol{W}^{\top}\boldsymbol{W})^{-1}\right\| \|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}} \le (1-\epsilon)^{-3}\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}.
\end{aligned}$$

Combine the above two bounds to reach

$$\begin{aligned}
|\mathfrak{R}_{\mathcal{S}}^{\mathrm{p},1}| &\le \delta_{2\boldsymbol{r}} \frac{1 + \frac{3}{2}\epsilon + \epsilon^2 + \frac{1}{4}\epsilon^3}{(1-\epsilon)^3} \|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\left(\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\right) \\
&\lesssim \delta_{2\boldsymbol{r}} \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_{\star})
\end{aligned}$$

as long as ϵ is a sufficiently small constant.

Step 2: bounding $\mathfrak{R}_{\mathcal{S}}^{\mathrm{p},2}$. Similarly, we can bound $\mathfrak{R}_{\mathcal{S}}^{\mathrm{p},2}$ by

$$\begin{aligned}
|\mathfrak{R}_{\mathcal{S}}^{\mathrm{p},2}| &\le \delta_{2\boldsymbol{r}} \left\| \left(\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-2}\boldsymbol{U}^\top, \boldsymbol{V}(\boldsymbol{V}^\top\boldsymbol{V})^{-2}\boldsymbol{V}^\top, \boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-2}\boldsymbol{W}^\top\right)\boldsymbol{\cdot}((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_\star - \boldsymbol{\mathcal{X}}_\star)\right\|_{\mathsf{F}} \\
&\qquad (1+\frac{3}{2}\epsilon+\epsilon^2+\frac{1}{4}\epsilon^3)\left(\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\right) \\
&\le \delta_{2\boldsymbol{r}} \frac{(1+\epsilon+\frac{1}{3}\epsilon^2)(1+\frac{3}{2}\epsilon+\epsilon^2+\frac{1}{4}\epsilon^3)}{(1-\epsilon)^6}\left(\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}\right) \\
&\qquad \left(\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\right) \\
&\lesssim \delta_{2\boldsymbol{r}}\,\mathrm{dist}^2(\boldsymbol{F}_t,\boldsymbol{F}_\star).
\end{aligned}$$

Step 3: bound of $\mathfrak{R}_{\mathcal{S}}^{\mathrm{p},3}$. Apply the variational representation of the Frobenius norm to write

$$\sqrt{\mathfrak{R}_{\mathcal{S}}^{\mathrm{p},3}} = \left\langle \left(\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}, \boldsymbol{V}(\boldsymbol{V}^\top\boldsymbol{V})^{-1}, \boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\right)\boldsymbol{\cdot}\widetilde{\boldsymbol{\mathcal{S}}}, (\mathcal{A}^*\mathcal{A}-\mathcal{I})((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\boldsymbol{\cdot}\boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_\star)\right\rangle$$

for some $\widetilde{\boldsymbol{\mathcal{S}}} \in \mathbb{R}^{r_1\times r_2\times r_3}$ obeying $\|\widetilde{\boldsymbol{\mathcal{S}}}\|_{\mathsf{F}} = 1$. Repeat the same argument as in bounding $\mathfrak{R}_{\boldsymbol{U}}^{\mathrm{p},3}$ to see

$$\begin{aligned}
\sqrt{\mathfrak{R}_{\mathcal{S}}^{\mathrm{p},3}} &\le \delta_{2\boldsymbol{r}} \left\| \left(\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}, \boldsymbol{V}(\boldsymbol{V}^\top\boldsymbol{V})^{-1}, \boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\right)\boldsymbol{\cdot}\widetilde{\boldsymbol{\mathcal{S}}}\right\|_{\mathsf{F}} \\
&\qquad (1+\frac{3}{2}\epsilon+\epsilon^2+\frac{1}{4}\epsilon^3)\left(\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\right) \\
&\le \delta_{2\boldsymbol{r}} \frac{1+\frac{3}{2}\epsilon+\epsilon^2+\frac{1}{4}\epsilon^3}{(1-\epsilon)^3}\left(\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\right).
\end{aligned}$$

Then take the square on both sides to conclude

$$\begin{aligned}
\mathfrak{R}_{\mathcal{S}}^{\mathrm{p},3} &\le \delta_{2\boldsymbol{r}}^2 \frac{(1+\frac{3}{2}\epsilon+\epsilon^2+\frac{1}{4}\epsilon^3)^2}{(1-\epsilon)^6}\left(\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\right)^2 \\
&\lesssim \delta_{2\boldsymbol{r}}^2\,\mathrm{dist}^2(\boldsymbol{F}_t,\boldsymbol{F}_\star).
\end{aligned}$$

D.2 Proof of spectral initialization (Lemma 5)

In view of Lemma 8, we can relate $\mathrm{dist}(\boldsymbol{F}_0,\boldsymbol{F}_\star)$ to $\|(\boldsymbol{U}_0,\boldsymbol{V}_0,\boldsymbol{W}_0)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_0 - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}$ as

$$\mathrm{dist}(\boldsymbol{F}_0,\boldsymbol{F}_\star) \le (\sqrt{2}+1)^{3/2}\,\|(\boldsymbol{U}_0,\boldsymbol{V}_0,\boldsymbol{W}_0)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_0 - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}.$$

To proceed, we need to control $\|(\boldsymbol{U}_0,\boldsymbol{V}_0,\boldsymbol{W}_0)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_0 - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}$, where $(\boldsymbol{U}_0,\boldsymbol{V}_0,\boldsymbol{W}_0)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_0$ is the output of HOSVD, which has been considered in [LZ21,HWZ20,ZLRY20]. Invoking the result in [HWZ20, Appendix D.2 Step 5], we obtain

$$\|(\boldsymbol{U}_0,\boldsymbol{V}_0,\boldsymbol{W}_0)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_0 - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \lesssim \sqrt{\frac{nr+r^3}{m}}\|\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}$$

as long as $m \gtrsim (\sqrt{n_1n_2n_3}+nr\kappa)\|\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}^2/\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star)$. Further notice that

$$\|\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}^2 \le r\kappa^2\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star).$$

Therefore, under the condition $m \gtrsim \epsilon_0^{-2}(\sqrt{n_1n_2n_3}r\kappa^2 + nr^2\kappa^3)$, with overwhelming probability, we conclude that

$$\mathrm{dist}(\boldsymbol{F}_0,\boldsymbol{F}_\star) \le (\sqrt{2}+1)^{3/2}\,\|(\boldsymbol{U}_0,\boldsymbol{V}_0,\boldsymbol{W}_0)\boldsymbol{\cdot}\boldsymbol{\mathcal{S}}_0 - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \le \epsilon_0\sigma_{\min}(\boldsymbol{\mathcal{X}}).$$