Scaled Gradient Methods for Ill-conditioned Low-rank Matrix and Tensor Estimation

Submitted in partial fulfillment of the requirements for

the degree of

Doctor of Philosophy

in

Electrical and Computer Engineering

Tian Tong

B.S., Electronic Engineering, Tsinghua University

Carnegie Mellon University

Pittsburgh, PA

February 2022

Acknowledgments

It is the earnest love to mathematics and science that solidified my dedication to pursue the truth. The journey is wonderful and memorable, brimming with plentiful guidance and supports.

First and foremost, I would like to express my deepest thanks to my advisor, Yuejie Chi, for offering numerous brilliant ideas and lighting my road. She is enthusiastic about every work and meticulous about every detail, as everything I aspire to be. I would like to thank my long-term collaborator, Cong Ma, for enormous helpful suggestions and discussions. I would like to extend my gratitude to my thesis committee (Yuejie Chi, Giulia Fanti, Sivaraman Balakrishnan, Yuxin Chen) for shaping this work. I am grateful for generous supports from Office of Naval Research (ONR), Army Research Office (ARO), Air Force Research Laboratory (AFRL), and National Science Foundation (NSF).

In addition, I would like to acknowledge Yuejie's group (Yuanxin Li, Maxime Ferreira Da Costa, Harlin Lee, Vincent Monardo, Boyue Li, Laixi Shi, Shicong Cen, Diogo Cardoso, Pedro Valdeira, Jiin Woo, Harry Dong, Lingjing Kong) for academic discussions and office camaraderie. I would like to thank Yuwei Qin, Ruizhou Ding, Jiaqi Liu for extensive helps during my qualifying exam and job search—the days we discussed mathematics and coding questions together are unforgettable. I would like to highlight Laixi Shi, Shuhua Yu, Ran Xin, Yingsi Qin, Yuhang Yao, Meiyi Li, Xiang Wang for accompanying me in Porter Hall and organizing outdoor activities such as tennis, skiing, hiking, etc.—my life has become so beautiful since you came. I appreciate Yuxing Zhang for being a great roommate and collaborating on course projects. I enjoy the time with my tennis coach Mark Paull and my partner Yuting Bu. Finally, I would like to thank my parents for their unconditional love.

Tian Tong

Abstract

Many problems encountered in machine learning and signal processing can be formulated as estimating a low-rank object from incomplete, and possibly corrupted, linear measurements; prominent examples include matrix completion and tensor completion. Through the lens of matrix and tensor factorization, one of the most popular approaches is to employ simple iterative algorithms such as gradient descent to recover the low-rank factors directly, which allow for small memory and computation footprints. However, the convergence rate of gradient descent depends linearly, and sometimes even quadratically, on the condition number of the low-rank object, and therefore, slows down painstakingly when the problem is ill-conditioned. This thesis introduces a new algorithm: scaled gradient descent (`ScaledGD`), which provably converges linearly at a constant rate independent of the condition number of the low-rank object, while maintaining the low per-iteration cost of gradient descent. In addition, a nonsmooth variant of `ScaledGD` provides further robustness to corruptions by optimizing the least absolute deviation loss. In total, `ScaledGD` 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 symmetry in low-rank factorization.

Contents

Acknowledgments iii

Abstract iv

List of Figures x

List of Tables xiii

List of Algorithms xv

Chapter 1 Introduction 1

1.1 Contributions and organization . . . 4

1.2 Notation . . . 6

1.3 Reproducible research . . . 7

Chapter 2 Low-rank Matrix Estimation 8

2.1 Introduction . . . 8

2.1.1 Related work . . . 11

2.1.2 Chapter organization . . . 12

2.2 Scaled Gradient Descent for Low-Rank Matrix Estimation . . . 12

2.2.1 Assumptions and error metric . . . 13

2.2.2 Matrix sensing . . . 15

2.2.3 Robust PCA . . . 17

2.2.4 Matrix completion . . . 20

2.2.5 Optimizing general loss functions . . . 22

2.3 Proof Sketch . . . 24

2.3.1 A warm-up analysis: matrix factorization . . . 24

2.3.2 Proof outline for matrix sensing . . . 26

2.3.3 Proof outline for robust PCA . . . 26
2.3.4 Proof outline for matrix completion . . . 28
2.4 Numerical Experiments . . . 30
2.4.1 Comparison with vanilla GD . . . 30
2.4.2 Run time comparisons . . . 34
2.5 Conclusions . . . 37

Chapter 3 Robust Low-rank Matrix Estimation **39**
3.1 Introduction . . . 39
3.1.1 Main contributions . . . 40
3.1.2 Related work . . . 41
3.1.3 Chapter organization . . . 43
3.2 Problem Formulation and Algorithms . . . 43
3.2.1 Problem formulation . . . 43
3.2.2 Scaled subgradient method . . . 44
3.3 Theoretical Guarantees . . . 47
3.3.1 Geometric assumptions . . . 48
3.3.2 Main results . . . 49
3.3.3 A case study: robust low-rank matrix recovery . . . 51
3.4 Numerical Experiments . . . 54
3.5 Discussions . . . 58

Chapter 4 Low-rank Tensor Estimation **61**
4.1 Introduction . . . 61
4.1.1 A gradient descent approach? . . . 63
4.1.2 A new algorithm: scaled gradient descent . . . 64
4.1.3 Additional related works . . . 68
4.1.4 A primer on tensor algebra and notation . . . 71
4.2 Main Results . . . 73

4.2.1 Models and assumptions . . . 73
4.2.2 ScaledGD for tensor completion . . . 75
4.2.3 ScaledGD for tensor regression . . . 78
4.3 Analysis . . . 80
4.3.1 A warm-up case: ScaledGD for tensor factorization . . . 81
4.3.2 Proof outline for tensor completion (Theorem 8) . . . 84
4.3.3 Proof outline for tensor regression (Theorem 9) . . . 86
4.4 Numerical Experiments . . . 88
4.5 Discussions . . . 92

Chapter 5 Robust Low-rank Tensor Estimation 94

5.1 Introduction . . . 94
5.1.1 Related works . . . 95
5.1.2 Chapter organization . . . 96
5.2 Formulation and Proposed Algorithms . . . 96
5.2.1 Truncated spectral initialization . . . 99
5.3 Theoretical Guarantees . . . 100
5.3.1 A general theory of local linear convergence . . . 100
5.3.2 Case study: Gaussian design . . . 102
5.4 Numerical Experiments . . . 102
5.5 Conclusions . . . 104

Appendix A Proofs for Low-rank Matrix Estimation 105

A.1 Technical Lemmas . . . 105
A.1.1 New distance metric . . . 105
A.1.2 Matrix perturbation bounds . . . 110
A.1.3 Partial Frobenius norm . . . 113
A.2 Proof for Low-Rank Matrix Factorization . . . 116
A.2.1 Proof of Proposition 2 . . . 116

A.2.2 Proof of Theorem 5 . . . 116
A.3 Proof for Low-Rank Matrix Sensing . . . 121
A.3.1 Proof of Lemma 1 . . . 122
A.3.2 Proof of Lemma 2 . . . 126
A.4 Proof for Robust PCA . . . 127
A.4.1 Proof of Lemma 3 . . . 132
A.4.2 Proof of Lemma 4 . . . 139
A.4.3 Proof of Lemma 5 . . . 141
A.5 Proof for Matrix Completion . . . 144
A.5.1 New projection operator . . . 144
A.5.2 Proof of Lemma 7 . . . 147
A.5.3 Proof of Lemma 8 . . . 155
A.6 Proof for General Loss Functions . . . 156
A.6.1 Proof of Theorem 4 . . . 157

Appendix B Proofs for Robust Low-rank Matrix Estimation 163

B.1 Proof of Theorem 6 . . . 164
B.1.1 Convergence with Polyak's stepsizes . . . 167
B.1.2 Convergence with geometrically decaying stepsizes . . . 168
B.2 Proof of Theorem 7 . . . 170
B.3 Proof of Proposition 3 . . . 172
B.4 Proof of Proposition 4 . . . 172

Appendix C Proofs for Low-rank Tensor Estimation 174

C.1 Preliminaries . . . 174
C.1.1 Understanding the scaled distance . . . 174
C.1.2 Several perturbation bounds . . . 178
C.2 Proof for Tensor Factorization (Theorem 10) . . . 184
C.2.1 Proof of Claim 8 . . . 188

C.2.2 Proof of Claim 9 . . . 191
C.2.3 Proof of Claim 10 . . . 193
C.2.4 Proof of Claim 11 . . . 197
C.3 Proof for Tensor Completion . . . 198
C.3.1 Proof of Lemma 10 . . . 199
C.3.2 Concentration inequalities . . . 202
C.3.3 Proof of spectral initialization (Lemma 9) . . . 208
C.3.4 Proof of local convergence (Lemma 11) . . . 211
C.4 Proof for Tensor Regression . . . 222
C.4.1 Proof of local convergence (Lemma 12) . . . 222
C.4.2 Proof of spectral initialization (Lemma 13) . . . 229

Appendix D Proofs for Robust Low-rank Tensor Estimation **235**
D.1 Proof of Theorem 11 . . . 236
D.2 Proof of Proposition 5 . . . 243

Bibliography **246**

List of Figures

1.1 Performance of `ScaledGD` and vanilla GD for completing a 1000×1000 incoherent matrix of rank 10 with different condition numbers $\kappa = 2, 10, 50$, where each entry is observed independently with probability 0.2. Here, both methods are initialized via the spectral method. It can be seen that `ScaledGD` converges much faster than vanilla GD even for moderately large condition numbers. 5

2.1 The relative errors of `ScaledGD` and vanilla GD with respect to the iteration count under different condition numbers $\kappa = 1, 5, 10, 20$ for (a) matrix sensing, (b) robust PCA, (c) matrix completion, and (d) Hankel matrix completion. 31

2.2 The relative errors of `ScaledGD` and vanilla GD with respect to the iteration count under the condition number $\kappa = 10$ and signal-to-noise ratios SNR $= 40, 60, 80$dB for (a) matrix sensing, (b) robust PCA, (c) matrix completion, and (d) Hankel matrix completion. 32

2.3 The relative errors of `ScaledGD` and vanilla GD after 80 iterations with respect to different step sizes η from 0.1 to 1.2, for matrix completion with $n = 1000, r = 10, p = 0.2$. 34

2.4 The relative errors of `ScaledGD`, vanilla GD and `AltMin` with respect to the iteration count and run time (in seconds) under different condition numbers $\kappa = 1, 5, 20$ for matrix sensing with $n = 200$, and $m = 5nr$. (a, b): $r = 10$; (c, d): $r = 20$. 35

2.5 The relative errors of `ScaledGD`, vanilla GD and `AltMin` with respect to the iteration count and run time (in seconds) under different condition numbers $\kappa = 1, 5, 20$ for matrix completion with $n = 1000$, and $p = 0.2$. (a, b): $r = 10$; (c, d): $r = 50$. 36

3.1 Performance comparisons of `ScaledSM` and `VanillaSM` for matrix sensing without or with outliers under different condition numbers $\kappa = 1, 5, 10, 20$, where $n = 100$, $r = 10$, and $m = 8nr$. 55

3.2 Performance comparisons of `ScaledSM` and `VanillaSM` for quadratic sampling without or with outliers under different condition numbers $\kappa = 1, 5, 10, 20$, where $n = 100$, $r = 5$, and $m = 8nr$. 56

3.3 Performance comparisons of `ScaledSM` and `VanillaSM` for matrix sensing under different noise and outlier models, where $n = 100$, $r = 10$, $m = 8nr$, and $\kappa = 10$. 57

3.4 Performance comparisons of `ScaledSM` and `VanillaSM` for quadratic sampling under different noise and outlier models, where $n = 100$, $r = 5$, $m = 8nr$, and $\kappa = 10$. . . . 57

3.5 Performance comparisons of `ScaledSM` for matrix sensing using geometrically decaying stepsizes with parameters (λ, q) and Polyak's stepsizes, where we fix $n = 100$, $r = 10$, $m = 8nr$, $\kappa = 10$, and $p_s = 0.2$: (a) the final relative error for various combinations of (λ, q), (b) the relative error versus iteration count for fixed $q = 0.91$ and varying λ, (c) the relative error versus iteration count for fixed $\lambda = 5$ and varying q, and (d) shows properly tuned geometrically decaying stepsizes with $\lambda = 1.85$ and $q = 0.91$ essentially match Polyak's stepsizes. 59

3.6 Performance comparisons of `ScaledSM` for quadratic sampling using geometrically decaying stepsizes with parameters (λ, q) and Polyak's stepsizes, where we fix $n = 100$, $r = 5$, $m = 8nr$, $\kappa = 10$, and $p_s = 0.2$: (a) the final relative error for various combinations of (λ, q), (b) the relative error versus iteration count for fixed $q = 0.92$ and varying λ, (c) the relative error versus iteration count for fixed $\lambda = 2$ and varying q, and (d) shows properly tuned geometrically decaying stepsizes with $\lambda = 1.36$ and $q = 0.88$ essentially match Polyak's stepsizes. 60

4.1 The iteration complexities of `ScaledGD` (this thesis) 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 $p = 0.1$. 68

4.2 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. 89

4.3 The relative errors of `ScaledGD` and GD after 80 iterations with respect to different step sizes η from 0.1 to 0.9 for tensor completion with $n = 100$, $r = 5$, $p = 0.1$. 90

4.4 The relative errors of `ScaledGD` and 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 $p = 0.1$. 90

4.5 The relative errors of random-initialized `ScaledGD` and GD with respect to the iteration count under different condition numbers $\kappa = 1, 2, 5, 10$ for tensor completion with $n = 100$, $r = 5$, $p = 0.1$. 91

4.6 The relative errors of `ScaledGD` and GD with respect to the iteration count under signal-to-noise ratios SNR $= 40, 60, 80$dB for tensor completion with $n = 100$, $r = 5$, and $p = 0.1$. 92

5.1 Performance comparisons of `ScaledSM` and the vanilla subgradient method (SM). (a) The reconstruction errors $\|\boldsymbol{\mathcal{X}}_t - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}/\|\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}$ w.r.t. the iteration count under different condition numbers $\kappa = 1, 2, 5, 10$ with $p_s = 0.2$. (b) The iteration complexities w.r.t. the condition number for achieving $\|\boldsymbol{\mathcal{X}}_t - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \le 10^{-3}\|\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}$ with $p_s = 0.2$. (c) The reconstruction errors w.r.t. the iteration count under different amounts of outliers $p_s = 0.1, 0.2, 0.3, 0.4$ with $\kappa = 5$. (d) The reconstruction errors w.r.t. the iteration count under different signal-to-noise ratios SNR $= 40, 60, 80$dB with $p_s = 0.2$. 103

List of Tables

2.1 Comparisons of `ScaledGD` with gradient descent (`GD`) when tailored to various problems (with spectral initialization) [TBS+16, YPCC16, ZL16], where they have comparable per-iteration costs. Here, we say that the output $\boldsymbol{X}$ of an algorithm reaches ϵ-accuracy, if it satisfies $\|\boldsymbol{X}-\boldsymbol{X}_\star\|_{\mathsf{F}} \le \epsilon\sigma_r(\boldsymbol{X}_\star)$. Here, $n \coloneqq n_1 \vee n_2 = \max\{n_1, n_2\}$, κ and μ are the condition number and incoherence parameter of $\boldsymbol{X}_\star$. 10

3.1 Local iteration complexities of the proposed scaled subgradient method (`ScaledSM`) in comparison with prior algorithms for matrix sensing and quadratic sampling. `ScaledSM` outperforms the vanilla subgradient method (`SM`) by a factor of κ in both problems, while outperforms scaled gradient descent (`ScaledGD`), and `GD` with additional robustness guarantees. Here, $n = \max\{n_1, n_2\}$, r is the rank, κ is the condition number of $\boldsymbol{X}_\star$, and $0 \le p_s < 1/2$ is the fraction of outliers. We say that the output $\boldsymbol{X}$ of an algorithm reaches ϵ-accuracy, if it satisfies $\|\boldsymbol{X}-\boldsymbol{X}_\star\|_{\mathsf{F}} \le \epsilon\sigma_r(\boldsymbol{X}_\star)$, where $\sigma_r(\boldsymbol{X}_\star)$ denotes the r-th largest singular value of $\boldsymbol{X}_\star$. 41

4.1 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 4.2.1). 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. 66

4.2 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 $\|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \leq \varepsilon\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$. 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 4.2.1). 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. 67

List of Algorithms

1 `ScaledGD` for low-rank matrix sensing with spectral initialization 16

2 `ScaledGD` for robust PCA with spectral initialization 18

3 `ScaledPGD` for matrix completion with spectral initialization 21

4 `ScaledGD` for low-rank tensor completion 77

5 `ScaledGD` for low-rank tensor regression 79

6 `ScaledSM` for low-rank tensor recovery 100

Chapter 1

Introduction

Most signal processing and machine learning tasks propose to solve a mathematical optimization problem, for which gradient descent and its variants such as stochastic gradient descent and momentum methods are the most popular algorithms. When the optimization problem is convex and smooth, which is often the case in classical models, gradient descent is guaranteed to work efficiently [Bec17]. On the other hand, modern machine learning models, like deep neural networks, often require solving a nonconvex and nonsmooth problem. This leads to a rapid paradigm shift in large-scale inference: heuristic nonconvex algorithms, instead of tractable convex approaches, become increasingly more popular due to their superior efficiency and scalability. In general, nonconvex optimization problems cannot be solved efficiently. However, in practice, many important nonconvex problems enjoy benign geometric landscape [ZQW20], thus gradient descent and its variants can solve them successfully. These competing facts indicate that often there are special structures such that the optimization problems are not as hard as they seem. This thesis studies a set of such problems categorized as follows, with specific questions called out to advance the state-of-the-art.

- Many problems encountered in data science can be formulated as *low-rank matrix estimation* [CLC19]. Examples include phase retrieval [SEC+15], blind deconvolution [ARR14], robust principal component analysis [CSPW11,CLMW11], low-rank matrix completion [CR09,DR16], and so on. A common goal is to develop reliable, scalable, and robust algorithms to estimate a low-rank matrix from highly incomplete, potentially corrupted and noisy observations. Broadly speaking, one aims to recover a rank-r matrix $\boldsymbol{X}_\star \in \mathbb{R}^{n_1 \times n_2}$ from a set of observations $\boldsymbol{y} = \mathcal{A}(\boldsymbol{X}_\star)$, where the operator $\mathcal{A}(\cdot)$ models the measurement process. It is natural to minimize the least-squares

loss function subject to a rank constraint:

$$\underset{\boldsymbol{X}\in\mathbb{R}^{n_1\times n_2}}{\text{minimize}}\ f(\boldsymbol{X}) \coloneqq \tfrac{1}{2}\|\mathcal{A}(\boldsymbol{X})-\boldsymbol{y}\|_2^2 \qquad \text{s.t.} \quad \operatorname{rank}(\boldsymbol{X})\le r, \tag{1.1}$$

which is, however, computationally intractable in general due to the rank constraint. In the last decades, convex relaxation approaches have been developed, where the basic idea is to replace the rank constraint by a convex surrogate, e.g. a nuclear norm [CR09, DR16]. Such convex relaxation approaches exhibit intriguing performance in many aspects, however, the parameter space is often much larger than the target space. As the size of the matrix increases, the costs involved in optimizing over the full matrix space (i.e. $\mathbb{R}^{n_1\times n_2}$) are prohibitive in terms of both memory and computation. To cope with these challenges, one popular approach is to parametrize $\boldsymbol{X}=\boldsymbol{L}\boldsymbol{R}^\top$ by two low-rank factors $\boldsymbol{L}\in\mathbb{R}^{n_1\times r}$ and $\boldsymbol{R}\in\mathbb{R}^{n_2\times r}$ that are more memory-efficient, and then to optimize over the factors instead:

$$\underset{\boldsymbol{L}\in\mathbb{R}^{n_1\times r},\boldsymbol{R}\in\mathbb{R}^{n_2\times r}}{\text{minimize}}\ \mathcal{L}(\boldsymbol{L},\boldsymbol{R}) \coloneqq f(\boldsymbol{L}\boldsymbol{R}^\top). \tag{1.2}$$

Although this leads to a nonconvex optimization problem over the factors, recent breakthroughs have shown that simple algorithms (e.g. gradient descent, alternating minimization), when properly initialized (e.g. via the spectral method), can provably converge to the true low-rank factors under mild statistical assumptions. These benign convergence guarantees hold for a growing number of problems such as low-rank matrix sensing, matrix completion, robust principal component analysis (robust PCA), phase synchronization, and so on.

However, upon closer examination, existing approaches such as gradient descent and alternating minimization are still computationally expensive, especially for ill-conditioned matrices. Take low-rank matrix sensing as an example: although the per-iteration cost is small, the iteration complexity of gradient descent scales linearly with respect to the condition number of the low-rank matrix $\boldsymbol{X}_\star$ [TBS+16]; on the other end, while the iteration complexity of alternating minimization [JNS13] is independent of the condition number, each iteration requires inverting a linear system whose size is proportional to the dimension of the matrix and thus the per-iteration cost is

prohibitive for large-scale problems. These together raise an important open question:

Can we design an algorithm with a comparable per-iteration cost as gradient descent, but converges much faster at a rate that is independent of the condition number as alternating minimization in a provable manner for a wide variety of low-rank matrix estimation tasks?

- In addition, due to the heavy-tailed nature of certain measurement operators, such as those encountered in phase retrieval [CLS15] and quadratic sampling [SWW17], the least-squares formulation mentioned above may suffer from a large smoothness parameter (and hence a large condition number of the loss function) that scales at least linearly with respect to the ambient dimension, where the iteration complexity of gradient descent scales poorly both with the dimension as well as the condition number of the low-rank matrix, leading to a conservative choice of stepsizes and a high iteration complexity when the problem dimension is large. Moreover, the smooth formulation is not robust to corruptions. While there have been encouraging activities [CCD+21, MWCC19, LMCC21, TMC21a] that try to alleviate these issues regarding ill-conditioning, none of the existing first-order approaches are able to simultaneously remove both sources of ill-conditioning and achieve fast convergence. In contrast, nonsmooth formulations yield better conditioning in such problems and exhibit apparent benefits over their smooth counterparts. This leads to the following important question:

  Can we develop first-order methods for nonsmooth formulations that are guaranteed to converge at a fast rate that is almost dimension-free and independent of the condition number, even in the presence of corruptions?

- Moving beyond matrix estimation, a natural higher-order generalization is 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. In many problems across science and engineering, the central task is *low-rank tensor estimation*, where the goal is to estimate a tensor $\boldsymbol{\mathcal{X}}_\star \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_K}$ from its observations $\boldsymbol{y} \in \mathbb{R}^m$ given

by $\boldsymbol{y} \approx \mathcal{A}(\boldsymbol{\mathcal{X}}_\star)$. Here, $\mathcal{A}(\cdot)$ represents a certain linear map modeling the data collection process. Examples include tensor completion [CLPC19, LM20] and tensor regression [HWZ20, ARB20]. There are intrinsic difficulties of estimating a tensor in many aspects, thus one hopes to answer the question:

> *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?*

1.1 Contributions and organization

In this thesis, we answer these questions affirmatively by proposing a nonconvex optimization framework—scaled gradient methods, whose variants designed for various low-rank matrix and tensor estimation tasks are described as follows.

- **Low-rank matrix estimation**. We set forth a competitive algorithmic approach dubbed *Scaled Gradient Descent* (`ScaledGD`) which can be viewed as preconditioned or diagonally-scaled gradient descent, where the preconditioners are adaptive and iteration-varying with a minimal computational overhead. We expect that the `ScaledGD` algorithm can accelerate the convergence for other low-rank matrix estimation problems, as well as facilitate the design and analysis of other quasi-Newton first-order algorithms. As a teaser, Figure 1.1 illustrates the relative error of completing a 1000×1000 incoherent matrix of rank 10 with varying condition numbers from 20% of its entries, using either `ScaledGD` or vanilla GD with spectral initialization. Even for moderately ill-conditioned matrices, the convergence rate of vanilla GD slows down dramatically, while it is evident that `ScaledGD` converges at a rate independent of the condition number and therefore is much more efficient.

  With tailored variants for low-rank matrix sensing, robust principal component analysis and matrix completion, we theoretically show that `ScaledGD` achieves the best of both worlds: it converges linearly at a rate independent of the condition number of the low-rank matrix similar as alternating minimization, while maintaining the low per-iteration cost of gradient descent.

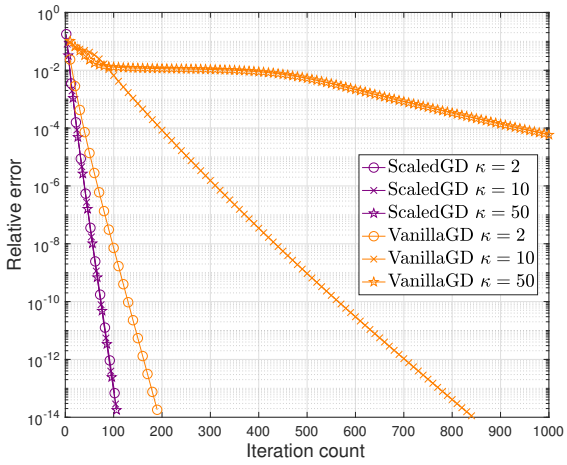


Figure 1.1: Performance of `ScaledGD` and vanilla GD for completing a 1000×1000 incoherent matrix of rank 10 with different condition numbers $\kappa = 2, 10, 50$, where each entry is observed independently with probability 0.2. Here, both methods are initialized via the spectral method. It can be seen that `ScaledGD` converges much faster than vanilla GD even for moderately large condition numbers.

Our analysis is also applicable to general loss functions that are restricted strongly convex and smooth over low-rank matrices. To the best of our knowledge, `ScaledGD` is the first algorithm that provably has such properties over a wide range of low-rank matrix estimation tasks. At the core of our analysis is the introduction of a new distance function that takes account of the preconditioners when measuring the distance between the iterates and the ground truth. Numerical examples are provided to demonstrate the effectiveness of `ScaledGD` in accelerating the convergence rate of ill-conditioned low-rank matrix estimation in a wide number of applications. Details are presented in Chapter 2, based on the paper [TMC21a].

- **Robust low-rank matrix estimation**. We propose *scaled subgradient methods* (`ScaledSM`) to minimize a family of nonsmooth and nonconvex formulations—in particular, the residual sum of absolute errors—which is guaranteed to converge at a fast rate that is almost dimension-free and independent of the condition number, even in the presence of corruptions. We illustrate the effectiveness of our approach when the observation operator satisfies certain mixed-norm restricted isometry properties, and derive state-of-the-art performance guarantees for a variety of problems such as robust low-rank matrix sensing and quadratic sampling. Details are presented

in Chapter 3, based on the paper [TMC21b].

- **Low-rank Tensor estimation**. We generalize `ScaledGD` to low-rank tensor estimation, and show 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. Details are presented in Chapter 4, based on the paper [TMPB+21].

- **Robust low-rank tensor estimation**. We generalize `ScaledSM` to estimate the tensor factors by solving a nonsmooth and nonconvex composite optimization problem that minimizes the least absolute deviation loss. The proposed algorithm—built on subgradient methods—harnesses preconditioners that are designed to be equivariant w.r.t. the low-rank parameterization, and is shown to achieve local linear convergence at a constant rate under the Gaussian design. Numerical experiments are provided to corroborate the superior performance of the proposed algorithm. Details are presented in Chapter 5.

1.2 Notation

Before continuing, we introduce several notation used throughout the thesis. First of all, we use boldfaced symbols (e.g. $\boldsymbol{x}$) to denote vectors, boldface capitalized letters (e.g. $\boldsymbol{X}$) to denote matrices, and boldface calligraphic letters (e.g. $\boldsymbol{\mathcal{X}}$) to denote tensors. For a vector $\boldsymbol{v}$, we use $\|\boldsymbol{v}\|_0$ to denote

its ℓ_0 counting norm, and $\|\boldsymbol{v}\|_2$ to denote the ℓ_2 norm. For any matrix $\boldsymbol{A}$, we use $\sigma_i(\boldsymbol{A})$ to denote its i-th largest singular value, and $\sigma_{\max}(\boldsymbol{A})$ (resp. $\sigma_{\min}(\boldsymbol{A})$) to denote its largest (resp. smallest) nonzero singular value. Let $\boldsymbol{A}_{i,\cdot}$ or $\boldsymbol{A}(i,:)$ (resp. $\boldsymbol{A}_{\cdot,j}$ or $\boldsymbol{A}(:,j)$) to denote its i-th row (resp. j-th column). In addition, $\|\boldsymbol{A}\|$, $\|\boldsymbol{A}\|_{\mathsf{F}}$, $\|\boldsymbol{A}\|_{1,\infty}$, $\|\boldsymbol{A}\|_{2,\infty}$, and $\|\boldsymbol{A}\|_{\infty}$ 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 $\boldsymbol{A}$. Let $\mathcal{P}_{\mathsf{diag}}(\boldsymbol{A})$ denote the projection that keeps only the diagonal entries of $\boldsymbol{A}$, and $\mathcal{P}_{\mathsf{off\text{-}diag}}(\boldsymbol{A}) = \boldsymbol{A} - \mathcal{P}_{\mathsf{diag}}(\boldsymbol{A})$, for a square matrix $\boldsymbol{A}$. We denote

$$\mathcal{P}_r(\boldsymbol{A}) = \min_{\widetilde{\boldsymbol{A}}:\mathrm{rank}(\widetilde{\boldsymbol{A}})\le r} \|\boldsymbol{A} - \widetilde{\boldsymbol{A}}\|_{\mathsf{F}}^2 \tag{1.3}$$

as the rank-r approximation of $\boldsymbol{A}$, which is given by the top-r SVD of $\boldsymbol{A}$ by the Eckart-Young-Mirsky theorem. We also use $\mathrm{vec}(\boldsymbol{A})$ to denote the vectorization of a matrix $\boldsymbol{A}$. For matrices $\boldsymbol{A}, \boldsymbol{B}$ of the same size, we use $\langle \boldsymbol{A}, \boldsymbol{B}\rangle = \sum_{i,j} \boldsymbol{A}_{i,j}\boldsymbol{B}_{i,j} = \mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{B})$ to denote their inner product. The set of invertible matrices in $\mathbb{R}^{r\times r}$ is denoted by $\mathrm{GL}(r)$.

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 $C > 0$, $f(n) \gtrsim g(n)$ means $|f(n)|/|g(n)| \ge C$ for some constant $C > 0$, and $f(n) \asymp g(n)$ means $C_1 \le |f(n)|/|g(n)| \le C_2$ for some constants $C_1, C_2 > 0$. Additionally, $f(n) \ll g(n)$ indicates $|f(n)|/|g(n)| \le c$ for some sufficient small constant $c > 0$, and $f(n) \gg g(n)$ indicates $|f(n)|/|g(n)| \ge C$ for some sufficient large constant $C > 0$. We use $C, C_1, C_2, c, c_1, c_2 \ldots$ 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 $1 - c_1 n^{-c_2}$.

1.3 Reproducible research

The simulations are performed in Matlab with a 3.6 GHz Intel Xeon Gold 6244 CPU. The codes are available at

https://github.com/Titan-Tong/ScaledGD.

Chapter 2

Low-rank Matrix Estimation

2.1 Introduction

Low-rank matrix estimation plays a critical role in fields such as machine learning, signal processing, imaging science, and many others. The goal is to recover a rank-r matrix $\boldsymbol{X}_\star \in \mathbb{R}^{n_1 \times n_2}$ from a set of observations $\boldsymbol{y} = \mathcal{A}(\boldsymbol{X}_\star)$, where the operator $\mathcal{A}(\cdot)$ models the measurement process. In consideration of memory and computation efficiency, we parametrize $\boldsymbol{X} = \boldsymbol{L}\boldsymbol{R}^\top$ by two low-rank factors $\boldsymbol{L} \in \mathbb{R}^{n_1 \times r}$ and $\boldsymbol{R} \in \mathbb{R}^{n_2 \times r}$ that are more memory-efficient, and then optimize over the factors:

$$\underset{\boldsymbol{L}\in\mathbb{R}^{n_1\times r},\boldsymbol{R}\in\mathbb{R}^{n_2\times r}}{\text{minimize}} \quad \mathcal{L}(\boldsymbol{L},\boldsymbol{R}) \coloneqq f(\boldsymbol{L}\boldsymbol{R}^\top). \tag{2.1}$$

In this chapter, we introduce scaled gradient descent (`ScaledGD`) algorithm for low-rank matrix estimation. Given an initialization $(\boldsymbol{L}_0, \boldsymbol{R}_0)$, `ScaledGD` proceeds as follows

$$\begin{aligned} \boldsymbol{L}_{t+1} &= \boldsymbol{L}_t - \eta \nabla_{\boldsymbol{L}} \mathcal{L}(\boldsymbol{L}_t, \boldsymbol{R}_t)(\boldsymbol{R}_t^\top \boldsymbol{R}_t)^{-1}, \\ \boldsymbol{R}_{t+1} &= \boldsymbol{R}_t - \eta \nabla_{\boldsymbol{R}} \mathcal{L}(\boldsymbol{L}_t, \boldsymbol{R}_t)(\boldsymbol{L}_t^\top \boldsymbol{L}_t)^{-1}, \end{aligned} \tag{2.2}$$

where $\eta > 0$ is the step size and $\nabla_{\boldsymbol{L}} \mathcal{L}(\boldsymbol{L}_t, \boldsymbol{R}_t)$ (resp. $\nabla_{\boldsymbol{R}} \mathcal{L}(\boldsymbol{L}_t, \boldsymbol{R}_t)$) is the gradient of the loss function $\mathcal{L}$ with respect to the factor $\boldsymbol{L}_t$ (resp. $\boldsymbol{R}_t$) at the t-th iteration. Comparing to vanilla gradient descent, the search directions of the low-rank factors $\boldsymbol{L}_t, \boldsymbol{R}_t$ in (2.2) are *scaled* by $(\boldsymbol{R}_t^\top \boldsymbol{R}_t)^{-1}$ and $(\boldsymbol{L}_t^\top \boldsymbol{L}_t)^{-1}$ respectively. Intuitively, the scaling serves as a preconditioner as in quasi-Newton type algorithms, with the hope of improving the quality of the search direction to allow larger step sizes. Since the computation of the Hessian is extremely expensive, it is necessary to design preconditioners

that are both theoretically sound and practically cheap to compute. Such requirements are met by `ScaledGD`, where the preconditioners are computed by inverting two $r \times r$ matrices, whose size is much smaller than the dimension of matrix factors. Therefore, each iteration of `ScaledGD` adds minimal overhead to the gradient computation and has the order-wise same per-iteration cost as gradient descent. Moreover, the preconditioners are adaptive and iteration-varying. Another key property of `ScaledGD` is that it ensures the iterates are covariant with respect to the parameterization of low-rank factors up to invertible transforms.

While `ScaledGD` and its alternating variants have been proposed in [MAS12,MS16,TW16] for a subset of the problems we studied, none of these prior art provides any theoretical validations to the empirical success. In this work, we confirm *theoretically* that `ScaledGD` achieves linear convergence at a rate *independent of* the condition number of the matrix when initialized properly, e.g. using the standard spectral method, for several canonical problems: low-rank matrix sensing, robust PCA, and matrix completion. Table 2.1 summarizes the performance guarantees of `ScaledGD` in terms of both statistical and computational complexities with comparisons to prior algorithms using the vanilla gradient method.

- *Low-rank matrix sensing.* As long as the measurement operator satisfies the standard restricted isometry property (RIP) with an RIP constant $\delta_{2r} \lesssim 1/(\sqrt{r}\kappa)$, where κ is the condition number of $\boldsymbol{X}_\star$, `ScaledGD` reaches ϵ-accuracy in $O(\log(1/\epsilon))$ iterations when initialized by the spectral method. This strictly improves the iteration complexity $O(\kappa \log(1/\epsilon))$ of gradient descent in [TBS+16] under the same sample complexity requirement.

- *Robust PCA.* Under the deterministic corruption model [CSPW11], as long as the fraction α of corruptions per row / column satisfies $\alpha \lesssim 1/(\mu r^{3/2}\kappa)$, where μ is the incoherence parameter of $\boldsymbol{X}_\star$, `ScaledGD` in conjunction with hard thresholding reaches ϵ-accuracy in $O(\log(1/\epsilon))$ iterations when initialized by the spectral method. This strictly improves the iteration complexity of projected gradient descent [YPCC16].

- *Matrix completion.* Under the random Bernoulli observation model, as long as the sample complexity satisfies $n_1 n_2 p \gtrsim (\mu\kappa^2 \vee \log n)\mu n r^2 \kappa^2$ with $n = n_1 \vee n_2$, `ScaledGD` in conjunction with

	Matrix sensing		Robust PCA		Matrix completion	
Algorithms	sample complexity	iteration complexity	corruption fraction	iteration complexity	sample complexity	iteration complexity
`GD`	$nr^2\kappa^2$	$\kappa \log \frac{1}{\epsilon}$	$\frac{1}{\mu r^{3/2}\kappa^{3/2} \vee \mu r \kappa^2}$	$\kappa \log \frac{1}{\epsilon}$	$(\mu \vee \log n)\mu n r^2 \kappa^2$	$\kappa \log \frac{1}{\epsilon}$
`ScaledGD` (this Chapter)	$nr^2\kappa^2$	$\log \frac{1}{\epsilon}$	$\frac{1}{\mu r^{3/2}\kappa}$	$\log \frac{1}{\epsilon}$	$(\mu\kappa^2 \vee \log n)\mu n r^2 \kappa^2$	$\log \frac{1}{\epsilon}$

Table 2.1: Comparisons of `ScaledGD` with gradient descent (`GD`) when tailored to various problems (with spectral initialization) [TBS+16, YPCC16, ZL16], where they have comparable per-iteration costs. Here, we say that the output $\boldsymbol{X}$ of an algorithm reaches ϵ-accuracy, if it satisfies $\|\boldsymbol{X}-\boldsymbol{X}_\star\|_{\mathsf{F}} \leq \epsilon\sigma_r(\boldsymbol{X}_\star)$. Here, $n \coloneqq n_1 \vee n_2 = \max\{n_1, n_2\}$, κ and μ are the condition number and incoherence parameter of $\boldsymbol{X}_\star$.

a properly designed projection operator reaches ϵ-accuracy in $O(\log(1/\epsilon))$ iterations when initialized by the spectral method. This improves the iteration complexity of projected gradient descent [ZL16] at the expense of requiring a larger sample size.

In addition, `ScaledGD` does not require any explicit regularizations that balance the norms of two low-rank factors as required in [TBS+16, YPCC16, ZL16], and removed the additional projection that maintains the incoherence properties in robust PCA [YPCC16], thus unveiling the implicit regularization property of `ScaledGD`. To the best of our knowledge, this is the first factored gradient descent algorithm that achieves a fast convergence rate that is independent of the condition number of the low-rank matrix at near-optimal sample complexities without increasing the per-iteration computational cost. Our analysis is also applicable to general loss functions that are restricted strongly convex and smooth over low-rank matrices.

At the core of our analysis, we introduce a new distance metric (i.e. Lyapunov function) that accounts for the preconditioners, and carefully show the contraction of the `ScaledGD` iterates under the new distance metric.

Remark 1 (`ScaledGD` for PSD matrices). When the low-rank matrix of interest is positive semi-definite (PSD), we factorize the matrix $\boldsymbol{X} \in \mathbb{R}^{n\times n}$ as $\boldsymbol{X} = \boldsymbol{L}\boldsymbol{L}^\top$, with $\boldsymbol{L} \in \mathbb{R}^{n\times r}$. The update rule of `ScaledGD` simplifies to

$$\boldsymbol{L}_{t+1} = \boldsymbol{L}_t - \eta \nabla_{\boldsymbol{L}} \mathcal{L}(\boldsymbol{L}_t)(\boldsymbol{L}_t^\top \boldsymbol{L}_t)^{-1}. \tag{2.3}$$

We focus on the asymmetric case since the analysis is more involved with two factors. Our theory applies to the PSD case without loss of generality.

2.1.1 Related work

Our work contributes to the growing literature of design and analysis of provable nonconvex optimization procedures for high-dimensional signal estimation; see e.g. [JK17, CC18, CLC19] for recent overviews. A growing number of problems have been demonstrated to possess benign geometry that is amenable for optimization [MBM18] either globally or locally under appropriate statistical models. On one end, it is shown that there are no spurious local minima in the optimization landscape of matrix sensing and completion [GLM16, BNS16, PKCS17, GJZ17], phase retrieval [SQW18, DDP20], dictionary learning [SQW15], kernel PCA [CL19] and linear neural networks [BH89, Kaw16]. Such landscape analysis facilitates the adoption of generic saddle-point escaping algorithms [NP06, GHJY15, JGN+17] to ensure global convergence. However, the resulting iteration complexity is typically high. On the other end, local refinements with carefully-designed initializations often admit fast convergence, for example in phase retrieval [CLS15, MWCC19], matrix sensing [JNS13, ZL15, WCCL16], matrix completion [SL16, CW15, MWCC19, CLL20, ZL16, CCF+20], blind deconvolution [LLSW19, MWCC19], and robust PCA [NNS+14, YPCC16, CFMY21], to name a few.

Existing approaches for asymmetric low-rank matrix estimation often require additional regularization terms to balance the two factors, either in the form of $\frac{1}{2}\|\boldsymbol{L}^\top\boldsymbol{L}-\boldsymbol{R}^\top\boldsymbol{R}\|_{\mathsf{F}}^2$ [TBS+16, PKCS17] or $\frac{1}{2}\|\boldsymbol{L}\|_{\mathsf{F}}^2+\frac{1}{2}\|\boldsymbol{R}\|_{\mathsf{F}}^2$ [ZLTW18, CCF+20, CFMY21], which ease the theoretical analysis but are often unnecessary for the practical success, as long as the initialization is balanced. Some recent work studies the unregularized gradient descent for low-rank matrix factorization and sensing including [CCD+21, DHL18, MLC21]. However, the iteration complexity of all these approaches scales at least linearly with respect to the condition number κ of the low-rank matrix, e.g. $O(\kappa\log(1/\epsilon))$, to reach ϵ-accuracy, therefore they converge slowly when the underlying matrix becomes ill-conditioned. In contrast, `ScaledGD` enjoys a local convergence rate of $O(\log(1/\epsilon))$, therefore incurring a much smaller computational footprint when κ is large. Last but not least, alternating minimization [JNS13, HW14]

(which alternatively updates $\boldsymbol{L}_t$ and $\boldsymbol{R}_t$) or singular value projection [NNS+14, JMD10] (which operates in the matrix space) also converge at the rate $O(\log(1/\epsilon))$, but the per-iteration cost is much higher than `ScaledGD`. Another notable algorithm is the Riemannian gradient descent algorithm in [WCCL16], which also converges at the rate $O(\log(1/\epsilon))$ under the same sample complexity for low-rank matrix sensing, but requires a higher memory complexity since it operates in the matrix space rather than the factor space.

From an algorithmic perspective, our approach is closely related to the alternating steepest descent (ASD) method in [TW16] for low-rank matrix completion, which performs the proposed updates (2.2) for the low-rank factors in an alternating manner. Furthermore, the scaled gradient updates were also introduced in [MAS12, MS16] for low-rank matrix completion from the perspective of Riemannian optimization. However, none of [TW16, MAS12, MS16] offered any statistical nor computational guarantees for global convergence. Our analysis of `ScaledGD` can be viewed as providing justifications to these precursors. Moreover, we have systematically extended the framework of `ScaledGD` to work in a large number of low-rank matrix estimation tasks such as robust PCA.

2.1.2 Chapter organization

The rest of this chapter is organized as follows. Section 2.2 describes the proposed `ScaledGD` method and details its application to low-rank matrix sensing, robust PCA and matrix completion with theoretical guarantees in terms of both statistical and computational complexities, highlighting the role of a new distance metric. The convergence guarantee of `ScaledGD` under the general loss function is also presented. In Section 2.3, we outline the proof for our main results. Section 2.4 illustrates the excellent empirical performance of `ScaledGD` in a variety of low-rank matrix estimation problems. Finally, we conclude in Section 2.5.

2.2 Scaled Gradient Descent for Low-Rank Matrix Estimation

This section is devoted to introducing `ScaledGD` and establishing its statistical and computational guarantees for various low-rank matrix estimation problems. Before we instantiate tailored versions of `ScaledGD` on concrete low-rank matrix estimation problems, we first pause to provide more

insights of the update rule of `ScaledGD`, by connecting it to the quasi-Newton method. Note that the update rule (2.2) for `ScaledGD` can be equivalently written in a vectorization form as

$$\begin{aligned}\mathrm{vec}(\boldsymbol{F}_{t+1}) &= \mathrm{vec}(\boldsymbol{F}_t) - \eta \begin{bmatrix} (\boldsymbol{R}_t^\top \boldsymbol{R}_t)^{-1} \otimes \boldsymbol{I}_{n_1} & \boldsymbol{0} \\ \boldsymbol{0} & (\boldsymbol{L}_t^\top \boldsymbol{L}_t)^{-1} \otimes \boldsymbol{I}_{n_2} \end{bmatrix} \mathrm{vec}(\nabla_{\boldsymbol{F}} \mathcal{L}(\boldsymbol{F}_t)) \\ &= \mathrm{vec}(\boldsymbol{F}_t) - \eta \boldsymbol{H}_t^{-1} \mathrm{vec}(\nabla_{\boldsymbol{F}} \mathcal{L}(\boldsymbol{F}_t)), \end{aligned} \tag{2.4}$$

where we denote $\boldsymbol{F}_t = [\boldsymbol{L}_t^\top, \boldsymbol{R}_t^\top]^\top \in \mathbb{R}^{(n_1+n_2)\times r}$, and by $\otimes$ the Kronecker product. Here, the block diagonal matrix $\boldsymbol{H}_t$ is set to be

$$\boldsymbol{H}_t \coloneqq \begin{bmatrix} (\boldsymbol{R}_t^\top \boldsymbol{R}_t) \otimes \boldsymbol{I}_{n_1} & \boldsymbol{0} \\ \boldsymbol{0} & (\boldsymbol{L}_t^\top \boldsymbol{L}_t) \otimes \boldsymbol{I}_{n_2} \end{bmatrix}.$$

The form (2.4) makes it apparent that `ScaledGD` can be interpreted as a quasi-Newton algorithm, where the inverse of $\boldsymbol{H}_t$ can be cheaply computed through inverting two rank-r matrices.

2.2.1 Assumptions and error metric

Denote by $\boldsymbol{U}_\star \boldsymbol{\Sigma}_\star \boldsymbol{V}_\star^\top$ the compact singular value decomposition (SVD) of the rank-r matrix $\boldsymbol{X}_\star \in \mathbb{R}^{n_1 \times n_2}$. Here $\boldsymbol{U}_\star \in \mathbb{R}^{n_1 \times r}$ and $\boldsymbol{V}_\star \in \mathbb{R}^{n_2 \times r}$ are composed of r left and right singular vectors, respectively, and $\boldsymbol{\Sigma}_\star \in \mathbb{R}^{r\times r}$ is a diagonal matrix consisting of r singular values of $\boldsymbol{X}_\star$ organized in a non-increasing order, i.e. $\sigma_1(\boldsymbol{X}_\star) \geq \cdots \geq \sigma_r(\boldsymbol{X}_\star) > 0$. Define

$$\kappa \coloneqq \sigma_1(\boldsymbol{X}_\star)/\sigma_r(\boldsymbol{X}_\star) \tag{2.5}$$

as the condition number of $\boldsymbol{X}_\star$. Define the ground truth low-rank factors as

$$\boldsymbol{L}_\star \coloneqq \boldsymbol{U}_\star \boldsymbol{\Sigma}_\star^{1/2}, \qquad \text{and} \qquad \boldsymbol{R}_\star \coloneqq \boldsymbol{V}_\star \boldsymbol{\Sigma}_\star^{1/2}, \tag{2.6}$$

so that $\boldsymbol{X}_\star = \boldsymbol{L}_\star \boldsymbol{R}_\star^\top$. Correspondingly, denote the stacked factor matrix as

$$\boldsymbol{F}_\star := \begin{bmatrix} \boldsymbol{L}_\star \\ \boldsymbol{R}_\star \end{bmatrix} \in \mathbb{R}^{(n_1+n_2)\times r}. \tag{2.7}$$

Next, we are in need of a right metric to measure the performance of the `ScaledGD` iterates $\boldsymbol{F}_t := [\boldsymbol{L}_t^\top, \boldsymbol{R}_t^\top]^\top$. Obviously, the factored representation is not unique in that for any invertible matrix $\boldsymbol{Q} \in \mathrm{GL}(r)$, one has $\boldsymbol{L}\boldsymbol{R}^\top = (\boldsymbol{L}\boldsymbol{Q})(\boldsymbol{R}\boldsymbol{Q}^{-\top})^\top$. Therefore, the reconstruction error metric needs to take into account this identifiability issue. More importantly, we need a diagonal scaling in the distance error metric to properly account for the effect of preconditioning. To provide intuition, note that the update rule (2.2) can be viewed as finding the best local quadratic approximation of $\mathcal{L}(\cdot)$ in the following sense:

$$\begin{aligned}(\boldsymbol{L}_{t+1}, \boldsymbol{R}_{t+1}) = \operatorname*{argmin}_{\boldsymbol{L},\boldsymbol{R}} \mathcal{L}(\boldsymbol{L}_t, \boldsymbol{R}_t) + \langle \nabla_{\boldsymbol{L}} \mathcal{L}(\boldsymbol{L}_t, \boldsymbol{R}_t), \boldsymbol{L} - \boldsymbol{L}_t \rangle + \langle \nabla_{\boldsymbol{R}} \mathcal{L}(\boldsymbol{L}_t, \boldsymbol{R}_t), \boldsymbol{R} - \boldsymbol{R}_t \rangle \\ + \frac{1}{2\eta} \left(\left\| (\boldsymbol{L} - \boldsymbol{L}_t)(\boldsymbol{R}_t^\top \boldsymbol{R}_t)^{1/2} \right\|_{\mathsf{F}}^2 + \left\| (\boldsymbol{R} - \boldsymbol{R}_t)(\boldsymbol{L}_t^\top \boldsymbol{L}_t)^{1/2} \right\|_{\mathsf{F}}^2 \right),\end{aligned}$$

where it is different from the common interpretation of gradient descent in the way the quadratic approximation is taken by a scaled norm. When $\boldsymbol{L}_t \approx \boldsymbol{L}_\star$ and $\boldsymbol{R}_t \approx \boldsymbol{R}_\star$ are approaching the ground truth, the additional scaling factors can be approximated by $\boldsymbol{L}_t^\top \boldsymbol{L}_t \approx \boldsymbol{\Sigma}_\star$ and $\boldsymbol{R}_t^\top \boldsymbol{R}_t \approx \boldsymbol{\Sigma}_\star$, leading to the following error metric

$$\mathrm{dist}^2(\boldsymbol{F}, \boldsymbol{F}_\star) := \inf_{\boldsymbol{Q}\in \mathrm{GL}(r)} \left\| (\boldsymbol{L}\boldsymbol{Q} - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2 + \left\| (\boldsymbol{R}\boldsymbol{Q}^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2. \tag{2.8}$$

Correspondingly, we define the optimal alignment matrix $\boldsymbol{Q}$ between $\boldsymbol{F}$ and $\boldsymbol{F}_\star$ as

$$\boldsymbol{Q} := \operatorname*{argmin}_{\boldsymbol{Q}\in \mathrm{GL}(r)} \left\| (\boldsymbol{L}\boldsymbol{Q} - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2 + \left\| (\boldsymbol{R}\boldsymbol{Q}^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2, \tag{2.9}$$

whenever the minimum is achieved.[1] It turns out that for the `ScaledGD` iterates $\{\boldsymbol{F}_t\}$, the optimal alignment matrices $\{\boldsymbol{Q}_t\}$ always exist (at least when properly initialized) and hence are well-defined.

[1] If there are multiple minimizers, we can arbitrarily take one to be $\boldsymbol{Q}$.

The design and analysis of this new distance metric are of crucial importance in obtaining the improved rate of `ScaledGD`; see Appendix A.1.1 for a collection of its properties. In comparison, the previously studied distance metrics (proposed mainly for GD) either do not include the diagonal scaling [MLC21, TBS+16], or only consider the ambiguity class up to orthonormal transforms [TBS+16], which fail to unveil the benefit of `ScaledGD`.

2.2.2 Matrix sensing

Assume that we have collected a set of linear measurements about a rank-r matrix $\boldsymbol{X}_\star \in \mathbb{R}^{n_1 \times n_2}$, given as

$$\boldsymbol{y} = \mathcal{A}(\boldsymbol{X}_\star) \in \mathbb{R}^m, \tag{2.10}$$

where $\mathcal{A}(\boldsymbol{X}) = \{\langle \boldsymbol{A}_k, \boldsymbol{X}\rangle\}_{k=1}^m : \mathbb{R}^{n_1 \times n_2} \mapsto \mathbb{R}^m$ is the linear map modeling the measurement process. The goal of low-rank matrix sensing is to recover $\boldsymbol{X}_\star$ from $\boldsymbol{y}$, especially when the number of measurements $m \ll n_1 n_2$, by exploiting the low-rank property. This problem has wide applications in medical imaging, signal processing, and data compression [CP11].

Algorithm. Writing $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$ into a factored form $\boldsymbol{X} = \boldsymbol{L}\boldsymbol{R}^\top$, we consider the following optimization problem:

$$\underset{\boldsymbol{F} \in \mathbb{R}^{(n_1+n_2)\times r}}{\text{minimize}} \quad \mathcal{L}(\boldsymbol{F}) = \frac{1}{2}\left\|\mathcal{A}(\boldsymbol{L}\boldsymbol{R}^\top) - \boldsymbol{y}\right\|_2^2. \tag{2.11}$$

Here as before, $\boldsymbol{F}$ denotes the stacked factor matrix $[\boldsymbol{L}^\top, \boldsymbol{R}^\top]^\top$. We suggest running `ScaledGD` (2.2) with the spectral initialization to solve (2.11), which performs the top-r SVD on $\mathcal{A}^*(\boldsymbol{y})$, where $\mathcal{A}^*(\cdot)$ is the adjoint operator of $\mathcal{A}(\cdot)$. The full algorithm is stated in Algorithm 1. The low-rank matrix can be estimated as $\boldsymbol{X}_T = \boldsymbol{L}_T \boldsymbol{R}_T^\top$ after running T iterations of `ScaledGD`.

Theoretical guarantees. To understand the performance of `ScaledGD` for low-rank matrix sensing, we adopt a standard assumption on the sensing operator $\mathcal{A}(\cdot)$, namely the Restricted Isometry Property (RIP).

Algorithm 1 `ScaledGD` for low-rank matrix sensing with spectral initialization

Spectral initialization: Let $\boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^\top$ be the top-r SVD of $\mathcal{A}^*(\boldsymbol{y})$, and set

$$\boldsymbol{L}_0 = \boldsymbol{U}_0\boldsymbol{\Sigma}_0^{1/2}, \quad \text{and} \quad \boldsymbol{R}_0 = \boldsymbol{V}_0\boldsymbol{\Sigma}_0^{1/2}. \tag{2.12}$$

Scaled gradient updates: **for** $t = 0, 1, 2, \ldots, T-1$ **do**

$$\begin{aligned} \boldsymbol{L}_{t+1} &= \boldsymbol{L}_t - \eta\mathcal{A}^*(\mathcal{A}(\boldsymbol{L}_t\boldsymbol{R}_t^\top) - \boldsymbol{y})\boldsymbol{R}_t(\boldsymbol{R}_t^\top\boldsymbol{R}_t)^{-1}, \\ \boldsymbol{R}_{t+1} &= \boldsymbol{R}_t - \eta\mathcal{A}^*(\mathcal{A}(\boldsymbol{L}_t\boldsymbol{R}_t^\top) - \boldsymbol{y})^\top\boldsymbol{L}_t(\boldsymbol{L}_t^\top\boldsymbol{L}_t)^{-1}. \end{aligned} \tag{2.13}$$

Definition 1 (RIP [RFP10])**.** The linear map $\mathcal{A}(\cdot)$ is said to obey the rank-r RIP with a constant $\delta_r \in [0, 1)$, if for all matrices $\boldsymbol{M} \in \mathbb{R}^{n_1\times n_2}$ of rank at most r, one has

$$(1-\delta_r)\|\boldsymbol{M}\|_{\mathsf{F}}^2 \le \|\mathcal{A}(\boldsymbol{M})\|_2^2 \le (1+\delta_r)\|\boldsymbol{M}\|_{\mathsf{F}}^2.$$

It is well-known that many measurement ensembles satisfy the RIP property [RFP10, CP11]. For example, if the entries of $\boldsymbol{A}_i$'s are composed of i.i.d. Gaussian entries $\mathcal{N}(0, 1/m)$, then the RIP is satisfied for a constant δ_r as long as m is on the order of $(n_1+n_2)r/\delta_r^2$. With the RIP condition in place, the following theorem demonstrates that `ScaledGD` converges linearly — in terms of the new distance metric (cf. (2.8)) — at a constant rate as long as the sensing operator $\mathcal{A}(\cdot)$ has a sufficiently small RIP constant.

Theorem 1. *Suppose that $\mathcal{A}(\cdot)$ obeys the $2r$-RIP with $\delta_{2r} \le 0.02/(\sqrt{r}\kappa)$. If the step size obeys $0 < \eta \le 2/3$, then for all $t \ge 0$, the iterates of the `ScaledGD` method in Algorithm 1 satisfy*

$$\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \le (1-0.6\eta)^t 0.1\sigma_r(\boldsymbol{X}_\star), \quad \text{and} \quad \left\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\right\|_{\mathsf{F}} \le (1-0.6\eta)^t 0.15\sigma_r(\boldsymbol{X}_\star).$$

Theorem 1 establishes that the distance $\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$ contracts linearly at a constant rate, as long as the sample size satisfies $m = O(nr^2\kappa^2)$ with Gaussian random measurements [RFP10], where we recall that $n = n_1 \vee n_2$. To reach ϵ-accuracy, i.e. $\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\|_{\mathsf{F}} \le \epsilon\sigma_r(\boldsymbol{X}_\star)$, `ScaledGD` takes at most $T = O(\log(1/\epsilon))$ iterations, which is *independent* of the condition number κ of $\boldsymbol{X}_\star$. In comparison, alternating minimization with spectral initialization (`AltMinSense`) converges in

$O(\log(1/\epsilon))$ iterations as long as $m = O(nr^3\kappa^4)$ [JNS13], where the per-iteration cost is much higher.[2] On the other end, gradient descent with spectral initialization in [TBS+16] converges in $O(\kappa \log(1/\epsilon))$ iterations as long as $m = O(nr^2\kappa^2)$. Therefore, `ScaledGD` converges at a much faster rate than GD at the same sample complexity while requiring a significantly lower per-iteration cost than `AltMinSense`.

Remark 2. [TBS+16] suggested that one can employ a more expensive initialization scheme, e.g. performing multiple projected gradient descent steps over the low-rank matrix, to reduce the sample complexity. By seeding `ScaledGD` with the output of updates of the form $\boldsymbol{X}_{\tau+1} = \mathcal{P}_r\left(\boldsymbol{X}_\tau - \mathcal{A}^*(\mathcal{A}(\boldsymbol{X}_\tau) - \boldsymbol{y})\right)$ after $T_0 \gtrsim \log(\sqrt{r}\kappa)$ iterations, where $\mathcal{P}_r(\cdot)$ is defined in (1.3), `ScaledGD` succeeds with the sample size $O(nr)$ which is information theoretically optimal.

2.2.3 Robust PCA

Assume that we have observed the data matrix

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

which is a superposition of a rank-r matrix $\boldsymbol{X}_\star$, modeling the clean data, and a sparse matrix $\boldsymbol{S}_\star$, modeling the corruption or outliers. The goal of robust PCA [CLMW11, CSPW11] is to separate the two matrices $\boldsymbol{X}_\star$ and $\boldsymbol{S}_\star$ from their mixture $\boldsymbol{Y}$. This problem finds numerous applications in video surveillance, image processing, and so on.

Following [CSPW11, NNS+14, YPCC16], we consider a deterministic sparsity model for $\boldsymbol{S}_\star$, in which $\boldsymbol{S}_\star$ contains at most α-fraction of nonzero entries per row and column for some $\alpha \in [0, 1)$, i.e. $\boldsymbol{S}_\star \in \mathcal{S}_\alpha$, where we denote

$$\mathcal{S}_\alpha \coloneqq \{\boldsymbol{S} \in \mathbb{R}^{n_1 \times n_2} : \|\boldsymbol{S}_{i,\cdot}\|_0 \le \alpha n_2 \text{ for all } i, \text{ and } \|\boldsymbol{S}_{\cdot,j}\|_0 \le \alpha n_1 \text{ for all } j\}. \tag{2.14}$$

[2] The exact per-iteration complexity of `AltMinSense` depends on how the least-squares subproblems are solved with m equations and nr unknowns; see [LHLZ20, Table 1] for detailed comparisons.

Algorithm. Writing $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$ into the factored form $\boldsymbol{X} = \boldsymbol{L}\boldsymbol{R}^\top$, we consider the following optimization problem:

$$\underset{\boldsymbol{F} \in \mathbb{R}^{(n_1+n_2)\times r}, \boldsymbol{S} \in \mathcal{S}_\alpha}{\text{minimize}} \ \mathcal{L}(\boldsymbol{F}, \boldsymbol{S}) = \frac{1}{2} \left\| \boldsymbol{L}\boldsymbol{R}^\top + \boldsymbol{S} - \boldsymbol{Y} \right\|_{\mathsf{F}}^2. \tag{2.15}$$

It is thus natural to alternatively update $\boldsymbol{F} = [\boldsymbol{L}^\top, \boldsymbol{R}^\top]^\top$ and $\boldsymbol{S}$, where $\boldsymbol{F}$ is updated via the proposed `ScaledGD` algorithm, and $\boldsymbol{S}$ is updated by hard thresholding, which trims the small entries of the residual matrix $\boldsymbol{Y} - \boldsymbol{L}\boldsymbol{R}^\top$. More specifically, for some truncation level $0 \le \bar{\alpha} \le 1$, we define the sparsification operator that only keeps $\bar{\alpha}$ fraction of largest entries in each row and column:

$$(\mathcal{T}_{\bar{\alpha}}[\boldsymbol{A}])_{i,j} = \begin{cases} \boldsymbol{A}_{i,j}, & \text{if } |\boldsymbol{A}|_{i,j} \ge |\boldsymbol{A}|_{i,(\bar{\alpha}n_2)}, \text{ and } |\boldsymbol{A}|_{i,j} \ge |\boldsymbol{A}|_{(\bar{\alpha}n_1),j} \\ 0, & \text{otherwise} \end{cases}, \tag{2.16}$$

where $|\boldsymbol{A}|_{i,(k)}$ (resp. $|\boldsymbol{A}|_{(k),j}$) denote the k-th largest element in magnitude in the i-th row (resp. j-th column).

The `ScaledGD` algorithm with the spectral initialization for solving robust PCA is formally stated in Algorithm 2. Note that, comparing with [YPCC16], we do not require a balancing term $\|\boldsymbol{L}^\top\boldsymbol{L} - \boldsymbol{R}^\top\boldsymbol{R}\|_{\mathsf{F}}^2$ in the loss function (2.15), nor the projection of the low-rank factors onto the $\ell_{2,\infty}$ ball in each iteration.

Algorithm 2 `ScaledGD` for robust PCA with spectral initialization

Spectral initialization: Let $\boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^\top$ be the top-r SVD of $\boldsymbol{Y} - \mathcal{T}_\alpha[\boldsymbol{Y}]$, and set

$$\boldsymbol{L}_0 = \boldsymbol{U}_0\boldsymbol{\Sigma}_0^{1/2}, \quad \text{and} \quad \boldsymbol{R}_0 = \boldsymbol{V}_0\boldsymbol{\Sigma}_0^{1/2}. \tag{2.17}$$

Scaled gradient updates: **for** $t = 0, 1, 2, \ldots, T-1$ **do**

$$\begin{aligned} \boldsymbol{S}_t &= \mathcal{T}_{2\alpha}[\boldsymbol{Y} - \boldsymbol{L}_t\boldsymbol{R}_t^\top], \\ \boldsymbol{L}_{t+1} &= \boldsymbol{L}_t - \eta(\boldsymbol{L}_t\boldsymbol{R}_t^\top + \boldsymbol{S}_t - \boldsymbol{Y})\boldsymbol{R}_t(\boldsymbol{R}_t^\top\boldsymbol{R}_t)^{-1}, \\ \boldsymbol{R}_{t+1} &= \boldsymbol{R}_t - \eta(\boldsymbol{L}_t\boldsymbol{R}_t^\top + \boldsymbol{S}_t - \boldsymbol{Y})^\top\boldsymbol{L}_t(\boldsymbol{L}_t^\top\boldsymbol{L}_t)^{-1}. \end{aligned} \tag{2.18}$$

Theoretical guarantee. Before stating our main result for robust PCA, we introduce the incoherence condition which is known to be crucial for reliable estimation of the low-rank matrix $\boldsymbol{X}_\star$ in robust PCA [Che15].

Definition 2 (Incoherence)**.** A rank-r matrix $\boldsymbol{X}_\star \in \mathbb{R}^{n_1 \times n_2}$ with compact SVD as $\boldsymbol{X}_\star = \boldsymbol{U}_\star \boldsymbol{\Sigma}_\star \boldsymbol{V}_\star^\top$ is said to be μ-incoherent if

$$\|\boldsymbol{U}_\star\|_{2,\infty} \le \sqrt{\frac{\mu}{n_1}}\|\boldsymbol{U}_\star\|_{\mathsf{F}} = \sqrt{\frac{\mu r}{n_1}}, \quad \text{and} \quad \|\boldsymbol{V}_\star\|_{2,\infty} \le \sqrt{\frac{\mu}{n_2}}\|\boldsymbol{V}_\star\|_{\mathsf{F}} = \sqrt{\frac{\mu r}{n_2}}.$$

The following theorem establishes that `ScaledGD` converges linearly at a constant rate as long as the fraction α of corruptions is sufficiently small.

Theorem 2. *Suppose that $\boldsymbol{X}_\star$ is μ-incoherent and that the corruption fraction α obeys $\alpha \le c/(\mu r^{3/2}\kappa)$ for some sufficiently small constant $c > 0$. If the step size obeys $0.1 \le \eta \le 2/3$, then for all $t \ge 0$, the iterates of `ScaledGD` in Algorithm 2 satisfy*

$$\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \le (1-0.6\eta)^t 0.02\sigma_r(\boldsymbol{X}_\star), \quad \textit{and} \quad \left\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\right\|_{\mathsf{F}} \le (1-0.6\eta)^t 0.03\sigma_r(\boldsymbol{X}_\star).$$

Theorem 2 establishes that the distance $\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$ contracts linearly at a constant rate, as long as the fraction of corruptions satisfies $\alpha \lesssim 1/(\mu r^{3/2}\kappa)$. To reach ϵ-accuracy, i.e. $\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\|_{\mathsf{F}} \le \epsilon\sigma_r(\boldsymbol{X}_\star)$, `ScaledGD` takes at most $T = O(\log(1/\epsilon))$ iterations, which is *independent* of κ. In comparison, the `AltProj` algorithm[3] with spectral initialization converges in $O(\log(1/\epsilon))$ iterations as long as $\alpha \lesssim 1/(\mu r)$ [NNS+14], where the per-iteration cost is much higher both in terms of computation and memory as it requires the computation of the low-rank SVD of the full matrix. On the other hand, projected gradient descent with spectral initialization in [YPCC16] converges in $O(\kappa\log(1/\epsilon))$ iterations as long as $\alpha \lesssim 1/(\mu r^{3/2}\kappa^{3/2} \vee \mu r\kappa^2)$. Therefore, `ScaledGD` converges at a much faster rate than GD while requesting a significantly lower per-iteration cost than `AltProj`. In addition, our theory suggests that `ScaledGD` maintains the incoherence and balancedness of the low-rank factors without imposing explicit regularizations, which is not captured in previous

[3] `AltProj` employs a multi-stage strategy to remove the dependence on κ in α, which we do not consider here. The same strategy might also improve the dependence on κ for `ScaledGD`, which we leave for future work.

analysis [YPCC16].

2.2.4 Matrix completion

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

$$(\mathcal{P}_\Omega(\boldsymbol{X}))_{i,j} = \begin{cases} \boldsymbol{X}_{i,j}, & \text{if } (i,j)\in\Omega \\ 0, & \text{otherwise} \end{cases}. \tag{2.19}$$

Here Ω is generated according to the Bernoulli model in the sense that each $(i,j)\in\Omega$ independent with probability p. The goal of matrix completion is to recover the matrix $\boldsymbol{X}_\star$ from its partial observation $\mathcal{P}_\Omega(\boldsymbol{X}_\star)$. This problem has many applications in recommendation systems, signal processing, sensor network localization, and so on [CR09].

Algorithm. Again, writing $\boldsymbol{X}\in\mathbb{R}^{n_1\times n_2}$ into the factored form $\boldsymbol{X}=\boldsymbol{L}\boldsymbol{R}^\top$, we consider the following optimization problem:

$$\underset{\boldsymbol{F}\in\mathbb{R}^{(n_1+n_2)\times r}}{\text{minimize}}\ \mathcal{L}(\boldsymbol{F}) \coloneqq \frac{1}{2p}\left\|\mathcal{P}_\Omega(\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star)\right\|_{\mathsf{F}}^2. \tag{2.20}$$

Similarly to robust PCA, the underlying low-rank matrix $\boldsymbol{X}_\star$ needs to be incoherent (cf. Definition 2) to avoid ill-posedness. One typical strategy to ensure the incoherence condition is to perform projection after the gradient update, by projecting the iterates to maintain small $\ell_{2,\infty}$ norms of the factor matrices. However, the standard projection operator [CW15] is not covariant with respect to invertible transforms, and consequently, needs to be modified when using scaled gradient updates. To that end, we introduce the following new projection operator: for every $\widetilde{\boldsymbol{F}}\in\mathbb{R}^{(n_1+n_2)\times r} = [\widetilde{\boldsymbol{L}}^\top,\widetilde{\boldsymbol{R}}^\top]^\top$,

$$\begin{aligned}\mathcal{P}_B(\widetilde{\boldsymbol{F}}) = \underset{\boldsymbol{F}\in\mathbb{R}^{(n_1+n_2)\times r}}{\operatorname{argmin}} & \left\|(\boldsymbol{L}-\widetilde{\boldsymbol{L}})(\widetilde{\boldsymbol{R}}^\top\widetilde{\boldsymbol{R}})^{1/2}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{R}-\widetilde{\boldsymbol{R}})(\widetilde{\boldsymbol{L}}^\top\widetilde{\boldsymbol{L}})^{1/2}\right\|_{\mathsf{F}}^2 \\ \text{s.t.}\quad & \sqrt{n_1}\left\|\boldsymbol{L}(\widetilde{\boldsymbol{R}}^\top\widetilde{\boldsymbol{R}})^{1/2}\right\|_{2,\infty} \vee \sqrt{n_2}\left\|\boldsymbol{R}(\widetilde{\boldsymbol{L}}^\top\widetilde{\boldsymbol{L}})^{1/2}\right\|_{2,\infty} \le B\end{aligned}, \tag{2.21}$$

which finds a factored matrix that is closest to $\widetilde{\boldsymbol{F}}$ and stays incoherent in a weighted sense. Luckily, the solution to the above scaled projection admits a simple closed-form solution, as stated below.

Proposition 1. *The solution to* (2.21) *is given by*

$$\mathcal{P}_B(\widetilde{\boldsymbol{F}}) \coloneqq \begin{bmatrix} \boldsymbol{L} \\ \boldsymbol{R} \end{bmatrix}, \quad \text{where} \quad \begin{aligned} \boldsymbol{L}_{i,\cdot} &\coloneqq \left(1 \wedge \frac{B}{\sqrt{n_1}\|\widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{R}}^\top\|_2}\right)\widetilde{\boldsymbol{L}}_{i,\cdot}, \quad 1 \le i \le n_1, \\ \boldsymbol{R}_{j,\cdot} &\coloneqq \left(1 \wedge \frac{B}{\sqrt{n_2}\|\widetilde{\boldsymbol{R}}_{j,\cdot}\widetilde{\boldsymbol{L}}^\top\|_2}\right)\widetilde{\boldsymbol{R}}_{j,\cdot}, \quad 1 \le j \le n_2. \end{aligned} \tag{2.22}$$

Proof. See Appendix A.5.1. □

With the new projection operator in place, we propose the scaled projected gradient descent (`ScaledPGD`) method with the spectral initialization for solving matrix completion, formally stated in Algorithm 3.

Algorithm 3 `ScaledPGD` for matrix completion with spectral initialization

Spectral initialization: Let $\boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^\top$ be the top-r SVD of $\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{X}_\star)$, and set

$$\begin{bmatrix} \boldsymbol{L}_0 \\ \boldsymbol{R}_0 \end{bmatrix} = \mathcal{P}_B\left(\begin{bmatrix} \boldsymbol{U}_0\boldsymbol{\Sigma}_0^{1/2} \\ \boldsymbol{V}_0\boldsymbol{\Sigma}_0^{1/2} \end{bmatrix}\right). \tag{2.23}$$

Scaled projected gradient updates: **for** $t = 0, 1, 2, \ldots, T-1$ **do**

$$\begin{bmatrix} \boldsymbol{L}_{t+1} \\ \boldsymbol{R}_{t+1} \end{bmatrix} = \mathcal{P}_B\left(\begin{bmatrix} \boldsymbol{L}_t - \frac{\eta}{p}\mathcal{P}_\Omega(\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star)\boldsymbol{R}_t(\boldsymbol{R}_t^\top\boldsymbol{R}_t)^{-1} \\ \boldsymbol{R}_t - \frac{\eta}{p}\mathcal{P}_\Omega(\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star)^\top\boldsymbol{L}_t(\boldsymbol{L}_t^\top\boldsymbol{L}_t)^{-1} \end{bmatrix}\right). \tag{2.24}$$

Theoretical guarantee. Consider a random observation model, where each index (i, j) belongs to the index set Ω independently with probability $0 < p \le 1$. The following theorem establishes that `ScaledPGD` converges linearly at a constant rate as long as the number of observations is sufficiently large.

Theorem 3. *Suppose that $\boldsymbol{X}_\star$ is μ-incoherent, and that p satisfies $p \ge C(\mu\kappa^2 \vee \log(n_1 \vee n_2))\mu r^2\kappa^2/(n_1 \wedge n_2)$ for some sufficiently large constant C. Set the projection radius as $B = C_B\sqrt{\mu r}\sigma_1(\boldsymbol{X}_\star)$ for*

some constant $C_B \geq 1.02$. *If the step size obeys* $0 < \eta \leq 2/3$, *then with probability at least* $1 - c_1(n_1 \vee n_2)^{-c_2}$, *for all* $t \geq 0$, *the iterates of* `ScaledPGD` *in* (2.24) *satisfy*

$$\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \leq (1-0.6\eta)^t 0.02\sigma_r(\boldsymbol{X}_\star), \quad \textit{and} \quad \left\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\right\|_{\mathsf{F}} \leq (1-0.6\eta)^t 0.03\sigma_r(\boldsymbol{X}_\star).$$

Here $c_1, c_2 > 0$ *are two universal constants.*

Theorem 3 establishes that the distance $\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$ contracts linearly at a constant rate, as long as the probability of observation satisfies $p \gtrsim (\mu\kappa^2 \vee \log(n_1 \vee n_2))\mu r^2\kappa^2/(n_1 \wedge n_2)$. To reach ϵ-accuracy, i.e. $\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\|_{\mathsf{F}} \leq \epsilon\sigma_r(\boldsymbol{X}_\star)$, `ScaledPGD` takes at most $T = O(\log(1/\epsilon))$ iterations, which is *independent* of κ. In comparison, projected gradient descent [ZL16] with spectral initialization converges in $O(\kappa\log(1/\epsilon))$ iterations as long as $p \gtrsim (\mu \vee \log(n_1 \vee n_2))\mu r^2\kappa^2/(n_1 \wedge n_2)$. Therefore, `ScaledPGD` achieves much faster convergence than its unscaled counterpart, at an expense of higher sample complexity. We believe this higher sample complexity is an artifact of our proof techniques, as numerically we do not observe a degradation in terms of sample complexity.

2.2.5 Optimizing general loss functions

Last but not least, we generalize our analysis of `ScaledGD` to minimize a general loss function in the form of (2.1), where the update rule of `ScaledGD` is given by

$$\begin{aligned}\boldsymbol{L}_{t+1} &= \boldsymbol{L}_t - \eta\nabla f(\boldsymbol{L}_t\boldsymbol{R}_t^\top)\boldsymbol{R}_t(\boldsymbol{R}_t^\top\boldsymbol{R}_t)^{-1},\\ \boldsymbol{R}_{t+1} &= \boldsymbol{R}_t - \eta\nabla f(\boldsymbol{L}_t\boldsymbol{R}_t^\top)^\top\boldsymbol{L}_t(\boldsymbol{L}_t^\top\boldsymbol{L}_t)^{-1}.\end{aligned} \tag{2.25}$$

Two important properties of the loss function $f : \mathbb{R}^{n_1\times n_2} \mapsto \mathbb{R}$ play a key role in the analysis.

Definition 3 (Restricted smoothness)**.** A differentiable function $f : \mathbb{R}^{n_1\times n_2} \mapsto \mathbb{R}$ is said to be rank-r restricted L-smooth for some $L > 0$ if

$$f(\boldsymbol{X}_2) \leq f(\boldsymbol{X}_1) + \langle\nabla f(\boldsymbol{X}_1), \boldsymbol{X}_2 - \boldsymbol{X}_1\rangle + \frac{L}{2}\|\boldsymbol{X}_2 - \boldsymbol{X}_1\|_{\mathsf{F}}^2,$$

for any $\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathbb{R}^{n_1\times n_2}$ with rank at most r.

Definition 4 (Restricted strong convexity)**.** A differentiable function $f : \mathbb{R}^{n_1 \times n_2} \mapsto \mathbb{R}$ is said to be rank-r restricted μ-strongly convex for some $\mu \geq 0$ if

$$f(\boldsymbol{X}_2) \geq f(\boldsymbol{X}_1) + \langle \nabla f(\boldsymbol{X}_1), \boldsymbol{X}_2 - \boldsymbol{X}_1 \rangle + \frac{\mu}{2} \|\boldsymbol{X}_2 - \boldsymbol{X}_1\|_{\mathsf{F}}^2,$$

for any $\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathbb{R}^{n_1 \times n_2}$ with rank at most r. When $\mu = 0$, we simply say $f(\cdot)$ is rank-r restricted convex.

Further, when $\mu > 0$, define the condition number of the loss function $f(\cdot)$ over rank-r matrices as

$$\kappa_f \coloneqq L/\mu. \tag{2.26}$$

Encouragingly, many problems can be viewed as a special case of optimizing this general loss (2.25), including but not limited to:

- *low-rank matrix factorization*, where the loss function $f(\boldsymbol{X}) = \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{X}_\star\|_{\mathsf{F}}^2$ in (2.27) satisfies $\kappa_f = 1$;
- *low-rank matrix sensing*, where the loss function $f(\boldsymbol{X}) = \frac{1}{2}\|\mathcal{A}(\boldsymbol{X} - \boldsymbol{X}_\star)\|_2^2$ in (2.11) satisfies $\kappa_f \approx 1$ when $\mathcal{A}(\cdot)$ obeys the rank-r RIP with a sufficiently small RIP constant;
- *quadratic sampling*, where the loss function $f(\boldsymbol{X}) = \frac{1}{2}\sum_{i=1}^m |\langle \boldsymbol{a}_i \boldsymbol{a}_i^\top, \boldsymbol{X} - \boldsymbol{X}_\star \rangle|^2$ satisfies restricted strong convexity and smoothness when $\boldsymbol{a}_i$'s are i.i.d. Gaussian vectors for sufficiently large m [SWW17, LMCC21];
- *exponential-family PCA*, where the loss function $f(\boldsymbol{X}) = -\sum_{i,j} \log p(\boldsymbol{Y}_{i,j}|\boldsymbol{X}_{i,j})$, where $p(\boldsymbol{Y}_{i,j}|\boldsymbol{X}_{i,j})$ is the probability density function of $\boldsymbol{Y}_{i,j}$ conditional on $\boldsymbol{X}_{i,j}$, following an exponential-family distribution such as Bernoulli and Poisson distributions. The resulting loss function satisfies restricted strong convexity and smoothness with a condition number $\kappa_f > 1$ depending on the property of the specific distribution [GRG14, Laf15].

Indeed, the treatment of a general loss function brings the condition number of $f(\cdot)$ under the

spotlight, since in our earlier case studies $\kappa_f \approx 1$. Our purpose is thus to understand the interplay of two types of conditioning numbers in the convergence of first-order methods. For simplicity, we assume that $f(\cdot)$ is minimized at the ground truth rank-r matrix $\boldsymbol{X}_\star$.[4] The following theorem establishes that as long as properly initialized, then `ScaledGD` converges linearly at a constant rate.

Theorem 4. *Suppose that $f(\cdot)$ is rank-$2r$ restricted L-smooth and μ-strongly convex, of which $\boldsymbol{X}_\star$ is a minimizer, and that the initialization $\boldsymbol{F}_0$ satisfies $\mathrm{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \le 0.1\sigma_r(\boldsymbol{X}_\star)/\sqrt{\kappa_f}$. If the step size obeys $0 < \eta \le 0.4/L$, then for all $t \ge 0$, the iterates of `ScaledGD` in (2.25) satisfy*

$$\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \le (1-0.7\eta\mu)^t 0.1\sigma_r(\boldsymbol{X}_\star)/\sqrt{\kappa_f}, \quad \text{and} \quad \left\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\right\|_{\mathsf{F}} \le (1-0.7\eta\mu)^t 0.15\sigma_r(\boldsymbol{X}_\star)/\sqrt{\kappa_f}.$$

Theorem 4 establishes that the distance $\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$ contracts linearly at a constant rate, as long as the initialization $\boldsymbol{F}_0$ is sufficiently close to $\boldsymbol{F}_\star$. To reach ϵ-accuracy, i.e. $\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\|_{\mathsf{F}} \le \epsilon\sigma_r(\boldsymbol{X}_\star)$, `ScaledGD` takes at most $T = O(\kappa_f \log(1/\epsilon))$ iterations, which depends only on the condition number κ_f of $f(\cdot)$, but is independent of the condition number κ of the matrix $\boldsymbol{X}_\star$. In contrast, prior theory of vanilla gradient descent [PKCS18, BKS16] requires $O(\kappa_f\kappa \log(1/\epsilon))$ iterations, which is worse than our rate by a factor of κ.

2.3 Proof Sketch

In this section, we sketch the proof of the main theorems, highlighting the role of the scaled distance metric (cf. (2.8)) in these analyses.

2.3.1 A warm-up analysis: matrix factorization

Let us consider the problem of factorizing a matrix $\boldsymbol{X}_\star$ into two low-rank factors:

$$\underset{\boldsymbol{F}\in\mathbb{R}^{(n_1+n_2)\times r}}{\text{minimize}} \quad \mathcal{L}(\boldsymbol{F}) = \frac{1}{2}\left\|\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star\right\|_{\mathsf{F}}^2. \tag{2.27}$$

[4] In practice, due to the presence of statistical noise, the minimizer of $f(\cdot)$ might be only approximately low-rank, to which our analysis can be extended in a straightforward fashion.

For this toy problem, the update rule of `ScaledGD` is given as

$$\begin{aligned} \boldsymbol{L}_{t+1} &= \boldsymbol{L}_t - \eta(\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star)\boldsymbol{R}_t(\boldsymbol{R}_t^\top\boldsymbol{R}_t)^{-1}, \\ \boldsymbol{R}_{t+1} &= \boldsymbol{R}_t - \eta(\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star)^\top\boldsymbol{L}_t(\boldsymbol{L}_t^\top\boldsymbol{L}_t)^{-1}. \end{aligned} \tag{2.28}$$

To shed light on why `ScaledGD` is robust to ill-conditioning, it is worthwhile to think of `ScaledGD` as a quasi-Newton algorithm: the following proposition (proven in Appendix A.2.1) reveals that `ScaledGD` is equivalent to approximating the Hessian of the loss function in (2.27) by only keeping its diagonal blocks.

Proposition 2. *For the matrix factorization problem (2.27), `ScaledGD` is equivalent to the following update rule*

$$\mathrm{vec}(\boldsymbol{F}_{t+1}) = \mathrm{vec}(\boldsymbol{F}_t) - \eta \begin{bmatrix} \nabla^2_{\boldsymbol{L},\boldsymbol{L}}\mathcal{L}(\boldsymbol{F}_t) & \boldsymbol{0} \\ \boldsymbol{0} & \nabla^2_{\boldsymbol{R},\boldsymbol{R}}\mathcal{L}(\boldsymbol{F}_t) \end{bmatrix}^{-1} \mathrm{vec}(\nabla_{\boldsymbol{F}}\mathcal{L}(\boldsymbol{F}_t)).$$

Here, $\nabla^2_{\boldsymbol{L},\boldsymbol{L}}\mathcal{L}(\boldsymbol{F}_t)$ (resp. $\nabla^2_{\boldsymbol{R},\boldsymbol{R}}\mathcal{L}(\boldsymbol{F}_t)$) denotes the second order derivative w.r.t. $\boldsymbol{L}$ (resp. $\boldsymbol{R}$) at $\boldsymbol{F}_t$.

The following theorem, whose proof can be found in Appendix A.2.2, formally establishes that as long as `ScaledGD` is initialized close to the ground truth, $\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$ will contract at a constant linear rate for the matrix factorization problem.

Theorem 5. *Suppose that the initialization $\boldsymbol{F}_0$ satisfies $\mathrm{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \le 0.1\sigma_r(\boldsymbol{X}_\star)$. If the step size obeys $0 < \eta \le 2/3$, then for all $t \ge 0$, the iterates of the `ScaledGD` method in (2.28) satisfy*

$$\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \le (1-0.7\eta)^t 0.1\sigma_r(\boldsymbol{X}_\star), \quad \text{and} \quad \left\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\right\|_{\mathsf{F}} \le (1-0.7\eta)^t 0.15\sigma_r(\boldsymbol{X}_\star).$$

Comparing to the rate of contraction $(1-1/\kappa)$ of gradient descent for matrix factorization [MLC21, CLC19], Theorem 5 demonstrates that the preconditioners indeed allow better search directions in the local neighborhood of the ground truth, and hence a faster convergence rate.

2.3.2 Proof outline for matrix sensing

It can be seen that the update rule (2.13) of `ScaledGD` in Algorithm 1 closely mimics (2.28) when $\mathcal{A}(\cdot)$ satisfies the RIP. Therefore, leveraging the RIP of $\mathcal{A}(\cdot)$ and Theorem 5, we can establish the following local convergence guarantee of Algorithm 1, which has a weaker requirement on δ_{2r} than the main theorem (cf. Theorem 1).

Lemma 1. *Suppose that $\mathcal{A}(\cdot)$ obeys the $2r$-RIP with $\delta_{2r} \le 0.02$. If the t-th iterate satisfies $\operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \le 0.1\sigma_r(\boldsymbol{X}_\star)$, then $\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\|_{\mathsf{F}} \le 1.5\operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$. In addition, if the step size obeys $0 < \eta \le 2/3$, then the $(t+1)$-th iterate $\boldsymbol{F}_{t+1}$ of the `ScaledGD` method in (2.13) of Algorithm 1 satisfies*

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

It then boils to down to finding a good initialization, for which we have the following lemma on the quality of the spectral initialization.

Lemma 2. *Suppose that $\mathcal{A}(\cdot)$ obeys the $2r$-RIP with a constant δ_{2r}. Then the spectral initialization in (2.12) for low-rank matrix sensing satisfies*

$$\operatorname{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \le 5\delta_{2r}\sqrt{r}\kappa\sigma_r(\boldsymbol{X}_\star).$$

Therefore, as long as δ_{2r} is small enough, say $\delta_{2r} \le 0.02/(\sqrt{r}\kappa)$ as specified in Theorem 1, the initial distance satisfies $\operatorname{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \le 0.1\sigma_r(\boldsymbol{X}_\star)$, allowing us to invoke Lemma 1 recursively. The proof of Theorem 1 is then complete. The proofs of Lemmas 1-2 can be found in Appendix A.3.

2.3.3 Proof outline for robust PCA

As before, we begin with the following local convergence guarantee of Algorithm 2, which has a weaker requirement on α than the main theorem (cf. Theorem 2). The difference with low-rank matrix sensing is that local convergence for robust PCA requires a further incoherence condition on the iterates (cf. (2.29)), where we recall from (2.9) that $\boldsymbol{Q}_t$ is the optimal alignment matrix between

$\boldsymbol{F}_t$ and $\boldsymbol{F}_\star$.

Lemma 3. *Suppose that $\boldsymbol{X}_\star$ is μ-incoherent and $\alpha \le 10^{-4}/(\mu r)$. If the t-th iterate satisfies $\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \le 0.02\sigma_r(\boldsymbol{X}_\star)$ and the incoherence condition*

$$\sqrt{n_1}\left\|(\boldsymbol{L}_t\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty} \vee \sqrt{n_2}\left\|(\boldsymbol{R}_t\boldsymbol{Q}_t^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty} \le \sqrt{\mu r}\sigma_r(\boldsymbol{X}_\star), \tag{2.29}$$

then $\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\|_{\mathsf{F}} \le 1.5\,\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$. In addition, if the step size obeys $0.1 \le \eta \le 2/3$, then the $(t+1)$-th iterate $\boldsymbol{F}_{t+1}$ of the `ScaledGD` method in (2.18) of Algorithm 2 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

$$\sqrt{n_1}\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_{t+1} - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty} \vee \sqrt{n_2}\left\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_{t+1}^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty} \le \sqrt{\mu r}\sigma_r(\boldsymbol{X}_\star).$$

As long as the initialization is close to the ground truth and satisfies the incoherence condition, Lemma 3 ensures that the iterates of `ScaledGD` remain incoherent and converge linearly. This allows us to remove the unnecessary projection step in [YPCC16], whose main objective is to ensure the incoherence of the iterates.

We are left with checking the initial conditions. The following lemma ensures that the spectral initialization in (2.17) is close to the ground truth as long as α is sufficiently small.

Lemma 4. *Suppose that $\boldsymbol{X}_\star$ is μ-incoherent. Then the spectral initialization (2.17) for robust PCA satisfies*

$$\mathrm{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \le 20\alpha\mu r^{3/2}\kappa\sigma_r(\boldsymbol{X}_\star).$$

As a result, setting $\alpha \le 10^{-3}/(\mu r^{3/2}\kappa)$, the spectral initialization satisfies $\mathrm{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \le 0.02\sigma_r(\boldsymbol{X}_\star)$. In addition, we need to make sure that the spectral initialization satisfies the incoherence condition, which is provided in the following lemma.

Lemma 5. *Suppose that $\boldsymbol{X}_\star$ is μ-incoherent and $\alpha \le 0.1/(\mu r\kappa)$, and that $\mathrm{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \le 0.02\sigma_r(\boldsymbol{X}_\star)$. Then the spectral initialization (2.17) satisfies the incoherence condition*

$$\sqrt{n_1}\left\|(\boldsymbol{L}_0\boldsymbol{Q}_0 - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty} \vee \sqrt{n_2}\left\|(\boldsymbol{R}_0\boldsymbol{Q}_0^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty} \le \sqrt{\mu r}\sigma_r(\boldsymbol{X}_\star).$$

Combining Lemmas 3-5 finishes the proof of Theorem 2. The proofs of the the three supporting lemmas can be found in Section A.4.

2.3.4 Proof outline for matrix completion

A key property of the new projection operator. We start with the following lemma that entails a key property of the scaled projection (2.22), which ensures the scaled projection satisfies both non-expansiveness and incoherence under the scaled metric.

Lemma 6. *Suppose that $\boldsymbol{X}_\star$ is μ-incoherent, and $\mathrm{dist}(\widetilde{\boldsymbol{F}}, \boldsymbol{F}_\star) \le \epsilon\sigma_r(\boldsymbol{X}_\star)$ for some $\epsilon < 1$. Set $B \ge (1+\epsilon)\sqrt{\mu r}\sigma_1(\boldsymbol{X}_\star)$, then $\mathcal{P}_B(\widetilde{\boldsymbol{F}})$ satisfies the non-expansiveness*

$$\mathrm{dist}(\mathcal{P}_B(\widetilde{\boldsymbol{F}}), \boldsymbol{F}_\star) \le \mathrm{dist}(\widetilde{\boldsymbol{F}}, \boldsymbol{F}_\star),$$

and the incoherence condition

$$\sqrt{n_1}\|\boldsymbol{L}\boldsymbol{R}^\top\|_{2,\infty} \vee \sqrt{n_2}\|\boldsymbol{R}\boldsymbol{L}^\top\|_{2,\infty} \le B.$$

It is worth noting that the incoherence condition adopts a slightly different form than that of robust PCA, which is more convenient for matrix completion. The next lemma guarantees the fast local convergence of Algorithm 3 as long as the sample complexity is large enough and the parameter B is set properly.

Lemma 7. *Suppose that $\boldsymbol{X}_\star$ is μ-incoherent, and $p \ge C(\mu r\kappa^4 \vee \log(n_1 \vee n_2))\mu r/(n_1 \wedge n_2)$ for some sufficiently large constant C. Set the projection radius as $B = C_B\sqrt{\mu r}\sigma_1(\boldsymbol{X}_\star)$ for some constant $C_B \ge 1.02$. Under an event $\mathcal{E}$ which happens with overwhelming probability (i.e. at least $1 - c_1(n_1 \vee n_2)^{-c_2}$), if the t-th iterate satisfies $\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \le 0.02\sigma_r(\boldsymbol{X}_\star)$, and the incoherence*

condition

$$\sqrt{n_1}\|\boldsymbol{L}_t\boldsymbol{R}_t^\top\|_{2,\infty} \vee \sqrt{n_1}\|\boldsymbol{R}_t\boldsymbol{L}_t^\top\|_{2,\infty} \le B,$$

then $\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\|_{\mathsf{F}} \le 1.5\,\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$. *In addition, if the step size obeys* $0 < \eta \le 2/3$, *then the* $(t+1)$*-th iterate* $\boldsymbol{F}_{t+1}$ *of the* `ScaledPGD` *method in* (2.24) *of Algorithm 3 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

$$\sqrt{n_1}\|\boldsymbol{L}_{t+1}\boldsymbol{R}_{t+1}^\top\|_{2,\infty} \vee \sqrt{n_2}\|\boldsymbol{R}_{t+1}\boldsymbol{L}_{t+1}^\top\|_{2,\infty} \le B.$$

As long as we can find an initialization that is close to the ground truth and satisfies the incoherence condition, Lemma 7 ensures that the iterates of `ScaledPGD` remain incoherent and converge linearly. The follow lemma ensures that such an initialization can be ensured via the spectral method.

Lemma 8. *Suppose that* $\boldsymbol{X}_\star$ *is* μ*-incoherent, then with overwhelming probability, the spectral initialization before projection* $\widetilde{\boldsymbol{F}}_0 \coloneqq \begin{bmatrix} \boldsymbol{U}_0\boldsymbol{\Sigma}_0^{1/2} \\ \boldsymbol{V}_0\boldsymbol{\Sigma}_0^{1/2} \end{bmatrix}$ *in* (2.23) *satisfies*

$$\mathrm{dist}(\widetilde{\boldsymbol{F}}_0, \boldsymbol{F}_\star) \le C_0 \left(\frac{\mu r \log(n_1 \vee n_2)}{p\sqrt{n_1 n_2}} + \sqrt{\frac{\mu r \log(n_1 \vee n_2)}{p(n_1 \wedge n_2)}} \right) 5\sqrt{r}\kappa\sigma_r(\boldsymbol{X}_\star).$$

Therefore, as long as $p \ge C\mu r^2\kappa^2 \log(n_1 \vee n_2)/(n_1 \wedge n_2)$ for some sufficiently large constant C, the initial distance satisfies $\mathrm{dist}(\widetilde{\boldsymbol{F}}_0, \boldsymbol{F}_\star) \le 0.02\sigma_r(\boldsymbol{X}_\star)$. One can then invoke Lemma 6 to see that $\boldsymbol{F}_0 = \mathcal{P}_B(\widetilde{\boldsymbol{F}}_0)$ meets the requirements of Lemma 7 due to the non-expansiveness and incoherence properties of the projection operator. The proofs of the the the supporting lemmas can be found in Section A.5.

2.4 Numerical Experiments

In this section, we provide numerical experiments to corroborate our theoretical findings.

2.4.1 Comparison with vanilla GD

To begin, we compare the iteration complexity of `ScaledGD` with vanilla gradient descent (GD). The update rule of vanilla GD for solving (2.1) is given as

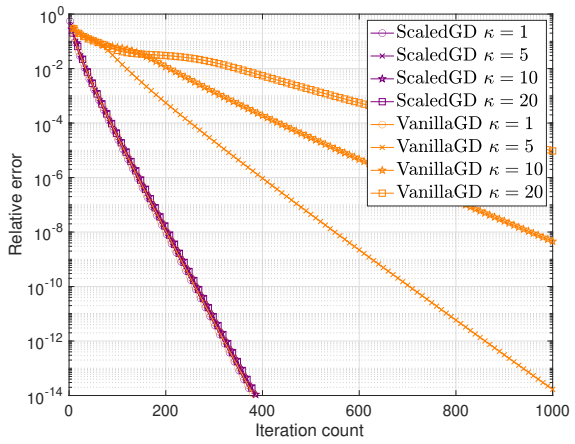
$$\begin{aligned} \boldsymbol{L}_{t+1} &= \boldsymbol{L}_t - \eta_{\texttt{GD}} \nabla_{\boldsymbol{L}} \mathcal{L}(\boldsymbol{L}_t, \boldsymbol{R}_t), \\ \boldsymbol{R}_{t+1} &= \boldsymbol{R}_t - \eta_{\texttt{GD}} \nabla_{\boldsymbol{R}} \mathcal{L}(\boldsymbol{L}_t, \boldsymbol{R}_t), \end{aligned} \tag{2.30}$$

where $\eta_{\texttt{GD}} = \eta/\sigma_1(\boldsymbol{X}_\star)$ stands for the step size for gradient descent. This choice is often recommended by the theory of vanilla GD [TBS+16, YPCC16, MWCC19] and the scaling by $\sigma_1(\boldsymbol{X}_\star)$ is needed for its convergence. For ease of comparison, we fix $\eta = 0.5$ for both `ScaledGD` and vanilla GD (see Figure 2.3 for justifications). Both algorithms start from the same spectral initialization. To avoid notational clutter, we work on square *asymmetric* matrices with $n_1 = n_2 = n$. We consider four low-rank matrix estimation tasks:

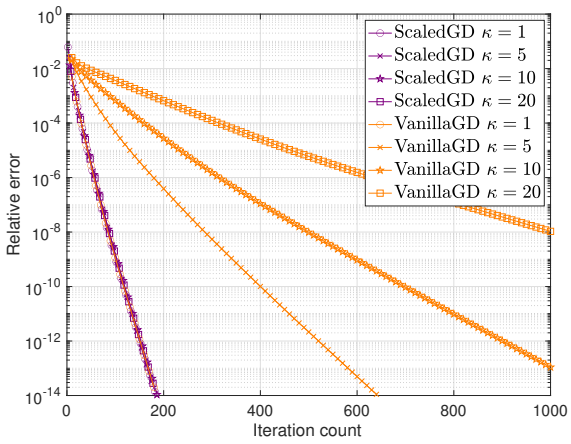
- *Low-rank matrix sensing.* The problem formulation is detailed in Section 2.2.2. Here, we collect $m = 5nr$ measurements in the form of $\boldsymbol{y}_k = \langle \boldsymbol{A}_k, \boldsymbol{X}_\star \rangle + \boldsymbol{w}_k$, in which the measurement matrices $\boldsymbol{A}_k$ are generated with i.i.d. Gaussian entries with zero mean and variance $1/m$, and $\boldsymbol{w}_k \sim \mathcal{N}(0, \sigma_w^2)$ are i.i.d. Gaussian noises.

- *Robust PCA.* The problem formulation is stated in Section 2.2.3. We generate the corruption with a sparse matrix $\boldsymbol{S}_\star \in \mathcal{S}_\alpha$ with $\alpha = 0.1$. More specifically, we generate a matrix with standard Gaussian entries and pass it through $\mathcal{T}_\alpha[\cdot]$ to obtain $\boldsymbol{S}_\star$. The observation is $\boldsymbol{Y} = \boldsymbol{X}_\star + \boldsymbol{S}_\star + \boldsymbol{W}$, where $\boldsymbol{W}_{i,j} \sim \mathcal{N}(0, \sigma_w^2)$ are i.i.d. Gaussian noises.

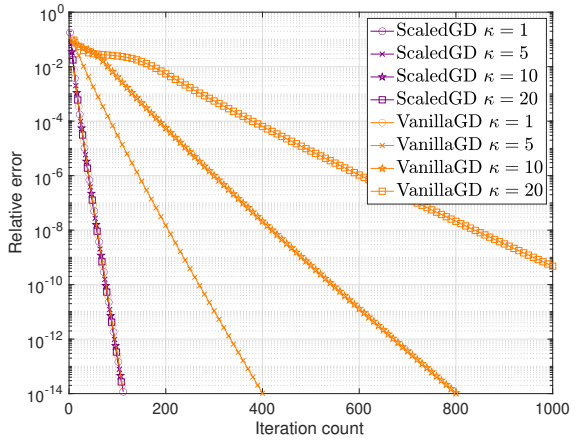
- *Matrix completion.* The problem formulation is stated in Section 2.2.4. We assume random Bernoulli observations, where each entry of $\boldsymbol{X}_\star$ is observed with probability $p = 0.2$ independently. The observation is $\boldsymbol{Y} = \mathcal{P}_\Omega(\boldsymbol{X}_\star + \boldsymbol{W})$, where $\boldsymbol{W}_{i,j} \sim \mathcal{N}(0, \sigma_w^2)$ are i.i.d. Gaussian noises. Moreover,



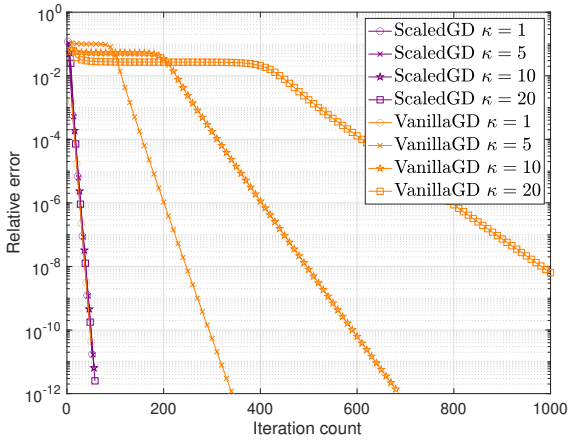
(a) Matrix sensing
$n = 200, r = 10, m = 5nr$



(b) Robust PCA
$n = 1000, r = 10, \alpha = 0.1$



(c) Matrix completion
$n = 1000, r = 10, p = 0.2$

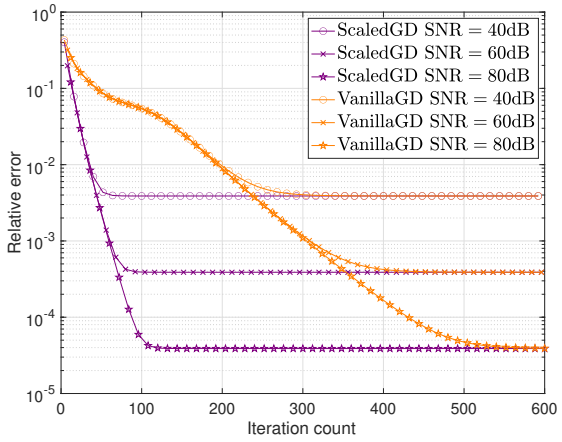


(d) Hankel matrix completion
$n = 1000, r = 10, p = 0.2$

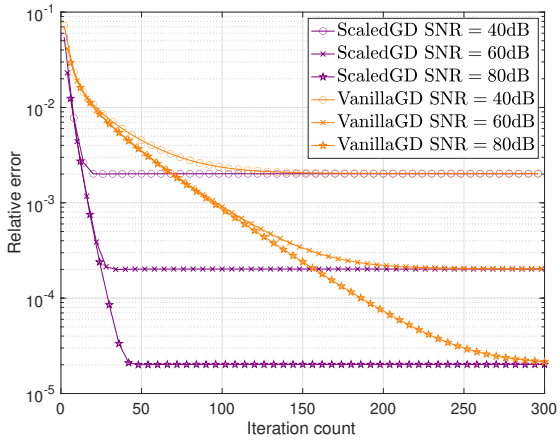
Figure 2.1: The relative errors of `ScaledGD` and vanilla GD with respect to the iteration count under different condition numbers $\kappa = 1, 5, 10, 20$ for (a) matrix sensing, (b) robust PCA, (c) matrix completion, and (d) Hankel matrix completion.

we perform the scaled gradient updates without projections.

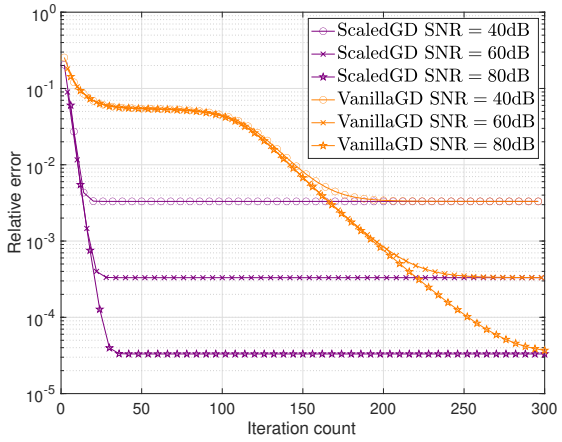
- *Hankel matrix completion.* Briefly speaking, a Hankel matrix shares the same value along each skew-diagonal, and we aim at recovering a low-rank Hankel matrix from observing a few skew-diagonals [CC14, CWW18]. We assume random Bernoulli observations, where each skew-diagonal



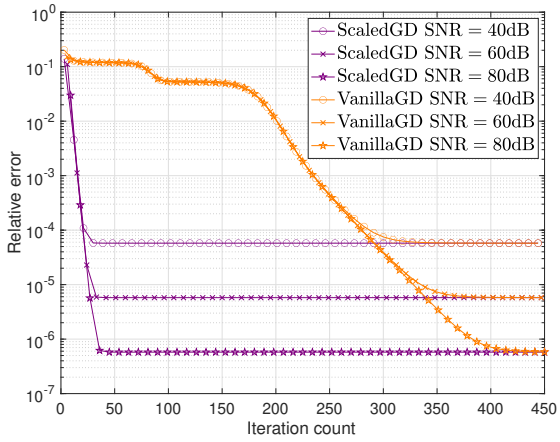
(a) Matrix Sensing
$n = 200, r = 10, m = 5nr$



(b) Robust PCA
$n = 1000, r = 10, \alpha = 0.1$



(c) Matrix completion
$n = 1000, r = 10, p = 0.2$



(d) Hankel matrix completion
$n = 1000, r = 10, p = 0.2$

Figure 2.2: The relative errors of `ScaledGD` and vanilla GD with respect to the iteration count under the condition number $\kappa = 10$ and signal-to-noise ratios SNR $= 40, 60, 80$dB for (a) matrix sensing, (b) robust PCA, (c) matrix completion, and (d) Hankel matrix completion.

of $\boldsymbol{X}_\star$ is observed with probability $p = 0.2$ independently. The loss function is

$$\mathcal{L}(\boldsymbol{L}, \boldsymbol{R}) = \frac{1}{2p} \left\| \mathcal{H}_\Omega(\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{Y}) \right\|_{\mathsf{F}}^2 + \frac{1}{2} \left\| (\mathcal{I} - \mathcal{H})(\boldsymbol{L}\boldsymbol{R}^\top) \right\|_{\mathsf{F}}^2, \tag{2.31}$$

where $\mathcal{I}(\cdot)$ denotes the identity operator, and the Hankel projection is defined as $\mathcal{H}(\boldsymbol{X}) := \sum_{k=1}^{2n-1} \langle \boldsymbol{H}_k, \boldsymbol{X} \rangle \boldsymbol{H}_k$, which maps $\boldsymbol{X}$ to its closest Hankel matrix. Here, the Hankel basis ma-

trix $\boldsymbol{H}_k$ is the $n \times n$ matrix with the entries in the k-th skew diagonal as $\frac{1}{\sqrt{\omega_k}}$, and all other entries as 0, where ω_k is the length of the k-th skew diagonal. Note that $\boldsymbol{X}$ is a Hankel matrix if and only if $(\mathcal{I} - \mathcal{H})(\boldsymbol{X}) = \boldsymbol{0}$. The Hankel projection on the observation index set Ω is defined as $\mathcal{H}_\Omega(\boldsymbol{X}) \coloneqq \sum_{k\in\Omega}\langle \boldsymbol{H}_k, \boldsymbol{X}\rangle \boldsymbol{H}_k$. The observation is $\boldsymbol{Y} = \mathcal{H}_\Omega(\boldsymbol{X}_\star + \boldsymbol{W})$, where $\boldsymbol{W}$ is a Hankel matrix whose entries along each skew-diagonal are i.i.d. Gaussian noises $\mathcal{N}(0, \sigma_w^2)$.

For the first three problems, we generate the ground truth matrix $\boldsymbol{X}_\star \in \mathbb{R}^{n\times n}$ in the following way. We first generate an $n \times r$ matrix with i.i.d. random signs, and take its r left singular vectors as $\boldsymbol{U}_\star$, and similarly for $\boldsymbol{V}_\star$. The singular values are set to be linearly distributed from 1 to $1/\kappa$. The ground truth is then defined as $\boldsymbol{X}_\star = \boldsymbol{U}_\star \boldsymbol{\Sigma}_\star \boldsymbol{V}_\star^\top$ which has the specified condition number κ and rank r. For Hankel matrix completion, we generate $\boldsymbol{X}_\star$ as an $n \times n$ Hankel matrix with entries given as

$$(\boldsymbol{X}_\star)_{i,j} = \sum_{\ell=1}^{r} \frac{\sigma_\ell}{n} e^{2\pi\imath(i+j-2)f_\ell}, \quad i,j = 1,\dots,n,$$

where f_ℓ, $\ell = 1,\dots,r$ are randomly chosen from $1/n, 2/n, \dots, 1$, and σ_ℓ are linearly distributed from 1 to $1/\kappa$. The Vandermonde decomposition lemma tells that $\boldsymbol{X}_\star$ has rank r and singular values σ_ℓ, $\ell = 1,\dots,r$.

We first illustrate the convergence performance under noise-free observations, i.e. $\sigma_w = 0$. We plot the relative reconstruction error $\|\boldsymbol{X}_t - \boldsymbol{X}_\star\|_{\mathsf{F}}/\|\boldsymbol{X}_\star\|_{\mathsf{F}}$ with respect to the iteration count t in Figure 2.1 for the four problems under different condition numbers $\kappa = 1, 5, 10, 20$. For all these models, we can see that `ScaledGD` has a convergence rate independent of κ, with all curves almost overlay on each other. Under good conditioning $\kappa = 1$, `ScaledGD` converges at the same rate as vanilla GD; under ill conditioning, i.e. when κ is large, `ScaledGD` converges much faster than vanilla GD and leads to significant computational savings.

We next move to demonstrate that `ScaledGD` is robust to small additive noises. Denote the signal-to-noise ratio as $\mathrm{SNR} \coloneqq 10\log_{10}\frac{\|\boldsymbol{X}_\star\|_{\mathsf{F}}^2}{n^2\sigma_w^2}$ in dB. We plot the reconstruction error $\|\boldsymbol{X}_t - \boldsymbol{X}_\star\|_{\mathsf{F}}/\|\boldsymbol{X}_\star\|_{\mathsf{F}}$ with respect to the iteration count t in Figure 2.2 under the condition number $\kappa = 10$ and various $\mathrm{SNR} = 40, 60, 80\mathrm{dB}$. We can see that `ScaledGD` and vanilla GD achieve the same

statistical error eventually, but `ScaledGD` converges much faster. In addition, the convergence speeds are not influenced by the noise levels.

Careful readers might wonder how sensitivity our comparisons are with respect to the choice of step sizes. To address this, we illustrate the convergence speeds of both `ScaledGD` and vanilla GD under different step sizes η for matrix completion (under the same setting as Figure 2.1 (c)), where similar plots can be obtained for other problems as well. We run both algorithms for at most 80 iterations, and terminate if the relative error exceeds 10^2 (which happens if the step size is too large and the algorithm diverges). Figure 2.3 plots the relative error with respect to the step size η for both algorithms, where we can see that `ScaledGD` outperforms vanilla GD over a large range of step sizes, even under optimized values for performance. Hence, our choice of $\eta = 0.5$ in previous experiments renders a typical comparison between `ScaledGD` and vanilla GD.

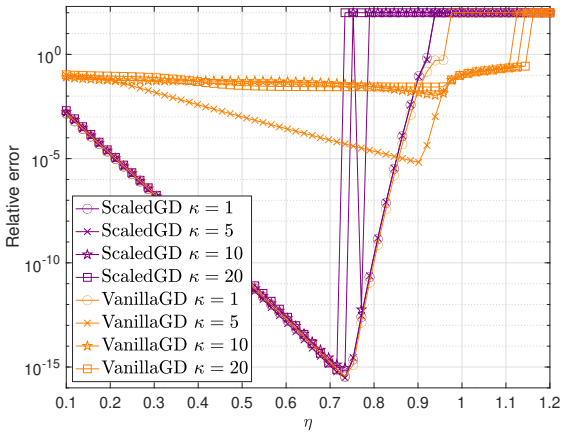
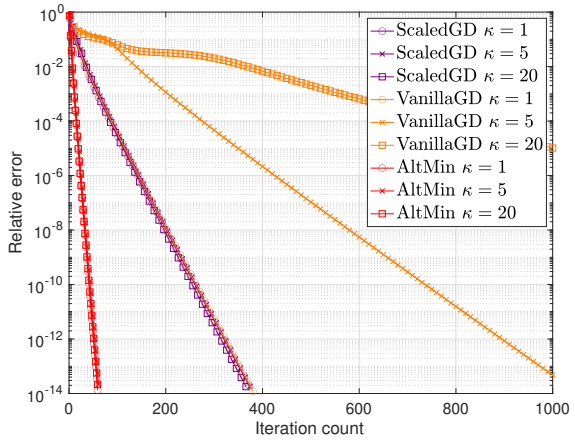


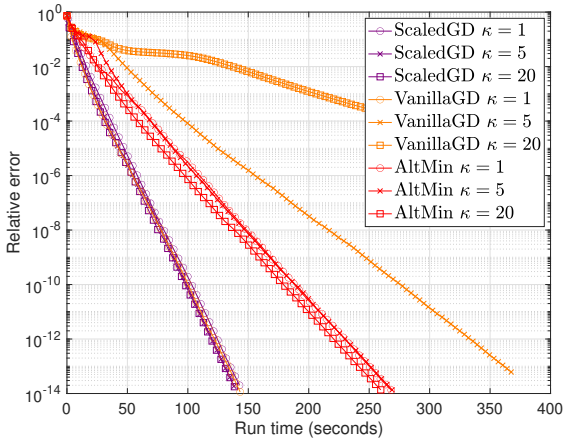
Figure 2.3: The relative errors of `ScaledGD` and vanilla GD after 80 iterations with respect to different step sizes η from 0.1 to 1.2, for matrix completion with $n = 1000, r = 10, p = 0.2$.

2.4.2 Run time comparisons

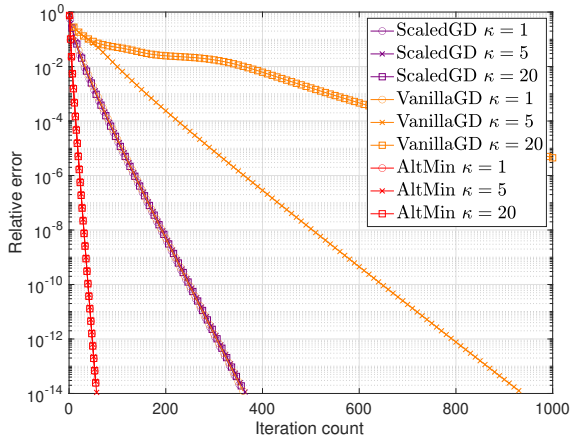
We now compare the run time of `ScaledGD` with vanilla GD and alternating minimization (`AltMin`) [JNS13]. Specifically, for matrix sensing, alternating minimization (`AltMinSense`) updates the fac-



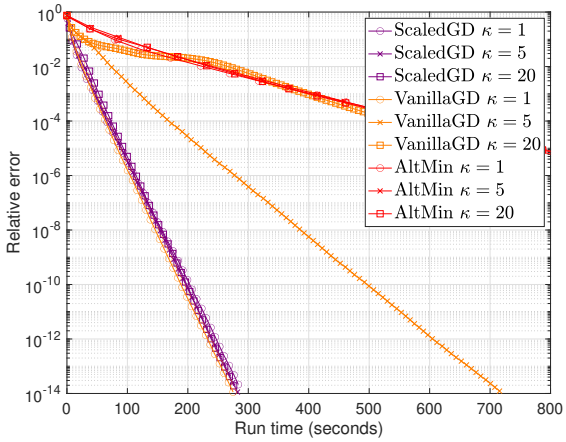
(a) iteration count with $r = 10$



(b) run time with $r = 10$



(c) iteration count with $r = 20$

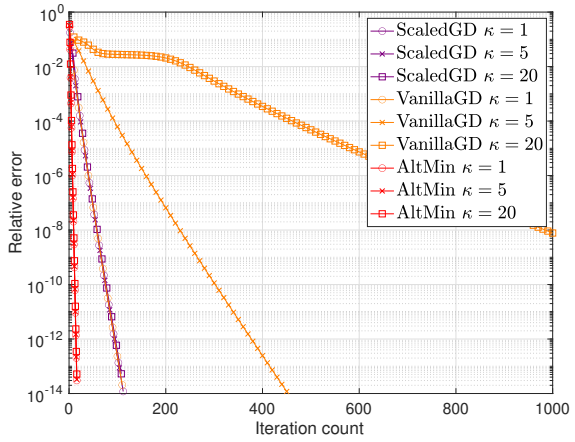


(d) run time with $r = 20$

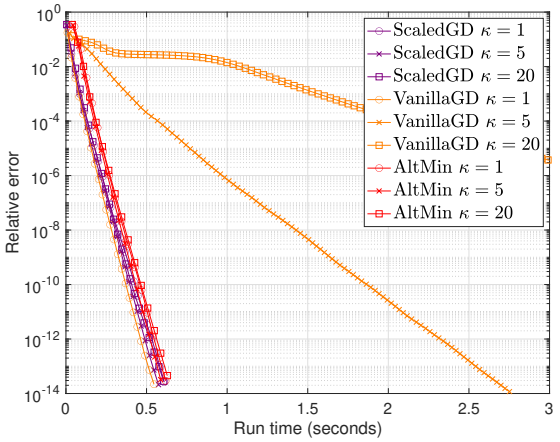
Figure 2.4: The relative errors of `ScaledGD`, vanilla GD and `AltMin` with respect to the iteration count and run time (in seconds) under different condition numbers $\kappa = 1, 5, 20$ for matrix sensing with $n = 200$, and $m = 5nr$. (a, b): $r = 10$; (c, d): $r = 20$.

tors alternatively as

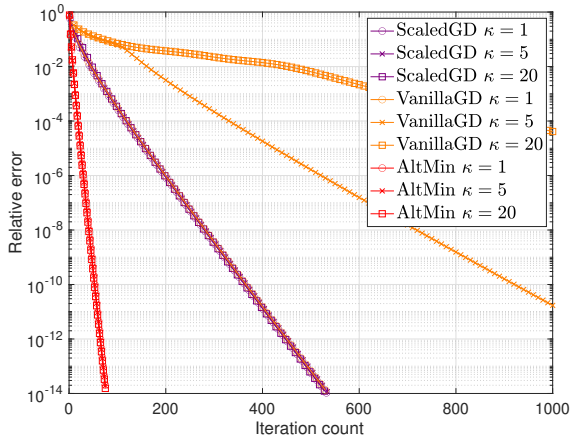
$$\boldsymbol{L}_{t+1} = \operatorname*{argmin}_{\boldsymbol{L}} \left\| \mathcal{A}(\boldsymbol{L}\boldsymbol{R}_t^\top) - \boldsymbol{y} \right\|_2^2,$$
$$\boldsymbol{R}_{t+1} = \operatorname*{argmin}_{\boldsymbol{R}} \left\| \mathcal{A}(\boldsymbol{L}_{t+1}\boldsymbol{R}^\top) - \boldsymbol{y} \right\|_2^2,$$



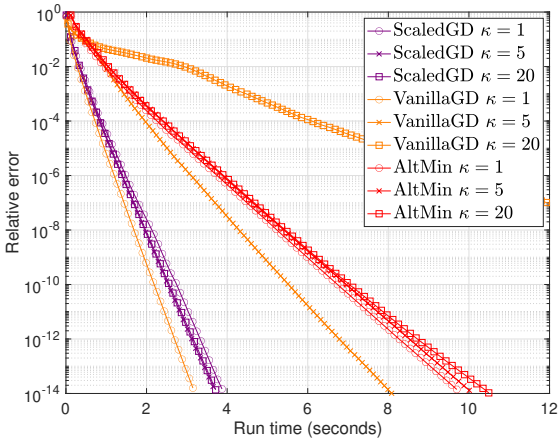
(a) iteration count with $r = 10$



(b) run time with $r = 10$



(c) iteration count with $r = 50$



(d) run time with $r = 50$

Figure 2.5: The relative errors of `ScaledGD`, vanilla GD and `AltMin` with respect to the iteration count and run time (in seconds) under different condition numbers $\kappa = 1, 5, 20$ for matrix completion with $n = 1000$, and $p = 0.2$. (a, b): $r = 10$; (c, d): $r = 50$.

which corresponds to solving two least-squares problems. For matrix completion, the update rule of alternating minimization proceeds as

$$\boldsymbol{L}_{t+1} = \operatorname*{argmin}_{\boldsymbol{L}} \left\| \mathcal{P}_\Omega(\boldsymbol{L}\boldsymbol{R}_t^\top - \boldsymbol{Y}) \right\|_2^2,$$
$$\boldsymbol{R}_{t+1} = \operatorname*{argmin}_{\boldsymbol{R}} \left\| \mathcal{P}_\Omega(\boldsymbol{L}_{t+1}\boldsymbol{R}^\top - \boldsymbol{Y}) \right\|_2^2,$$

which can be implemented more efficiently since each row of $\boldsymbol{L}$ (resp. $\boldsymbol{R}$) can be updated independently via solving a much smaller least-squares problem due to the decomposable structure of the objective function. It is worth noting that, to the best of our knowledge, this most natural variant of alternating minimization for matrix completion still eludes from a provable performance guarantee, nonetheless, we choose it to compare against due to its popularity and excellent empirical performance.

Figure 2.4 plots the relative errors of `ScaledGD`, vanilla GD and alternating minimization (`AltMin`) with respect to the iteration count and run time (in seconds) under different condition numbers $\kappa = 1, 5, 20$; and similarly, Figure 2.5 plots the corresponding results for matrix completion. It can be seen that, both `ScaledGD` and `AltMin` admit a convergence rate that is independent of the condition number, where the per-iteration complexity of `AltMin` is much higher than that of `ScaledGD`. As expected, the run time of `ScaledGD` only adds a minimal overhead to vanilla GD while being much more robust to ill-conditioning. Noteworthily, `AltMin` takes much more time and becomes significantly slower than `ScaledGD` when the rank r is larger. Nonetheless, we emphasize that since the run time is impacted by many factors in terms of problem parameters as well as implementation details, our purpose is to demonstrate the competitive performance of `ScaledGD` over alternatives, rather than claiming it as the state-of-the-art.

2.5 Conclusions

This chapter proposes scaled gradient descent (`ScaledGD`) for factored low-rank matrix estimation, which maintains the low per-iteration computational complexity of vanilla gradient descent, but offers significant speed-up in terms of the convergence rate with respect to the condition number κ of the low-rank matrix. In particular, we rigorously establish that for low-rank matrix sensing, robust PCA, and matrix completion, to reach ϵ-accuracy, `ScaledGD` only takes $O(\log(1/\epsilon))$ iterations without the dependency on the condition number when initialized via the spectral method, under standard assumptions. The key to our analysis is the introduction of a new distance metric that takes into account the preconditioning and unbalancedness of the low-rank factors, and we have developed new tools to analyze the trajectory of `ScaledGD` under this new metric. This work opens

up many venues for future research, as we discuss below.

- *Improved analysis.* In this chapter, we have focused on establishing the fast local convergence rate. It is interesting to study if the theory developed herein can be further strengthened in terms of sample complexity and the size of basin of attraction. For matrix completion, it will be interesting to see if a similar guarantee continues to hold in the absence of the projection, which will generalize recent works [MWCC19, CLL20] that successfully removed these projections for vanilla gradient descent.

- *Other low-rank recovery problems.* Besides the problems studied herein, there are many other applications involving the recovery of an ill-conditioned low-rank matrix, such as robust PCA with missing data, quadratic sampling, and so on. It is of interest to establish fast convergence rates of `ScaledGD` that are independent of the condition number for these problems as well. In addition, it is worthwhile to explore if a similar preconditioning trick can be useful to problems beyond low-rank matrix estimation.

- *Acceleration schemes?* As it is evident from our analysis of the general loss case, `ScaledGD` may still converge slowly when the loss function is ill-conditioned over low-rank matrices, i.e. κ_f is large. In this case, it might be of interest to combine techniques such as momentum [KC12] from the optimization literature to further accelerate the convergence.

Chapter 3

Robust Low-rank Matrix Estimation

3.1 Introduction

Many problems in data science can be treated as estimating a low-rank matrix $\boldsymbol{X}_\star \in \mathbb{R}^{n_1 \times n_2}$ from highly incomplete, sometimes even corrupted, observations $\boldsymbol{y} = \{y_i\}_{i=1}^m$ given by

$$y_i \approx \mathcal{A}_i(\boldsymbol{X}_\star), \qquad 1 \le i \le m. \tag{3.1}$$

Here, $\mathcal{A}(\cdot) = \{\mathcal{A}_i(\cdot)\}_{i=1}^m : \mathbb{R}^{n_1 \times n_2} \mapsto \mathbb{R}^m$ is the observation operator that models the measurement process. Instead of operating in the full matrix space, i.e. $\mathbb{R}^{n_1 \times n_2}$, a memory-efficient approach is to resort to low-rank matrix factorization, by writing $\boldsymbol{X}_\star = \boldsymbol{L}_\star \boldsymbol{R}_\star^\top$, if the rank r of $\boldsymbol{X}_\star$ is known *a priori*, where $\boldsymbol{L}_\star \in \mathbb{R}^{n_1 \times r}$ and $\boldsymbol{R}_\star \in \mathbb{R}^{n_2 \times r}$ are of a size that is proportional to the degrees of freedom of the low-rank matrix. Furthermore, the low-rank factors can be found by optimizing a smooth loss function, such as the residual sum of squares

$$\underset{\boldsymbol{L} \in \mathbb{R}^{n_1 \times r}, \boldsymbol{R} \in \mathbb{R}^{n_2 \times r}}{\text{minimize}} \quad \sum_{i=1}^m \Big(\mathcal{A}_i(\boldsymbol{L}\boldsymbol{R}^\top) - y_i \Big)^2, \tag{3.2}$$

using first-order methods (e.g. gradient descent). While tremendous progress has been made in recent years [CLC19], applying vanilla gradient descent to the above smooth formulation suffers from ill-conditioning originated from both the measurement operator and the underlying low-rank matrix $\boldsymbol{X}_\star$, which preclude a desirable computational efficiency from classical optimization principles.

3.1.1 Main contributions

In this chapter, we propose to minimize the following *nonsmooth and nonconvex* loss function known as the *least absolute deviations*, which measures the residual sum of absolute errors

$$\underset{\boldsymbol{L}\in\mathbb{R}^{n_1\times r},\boldsymbol{R}\in\mathbb{R}^{n_2\times r}}{\text{minimize}} \quad f(\boldsymbol{L}\boldsymbol{R}^\top) \coloneqq \sum_{i=1}^{m} \left|\mathcal{A}_i(\boldsymbol{L}\boldsymbol{R}^\top) - y_i\right|, \tag{3.3}$$

via a *scaled subgradient method*:

$$\begin{aligned} \boldsymbol{L}_{t+1} &\coloneqq \boldsymbol{L}_t - \eta_t \boldsymbol{S}_t \boldsymbol{R}_t (\boldsymbol{R}_t^\top \boldsymbol{R}_t)^{-1}, \\ \boldsymbol{R}_{t+1} &\coloneqq \boldsymbol{R}_t - \eta_t \boldsymbol{S}_t^\top \boldsymbol{L}_t (\boldsymbol{L}_t^\top \boldsymbol{L}_t)^{-1}. \end{aligned} \tag{3.4}$$

Here, $\boldsymbol{S}_t \in \partial f(\boldsymbol{L}_t\boldsymbol{R}_t^\top)$ is a subgradient of $f(\boldsymbol{X}) \coloneqq \sum_{i=1}^m |\mathcal{A}_i(\boldsymbol{X}) - y_i|$ at $\boldsymbol{L}_t\boldsymbol{R}_t^\top$, and $\eta_t > 0$ is a sequence of carefully-chosen stepsizes. Compared with vanilla subgradient methods, our new method (3.4) scales or preconditions the search directions $\boldsymbol{S}_t\boldsymbol{R}_t$ and $\boldsymbol{S}_t^\top\boldsymbol{L}_t$ by $(\boldsymbol{R}_t^\top\boldsymbol{R}_t)^{-1}$ and $(\boldsymbol{L}_t^\top\boldsymbol{L}_t)^{-1}$, respectively.[1] As explained in Chapter 2 where a similar preconditioning trick was employed for smooth formulations, the scaled subgradient enables better search directions and therefore larger stepsizes. Our main results can be summarized as follows:

- Under general geometric assumptions on $f(\cdot)$ such as restricted rank-r Lipschitz continuity and sharpness conditions, we demonstrate that the convergence rate of scaled subgradient methods using both Polyak's and geometrically decaying stepsizes is *independent* of the condition number of $\boldsymbol{X}_\star$.

- Instantiating our theory under the mixed-norm restricted isometry property (RIP) of the measurement operator, we demonstrate state-of-the-art computational guarantees for low-rank matrix sensing and quadratic sampling even when the observations are noisy and corrupted by outliers. This leads to improvements over the computational complexity of scaled gradient methods in Chapter 2 for heavy-tailed measurement ensembles, as well as of vanilla subgradient methods in [CCD+21]. Table 3.1 provides a detailed comparison of the local iteration complexities of the

[1] Under appropriate conditions, the inverse matrices always exist; in practice, one can use the pseudo-inverse matrices to avoid numerical instabilities.

	matrix sensing		quadratic sampling	
Algorithms	without corruptions	with corruptions	without corruptions	with corruptions
`GD` [TBS+16, LMCC21]	$\kappa \log \frac{1}{\epsilon}$	N/A	$r^2\kappa^2 \log \frac{1}{\epsilon}$	N/A
`ScaledGD` (Chapter 2)	$\log \frac{1}{\epsilon}$	N/A	$\mathrm{poly}(n) \log \frac{1}{\epsilon}$	N/A
`SM` [CCD+21, LZMCSV20]	$\kappa \log \frac{1}{\epsilon}$	$\frac{\kappa}{(1-2p_s)^2} \log \frac{1}{\epsilon}$	$r\kappa \log \frac{1}{\epsilon}$	$\frac{r\kappa}{(1-2p_s)^2} \log \frac{1}{\epsilon}$
`ScaledSM` (this Chapter)	$\log \frac{1}{\epsilon}$	$\frac{1}{(1-2p_s)^2} \log \frac{1}{\epsilon}$	$r \log \frac{1}{\epsilon}$	$\frac{r}{(1-2p_s)^2} \log \frac{1}{\epsilon}$

Table 3.1: Local iteration complexities of the proposed scaled subgradient method (`ScaledSM`) in comparison with prior algorithms for matrix sensing and quadratic sampling. `ScaledSM` outperforms the vanilla subgradient method (`SM`) by a factor of κ in both problems, while outperforms scaled gradient descent (`ScaledGD`), and `GD` with additional robustness guarantees. Here, $n = \max\{n_1, n_2\}$, r is the rank, κ is the condition number of $\boldsymbol{X}_\star$, and $0 \le p_s < 1/2$ is the fraction of outliers. We say that the output $\boldsymbol{X}$ of an algorithm reaches ϵ-accuracy, if it satisfies $\|\boldsymbol{X} - \boldsymbol{X}_\star\|_{\mathsf{F}} \le \epsilon \sigma_r(\boldsymbol{X}_\star)$, where $\sigma_r(\boldsymbol{X}_\star)$ denotes the r-th largest singular value of $\boldsymbol{X}_\star$.

proposed scaled subgradient method in comparison with these prior algorithms, highlighting its robustness to heavy-tailed observations, outliers, as well as a large condition number of the true matrix $\boldsymbol{X}_\star$.

Our work leverages exciting advances in nonsmooth optimization [CCD+21] and scaled first-order methods in Chapter 2 for low-rank matrix recovery. Our arguments are concise, which avoid the need of sophisticated trajectory-dependent analysis as have been used in [MWCC19, LMCC21] to achieve rapid and robust convergence guarantees.

3.1.2 Related work

Low-rank matrix recovery has been a target of intense interest in the last decade; we invite the readers to [DR16, CC18, CLC19] for recent overviews, and limit our discussions to the most relevant literature in the sequel.

Nonsmooth formulations for low-rank matrix recovery. Nonsmooth objective functions, such as the least absolute deviations, have been adopted earlier in both convex and nonconvex formulations of low-rank matrix recovery, including phase retrieval [Han17, DDP20, QZEW19, ZZLC17, DR19], blind deconvolution [Día19], quadratic sampling [LSC17, CL16, CCD+21, BL20], low-rank

matrix sensing [CCD+21, Li13, WGMM13, LZMCSV20], robust synchronization [WS13], to name a few. Our work is most closely related to and generalizes the vanilla subgradient method in [CCD+21], by establishing novel performance guarantees of *scaled* subgradient methods for robust low-rank matrix recovery.

Scaled first-order methods for low-rank matrix recovery. Variants of the scaled gradient methods are proposed in [MAS12, TW16, TMC21a] for minimizing the least-squares formulation (3.2), where strong statistical and computational complexities are first established in Chapter 2. To the best of our knowledge, this is the first work that provides rigorous statistical and computational guarantees for scaled subgradient methods for addressing nonsmooth formulations. When it comes to problems with heavy-tail observations such as quadratic sampling, while it is possible to establish faster convergence rates of vanilla gradient descent over the smooth least-squares loss function through a tailored analysis [MWCC19, LMCC21] via leave-one-out arguments, it is unclear if similar analyses are viable for scaled gradient methods (`ScaledGD`) in Chapter 2. Unfortunately, a direct application of the performance guarantee of `ScaledGD` on minimizing the smooth least-squares loss function leads to a much slower rate in terms of the problem dimension (see Table 3.1) for quadratic sampling. In contrast, our analysis for scaled subgradient methods yields strong guarantees in a more straightforward manner since the nonsmooth loss function has much better geometric properties [CCD+21].

Robust low-rank matrix recovery via nonconvex optimization. A pleasant side benefit of nonsmooth formulations is the added robustness to adversarial outliers under a simple algorithm design – the low-rank factors are updated essentially in the same manner regardless of the presence of outliers. In comparison, other nonconvex methods based on smooth formulations often need to introduce some special treatments to mitigate outliers before updating the low-rank factors, e.g. truncation or thresholding [ZCL16, LCZL20, LZMCSV20], which can be cumbersome to tune properly.

Condition number independent rate of convergence. It is well-known that first-order methods such as gradient descent exhibit poor scaling with respect to the condition number of the low-rank matrix. Possible remedies include alternating least-squares in the factored space [JNS13, HW14], or spectral methods over the matrix space [JMD10]. However, these approaches either require the inversion of a large matrix or a higher memory footprint, compared with the scaled first-order methods adopted herein.

3.1.3 Chapter organization

The rest of this chapter is organized as follows. Section 3.2 describes the proposed scaled subgradient method and its connections to existing methods. Section 3.3 provides the theoretical guarantees for the scaled subgradient method in terms of both statistical and computational complexities, which are then instantiated to robust low-rank matrix sensing and quadratic sampling. Section 3.4 illustrates the superior empirical performance of the proposed method. Finally, we conclude in Section 3.5. The proofs are deferred to the appendix.

3.2 Problem Formulation and Algorithms

In this section, we formulate the low-rank matrix recovery problem, followed by a detailed description of the proposed scaled subgradient method.

3.2.1 Problem formulation

Let $\boldsymbol{X}_\star \in \mathbb{R}^{n_1 \times n_2}$ be the ground truth rank-r matrix, whose compact singular value decomposition (SVD) is given by

$$\boldsymbol{X}_\star = \boldsymbol{U}_\star \boldsymbol{\Sigma}_\star \boldsymbol{V}_\star^\top, \tag{3.5}$$

where $\boldsymbol{U}_\star \in \mathbb{R}^{n_1 \times r}$ and $\boldsymbol{V}_\star \in \mathbb{R}^{n_2 \times r}$ are composed of r left and right singular vectors, respectively, and $\boldsymbol{\Sigma}_\star \in \mathbb{R}^{r \times r}$ is a diagonal matrix consisting of r singular values of $\boldsymbol{X}_\star$ organized in a non-increasing order, i.e. $\sigma_1(\boldsymbol{X}_\star) \geq \cdots \geq \sigma_r(\boldsymbol{X}_\star) > 0$. The condition number of $\boldsymbol{X}_\star$ is thus defined

as

$$\kappa \coloneqq \sigma_1(\boldsymbol{X}_\star)/\sigma_r(\boldsymbol{X}_\star). \tag{3.6}$$

Without loss of generality, we define the ground truth low-rank factors as

$$\boldsymbol{L}_\star \coloneqq \boldsymbol{U}_\star \boldsymbol{\Sigma}_\star^{1/2}, \quad \text{and} \quad \boldsymbol{R}_\star \coloneqq \boldsymbol{V}_\star \boldsymbol{\Sigma}_\star^{1/2}, \tag{3.7}$$

so that $\boldsymbol{X}_\star = \boldsymbol{L}_\star \boldsymbol{R}_\star^\top$. Moreover, we denote the ground truth stacked factor matrix as

$$\boldsymbol{F}_\star \coloneqq [\boldsymbol{L}_\star^\top, \boldsymbol{R}_\star^\top]^\top \in \mathbb{R}^{(n_1+n_2)\times r}. \tag{3.8}$$

Assume that we have access to a number of observations $\boldsymbol{y} = \{y_i\}_{i=1}^m$ of $\boldsymbol{X}_\star$, given as

$$y_i = \mathcal{A}_i(\boldsymbol{X}_\star) + w_i + s_i, \qquad 1 \le i \le m, \tag{3.9}$$

or equivalently,

$$\boldsymbol{y} = \mathcal{A}(\boldsymbol{X}_\star) + \boldsymbol{w} + \boldsymbol{s}, \tag{3.10}$$

where $\mathcal{A}(\boldsymbol{X}_\star) = \{\mathcal{A}_i(\boldsymbol{X}_\star)\}_{i=1}^m$ is the measurement ensemble, $\boldsymbol{w} = \{w_i\}_{i=1}^m$ denotes the bounded noise, and $\boldsymbol{s} = \{s_i\}_{i=1}^m$ models arbitrary corruptions. The goal of low-rank matrix recovery is to reconstruct $\boldsymbol{X}_\star$ from the noisy and corrupted observations $\boldsymbol{y}$ in a statistically and computationally efficient manner.

3.2.2 Scaled subgradient method

Consider the following nonsmooth and nonconvex optimization problem over the factors

$$\underset{\boldsymbol{L}\in\mathbb{R}^{n_1\times r}, \boldsymbol{R}\in\mathbb{R}^{n_2\times r}}{\text{minimize}} \quad \mathcal{L}(\boldsymbol{L},\boldsymbol{R}) \coloneqq f(\boldsymbol{L}\boldsymbol{R}^\top), \tag{3.11}$$

where $f(\cdot)$ is a nonsmooth surrogate of the observation residuals. Of particular interest is the residual sum of absolute errors, defined as

$$f(\boldsymbol{X}) \coloneqq \|\mathcal{A}(\boldsymbol{X}) - \boldsymbol{y}\|_1. \tag{3.12}$$

Correspondingly, the minimizer is called the least absolute deviations (LAD) solution.

Let us denote the stacked factor matrix in the t-th iterate as $\boldsymbol{F}_t \coloneqq [\boldsymbol{L}_t^\top, \boldsymbol{R}_t^\top]^\top$. Given an initialization $\boldsymbol{F}_0 = [\boldsymbol{L}_0^\top, \boldsymbol{R}_0^\top]^\top$, the proposed scaled subgradient method (`ScaledSM`) proceeds as

$$\begin{aligned} \boldsymbol{L}_{t+1} &\coloneqq \boldsymbol{L}_t - \eta_t \boldsymbol{S}_t \boldsymbol{R}_t (\boldsymbol{R}_t^\top \boldsymbol{R}_t)^{-1}, \\ \boldsymbol{R}_{t+1} &\coloneqq \boldsymbol{R}_t - \eta_t \boldsymbol{S}_t^\top \boldsymbol{L}_t (\boldsymbol{L}_t^\top \boldsymbol{L}_t)^{-1}, \end{aligned} \tag{3.13}$$

where $\boldsymbol{S}_t \in \partial f(\boldsymbol{L}_t \boldsymbol{R}_t^\top)$ is a subgradient of $f(\cdot)$ at $\boldsymbol{L}_t \boldsymbol{R}_t^\top$ (and hence $\boldsymbol{S}_t \boldsymbol{R}_t \in \partial_{\boldsymbol{L}} \mathcal{L}(\boldsymbol{L}_t, \boldsymbol{R}_t)$ and $\boldsymbol{S}_t^\top \boldsymbol{L}_t \in \partial_{\boldsymbol{R}} \mathcal{L}(\boldsymbol{L}_t, \boldsymbol{R}_t)$), and $\eta_t > 0$ is some properly selected stepsize, which we discuss next.

Stepsize schedules. We consider the following two choices of stepsize schedules:

- If we know the optimal value $f(\boldsymbol{X}_\star)$, we can invoke the following Polyak's stepsize, given by

$$\eta_t^{\mathsf{P}} \coloneqq \frac{f(\boldsymbol{L}_t \boldsymbol{R}_t^\top) - f(\boldsymbol{X}_\star)}{\|\boldsymbol{S}_t \boldsymbol{R}_t (\boldsymbol{R}_t^\top \boldsymbol{R}_t)^{-1/2}\|_{\mathsf{F}}^2 + \|\boldsymbol{S}_t^\top \boldsymbol{L}_t (\boldsymbol{L}_t^\top \boldsymbol{L}_t)^{-1/2}\|_{\mathsf{F}}^2}, \tag{3.14}$$

  where the denominator is the squared norm of the subgradient under a scaled metric concerted with the preconditioners. This schedule is implementable, for example, when the observations are noise-free, leading to $f(\boldsymbol{X}_\star) = 0$. However, when the observations are noisy and corrupted, it is not viable to know $f(\boldsymbol{X}_\star)$ beforehand.

- In general, we can apply the geometrically decaying stepsize originally introduced in [Gof77], given by

$$\eta_t^{\mathsf{G}} \coloneqq \frac{\lambda q^t}{\sqrt{\|\boldsymbol{S}_t \boldsymbol{R}_t (\boldsymbol{R}_t^\top \boldsymbol{R}_t)^{-1/2}\|_{\mathsf{F}}^2 + \|\boldsymbol{S}_t^\top \boldsymbol{L}_t (\boldsymbol{L}_t^\top \boldsymbol{L}_t)^{-1/2}\|_{\mathsf{F}}^2}}, \tag{3.15}$$

where the denominator is similarly scaled as (3.14), and $\lambda > 0$ and $q \in (0, 1)$ are some parameters to be specified. This choice is broadly applicable when dealing with noisy and corrupted observations.

Compared with the vanilla subgradient method, which proceeds according to

$$\begin{aligned} \boldsymbol{L}_{t+1} &\coloneqq \boldsymbol{L}_t - \eta_t \boldsymbol{S}_t \boldsymbol{R}_t, \\ \boldsymbol{R}_{t+1} &\coloneqq \boldsymbol{R}_t - \eta_t \boldsymbol{S}_t^\top \boldsymbol{L}_t, \end{aligned} \tag{3.16}$$

the update rule (3.13) scales the subgradient $\boldsymbol{S}_t \boldsymbol{R}_t$ and $\boldsymbol{S}_t^\top \boldsymbol{L}_t$ by $(\boldsymbol{R}_t^\top \boldsymbol{R}_t)^{-1}$ and $(\boldsymbol{L}_t^\top \boldsymbol{L}_t)^{-1}$, respectively; see Chapter 2 for its counterpart in smooth problems. An important highlight of the scaled subgradient method is that the update rule is covariant with respect to the ambiguity of low-rank matrix factorization. To see this, imagine that we modify the t-th updates as

$$\widetilde{\boldsymbol{L}}_t \coloneqq \boldsymbol{L}_t \boldsymbol{Q}, \qquad \widetilde{\boldsymbol{R}}_t \coloneqq \boldsymbol{R}_t \boldsymbol{Q}^{-\top} \tag{3.17}$$

for some invertible matrix $\boldsymbol{Q} \in \mathrm{GL}(r)$. It is easy to check:

(i) both the Polyak's stepsize (3.14) and the geometrically decaying stepsize (3.15) do not change, since

$$\|\boldsymbol{S}_t \boldsymbol{R}_t (\boldsymbol{R}_t^\top \boldsymbol{R}_t)^{-1/2}\|_{\mathsf{F}}^2 = \langle \boldsymbol{S}_t, \boldsymbol{S}_t \boldsymbol{R}_t (\boldsymbol{R}_t^\top \boldsymbol{R}_t)^{-1} \boldsymbol{R}_t^\top \rangle = \langle \boldsymbol{S}_t, \boldsymbol{S}_t \widetilde{\boldsymbol{R}}_t (\widetilde{\boldsymbol{R}}_t^\top \widetilde{\boldsymbol{R}}_t)^{-1} \widetilde{\boldsymbol{R}}_t^\top \rangle = \|\boldsymbol{S}_t \widetilde{\boldsymbol{R}}_t (\widetilde{\boldsymbol{R}}_t^\top \widetilde{\boldsymbol{R}}_t)^{-1/2}\|_{\mathsf{F}}^2,$$

which holds similarly for $\|\boldsymbol{S}_t^\top \boldsymbol{L}_t (\boldsymbol{L}_t^\top \boldsymbol{L}_t)^{-1/2}\|_{\mathsf{F}}^2$;

(ii) The next $(t+1)$-th iterate can be written as

$$\widetilde{\boldsymbol{L}}_{t+1} = \widetilde{\boldsymbol{L}}_t - \eta_t \boldsymbol{S}_t \widetilde{\boldsymbol{R}}_t (\widetilde{\boldsymbol{R}}_t^\top \widetilde{\boldsymbol{R}}_t)^{-1} = \Big[\boldsymbol{L}_t - \eta_t \boldsymbol{S}_t \boldsymbol{R}_t (\boldsymbol{R}_t^\top \boldsymbol{R}_t)^{-1}\Big] \boldsymbol{Q} = \boldsymbol{L}_{t+1} \boldsymbol{Q},$$

and similarly $\widetilde{\boldsymbol{R}}_{t+1} = \boldsymbol{R}_{t+1} \boldsymbol{Q}^{-\top}$. Therefore, all the iterates are covariant with respect to the invertible transform (3.17).

Remark 3 (Comparison with `ScaledGD`). Although not our focus, it is instructive to consider the

resulting update rule using the nonsmooth ℓ_2-loss function $f(\boldsymbol{X}) = \|\mathcal{A}(\boldsymbol{X}) - \boldsymbol{y}\|_2$ (which has been studied in [CCD+21]), whose subgradient is given by

$$\boldsymbol{S}_t = \frac{\mathcal{A}^*(\boldsymbol{r}_t)}{\|\boldsymbol{r}_t\|_2},$$

where $\mathcal{A}^*(\cdot)$ is the adjoint operator of $\mathcal{A}(\cdot)$, and $\boldsymbol{r}_t \coloneqq \mathcal{A}(\boldsymbol{L}_t\boldsymbol{R}_t^\top) - \boldsymbol{y}$ is the residual using the t-th iterate. Consequently, the scaled subgradient method follows the update rule

$$\begin{aligned}
\boldsymbol{L}_{t+1} &= \boldsymbol{L}_t - \frac{\eta_t}{\|\boldsymbol{r}_t\|_2}\mathcal{A}^*(\boldsymbol{r}_t)\boldsymbol{R}_t(\boldsymbol{R}_t^\top\boldsymbol{R}_t)^{-1}, \\
\boldsymbol{R}_{t+1} &= \boldsymbol{R}_t - \frac{\eta_t}{\|\boldsymbol{r}_t\|_2}\mathcal{A}^*(\boldsymbol{r}_t)^\top\boldsymbol{L}_t(\boldsymbol{L}_t^\top\boldsymbol{L}_t)^{-1},
\end{aligned}$$

for some stepsize η_t. Careful readers might realize that this coincides with the update rule of `ScaledGD` in Chapter 2 when optimizing the smooth squared ℓ_2-loss function $g(\boldsymbol{X}) = \frac{1}{2}\|\mathcal{A}(\boldsymbol{X})-\boldsymbol{y}\|_2^2$, except with an adaptive stepsize $\frac{\eta_t}{\|\boldsymbol{r}_t\|_2}$. Under the same assumption on $\mathcal{A}(\cdot)$ in Chapter 2, the convergence behaviors of `ScaledSM` applied on $f(\boldsymbol{X})$ match that of `ScaledGD` on $g(\boldsymbol{X})$.

Remark 4 (`ScaledSM` for PSD matrices). When the low-rank matrix of interest is positive semi-definite (PSD), we factorize the matrix $\boldsymbol{X} \in \mathbb{R}^{n\times n}$ as $\boldsymbol{X} = \boldsymbol{L}\boldsymbol{L}^\top$, with $\boldsymbol{L} \in \mathbb{R}^{n\times r}$. The update rule of `ScaledSM` simplifies to

$$\boldsymbol{L}_{t+1} = \boldsymbol{L}_t - \eta_t\boldsymbol{S}_t\boldsymbol{L}_t(\boldsymbol{L}_t^\top\boldsymbol{L}_t)^{-1},$$

where $\boldsymbol{S}_t \in \partial f(\boldsymbol{L}_t\boldsymbol{L}_t^\top)$ is a subgradient of $f(\cdot)$ at $\boldsymbol{L}_t\boldsymbol{L}_t^\top$. Our theory applies to this PSD case in a straightforward manner.

3.3 Theoretical Guarantees

In this section, we first provide the theoretical guarantees of the scaled subgradient method under general geometric assumptions on $f(\cdot)$, and then instantiate them to concrete problems including robust low-rank matrix sensing and quadratic sampling.

3.3.1 Geometric assumptions

We start by introducing the following geometric properties of the loss function $f(\cdot)$, which play a key role in the convergence analysis.

The first condition is similar to the usual Lipschitz property of a function.

Definition 5 (Restricted Lipschitz continuity)**.** A function $f : \mathbb{R}^{n_1 \times n_2} \mapsto \mathbb{R}$ is said to be rank-r restricted L-Lipschitz continuous for some quantity $L > 0$ if

$$|f(\boldsymbol{X}_1) - f(\boldsymbol{X}_2)| \leq L\|\boldsymbol{X}_1 - \boldsymbol{X}_2\|_{\mathsf{F}}$$

holds for any $\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathbb{R}^{n_1 \times n_2}$ such that $\boldsymbol{X}_1 - \boldsymbol{X}_2$ has rank at most $2r$.

The second geometric condition is akin to the (one-point) strong convexity of a function, with the key difference that strong convexity adopts the squared Euclidean norm whereas the following one uses the plain Euclidean norm.

Definition 6 (Restricted sharpness)**.** A function $f : \mathbb{R}^{n_1 \times n_2} \mapsto \mathbb{R}$ is said to be rank-r restricted μ-sharp w.r.t. $\boldsymbol{X}_\star$ for some $\mu > 0$ if

$$f(\boldsymbol{X}) - f(\boldsymbol{X}_\star) \geq \mu\|\boldsymbol{X} - \boldsymbol{X}_\star\|_{\mathsf{F}}$$

holds for any $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$ with rank at most r.

For notational simplicity, if a function $f(\cdot)$ is both restricted L-Lipschitz continuous and μ-sharp, we denote

$$\chi_f \coloneqq L/\mu. \tag{3.18}$$

In some cases, e.g. in the presence of noise, the loss function $f(\cdot)$ only satisfies an approximate restricted sharpness property, which is detailed below.

Definition 7 (Approximate restricted sharpness)**.** A function $f : \mathbb{R}^{n_1 \times n_2} \mapsto \mathbb{R}$ is said to be ξ-approximate rank-r restricted μ-sharp for some $\mu, \xi > 0$ if

$$f(\boldsymbol{X}) - f(\boldsymbol{X}_\star) \geq \mu \|\boldsymbol{X} - \boldsymbol{X}_\star\|_{\mathsf{F}} - \xi$$

holds for any $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$ with rank at most r.

As shall be seen in Section 3.3.3, these conditions can be ensured for proper choices of the loss function as long as the observation operator $\mathcal{A}(\cdot)$ satisfies certain mixed-norm RIP, which holds for a wide number of practical problems.

3.3.2 Main results

Motivated by Chapter 2, we measure the performance of $\boldsymbol{F} = [\boldsymbol{L}^\top, \boldsymbol{R}^\top]^\top$ using the following error metric

$$\operatorname{dist}^2(\boldsymbol{F}, \boldsymbol{F}_\star) \coloneqq \inf_{\boldsymbol{Q} \in \mathrm{GL}(r)} \left\|(\boldsymbol{L}\boldsymbol{Q} - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{R}\boldsymbol{Q}^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2, \tag{3.19}$$

which takes into consideration both the representational ambiguity of the factorization up to invertible transforms and the scaling effect of preconditioners. In comparison, the more standard distance metric [MLC21] in the analysis of vanilla gradient methods reads as follows

$$\operatorname{dist}_{\mathrm{u}}^2(\boldsymbol{F}, \boldsymbol{F}_\star) \coloneqq \inf_{\boldsymbol{Q} \in \mathrm{GL}(r)} \|\boldsymbol{L}\boldsymbol{Q} - \boldsymbol{L}_\star\|_{\mathsf{F}}^2 + \left\|\boldsymbol{R}\boldsymbol{Q}^{-\top} - \boldsymbol{R}_\star\right\|_{\mathsf{F}}^2,$$

which is inadequate to delineate the power of preconditioning. See Chapter 2 for more discussions.

We start with stating the linear convergence of the scaled subgradient method when $f(\cdot)$ satisfies both the rank-r restricted L-Lipschitz continuity and μ-sharpness. The proof is deferred to Appendix B.1.

Theorem 6 (Scaled subgradient method with exact convergence)**.** *Suppose that $f(\boldsymbol{X}) : \mathbb{R}^{n_1 \times n_2} \mapsto \mathbb{R}$ is convex in $\boldsymbol{X}$, and satisfies rank-r restricted L-Lipschitz continuity and μ-sharpness (cf. Defini-*

tions 5 and 6). In addition, suppose that the initialization $\boldsymbol{F}_0$ satisfies

$$\operatorname{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \le 0.02\sigma_r(\boldsymbol{X}_\star)/\chi_f, \tag{3.20}$$

and the scaled subgradient method in (3.13) adopts either Polyak's stepsizes in (3.14) or geometrically decaying stepsizes in (3.15) with $\lambda = \sqrt{\frac{\sqrt{2}-1}{2}}0.02\sigma_r(\boldsymbol{X}_\star)/\chi_f^2$ and $q = \sqrt{1-0.16/\chi_f^2}$. Then for all $t \ge 0$, the iterates satisfy

$$\begin{aligned}\operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) &\le (1-0.16/\chi_f^2)^{t/2}0.02\sigma_r(\boldsymbol{X}_\star)/\chi_f, \quad \text{and} \\ \left\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\right\|_{\mathsf{F}} &\le (1-0.16/\chi_f^2)^{t/2}0.03\sigma_r(\boldsymbol{X}_\star)/\chi_f.\end{aligned}$$

Theorem 6 shows that the iterates of the scaled subgradient method converges at a linear rate; to reach ϵ-accuracy, i.e. $\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\|_{\mathsf{F}} \le \epsilon\sigma_r(\boldsymbol{X}_\star)$, it takes at most $O(\chi_f^2 \log\frac{1}{\epsilon})$ iterations, which, importantly, is independent of the condition number κ of $\boldsymbol{X}_\star$. In addition, it is still possible to ensure approximate reconstruction when only the approximate restricted sharpness property holds, as shown in the next theorem. Again, we postpone the proof to Appendix B.2.

Theorem 7 (Scaled subgradient method with approximate convergence)**.** *Suppose that $f : \mathbb{R}^{n_1\times n_2} \mapsto \mathbb{R}$ is convex, and satisfies rank-r restricted L-Lipschitz continuity and ξ-approximate μ-sharpness (cf. Definitions 5 and 7) for some $\xi \le 10^{-3}\sigma_r(\boldsymbol{X}_\star)\mu/\chi_f$. Suppose that the initialization $\boldsymbol{F}_0$ satisfies $\operatorname{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \le 0.02\sigma_r(\boldsymbol{X}_\star)/\chi_f$, and the scaled subgradient method adopts geometrically decaying stepsizes (3.15) with $\lambda = \sqrt{\frac{\sqrt{2}-1}{2}}0.02\sigma_r(\boldsymbol{X}_\star)/\chi_f^2$ and $q = \sqrt{1-0.13/\chi_f^2}$. Then for all $t \ge 0$, the iterates satisfy*

$$\begin{aligned}\operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) &\le \max\left\{(1-0.13/\chi_f^2)^{t/2}0.02\sigma_r(\boldsymbol{X}_\star)/\chi_f, 20\xi/\mu\right\}, \quad \text{and} \\ \left\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\right\|_{\mathsf{F}} &\le \max\left\{(1-0.13/\chi_f^2)^{t/2}0.03\sigma_r(\boldsymbol{X}_\star)/\chi_f, 30\xi/\mu\right\}.\end{aligned}$$

Theorem 7 shows that as long as the relaxation parameter ξ is sufficiently small, i.e. $\xi \lesssim \sigma_r(\boldsymbol{X}_\star)\mu/\chi_f$, then the scaled subgradient method with geometrically decaying stepsizes converges at a linear rate until an error floor is hit. In particular, the iterates satisfy $\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\|_{\mathsf{F}} \le 30\xi/\mu$

after at most $O(\chi_f^2)$ iterations up to logarithmic factors.

Remark 5. For simplicity of exposition, we have fixed the values of λ and q for the geometrically decaying stepsizes in the above theorems. It is possible to allow a wider range of λ and q by slightly modifying the arguments without sacrificing the linear convergence. In practice, these parameters should be tuned in order to yield optimal performance.

3.3.3 A case study: robust low-rank matrix recovery

We now apply the above theorems to robust low-rank matrix recovery, which showcases the superior performance of the scaled subgradient method.

Noise-free case. We start with the observation model (3.10) with clean measurements, i.e. $\boldsymbol{w} = \boldsymbol{0}$ and $\boldsymbol{s} = \boldsymbol{0}$. To proceed, we assume that the observation operator $\mathcal{A}(\cdot)$ satisfies the following mixed-norm RIP.

Definition 8 (mixed-norm RIP [RFP10, CCG15, CCD+21])**.** The linear map $\mathcal{A}(\cdot)$ is said to obey the rank-$2r$ mixed-norm RIP with constants δ_1, δ_2 if for all matrices $\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2}$ of rank at most $2r$, one has

$$\delta_1 \|\boldsymbol{M}\|_{\mathsf{F}} \le \|\mathcal{A}(\boldsymbol{M})\|_1 \le \delta_2 \|\boldsymbol{M}\|_{\mathsf{F}}.$$

The next proposition verifies that the loss function (3.12) satisfies restricted Lipschitz continuity and sharpness properties under the mixed-norm RIP.

Proposition 3. *If $\mathcal{A}(\cdot)$ satisfies rank-$2r$ mixed-norm RIP with constants (δ_1, δ_2), then $f(\boldsymbol{X}) = \|\mathcal{A}(\boldsymbol{X}) - \boldsymbol{y}\|_1 = \|\mathcal{A}(\boldsymbol{X} - \boldsymbol{X}_\star)\|_1$ in (3.12) satisfies the rank-r restricted L-Lipschitz continuity and μ-sharpness with*

$$L = \delta_2, \quad \text{and} \quad \mu = \delta_1.$$

Proof. See Appendix B.3. □

With the geometric characterization of $f(\cdot)$ in place, we immediately have the following corollary that captures the performance of the scaled subgradient method when $\mathcal{A}(\cdot)$ satisfies the mixed-norm RIP.

Corollary 1. *If $\mathcal{A}(\cdot)$ satisfies rank-$2r$ mixed-norm RIP with (δ_1, δ_2), then the scaled subgradient method over the loss function $f(\boldsymbol{X}) = \|\mathcal{A}(\boldsymbol{X}) - \boldsymbol{y}\|_1$ using either Polyak's or geometrically decaying stepsizes achieves $\left\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\right\|_{\mathsf{F}} \le \epsilon\sigma_r(\boldsymbol{X}_\star)$ in $O\left(\frac{\delta_2^2}{\delta_1^2}\log\frac{1}{\epsilon}\right)$ iterations as long as the initialization satisfies $\operatorname{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \le \frac{0.02\delta_1}{\delta_2}\sigma_r(\boldsymbol{X}_\star)$.*

Noisy and corrupted case. We now consider the observation model (3.10) where the noise $\boldsymbol{w}$ is bounded with $\|\boldsymbol{w}\|_1 \le \sigma_w$ and $\|\boldsymbol{s}\|_0 = p_s m$, where $p_s \in [0, 1/2)$ is the fraction of outliers. Following [CCD+21], we further introduce another important property of $\mathcal{A}(\cdot)$.

Definition 9 ($\mathcal{S}$-outlier bound [CCD+21])**.** The linear map $\mathcal{A}(\cdot)$ is said to obey the rank-$2r$ $\mathcal{S}$-outlier bound w.r.t. a set $\mathcal{S}$ with a constant δ_3 if for all matrices $\boldsymbol{M} \in \mathbb{R}^{n_1\times n_2}$ of rank at most $2r$, one has

$$\delta_3\|\boldsymbol{M}\|_{\mathsf{F}} \le \|\mathcal{A}_{\mathcal{S}^c}(\boldsymbol{M})\|_1 - \|\mathcal{A}_{\mathcal{S}}(\boldsymbol{M})\|_1,$$

where $\mathcal{A}_{\mathcal{S}}(\boldsymbol{M}) = \{\mathcal{A}_i(\boldsymbol{M})\}_{i\in\mathcal{S}}$ and $\mathcal{A}_{\mathcal{S}^c}(\boldsymbol{M}) = \{\mathcal{A}_i(\boldsymbol{M})\}_{i\in\mathcal{S}^c}$.

The next proposition verifies that the loss function in (3.12) satisfies restricted Lipschitz continuity and approximate sharpness properties under the mixed-norm RIP (cf. Definition 8) and the $\mathcal{S}$-outlier bound (cf. Definition 9).

Proposition 4 (Matrix sensing with outliers)**.** *Denote the support of the outlier $\boldsymbol{s}$ as $\mathcal{S}$. Suppose that $\mathcal{A}(\cdot)$ satisfies rank-$2r$ mixed-norm RIP with (δ_1, δ_2) and $\mathcal{S}$-outlier bound with δ_3, then $f(\boldsymbol{X})$ in (3.12) satisfies rank-r restricted L-Lipschitz continuity and ξ-approximate μ-sharpness with*

$$L = \delta_2, \quad \mu = \delta_3, \quad \text{and} \quad \xi = 2\sigma_w. \tag{3.21}$$

Proof. See Appendix B.4. □

Similar to the previous noise-free case, this immediately leads to performance guarantees of the scaled subgradient method when $\mathcal{A}(\cdot)$ satisfies both the mixed-norm RIP and the $\mathcal{S}$-outlier bound.

Corollary 2. *If $\mathcal{A}(\cdot)$ satisfies rank-$2r$ mixed-norm RIP with (δ_1, δ_2) and $\mathcal{S}$-outlier bound with δ_3, and $\|\boldsymbol{w}\|_1 \le \sigma_w \le 10^{-3}\sigma_r(\boldsymbol{X}_\star)\delta_3^2/\delta_2$, then the scaled subgradient method over the loss function $f(\boldsymbol{X}) = \|\mathcal{A}(\boldsymbol{X}) - \boldsymbol{y}\|_1$ using the geometrically decaying stepsizes achieves $\left\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\right\|_{\mathsf{F}} \le \max\{\epsilon\sigma_r(\boldsymbol{X}_\star), 60\sigma_w/\delta_3\}$ in $O\left(\frac{\delta_2^2}{\delta_3^2}\log\frac{1}{\epsilon}\right)$ iterations as long as the initialization satisfies $\operatorname{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \le \frac{0.02\delta_3}{\delta_2}\sigma_r(\boldsymbol{X}_\star)$.*

We now instantiate the above general guarantee to the following two types of observation operators. For simplicity, we assume there is no dense noise, i.e. $\sigma_w = 0$; see Table 3.1 for a summary of the comparisons.

- *matrix sensing:* the measurement operator $\mathcal{A}_i(\cdot)$ is defined as $\mathcal{A}_i(\boldsymbol{X}_\star) = \frac{1}{m}\langle \boldsymbol{A}_i, \boldsymbol{X}_\star\rangle$, where the matrix $\boldsymbol{A}_i$ is composed of i.i.d. Gaussian entries $\mathcal{N}(0,1)$.[2] It is shown in [CCD+21] (see also [LZMCSV20]) that $\mathcal{A}(\cdot)$ satisfies the mixed-norm RIP and $\mathcal{S}$-outlier bound with

$$\delta_1 \gtrsim 1, \quad \delta_2 \lesssim 1, \quad \delta_3 \gtrsim 1 - 2p_s,$$

  as long as $m \gtrsim \frac{(n_1+n_2)r}{(1-2p_s)^2}\log(\frac{1}{1-2p_s})$. Hence, the scaled subgradient method converges linearly to ϵ-accuracy in $O\left(\frac{1}{(1-2p_s)^2}\log\frac{1}{\epsilon}\right)$ iterations provided that it is initialized properly, making it robust simultaneously to ill-conditioning of the matrix $\boldsymbol{X}_\star$ and the presence of the outliers.

- *quadratic sampling:* the measurement operator $\mathcal{A}_i(\cdot)$ is defined as $\mathcal{A}_i(\boldsymbol{X}_\star) = \frac{1}{m}\langle \boldsymbol{a}_i\boldsymbol{a}_i^\top, \boldsymbol{X}_\star\rangle$, where $\boldsymbol{X}_\star \in \mathbb{R}^{n\times n}$ is PSD and the vector $\boldsymbol{a}_i$ is composed of i.i.d. Gaussian entries $\mathcal{N}(0,1)$. It is shown in [CCD+21] that $\mathcal{A}(\cdot)$ satisfies the mixed-norm RIP and $\mathcal{S}$-outlier bound with

$$\delta_1 \gtrsim 1, \quad \delta_2 \lesssim \sqrt{r}, \quad \delta_3 \gtrsim 1 - 2p_s,$$

  as long as $m \gtrsim \frac{nr^2}{(1-2p_s)^2}\log(\frac{\sqrt{r}}{1-2p_s})$. Hence, the scaled subgradient method converges linearly

[2] The same guarantee also holds for sub-Gaussian measurements.

to ϵ-accuracy in $O\left(\frac{r}{(1-2p_s)^2}\log\frac{1}{\epsilon}\right)$ iterations, as long as it is seeded with a good initialization. In comparison, the iteration complexity of the scaled gradient descent method over the least-squares loss function depends polynomially with respect to n, due to the heavy-tailed nature of the observation operator, let alone its sensitivity to the outliers.

Remark 6 (Initialization). The above discussions are limited to the local iteration complexity, assuming a good initialization satisfying (3.20) is available. In the absence of outliers, a standard spectral method can be used, as shown in Chapter 2. In the presence of outliers, a truncated spectral method could be used; see e.g. [ZCL16, LCZL20].

3.4 Numerical Experiments

In this section, we conduct numerical experiments to corroborate our theory.

Comparisons of `ScaledSM` and `VanillaSM`. Since the vanilla subgradient method (`VanillaSM`) has been extensively benchmarked against other methods and established as state-of-the-art in [CCD+21], we focus on comparing the proposed scaled subgradient method (`ScaledSM`) to `VanillaSM` in the sequel. In general, the geometrically decaying stepsize (3.15) is a more practical choice than the Polyak's stepsize (3.14), especially in the presence of noise and outliers. Nonetheless, using properly tuned geometrically decaying stepsizes essentially matches the performance of using Polyak's stepsizes, for both `VanillaSM` [LZMCSV20] and `ScaledSM`, the latter of which we shall illustrate in the ensuing experiments. As such, we adopt Polyak's stepsizes in the comparisons below, to emulate the scenario where both methods are tuned to operate under its largest allowable stepsizes and achieve the fastest convergence. In addition, both algorithms start from the same initialization.

We consider two low-rank matrix estimation tasks discussed in Section 3.3.3. Recall the observation model in (3.10) and its entrywise version in (3.9), which we repeat below for convenience:

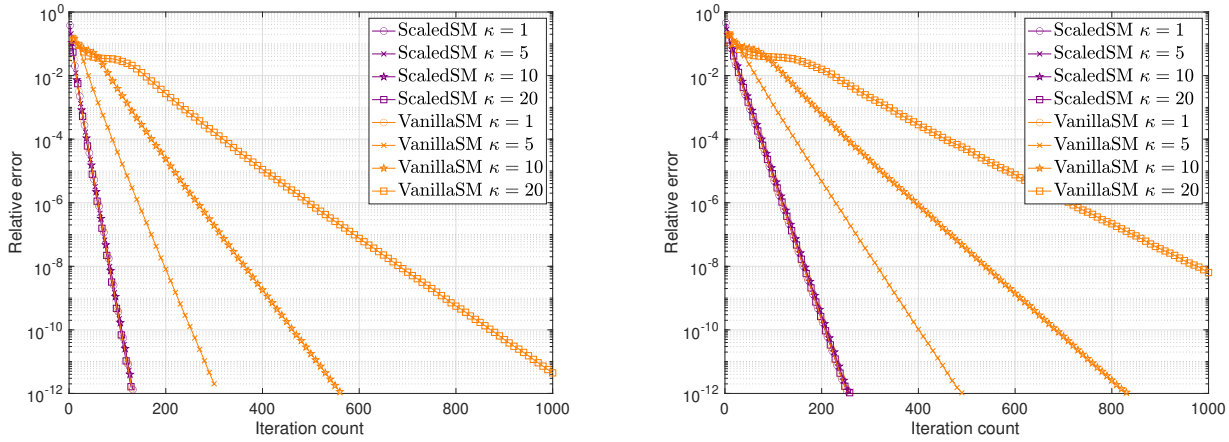
$$y_i = \mathcal{A}_i(\boldsymbol{X}_\star) + w_i + s_i, \qquad 1 \le i \le m.$$

In both tasks, the noise entry w_i is composed of i.i.d. entries uniformly drawn from $[-\frac{\sigma_w}{m}, \frac{\sigma_w}{m}]$.

The outlier $s_i = \bar{s}_i \Omega_i$ is a sparse vector where Ω_i is a Bernoulli random variable with probability $p_s \in [0, 1/2)$, and $\bar{s}_i$ is drawn uniformly at random from $[-10\|\mathcal{A}(\boldsymbol{X}_\star)\|_\infty, 10\|\mathcal{A}(\boldsymbol{X}_\star)\|_\infty]$. For ease of presentation, we assume that $\boldsymbol{X}_\star \in \mathbb{R}^{n\times n}$ is a square matrix with rank as r. We collect $m = 8nr$ measurements using the following respective measurement models. The signal-to-noise ratio is defined as $\mathrm{SNR} := 20\log_{10}\frac{\|\mathcal{A}(\boldsymbol{X}_\star)\|_1}{\sigma_w}$ in dB.

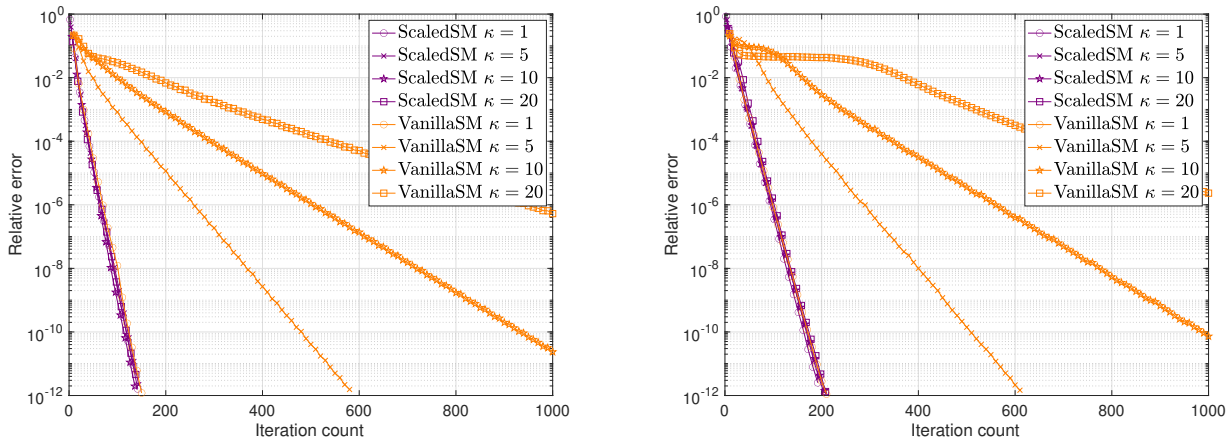
- *Matrix sensing.* Here, the measurement operator $\mathcal{A}_i(\cdot)$ is defined as $\mathcal{A}_i(\boldsymbol{X}_\star) = \frac{1}{m}\langle \boldsymbol{A}_i, \boldsymbol{X}_\star\rangle$, where the matrix $\boldsymbol{A}_i$ is composed of i.i.d. Gaussian entries $\mathcal{N}(0,1)$. The ground truth matrix $\boldsymbol{X}_\star$ is generated via its compact SVD $\boldsymbol{X}_\star = \boldsymbol{U}_\star\boldsymbol{\Sigma}_\star\boldsymbol{V}_\star^\top$, where $\boldsymbol{U}_\star \in \mathbb{R}^{n\times r}$ is generated as the orthonormal basis vectors of an $n \times r$ matrix with i.i.d. Rademacher entries, $\boldsymbol{\Sigma}_\star$ is a diagonal matrix with the diagonal entries linearly distributed from 1 to κ, and $\boldsymbol{V}_\star \in \mathbb{R}^{n\times r}$ is generated in a similar fashion to $\boldsymbol{U}_\star$.

- *Quadratic sampling.* Here, the measurement operator $\mathcal{A}_i(\cdot)$ is defined as $\mathcal{A}_i(\boldsymbol{X}_\star) = \frac{1}{m}\langle \boldsymbol{a}_i\boldsymbol{a}_i^\top, \boldsymbol{X}_\star\rangle$, where $\boldsymbol{a}_i$ is composed of i.i.d. Gaussian entries $\mathcal{N}(0,1)$. The ground truth matrix $\boldsymbol{X}_\star$ is positive semi-definite, and is generated via its compact SVD $\boldsymbol{X}_\star = \boldsymbol{U}_\star\boldsymbol{\Sigma}_\star\boldsymbol{U}_\star^\top$, where $\boldsymbol{U}_\star$ and $\boldsymbol{\Sigma}_\star$ are generated in the same manner described above.



(a) without outliers (b) with 20% outliers

Figure 3.1: Performance comparisons of `ScaledSM` and `VanillaSM` for matrix sensing without or with outliers under different condition numbers $\kappa = 1, 5, 10, 20$, where $n = 100$, $r = 10$, and $m = 8nr$.



(a) without outliers

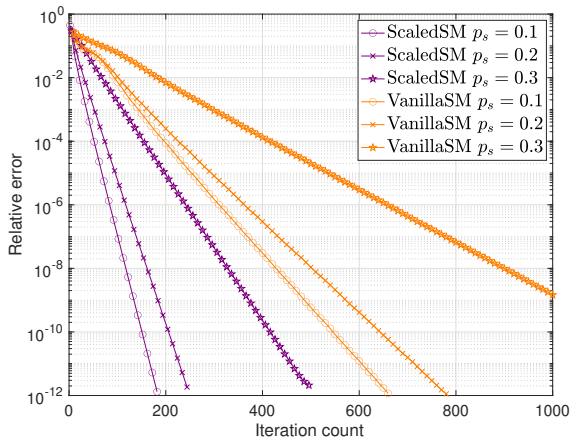
(b) with 20% outliers

Figure 3.2: Performance comparisons of `ScaledSM` and `VanillaSM` for quadratic sampling without or with outliers under different condition numbers $\kappa = 1, 5, 10, 20$, where $n = 100$, $r = 5$, and $m = 8nr$.

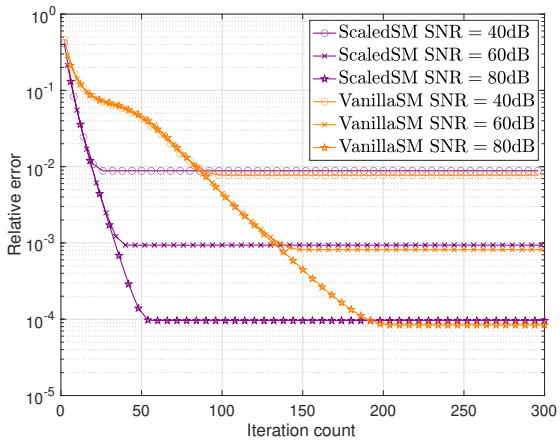
Denote the index set of the remaining measurements after discarding p_s fraction with largest amplitudes as $\mathcal{I} = \{i : |y_i| \leq |\boldsymbol{y}|_{(\lceil p_s m \rceil)}\}$, where $|\boldsymbol{y}|_{(k)}$ denotes the kth largest amplitude of $\boldsymbol{y}$. The truncated spectral method in [ZCL16, LCZL20] is used for initialization, where we apply the standard spectral method only on the subset $\mathcal{I}$ of the measurements. For matrix sensing, it follows the prescription in [LCZL20], and for quadratic sampling, it follows [LMCC21].

Fig. 3.1 shows the relative reconstruction error $\|\boldsymbol{X}_t - \boldsymbol{X}_\star\|_{\mathsf{F}} / \|\boldsymbol{X}_\star\|_{\mathsf{F}}$ for matrix sensing without outliers (in (a)) and with 20% outliers (i.e. $p_s = 0.2$ in (b)) under different condition numbers κ, where $\boldsymbol{X}_t$ is the estimated low-rank matrix at the t-th iteration. Fig. 3.2 shows the relative reconstruction error for quadratic sampling under the same setting. It can be seen that `ScaledSM` is insensitive to κ and converges as a fast rate that is independent with κ, while the convergence of `VanillaSM` slows down dramatically with the increase of κ. In addition, both algorithms still converge linearly in the presence of outliers, thanks to the robustness of the least absolute deviations.

Fig. 3.3 further examines the impact of the amount of outliers and noise on the convergence speed in matrix sensing with a fixed condition number $\kappa = 10$, where Fig. 3.3 (a) illustrates the convergence speed at varying amounts of outliers $p_s = 0.1, 0.2, 0.3$ respectively, and Fig. 3.3 (b) illustrates the convergence with $p_s = 0.1$ and additional bounded noise with varying SNR $= 40, 60, 80$dB.



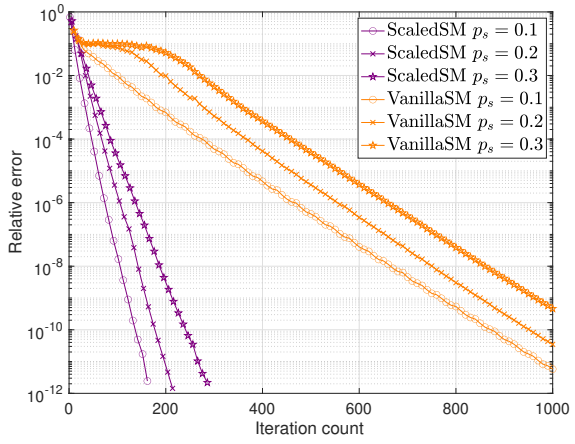
(a) with varying outliers



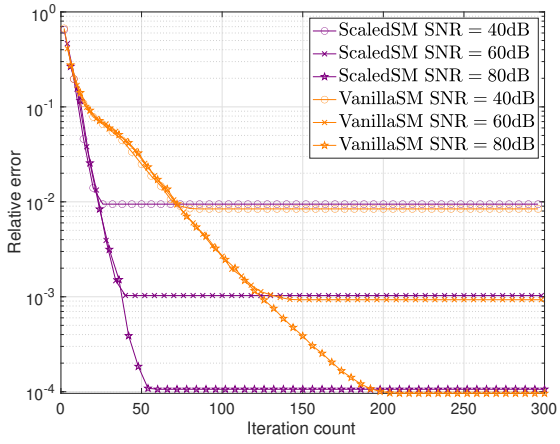
(b) with 10% outliers and noise

Figure 3.3: Performance comparisons of `ScaledSM` and `VanillaSM` for matrix sensing under different noise and outlier models, where $n = 100$, $r = 10$, $m = 8nr$, and $\kappa = 10$.

Similarly, Fig. 3.4 shows the same plots for quadratic sampling under the same setting. It can be seen that the convergence rate of `ScaledSM` slows down with the increase of outliers, which is again, consistent with the theory. Furthermore, the reconstruction is robust in the presence of additional bounded noise, where both `ScaledSM` and `VanillaSM` converge to the same accuracy that is proportional to the noise level, with `ScaledSM` converging at a faster speed.



(a) with varying outliers



(b) with 10% outliers and noises

Figure 3.4: Performance comparisons of `ScaledSM` and `VanillaSM` for quadratic sampling under different noise and outlier models, where $n = 100$, $r = 5$, $m = 8nr$, and $\kappa = 10$.

Comparisons of stepsize schedules. We now compare the geometrically decaying stepsize with the Polyak's stepsize for `ScaledSM`, which essentially mirrors similar experiments conducted in [LZMCSV20] for `VanillaSM`. We run `ScaledSM` for at most $T = 1000$ iterations, and stop early if the relative error achieves 10^{-12}. Fig. 3.5 and Fig. 3.6 show the performance comparisons of `ScaledSM` under various stepsize schedules for matrix sensing and quadratic sampling, respectively. For both figures, (a) shows the final relative error of `ScaledSM` using geometrically decaying stepsizes under various (λ, q), where we see that `ScaledSM` converges as long as λ is not too large and q is not too small. We further plot the relative error versus the iteration count for `ScaledSM` using geometrically decaying stepsizes with a fixed q and various λ in (b), and with a fixed λ and various q in (c), where the performance using Polyak's stepsizes is plotted for comparison. It can be seen that using Polyak's stepsizes yields the fastest convergence. Indeed, if properly tuned, geometrically decaying stepsizes match Polyak's stepsizes, as shown in (d). In general, we find that there is a wide range of parameters for geometrically decaying stepsizes where `ScaledSM` converges in a fast speed comparable to that of using Polyak's stepsizes, as long as λ is not too large and q is not too small.

3.5 Discussions

This chapter proposes scaled subgradient methods to minimize a family of nonsmooth and nonconvex formulations for low-rank matrix recovery—in particular, the residual sum of absolute errors—and guarantees its convergence at a rate that is almost dimension-free and independent of the condition number, even in the presence of corruptions. We illustrate the effectiveness of our approach by providing state-of-the-art performance guarantees for robust low-rank matrix sensing and quadratic sampling. In the future, it is of interest to study the performance of scaled subgradient methods for other signal estimation and statistical inference tasks, such as training student-teacher neural networks [DDKL20], as well as using random initializations [CCFM19].

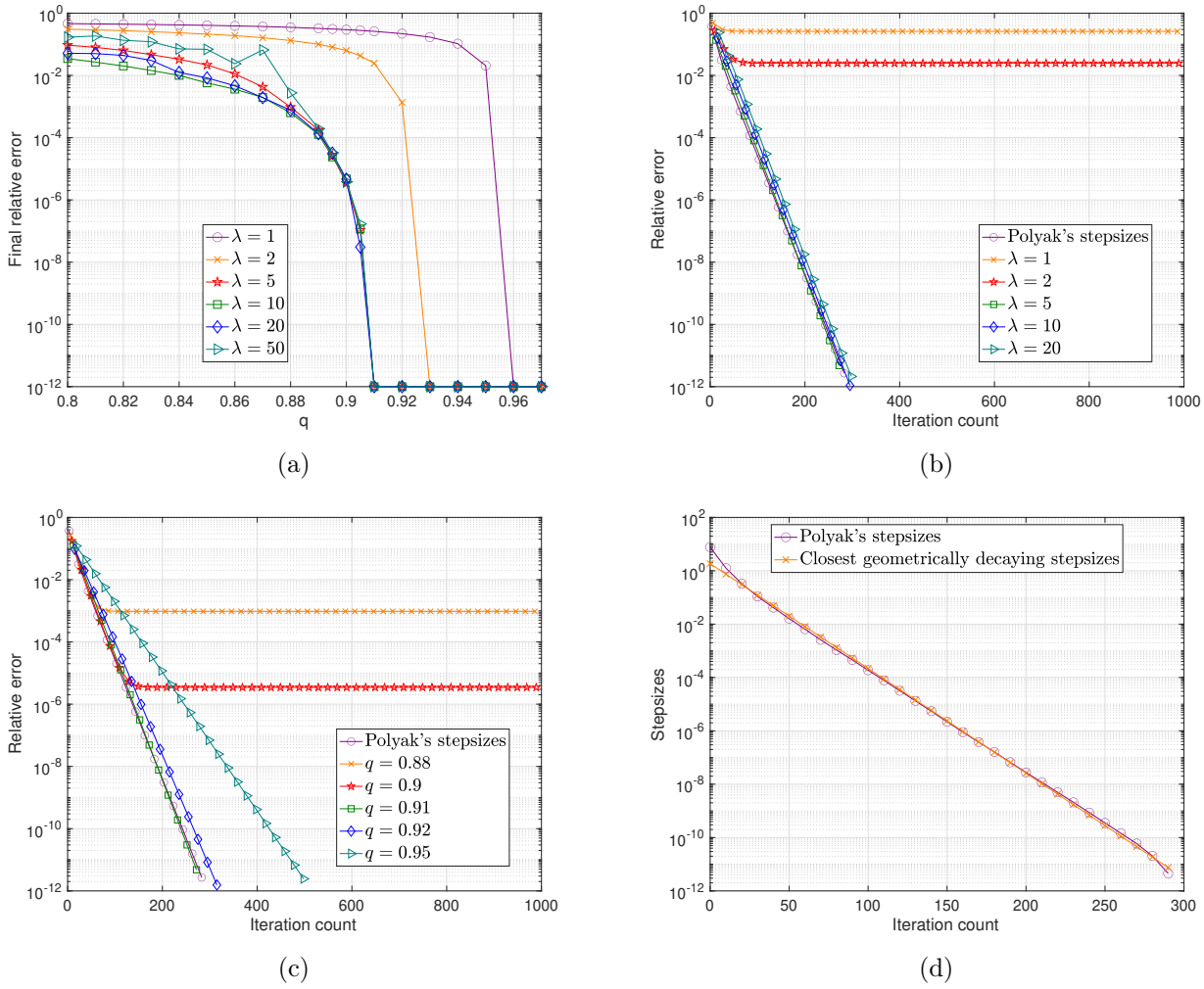


Figure 3.5: Performance comparisons of `ScaledSM` for matrix sensing using geometrically decaying stepsizes with parameters (λ, q) and Polyak's stepsizes, where we fix $n = 100$, $r = 10$, $m = 8nr$, $\kappa = 10$, and $p_s = 0.2$: (a) the final relative error for various combinations of (λ, q), (b) the relative error versus iteration count for fixed $q = 0.91$ and varying λ, (c) the relative error versus iteration count for fixed $\lambda = 5$ and varying q, and (d) shows properly tuned geometrically decaying stepsizes with $\lambda = 1.85$ and $q = 0.91$ essentially match Polyak's stepsizes.

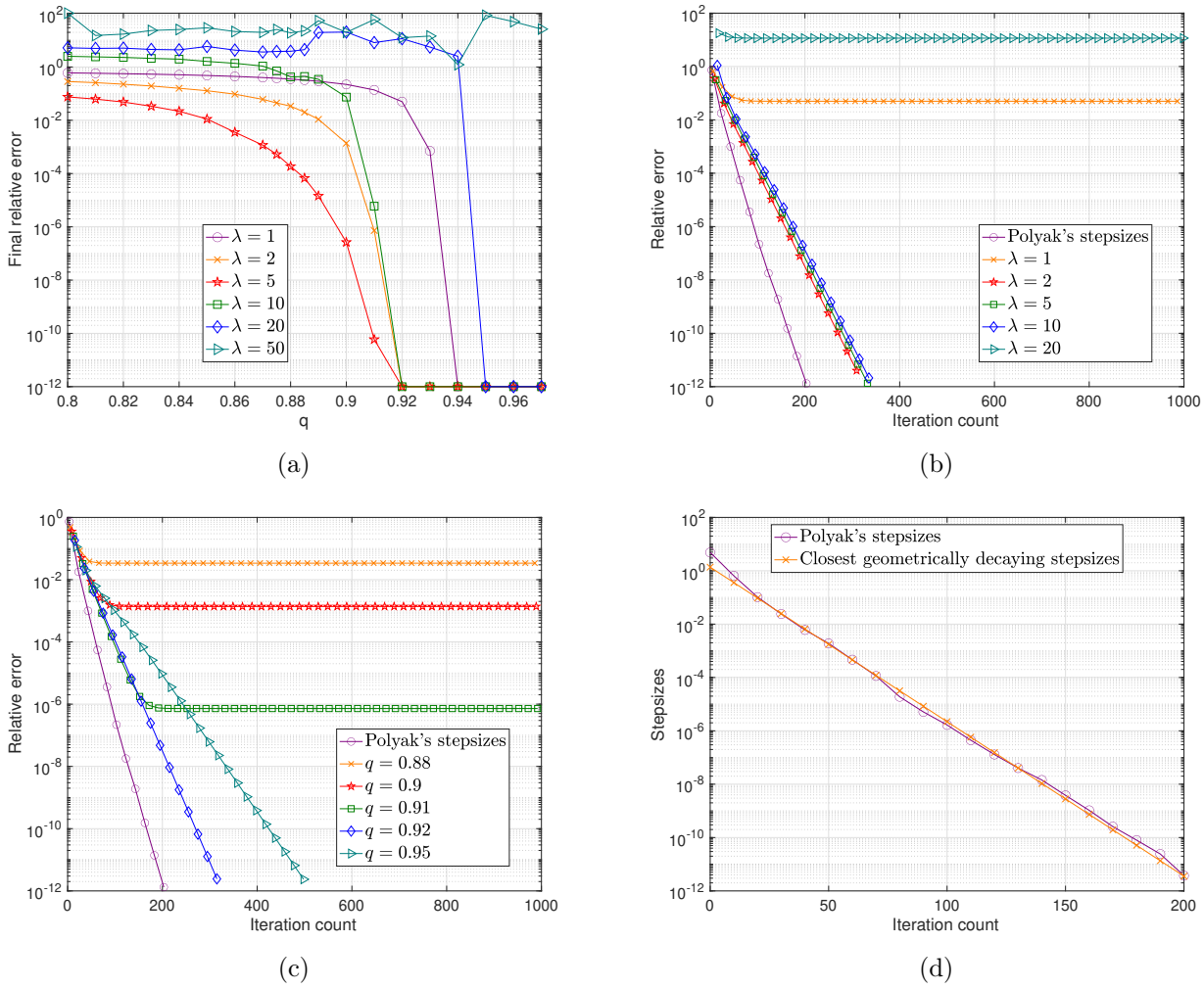


Figure 3.6: Performance comparisons of `ScaledSM` for quadratic sampling using geometrically decaying stepsizes with parameters (λ, q) and Polyak's stepsizes, where we fix $n = 100$, $r = 5$, $m = 8nr$, $\kappa = 10$, and $p_s = 0.2$: (a) the final relative error for various combinations of (λ, q), (b) the relative error versus iteration count for fixed $q = 0.92$ and varying λ, (c) the relative error versus iteration count for fixed $\lambda = 2$ and varying q, and (d) shows properly tuned geometrically decaying stepsizes with $\lambda = 1.36$ and $q = 0.88$ essentially match Polyak's stepsizes.

Chapter 4

Low-rank Tensor Estimation

4.1 Introduction

In this chapter, we generalize `ScaledGD` to 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 $n_1 n_2 n_3$ 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 chapter—is the low-rank structure under the *Tucker* decomposition [Tuc66]. Specifically, we assume that the ground truth tensor $\boldsymbol{\mathcal{X}}_\star$ admits the following Tucker decomposition[2]

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

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

[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.

[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 $(\boldsymbol{U}_\star, \boldsymbol{V}_\star, \boldsymbol{W}_\star) \cdot \boldsymbol{\mathcal{S}}_\star$ as in [XY19] for convenience of our theoretical developments.

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 $\boldsymbol{y} = \{y_i\}_{i=1}^m$.

4.1.1 A gradient descent approach?

Recent years remarkable successes have emerged 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] within 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 4.1.3 for further discussions.

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

$$\min_{\boldsymbol{F}} \quad \mathcal{L}(\boldsymbol{F}) := \frac{1}{2} \left\| \mathcal{A}\left((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \cdot \boldsymbol{\mathcal{S}}\right) - \boldsymbol{y} \right\|_2^2, \tag{4.1}$$

where $\boldsymbol{F} := (\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}, \boldsymbol{\mathcal{S}})$ 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 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 (4.1) via gradient descent (GD), which updates the factors according to

$$\boldsymbol{F}_{t+1} = \boldsymbol{F}_t - \eta \nabla \mathcal{L}(\boldsymbol{F}_t), \qquad t = 0, 1, \ldots, \tag{4.2}$$

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

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 $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}$:

$$\mathcal{L}_{\mathsf{reg}}(\boldsymbol{F}) \coloneqq \mathcal{L}(\boldsymbol{F}) + \frac{\alpha}{4}\left(\|\boldsymbol{U}^\top \boldsymbol{U} - \beta \boldsymbol{I}_{r_1}\|_{\mathsf{F}}^2 + \|\boldsymbol{V}^\top \boldsymbol{V} - \beta \boldsymbol{I}_{r_2}\|_{\mathsf{F}}^2 + \|\boldsymbol{W}^\top \boldsymbol{W} - \beta \boldsymbol{I}_{r_3}\|_{\mathsf{F}}^2\right), \tag{4.3}$$

and perform GD on the regularized loss. Here, $\alpha, \beta > 0$ 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 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.

4.1.2 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

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

minimize the loss function (4.1):

$$\begin{aligned}
\boldsymbol{U}_{t+1} &= \boldsymbol{U}_t - \eta \nabla_{\boldsymbol{U}} \mathcal{L}(\boldsymbol{F}_t) \big(\breve{\boldsymbol{U}}_t^\top \breve{\boldsymbol{U}}_t\big)^{-1}, \\
\boldsymbol{V}_{t+1} &= \boldsymbol{V}_t - \eta \nabla_{\boldsymbol{V}} \mathcal{L}(\boldsymbol{F}_t) \big(\breve{\boldsymbol{V}}_t^\top \breve{\boldsymbol{V}}_t\big)^{-1}, \\
\boldsymbol{W}_{t+1} &= \boldsymbol{W}_t - \eta \nabla_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{F}_t) \big(\breve{\boldsymbol{W}}_t^\top \breve{\boldsymbol{W}}_t\big)^{-1}, \\
\boldsymbol{\mathcal{S}}_{t+1} &= \boldsymbol{\mathcal{S}}_t - \eta \left((\boldsymbol{U}_t^\top \boldsymbol{U}_t)^{-1}, (\boldsymbol{V}_t^\top \boldsymbol{V}_t)^{-1}, (\boldsymbol{W}_t^\top \boldsymbol{W}_t)^{-1} \right) \cdot \nabla_{\boldsymbol{\mathcal{S}}} \mathcal{L}(\boldsymbol{F}_t),
\end{aligned} \tag{4.4}$$

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

$$\begin{aligned}
\breve{\boldsymbol{U}}_t &\coloneqq \mathcal{M}_1 \left((\boldsymbol{I}_{r_1}, \boldsymbol{V}_t, \boldsymbol{W}_t) \cdot \boldsymbol{\mathcal{S}}_t \right)^\top = (\boldsymbol{W}_t \otimes \boldsymbol{V}_t) \mathcal{M}_1(\boldsymbol{\mathcal{S}}_t)^\top, \\
\breve{\boldsymbol{V}}_t &\coloneqq \mathcal{M}_2 \left((\boldsymbol{U}_t, \boldsymbol{I}_{r_2}, \boldsymbol{W}_t) \cdot \boldsymbol{\mathcal{S}}_t \right)^\top = (\boldsymbol{W}_t \otimes \boldsymbol{U}_t) \mathcal{M}_2(\boldsymbol{\mathcal{S}}_t)^\top, \\
\breve{\boldsymbol{W}}_t &\coloneqq \mathcal{M}_3 \left((\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{I}_{r_3}) \cdot \boldsymbol{\mathcal{S}}_t \right)^\top = (\boldsymbol{V}_t \otimes \boldsymbol{U}_t) \mathcal{M}_3(\boldsymbol{\mathcal{S}}_t)^\top.
\end{aligned} \tag{4.5}$$

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 variant in the matrix setting in Chapter 2, the `ScaledGD` algorithm (4.4) exploits the structures of Tucker decomposition and possesses many desirable properties:

- *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 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

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

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 Chapter)	$n^{3/2} r^{5/2} \kappa^3 \log^3 n$	$\log \frac{1}{\varepsilon}$	factor

Table 4.1: 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 4.2.1). 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.

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 $\boldsymbol{\mathcal{X}}_t = (\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t) \cdot \boldsymbol{\mathcal{S}}_t$ regardless of the parameterization being adopted.

- *Implicit balancing:* `ScaledGD` optimizes the natural loss function (4.1) 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].

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

[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 $r\kappa^2$.

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 Chapter)	$n^{3/2} r^{3/2} \kappa^2$	$\log \frac{1}{\varepsilon}$	factor

Table 4.2: 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 $\|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \leq \varepsilon \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$. 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 4.2.1). 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.

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 $n^{3/2} r^{5/2} \kappa^3 \log^3 n$. 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`. Compared with existing approaches (cf. Table 4.1), `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 $n^{3/2} r^{3/2} \kappa^2$. 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 4.2), `ScaledGD` achieves competitive performance guarantees in terms of sample and iteration complexities with a low per-iteration cost in the factor space.

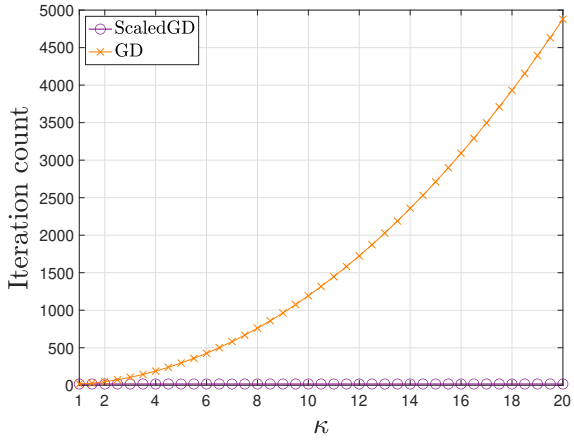


Figure 4.1: The iteration complexities of `ScaledGD` (this thesis) and regularized GD to achieve $\|\boldsymbol{\mathcal{X}}-\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \le 10^{-3}\|\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}$ with respect to different condition numbers for low-rank tensor completion with $n_1 = n_2 = n_3 = 100$, $r_1 = r_2 = r_3 = 5$, and the probability of observation $p = 0.1$.

Figure 4.1 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 4.4 for experimental settings). Clearly, the iteration complexity of GD 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 GD takes thousands of iterations even with a moderate condition number!

4.1.3 Additional related works

Comparison with Chapter 2. While the proposed `ScaledGD` algorithm is inspired by its matrix variant in Chapter 2 by utilizing the same principle of preconditioning, the exact form of preconditioning for tensor factorization needs to be designed carefully and is not trivially obtainable. There are many technical novelty in our analysis compared to Chapter 2. In the matrix case, the low-rank matrix is factorized as $\boldsymbol{L}\boldsymbol{R}^\top$, and only two factors are needed to be estimated. In contrast, in the tensor case, the low-rank tensor is factorized as $(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\cdot\boldsymbol{\mathcal{S}}$, and four factors are needed to be estimated, leading to a much more complicated nonconvex landscape than the matrix case.

In fact, when specialized to matrix completion, our `ScaledGD` algorithm does not degenerate to the same matrix variant in Chapter 2, due to overparamterization and estimating four factors at once, but still maintains the near-optimal performance guarantees. In addition, the tensor algebra possesses unique algebraic properties that requires much more delicate treatments in the analysis. For the local convergence, we establish new concentration properties regarding tensors, which are more challenging compared to the matrix counterparts; for spectral initialization, we establish the effectiveness of a second-order spectral method in the Tucker setting for the first time.

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 [ZX18, XYZ21], 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 thesis. [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 4.2). 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, CCFM21, 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, MWCC19, 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,MWCC19,SC21]; the readers are referred to the overviews [CLC19, CC18, ZQW20] for further references.

4.1.4 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]. We define the unfolding (i.e. flattening) operations of tensors and matrices as following.

- 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 $[\mathcal{M}_1(\boldsymbol{\mathcal{X}})]\big(i_1, i_2 + (i_3 - 1)n_2\big) = \boldsymbol{\mathcal{X}}(i_1, i_2, i_3)$, for $1 \le i_k \le n_k$, $k = 1, 2, 3$; $\mathcal{M}_2(\boldsymbol{\mathcal{X}})$ 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 $[\mathrm{vec}(\boldsymbol{\mathcal{X}})]\big(i_1 + (i_2 - 1)n_1 + (i_3 - 1)n_1 n_2\big) = \boldsymbol{\mathcal{X}}(i_1, i_2, i_3)$ for $1 \le i_k \le n_k$, $k = 1, 2, 3$.

- The 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 $[\mathrm{vec}(\boldsymbol{M})]\big(i_1 + (i_2 - 1)n_1\big) = \boldsymbol{M}(i_1, i_2)$ for $1 \le i_k \le n_k$, $k = 1, 2$.

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

$$\mathrm{vec}((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \cdot \boldsymbol{\mathcal{S}}) = \mathrm{vec}\left(\boldsymbol{U}\mathcal{M}_1(\boldsymbol{\mathcal{S}})(\boldsymbol{W} \otimes \boldsymbol{V})^\top\right) = (\boldsymbol{W} \otimes \boldsymbol{V} \otimes \boldsymbol{U})\,\mathrm{vec}(\boldsymbol{\mathcal{S}}). \tag{4.6a}$$

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

$$\langle \boldsymbol{\mathcal{X}}_1, \boldsymbol{\mathcal{X}}_2 \rangle = \langle \mathcal{M}_k(\boldsymbol{\mathcal{X}}_1), \mathcal{M}_k(\boldsymbol{\mathcal{X}}_2) \rangle, \quad k = 1, 2, 3, \tag{4.6b}$$

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:

$$(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\cdot\big((\boldsymbol{Q}_1,\boldsymbol{Q}_2,\boldsymbol{Q}_3)\cdot\boldsymbol{\mathcal{S}}\big)=(\boldsymbol{U}\boldsymbol{Q}_1,\boldsymbol{V}\boldsymbol{Q}_2,\boldsymbol{W}\boldsymbol{Q}_3)\cdot\boldsymbol{\mathcal{S}}, \tag{4.6c}$$

$$\langle(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\cdot\boldsymbol{\mathcal{S}},\boldsymbol{\mathcal{X}}\rangle=\Big\langle\boldsymbol{\mathcal{S}},(\boldsymbol{U}^\top,\boldsymbol{V}^\top,\boldsymbol{W}^\top)\cdot\boldsymbol{\mathcal{X}}\Big\rangle, \tag{4.6d}$$

$$\|(\boldsymbol{Q}_1,\boldsymbol{Q}_2,\boldsymbol{Q}_3)\cdot\boldsymbol{\mathcal{S}}\|_{\mathsf{F}}\le\|\boldsymbol{Q}_1\|\|\boldsymbol{Q}_2\|\|\boldsymbol{Q}_3\|\|\boldsymbol{\mathcal{S}}\|_{\mathsf{F}}, \tag{4.6e}$$

where $\boldsymbol{Q}_k\in\mathbb{R}^{r_k\times r_k}$, $k=1,2,3$. Define the ℓ_∞ norm of $\boldsymbol{\mathcal{X}}$ 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

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

as the maximum and minimum nonzero singular values of $\boldsymbol{\mathcal{X}}$. In addition, define the spectral norm of a tensor $\boldsymbol{\mathcal{X}}$ as

$$\|\boldsymbol{\mathcal{X}}\|=\sup_{\boldsymbol{u}_k\in\mathbb{R}^{n_k}:\,\|\boldsymbol{u}_k\|_2\le1}\left|\langle\boldsymbol{\mathcal{X}},(\boldsymbol{u}_1,\boldsymbol{u}_2,\boldsymbol{u}_3)\cdot1\rangle\right|.$$

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]

$$\|\boldsymbol{\mathcal{X}}\|_{\mathsf{F}}\le\sqrt{\frac{r_1r_2r_3}{r}}\|\boldsymbol{\mathcal{X}}\|,\qquad\text{where } r=\max_{k=1,2,3}r_k. \tag{4.7}$$

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

$$\mathcal{H}_{\boldsymbol{r}}(\boldsymbol{\mathcal{X}})=(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\cdot\boldsymbol{\mathcal{S}}, \tag{4.8a}$$

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 $\boldsymbol{\mathcal{S}}=(\boldsymbol{U}^\top,\boldsymbol{V}^\top,\boldsymbol{W}^\top)\cdot\boldsymbol{\mathcal{X}}$ is the core tensor.

Equivalently, we denote

$$(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}, \boldsymbol{\mathcal{S}}) = \mathrm{HOSVD}_{\boldsymbol{r}}(\boldsymbol{\mathcal{X}}) \tag{4.8b}$$

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

$$\|\boldsymbol{\mathcal{X}} - \mathcal{H}_{\boldsymbol{r}}(\boldsymbol{\mathcal{X}})\|_{\mathsf{F}} \le \sqrt{3} \inf_{\widetilde{\boldsymbol{\mathcal{X}}}:\, \mathrm{rank}(\mathcal{M}_k(\widetilde{\boldsymbol{\mathcal{X}}})) \le r_k} \|\boldsymbol{\mathcal{X}} - \widetilde{\boldsymbol{\mathcal{X}}}\|_{\mathsf{F}}. \tag{4.9}$$

There are many variants or alternatives of HOSVD in the literature, e.g. successive HOSVD, alternating least squares (ALS), higher-order orthogonal iteration (HOOI) [DLDMV00a, DLDMV00b], 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].

4.2 Main Results

4.2.1 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

$$\boldsymbol{\mathcal{X}}_\star(i_1, i_2, i_3) = \sum_{j_1=1}^{r_1} \sum_{j_2=1}^{r_2} \sum_{j_3=1}^{r_3} \boldsymbol{U}_\star(i_1, j_1) \boldsymbol{V}_\star(i_2, j_2) \boldsymbol{W}_\star(i_3, j_3) \boldsymbol{\mathcal{S}}_\star(j_1, j_2, j_3), \quad 1 \le i_k \le n_k, \tag{4.10}$$

or more compactly,

$$\boldsymbol{\mathcal{X}}_\star = (\boldsymbol{U}_\star, \boldsymbol{V}_\star, \boldsymbol{W}_\star) \cdot \boldsymbol{\mathcal{S}}_\star, \tag{4.11}$$

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

$$\mathcal{M}_1(\boldsymbol{\mathcal{X}}_\star) = \boldsymbol{U}_\star \mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{W}_\star \otimes \boldsymbol{V}_\star)^\top, \tag{4.12a}$$

$$\mathcal{M}_2(\boldsymbol{\mathcal{X}}_\star) = \boldsymbol{V}_\star \mathcal{M}_2(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{W}_\star \otimes \boldsymbol{U}_\star)^\top, \tag{4.12b}$$

$$\mathcal{M}_3(\boldsymbol{\mathcal{X}}_\star) = \boldsymbol{W}_\star \mathcal{M}_3(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{V}_\star \otimes \boldsymbol{U}_\star)^\top. \tag{4.12c}$$

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

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

We shall fix the ground truth factors 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

$$\mathcal{M}_k(\boldsymbol{\mathcal{S}}_\star)\mathcal{M}_k(\boldsymbol{\mathcal{S}}_\star)^\top = \boldsymbol{\Sigma}_{\star,k}^2, \qquad k = 1, 2, 3, \tag{4.13}$$

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

$$\kappa \coloneqq \frac{\sigma_{\max}(\boldsymbol{\mathcal{X}}_\star)}{\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)} = \frac{\max_{k=1,2,3} \sigma_{\max}(\mathcal{M}_k(\boldsymbol{\mathcal{X}}_\star))}{\min_{k=1,2,3} \sigma_{\min}(\mathcal{M}_k(\boldsymbol{\mathcal{X}}_\star))}. \tag{4.14}$$

Another parameter is the incoherence parameter, which plays an important role in governing

the well-posedness of low-rank tensor completion.

Definition 11 (Incoherence)**.** The incoherence parameter of $\boldsymbol{\mathcal{X}}_\star$ is defined as

$$\mu := \max\left\{\frac{n_1}{r_1}\|\boldsymbol{U}_\star\|_{2,\infty}^2, \frac{n_2}{r_2}\|\boldsymbol{V}_\star\|_{2,\infty}^2, \frac{n_3}{r_3}\|\boldsymbol{W}_\star\|_{2,\infty}^2\right\}. \tag{4.15}$$

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.

4.2.2 `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

$$[\mathcal{P}_\Omega(\boldsymbol{\mathcal{X}}_\star)](i_1,i_2,i_3) = \begin{cases} \boldsymbol{\mathcal{X}}_\star(i_1,i_2,i_3), & \text{if } (i_1,i_2,i_3)\in\Omega, \\ 0, & \text{otherwise.}\end{cases} \tag{4.16}$$

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

$$(i_1,i_2,i_3)\in\Omega \quad \text{independently with probability } p\in(0,1]. \tag{4.17}$$

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

$$\min_{\boldsymbol{F}=(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W},\boldsymbol{\mathcal{S}})} \mathcal{L}(\boldsymbol{F}) := \frac{1}{2p}\left\|\mathcal{P}_\Omega\big((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\cdot\boldsymbol{\mathcal{S}}\big)-\boldsymbol{\mathcal{Y}}\right\|_{\mathsf{F}}^2. \tag{4.18}$$

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 11) 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 Chapter 2, we introduce the scaled projection as follows,

$$(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}, \boldsymbol{\mathcal{S}}) = \mathcal{P}_B(\boldsymbol{U}_+, \boldsymbol{V}_+, \boldsymbol{W}_+, \boldsymbol{\mathcal{S}}_+), \tag{4.19}$$

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

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

Here, we recall $\breve{\boldsymbol{U}}_+$, $\breve{\boldsymbol{V}}_+$, $\breve{\boldsymbol{W}}_+$ are analogously defined in (4.5) 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 $(\boldsymbol{U}_+, \boldsymbol{V}_+, \boldsymbol{W}_+) \cdot \boldsymbol{\mathcal{S}}_+$. 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 4. It consists of two stages: spectral initialization followed by iterative refinements using the scaled projected gradient updates in (4.20). 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,XYZ21] for highly unbalanced matrices. Specifically, we estimate the subspace spanned by $\boldsymbol{U}_\star$ by that spanned by top-r_1 eigenvectors $\boldsymbol{U}_+$ of the diagonally-deleted Gram matrix of $p^{-1}\mathcal{M}_1(\boldsymbol{\mathcal{Y}})$, denoted as

$$\mathcal{P}_{\mathsf{off\text{-}diag}}(p^{-2}\mathcal{M}_1(\boldsymbol{\mathcal{Y}})\mathcal{M}_1(\boldsymbol{\mathcal{Y}})^\top),$$

Algorithm 4 `ScaledGD` for low-rank tensor completion

Input parameters: step size η, multilinear rank $\boldsymbol{r} = (r_1, r_2, r_3)$, probability of observation p, projection radius B.

Spectral initialization: Let $\boldsymbol{U}_+$ be the top-r_1 eigenvectors 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 $\boldsymbol{\mathcal{S}}_+ = p^{-1}(\boldsymbol{U}_+^\top, \boldsymbol{V}_+^\top, \boldsymbol{W}_+^\top)\cdot\boldsymbol{\mathcal{Y}}$. 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 $t = 0, 1, 2, \ldots, T-1$ **do**

$$\begin{aligned}
\boldsymbol{U}_{t+} &= \boldsymbol{U}_t - \eta\mathcal{M}_1(\boldsymbol{\mathcal{G}}_t)\breve{\boldsymbol{U}}_t\big(\breve{\boldsymbol{U}}_t^\top\breve{\boldsymbol{U}}_t\big)^{-1},\\
\boldsymbol{V}_{t+} &= \boldsymbol{V}_t - \eta\mathcal{M}_2(\boldsymbol{\mathcal{G}}_t)\breve{\boldsymbol{V}}_t\big(\breve{\boldsymbol{V}}_t^\top\breve{\boldsymbol{V}}_t\big)^{-1},\\
\boldsymbol{W}_{t+} &= \boldsymbol{W}_t - \eta\mathcal{M}_3(\boldsymbol{\mathcal{G}}_t)\breve{\boldsymbol{W}}_t\big(\breve{\boldsymbol{W}}_t^\top\breve{\boldsymbol{W}}_t\big)^{-1},\\
\boldsymbol{\mathcal{S}}_{t+} &= \boldsymbol{\mathcal{S}}_t - \eta\left((\boldsymbol{U}_t^\top\boldsymbol{U}_t)^{-1}\boldsymbol{U}_t^\top, (\boldsymbol{V}_t^\top\boldsymbol{V}_t)^{-1}\boldsymbol{V}_t^\top, (\boldsymbol{W}_t^\top\boldsymbol{W}_t)^{-1}\boldsymbol{W}_t^\top\right)\cdot\boldsymbol{\mathcal{G}}_t,
\end{aligned} \tag{4.20}$$

where $\boldsymbol{\mathcal{G}}_t \coloneqq \frac{1}{p}\left(\mathcal{P}_\Omega\left((\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t)\cdot\boldsymbol{\mathcal{S}}_t\right) - \boldsymbol{\mathcal{Y}}\right)$, $\breve{\boldsymbol{U}}_t$, $\breve{\boldsymbol{V}}_t$, and $\breve{\boldsymbol{W}}_t$ are defined in (4.5). Set $(\boldsymbol{U}_{t+1}, \boldsymbol{V}_{t+1}, \boldsymbol{W}_{t+1}, \boldsymbol{\mathcal{S}}_{t+1}) = \mathcal{P}_B(\boldsymbol{U}_{t+}, \boldsymbol{V}_{t+}, \boldsymbol{W}_{t+}, \boldsymbol{\mathcal{S}}_{t+})$.

and the other two factors $\boldsymbol{V}_+$ and $\boldsymbol{W}_+$ are estimated similarly. The core tensor is then estimated as

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

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 8 (`ScaledGD` for tensor completion)**.** *Suppose that $\boldsymbol{\mathcal{X}}_\star$ is μ-incoherent, $n_k \gtrsim \epsilon_0^{-1}\mu r_k^{3/2}\kappa^2$ for $k = 1, 2, 3$, and that p satisfies*

$$pn_1n_2n_3 \gtrsim \epsilon_0^{-1}\sqrt{n_1n_2n_3}\mu^{3/2}r^{5/2}\kappa^3\log^3 n + \epsilon_0^{-2}n\mu^3r^4\kappa^6\log^5 n$$

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 $C_B \geq (1+\epsilon_0)^3$. *If the step size obeys* $0 < \eta \leq 2/5$, *then with probability at least* $1 - c_1 n^{-c_2}$ *for universal constants* $c_1, c_2 > 0$, *for all* $t \geq 0$, *the iterates of Algorithm 4 satisfy*

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

Theorem 8 ensures that `ScaledGD` finds an ε-accurate estimate, i.e. $\|(\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t) \cdot \boldsymbol{\mathcal{S}}_t - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \leq \varepsilon \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$, 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

$$p n_1 n_2 n_3 \gtrsim n^{3/2} r^{5/2} \kappa^3 \log^3 n,$$

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 4.1, `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 in [XY19] does not offer an explicit iteration complexity, except that each iteration can be computed in a polynomial time.

4.2.3 `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

$$y_i = \langle \boldsymbol{\mathcal{A}}_i, \boldsymbol{\mathcal{X}}_\star \rangle, \quad i = 1, \ldots, m, \qquad \text{or concisely,} \qquad \boldsymbol{y} = \mathcal{A}(\boldsymbol{\mathcal{X}}_\star), \tag{4.21}$$

Algorithm 5 `ScaledGD` for low-rank tensor regression

Input parameters: step size η, 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 (4.8b).
Scaled gradient updates: for $t = 0, 1, 2, \ldots, T-1$

$$\begin{aligned}
\boldsymbol{U}_{t+1} &= \boldsymbol{U}_t - \eta \mathcal{M}_1(\boldsymbol{\mathcal{G}}_t)\breve{\boldsymbol{U}}_t^\top \big(\breve{\boldsymbol{U}}_t^\top \breve{\boldsymbol{U}}_t\big)^{-1}, \\
\boldsymbol{V}_{t+1} &= \boldsymbol{V}_t - \eta \mathcal{M}_2(\boldsymbol{\mathcal{G}}_t)\breve{\boldsymbol{V}}_t^\top \big(\breve{\boldsymbol{V}}_t^\top \breve{\boldsymbol{V}}_t\big)^{-1}, \\
\boldsymbol{W}_{t+1} &= \boldsymbol{W}_t - \eta \mathcal{M}_3(\boldsymbol{\mathcal{G}}_t)\breve{\boldsymbol{W}}_t^\top \big(\breve{\boldsymbol{W}}_t^\top \breve{\boldsymbol{W}}_t\big)^{-1}, \\
\boldsymbol{\mathcal{S}}_{t+1} &= \boldsymbol{\mathcal{S}}_t - \eta \Big((\boldsymbol{U}_t^\top \boldsymbol{U}_t)^{-1}\boldsymbol{U}_t^\top, (\boldsymbol{V}_t^\top \boldsymbol{V}_t)^{-1}\boldsymbol{V}_t^\top, (\boldsymbol{W}_t^\top \boldsymbol{W}_t)^{-1}\boldsymbol{W}_t^\top\Big) \cdot \boldsymbol{\mathcal{G}}_t,
\end{aligned} \tag{4.23}$$

where $\boldsymbol{\mathcal{G}}_t \coloneqq \mathcal{A}^*(\mathcal{A}((\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t) \cdot \boldsymbol{\mathcal{S}}_t) - \boldsymbol{y})$, $\breve{\boldsymbol{U}}_t$, $\breve{\boldsymbol{V}}_t$, and $\breve{\boldsymbol{W}}_t$ are defined in (4.5).

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

$$\min_{\boldsymbol{F}=(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W},\boldsymbol{\mathcal{S}})} \mathcal{L}(\boldsymbol{F}) \coloneqq \frac{1}{2} \left\| \mathcal{A}((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \cdot \boldsymbol{\mathcal{S}}) - \boldsymbol{y} \right\|_2^2. \tag{4.22}$$

The proposed `ScaledGD` algorithm to minimize (4.22) is described in Algorithm 5, where the algorithm is initialized by applying HOSVD to $\mathcal{A}^*(\boldsymbol{y})$, followed by scaled gradient updates given in (4.23).

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

$$m \gtrsim \epsilon_0^{-1} \sqrt{n_1 n_2 n_3}\, r^{3/2} \kappa^2 + \epsilon_0^{-2} (n r^2 \kappa^4 \log n + r^4 \kappa^2)$$

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 5 satisfy*

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

Theorem 9 ensures that `ScaledGD` finds an ε-accurate estimate, i.e. $\|(\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t) \cdot \boldsymbol{\mathcal{S}}_t - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \le \varepsilon \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$, 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

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

where again we keep only terms with dominating orders of n. Compared with the regularized GD [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.

4.3 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

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

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 (4.4) it is easy to check that the next iterate of `ScaledGD` is covariant with respect to invertible transforms, meaning

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

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

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

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 imbalance.

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

$$\begin{aligned}\operatorname{dist}^2(\boldsymbol{F}, \boldsymbol{F}_\star) \coloneqq \inf_{\boldsymbol{Q}_k \in \mathrm{GL}(r_k)} \|(\boldsymbol{U}\boldsymbol{Q}_1 - \boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 + \|(\boldsymbol{V}\boldsymbol{Q}_2 - \boldsymbol{V}_\star)\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}^2 + \|(\boldsymbol{W}\boldsymbol{Q}_3 - \boldsymbol{W}_\star)\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}^2 \\ + \left\|(\boldsymbol{Q}_1^{-1}, \boldsymbol{Q}_2^{-1}, \boldsymbol{Q}_3^{-1}) \cdot \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{S}}_\star\right\|_{\mathsf{F}}^2. \end{aligned} \tag{4.24}$$

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 $\operatorname{dist}(\boldsymbol{F}, \boldsymbol{F}_\star) \asymp \|(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \cdot \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}$ in the sense that (see Appendix C.1.1 for proofs):

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

4.3.1 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:

$$\mathcal{L}(\boldsymbol{F}) \coloneqq \frac{1}{2}\|(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \cdot \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}^2 = \frac{1}{2}\|\mathcal{M}_k\left((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \cdot \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_\star\right)\|_{\mathsf{F}}^2, \quad k = 1, 2, 3, \tag{4.25}$$

where the last equality follows from (4.6b). Recalling the update rule (4.4), `ScaledGD` proceeds as

$$\begin{aligned}
\boldsymbol{U}_{t+1} &= \boldsymbol{U}_t - \eta \mathcal{M}_1\left(\boldsymbol{\mathcal{X}}_t - \boldsymbol{\mathcal{X}}_\star\right) \breve{\boldsymbol{U}}_t^\top \left(\breve{\boldsymbol{U}}_t^\top \breve{\boldsymbol{U}}_t\right)^{-1}, \\
\boldsymbol{V}_{t+1} &= \boldsymbol{V}_t - \eta \mathcal{M}_2\left(\boldsymbol{\mathcal{X}}_t - \boldsymbol{\mathcal{X}}_\star\right) \breve{\boldsymbol{V}}_t^\top \left(\breve{\boldsymbol{V}}_t^\top \breve{\boldsymbol{V}}_t\right)^{-1}, \\
\boldsymbol{W}_{t+1} &= \boldsymbol{W}_t - \eta \mathcal{M}_3\left(\boldsymbol{\mathcal{X}}_t - \boldsymbol{\mathcal{X}}_\star\right) \breve{\boldsymbol{W}}_t^\top \left(\breve{\boldsymbol{W}}_t^\top \breve{\boldsymbol{W}}_t\right)^{-1}, \\
\boldsymbol{\mathcal{S}}_{t+1} &= \boldsymbol{\mathcal{S}}_t - \eta \left((\boldsymbol{U}_t^\top \boldsymbol{U}_t)^{-1}\boldsymbol{U}_t^\top, (\boldsymbol{V}_t^\top \boldsymbol{V}_t)^{-1}\boldsymbol{V}_t^\top, (\boldsymbol{W}_t^\top \boldsymbol{W}_t)^{-1}\boldsymbol{W}_t^\top\right) \cdot \left(\boldsymbol{\mathcal{X}}_t - \boldsymbol{\mathcal{X}}_\star\right),
\end{aligned} \tag{4.26}$$

where $\boldsymbol{\mathcal{X}}_t = (\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t) \cdot \boldsymbol{\mathcal{S}}_t$, with $\breve{\boldsymbol{U}}_t$, $\breve{\boldsymbol{V}}_t$, and $\breve{\boldsymbol{W}}_t$ defined in (4.5).

`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 (4.26) as

$$\operatorname{vec}(\boldsymbol{F}_{t+1}) = \operatorname{vec}(\boldsymbol{F}_t) - \eta \boldsymbol{H}_t^{-1} \nabla_{\operatorname{vec}(\boldsymbol{F})} \mathcal{L}(\boldsymbol{F}_t), \tag{4.27}$$

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 (4.25) are given precisely as

$$\begin{aligned}
\nabla^2_{\operatorname{vec}(\boldsymbol{U}),\operatorname{vec}(\boldsymbol{U})} \mathcal{L}(\boldsymbol{F}_t) &= (\breve{\boldsymbol{U}}_t^\top \breve{\boldsymbol{U}}_t) \otimes \boldsymbol{I}_{n_1}, \\
\nabla^2_{\operatorname{vec}(\boldsymbol{V}),\operatorname{vec}(\boldsymbol{V})} \mathcal{L}(\boldsymbol{F}_t) &= (\breve{\boldsymbol{V}}_t^\top \breve{\boldsymbol{V}}_t) \otimes \boldsymbol{I}_{n_2}, \\
\nabla^2_{\operatorname{vec}(\boldsymbol{W}),\operatorname{vec}(\boldsymbol{W})} \mathcal{L}(\boldsymbol{F}_t) &= (\breve{\boldsymbol{W}}_t^\top \breve{\boldsymbol{W}}_t) \otimes \boldsymbol{I}_{n_3}, \\
\nabla^2_{\operatorname{vec}(\boldsymbol{\mathcal{S}}),\operatorname{vec}(\boldsymbol{\mathcal{S}})} \mathcal{L}(\boldsymbol{F}_t) &= (\boldsymbol{W}_t^\top \boldsymbol{W}_t) \otimes (\boldsymbol{V}_t^\top \boldsymbol{V}_t) \otimes (\boldsymbol{U}_t^\top \boldsymbol{U}_t).
\end{aligned} \tag{4.28}$$

Therefore, by vectorization of (4.26), `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 10, whose proof can be found in Appendix C.2.

Theorem 10. *For tensor factorization* (4.25)*, 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* $t \geq 0$*, the iterates of* `ScaledGD` *in* (4.26) *satisfy*

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

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

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

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

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 (4.26) and the scaled distance in (4.24) reduce respectively to

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

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

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

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

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

By the optimality condition for $\boldsymbol{Q}_t$ (see Lemma 32), it follows that $\mathcal{P}_{\mathsf{diag}}{}^{1/2}(\nabla^2 \mathcal{L}(\boldsymbol{f}_\star))(\boldsymbol{Q}_t \boldsymbol{f}_t - \boldsymbol{f}_\star)$ is approximately parallel to $\boldsymbol{1}$. Thus, $\mathrm{dist}(\boldsymbol{f}_{t+1}, \boldsymbol{f}_\star)$ contracts at a constant rate as long as the step size η is set as a small constant obeying $0 < \eta \leq 2/5$.

4.3.2 Proof outline for tensor completion (Theorem 8)

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 9 (Initialization for tensor completion). *Suppose that $\boldsymbol{\mathcal{X}}_\star$ is μ-incoherent, $n_k \gtrsim \epsilon_0^{-1} \mu r_k^{3/2} \kappa^2$ for $k = 1, 2, 3$, and that p satisfies*

$$p n_1 n_2 n_3 \gtrsim \epsilon_0^{-1} \sqrt{n_1 n_2 n_3}\, \mu^{3/2} r^{5/2} \kappa^2 \log^3 n + \epsilon_0^{-2} n \mu^2 r^4 \kappa^4 \log^5 n$$

for some small constant $\epsilon_0 > 0$. Then with overwhelming probability (i.e. at least $1 - c_1 n^{-c_2}$), the spectral initialization before projection $\boldsymbol{F}_+ = (\boldsymbol{U}_+, \boldsymbol{V}_+, \boldsymbol{W}_+, \boldsymbol{S}_+)$ for low-rank tensor completion in Algorithm 4 satisfies

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

Under a suitable sample size condition, Lemma 9 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 10 (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 $\boldsymbol{F} = (\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}, \boldsymbol{\mathcal{S}}) := \mathcal{P}_B(\boldsymbol{F}_+)$ satisfies the non-expansiveness property*

$$\mathrm{dist}(\boldsymbol{F}, \boldsymbol{F}_\star) \le \mathrm{dist}(\boldsymbol{F}_+, \boldsymbol{F}_\star),$$

and the incoherence condition

$$\sqrt{n_1}\|\boldsymbol{U}\breve{\boldsymbol{U}}^\top\|_{2,\infty} \vee \sqrt{n_2}\|\boldsymbol{V}\breve{\boldsymbol{V}}^\top\|_{2,\infty} \vee \sqrt{n_3}\|\boldsymbol{W}\breve{\boldsymbol{W}}^\top\|_{2,\infty} \le B. \tag{4.29}$$

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

Lemma 11 (Local refinements for tensor completion)**.** *Suppose that $\boldsymbol{\mathcal{X}}_\star$ is μ-incoherent, and that p satisfies*

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

Under an event $\mathcal{E}$ which happens with overwhelming probability, 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* $\|(\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t)\cdot\boldsymbol{\mathcal{S}}_t - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \le 3\,\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$*. 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 4 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 9 and Lemma 10, we can ensure that the spectral initialization $\boldsymbol{F}_0 = \mathcal{P}_B(\boldsymbol{F}_+)$ satisfies the conditions required in Lemma 11, which further enables us to repetitively apply Lemma 11 to finish the proof of Theorem 8. The proofs of the above three lemmas are provided in Appendix C.3.

4.3.3 Proof outline for tensor regression (Theorem 9)

Now we turn to the proof outline for tensor regression (cf. Theorem 9). 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 12 (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 2] 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 12 (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 $\|(\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t)\cdot\boldsymbol{\mathcal{S}}_t - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \le 3\operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$. In addition, if the step size obeys $0 < \eta < 2/5$, then the $(t+1)$-th iterate of Algorithm 5 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 [LZ21, HWZ20, ZLRY20]. Under the Gaussian design, we have the following guarantee for the spectral initialization scheme that invokes HOSVD in Algorithm 5.

Lemma 13 (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*

$$m \gtrsim \epsilon_0^{-1}\sqrt{n_1 n_2 n_3}\, r^{3/2}\kappa^2 + \epsilon_0^{-2}(nr^2\kappa^4 \log n + r^4\kappa^2)$$

for some small constant $\epsilon_0 > 0$. Then with overwhelming probability, the spectral initialization for low-rank tensor regression in Algorithm 5 satisfies

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

Combining Lemma 12 and Lemma 13 finishes the proof of Theorem 9. Their proofs can be found in Appendix C.4.

4.4 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. Their algorithm was originally proposed for tenser regression, nevertheless, it naturally applies to tensor completion and exhibits similar results. Since the scaled projection does not visibly impact the performance, we implement `ScaledGD` without performing the projection. Also, we empirically find that the regularization used in [HWZ20] has no visible benefits, hence we implement GD without the regularization. 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 $\boldsymbol{\mathcal{X}}_\star = (\boldsymbol{U}_\star, \boldsymbol{V}_\star, \boldsymbol{W}_\star) \cdot \boldsymbol{\mathcal{S}}_\star$ 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 4.2 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 GD. We next compare the performance of `ScaledGD` with GD. For a fair comparison, both `ScaledGD` and GD start from the same spectral initialization, and we use the

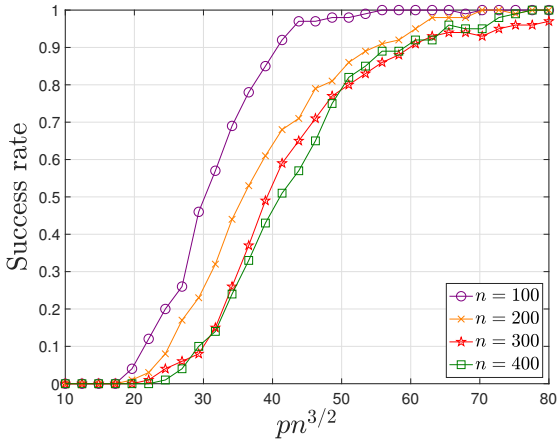


Figure 4.2: The success rate of `ScaledGD` with respect to the scaled sample size for tensor completion with $r = 5$, when the core tensor is composed of i.i.d. standard Gaussian entries, for various tensor size n.

following update rule of GD as

$$\begin{aligned} \boldsymbol{U}_{t+1} &= \boldsymbol{U}_t - \eta\sigma_{\max}^{-2}(\boldsymbol{\mathcal{X}}_\star)\nabla_{\boldsymbol{U}}\mathcal{L}(\boldsymbol{F}_t), \\ \boldsymbol{V}_{t+1} &= \boldsymbol{V}_t - \eta\sigma_{\max}^{-2}(\boldsymbol{\mathcal{X}}_\star)\nabla_{\boldsymbol{V}}\mathcal{L}(\boldsymbol{F}_t), \\ \boldsymbol{W}_{t+1} &= \boldsymbol{W}_t - \eta\sigma_{\max}^{-2}(\boldsymbol{\mathcal{X}}_\star)\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{F}_t), \\ \boldsymbol{\mathcal{S}}_{t+1} &= \boldsymbol{\mathcal{S}}_t - \eta\nabla_{\boldsymbol{\mathcal{S}}}\mathcal{L}(\boldsymbol{F}_t). \end{aligned} \tag{4.30}$$

Throughout the experiments, we used the ground truth value $\sigma_{\max}(\boldsymbol{\mathcal{X}}_\star)$ in running (4.30), while in practice, this parameter needs to estimated; to put it differently, the step size of 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 $\{\sigma_{j_1}\}_{1\le j_1\le r}$ take values spaced equally from 1 to $1/\kappa$. It then follows that $\sigma_{\max}(\boldsymbol{\mathcal{X}}_\star) = 1$, $\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star) = 1/\kappa$, and the condition number of $\boldsymbol{\mathcal{X}}_\star$ is exactly κ.

Figure 4.3 illustrates the convergence speed of `ScaledGD` and 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 GD

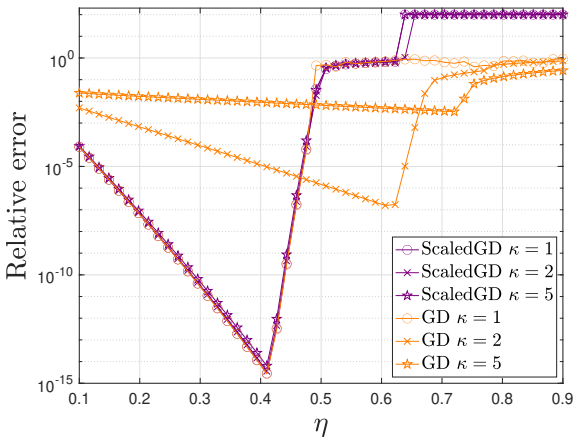
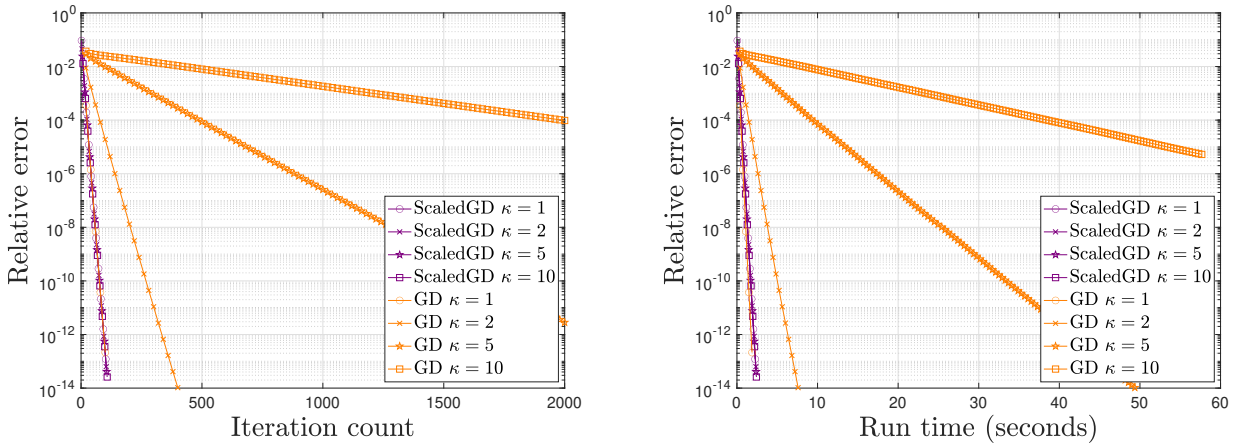


Figure 4.3: The relative errors of `ScaledGD` and GD after 80 iterations with respect to different step sizes η from 0.1 to 0.9 for tensor completion with $n = 100$, $r = 5$, $p = 0.1$.

quite significantly even when the step size of GD is optimized for its performance. Hence, we will fix $\eta = 0.3$ for the rest of the comparisons for both `ScaledGD` and GD without hurting the conclusions.



(a) (b)

Figure 4.4: The relative errors of `ScaledGD` and 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 $p = 0.1$.

Figure 4.4 compares the relative errors of `ScaledGD` and 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 matches the fastest rate of GD with perfect conditioning $\kappa = 1$. In contrast, the convergence rate of 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.

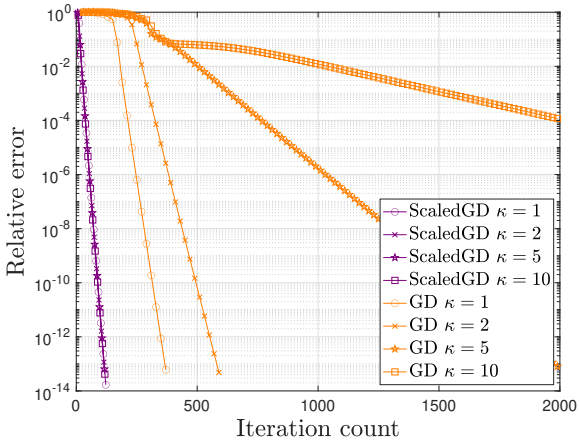


Figure 4.5: The relative errors of random-initialized `ScaledGD` and GD with respect to the iteration count under different condition numbers $\kappa = 1, 2, 5, 10$ for tensor completion with $n = 100$, $r = 5$, $p = 0.1$.

We next examine the performance of `ScaledGD` and GD when randomly initialized. Here, we initialize $\boldsymbol{U}_0, \boldsymbol{V}_0, \boldsymbol{W}_0$ composed of i.i.d. entries sampled from $\mathcal{N}(0, 1/n)$, and $\boldsymbol{\mathcal{S}}_0$ composed of i.i.d. entries sampled from $\mathcal{N}(0, \|\boldsymbol{\mathcal{Y}}\|_{\mathsf{F}}^2/(pr^3))$. Figure 4.5 plots the relative errors of `ScaledGD` and GD under different condition numbers $\kappa = 1, 2, 5, 10$, using the same random initialization. Surprisingly, while GD gets 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.

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 $\boldsymbol{\mathcal{W}}(i_1, i_2, i_3) \sim \mathcal{N}(0, \sigma_w^2)$ composed of i.i.d. Gaussian entries. Denote the signal-to-noise ratio as $\mathrm{SNR} \coloneqq 10 \log_{10} \frac{\|\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}^2}{n^3 \sigma_w^2}$ in dB. Figure 4.6 demonstrates the robustness of `ScaledGD`, by plotting the relative errors with respect to the iteration count under $\mathrm{SNR} = 40, 60, 80$dB. Here, the ground truth tensor $\boldsymbol{\mathcal{X}}_\star$ is constructed in the same manner as Figure 4.2, where its condition number is

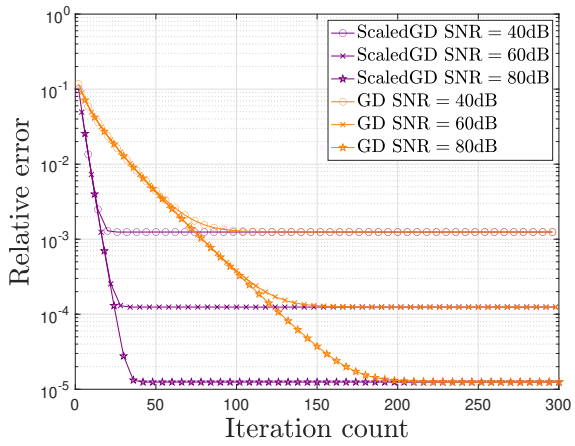


Figure 4.6: The relative errors of `ScaledGD` and GD with respect to the iteration count under signal-to-noise ratios SNR $= 40, 60, 80$dB for tensor completion with $n = 100$, $r = 5$, and $p = 0.1$.

approximately $\kappa \approx 2.6$. It can been seen that `ScaledGD` reaches the same statistical error as GD, but at a much faster rate. In addition, the convergence speeds are not impacted by the noise levels.

4.5 Discussions

This chapter develops `ScaledGD` 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 chapter, 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 [MWCC19] 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.

- *Random initialization?* As evident from the numerical experiment in Figure 4.5, `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.

Chapter 5

Robust Low-rank Tensor Estimation

5.1 Introduction

The modern data deluge has created a growing number of applications involving multi-dimensional or multi-attribute datasets, examples including video surveillance, hyperspectral imaging, neuroimaging, social network analysis, and so on. Tensors arise naturally as a suitable data structure that captures the underlying multi-way interactions, offering advantages over the matrix counterpart [KB09, SDLF+17]. An important problem, known as tensor regression, that arises frequently across different applications is to recover a tensor from a small number of its linear measurements, given by

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

where $\boldsymbol{\mathcal{X}}_\star \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_K}$ is a K-way tensor, $\boldsymbol{y} \in \mathbb{R}^m$ is the collected measurements, and $\mathcal{A}(\cdot)$ is a linear map that models the data collection process. For ease of presentation, we consider the case $K = 3$ throughout the paper, while our results hold for the general case without difficulty.

In practice, due to sensor failures and malicious attacks, it is common that the collected measurements may suffer from undesirable and unknown corruptions, which are possibly adversarial. Consequently, there is an imminent need to develop low-rank tensor recovery algorithms that are provably robust and efficient, which are still lacking. To fill the gap, instead of minimizing the smooth loss function in (5.3), which is known to be vulnerable to outliers, we resort to the least absolute deviations (LAD) loss, which measures the residual sum of absolute errors:

$$\min_{\boldsymbol{F}=(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W},\boldsymbol{\mathcal{C}})} \left\|\mathcal{A}((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\cdot\boldsymbol{\mathcal{C}}) - \boldsymbol{y}\right\|_1. \tag{5.1}$$

Leveraging recent insights in preconditioning for ill-conditioned low-rank matrix and tensor estimation [TMC21a, TMC21b, TMPB+21], we propose an efficient algorithm for solving the nonconvex composite optimization problem in (5.3), namely the scaled subgradient method (`ScaledSM`), which incorporates carefully-designed preconditioners in the local updates to preserve the equivariance of the low-rank parameterization. Under the Gaussian design, the proposed method provably finds the ground truth at a constant linear rate that is *independent of the condition number* even under a constant fraction of outliers, as long as it is initialized properly. The algorithm is much more scalable than its counterpart without the preconditioners, especially when the ground truth tensor is ill-conditioned. To the best of our knowledge, our work provides the first provable algorithm that achieves robust low-rank tensor regression from corrupted measurements, together with a fast rate of convergence independent of the condition number of the ground truth tensor.

5.1.1 Related works

Low-rank tensor recovery has attracted significant research interest in recent years, where many algorithms have been developed with provable performance guarantees, e.g. [RYC19, HMGW15, BM16, RSS17, CRY19, ZLRY20, HWZ20, TMPB+21, CLPC19, LM20]. Moreover, spectral methods [MS18, CLC+21, CCFM21] are often applied to provide a smart initialization from which iterative algorithms refine locally to enable global convergence despite the presence of nonconvexity. However, a majority of these algorithms are designed with respect to the smooth least-squares loss and therefore their performance is very sensitive to the existence of outliers.

Motivated by the success of robust principal component analysis for the matrix setting [CLMW11], convex relaxation approaches are proposed in [GQ14, HMGW15, LFC+16] via unfolding the tensor of interest and invoking matrix-based algorithms. However, their computational complexity is often prohibitive for large-scale problems. On the other end, the LAD loss is not new to handle outliers, and has been adopted for high-dimensional signal recovery [LSC17, DR19, CDDD19, LZMCSV20, TMC21b, CCD+21, MF21], where the subgradient method has been analyzed in [DR19, CCD+21, MF21]. Another popular strategy is to adaptively truncate or prune outliers in an iterative manner guided by quantile statistics, as done in [ZCL16, LCZL20, ZCL18, YPCC16].

The preconditioner design in our approach is directly inspired by `ScaledGD` method to optimize the smooth loss function (5.3) for low-rank tensor regression. In particular, the proposed subgradient method can be viewed as the tensor counterpart of Chapter 3, which generalizes the preconditioner designs to the nonsmooth setting.

5.1.2 Chapter organization

The rest of this chapter is organized as follows. Section 5.2 describes the problem formulation as well as the proposed algorithms. Section 5.3 provides the theoretical guarantees in terms of both statistical and computational complexities. Section 5.4 illustrates the performance of the proposed algorithms through numerical examples. Finally, we conclude in Section 5.5.

5.2 Formulation and Proposed Algorithms

Let $\boldsymbol{\mathcal{X}}_\star := [\boldsymbol{\mathcal{X}}_\star(i_1, i_2, i_3)] \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ be the ground truth tensor that satisfies the Tucker decomposition in (4.10). Consider the robust low-rank tensor regression problem, in which the measurements are corrupted by sparse outliers. Specifically, assume that we have access to a set of linear observations of $\boldsymbol{\mathcal{X}}_\star$, where the measurement vector $\boldsymbol{y} = \{y_i\}_{i=1}^m$ is given as

$$\boldsymbol{y} = \mathcal{A}(\boldsymbol{\mathcal{X}}_\star) + \boldsymbol{s}, \tag{5.2}$$

where $\mathcal{A}(\boldsymbol{\mathcal{X}}_\star) = \{\langle \boldsymbol{\mathcal{A}}_i, \boldsymbol{\mathcal{X}}_\star \rangle\}_{i=1}^m$ is the measurement operator, with $\boldsymbol{\mathcal{A}}_i \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ denoting the i-th sensing tensor, and $\boldsymbol{s} = \{s_i\}_{i=1}^m$ corresponds to the outlier vector. We assume the outlier $\boldsymbol{s}$ is a sparse vector obeying $\|\boldsymbol{s}\|_0 = p_s m$ for some $0 \le p_s \le 1$, which means that $\|\boldsymbol{s}\|_0$ is much smaller than its ambient dimension m, so that only a small fraction p_s of the measurements are corrupted. However, the corrupted entries can take arbitrary or adversarial magnitudes. The goal is to recover the low-rank tensor $\boldsymbol{\mathcal{X}}_\star$ from $\boldsymbol{y}$ in a robust and scalable manner.

To cope with the outliers, it is natural to minimize the least absolute deviation (LAD) loss of

the measurements, given by

$$f(\boldsymbol{\mathcal{X}}) \coloneqq \|\mathcal{A}(\boldsymbol{\mathcal{X}}) - \boldsymbol{y}\|_1 = \sum_{i=1}^{m} |\langle \boldsymbol{\mathcal{A}}_i, \boldsymbol{\mathcal{X}} \rangle - \boldsymbol{y}_i| \,. \tag{5.3}$$

In addition, to take advantage of the low-rank structure and minimize complexity, we factorize the tensor $\boldsymbol{\mathcal{X}} = (\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \cdot \boldsymbol{\mathcal{C}}$ with $\boldsymbol{U} \in \mathbb{R}^{n_1 \times r_1}$, $\boldsymbol{V} \in \mathbb{R}^{n_2 \times r_2}$, $\boldsymbol{W} \in \mathbb{R}^{n_3 \times r_3}$ and $\boldsymbol{\mathcal{C}} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$, and optimize the factors directly via the following unconstrained composite optimization problem:

$$\min_{\boldsymbol{F}=(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W},\boldsymbol{\mathcal{C}})} \mathcal{L}(\boldsymbol{F}) \coloneqq f\big((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \cdot \boldsymbol{\mathcal{C}}\big), \tag{5.4}$$

which is nonconvex and nonsmooth.

A natural idea to optimize (5.4) is via subgradient descent, which updates the factor quadruple iteratively according to

$$\begin{aligned}
\boldsymbol{U}_{t+1} &= \boldsymbol{U}_t - \eta_t \mathcal{M}_1(\boldsymbol{\mathcal{G}}_t) \breve{\boldsymbol{U}}_t, \\
\boldsymbol{V}_{t+1} &= \boldsymbol{V}_t - \eta_t \mathcal{M}_2(\boldsymbol{\mathcal{G}}_t) \breve{\boldsymbol{V}}_t, \\
\boldsymbol{W}_{t+1} &= \boldsymbol{W}_t - \eta_t \mathcal{M}_3(\boldsymbol{\mathcal{G}}_t) \breve{\boldsymbol{W}}_t, \\
\boldsymbol{\mathcal{C}}_{t+1} &= \boldsymbol{\mathcal{C}}_t - \eta_t \Big(\boldsymbol{U}_t^\top, \boldsymbol{V}_t^\top, \boldsymbol{W}_t^\top\Big) \cdot \boldsymbol{\mathcal{G}}_t.
\end{aligned} \tag{5.5}$$

where $\eta_t > 0$ is the step size, $\boldsymbol{\mathcal{G}}_t = \mathcal{A}^*(\operatorname{sgn}(\mathcal{A}(\boldsymbol{\mathcal{X}}_t)) - \boldsymbol{y}) \in \partial_{\boldsymbol{\mathcal{X}}} f(\boldsymbol{\mathcal{X}}_t)$ is a subgradient of $f(\boldsymbol{\mathcal{X}})$ with respect to $\boldsymbol{\mathcal{X}}$ at $\boldsymbol{\mathcal{X}}_t = (\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t) \cdot \boldsymbol{\mathcal{C}}_t$, and $\mathcal{A}^*(\cdot)$ is the adjoint operator of $\mathcal{A}(\cdot)$. Furthermore, the following short-hand notation is introduced:

$$\breve{\boldsymbol{U}}_t \coloneqq (\boldsymbol{W}_t \otimes \boldsymbol{V}_t) \mathcal{M}_1(\boldsymbol{\mathcal{C}}_t)^\top, \tag{5.6a}$$

$$\breve{\boldsymbol{V}}_t \coloneqq (\boldsymbol{W}_t \otimes \boldsymbol{U}_t) \mathcal{M}_2(\boldsymbol{\mathcal{C}}_t)^\top, \tag{5.6b}$$

$$\breve{\boldsymbol{W}}_t \coloneqq (\boldsymbol{V}_t \otimes \boldsymbol{U}_t) \mathcal{M}_3(\boldsymbol{\mathcal{C}}_t)^\top. \tag{5.6c}$$

While simple and straightforward, this approach tends to converge very slowly when the tensor is ill-conditioned. Inspired by Chapter 4, we propose to update the iterate along a preconditioned or

scaled direction of the subgradient, leading to the following scaled subgradient method (`ScaledSM`):

$$\begin{aligned}
\boldsymbol{U}_{t+1} &= \boldsymbol{U}_t - \eta_t \mathcal{M}_1(\boldsymbol{\mathcal{G}}_t)\breve{\boldsymbol{U}}_t(\breve{\boldsymbol{U}}_t^\top \breve{\boldsymbol{U}}_t)^{-1}, \\
\boldsymbol{V}_{t+1} &= \boldsymbol{V}_t - \eta_t \mathcal{M}_2(\boldsymbol{\mathcal{G}}_t)\breve{\boldsymbol{V}}_t(\breve{\boldsymbol{V}}_t^\top \breve{\boldsymbol{V}}_t)^{-1}, \\
\boldsymbol{W}_{t+1} &= \boldsymbol{W}_t - \eta_t \mathcal{M}_3(\boldsymbol{\mathcal{G}}_t)\breve{\boldsymbol{W}}_t(\breve{\boldsymbol{W}}_t^\top \breve{\boldsymbol{W}}_t)^{-1}, \\
\boldsymbol{\mathcal{C}}_{t+1} &= \boldsymbol{\mathcal{C}}_t - \eta_t \Big((\boldsymbol{U}_t^\top \boldsymbol{U}_t)^{-1}\boldsymbol{U}_t^\top, (\boldsymbol{V}_t^\top \boldsymbol{V}_t)^{-1}\boldsymbol{V}_t^\top, (\boldsymbol{W}_t^\top \boldsymbol{W}_t)^{-1}\boldsymbol{W}_t^\top \Big) \cdot \boldsymbol{\mathcal{G}}_t.
\end{aligned} \tag{5.7}$$

Step size schedules. We still need to specify the choice of the step size $\eta_t > 0$, which needs to be carefully scheduled in accordance with the scaled update. Specifically, we apply a geometrically decaying learning rate schedule [Gof77] with proper scaling,

$$\eta_t \coloneqq \frac{\lambda q^t}{N_t}, \tag{5.8}$$

where $q \in (0,1)$, $\lambda > 0$ and

$$\begin{aligned}
N_t^2 \coloneqq \Big\| \mathcal{M}_1(\boldsymbol{\mathcal{G}}_t)\breve{\boldsymbol{U}}_t(\breve{\boldsymbol{U}}_t^\top \breve{\boldsymbol{U}}_t)^{-1/2} \Big\|_{\mathsf{F}}^2 + \Big\| \mathcal{M}_2(\boldsymbol{\mathcal{G}}_t)\breve{\boldsymbol{V}}_t(\breve{\boldsymbol{V}}_t^\top \breve{\boldsymbol{V}}_t)^{-1/2} \Big\|_{\mathsf{F}}^2 + \Big\| \mathcal{M}_3(\boldsymbol{\mathcal{G}}_t)\breve{\boldsymbol{W}}_t(\breve{\boldsymbol{W}}_t^\top \breve{\boldsymbol{W}}_t)^{-1/2} \Big\|_{\mathsf{F}}^2 \\
+ \Big\| \Big((\boldsymbol{U}_t^\top \boldsymbol{U}_t)^{-1/2}\boldsymbol{U}_t^\top, (\boldsymbol{V}_t^\top \boldsymbol{V}_t)^{-1/2}\boldsymbol{V}_t^\top, (\boldsymbol{W}_t^\top \boldsymbol{W}_t)^{-1/2}\boldsymbol{W}_t^\top \Big) \cdot \boldsymbol{\mathcal{G}}_t \Big\|_{\mathsf{F}}^2.
\end{aligned} \tag{5.9}$$

In fact, N_t can be viewed as the norm of the subgradient under a scaled metric compatible with our preconditioners. This choice is informed by our theory.

Remark 7. Ideally, one might be tempted to apply the Polyak's step size, given by $\eta_t \coloneqq \frac{f(\boldsymbol{\mathcal{X}}_t) - f(\boldsymbol{\mathcal{X}}_\star)}{N_t^2}$. However, it is impractical due to the unknown optimal function value $f(\boldsymbol{\mathcal{X}}_\star)$. As illustrated in Chapter 3, geometric step size achieves the same performance as Polyak's step size when parameters λ, q are tuned appropriately.

Equivariance to low-rank parameterization. A crucial property of `ScaledSM` is that the update of the low-rank tensor $\boldsymbol{\mathcal{X}}_t$ is invariant w.r.t. the low-rank parameterization. Suppose that at the t-th iteration, we reparameterize the factor $\boldsymbol{F}_t = (\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t, \boldsymbol{\mathcal{C}}_t)$ by

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

via any invertible matrices $\boldsymbol{Q}_k \in \mathrm{GL}(r_k)$, $k = 1, 2, 3$, where both $\boldsymbol{F}_t$ and $\widetilde{\boldsymbol{F}}_t$ correspond to the same low-rank tensor $\boldsymbol{\mathcal{X}}_t = (\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t) \cdot \boldsymbol{\mathcal{C}}_t$. By checking (5.7) and (5.8), it is straightforward to verify that the next iterate from $\widetilde{\boldsymbol{F}}_t$ follow the same change of parameterization, i.e.

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

which ensures the update rule of `ScaledSM` is insensitive to the imbalance of the factors in the low-rank parameterization—a key property that is absent in the vanilla subgradient method and contributes to the performance gain.

5.2.1 Truncated spectral initialization

Inspired by the median-truncated spectral initialization in [ZCL16, LCZL20, ZCL18], we propose a tensor counterpart that is tailored to our problem to initialize `ScaledSM`. Denote $\boldsymbol{y}_{\mathsf{trunc}}$ as the vector after discarding p_s fraction of measurements with largest magnitudes:

$$[\boldsymbol{y}_{\mathsf{trunc}}]_i = \begin{cases} \frac{y_i}{1-p_s}, & \text{if} \quad |y_i| \le |\boldsymbol{y}|_{(\lceil p_s m \rceil)} \\ 0, & \text{otherwise} \end{cases}, \tag{5.10}$$

where $|\boldsymbol{y}|_{(k)}$ denotes the k-th largest amplitude of $\boldsymbol{y}$. Let $\mathcal{A}^*(\cdot)$ be the adjoint operator of $\mathcal{A}(\cdot)$. The truncated spectral initialization $\boldsymbol{F}_0 = (\boldsymbol{U}_0, \boldsymbol{V}_0, \boldsymbol{W}_0, \boldsymbol{\mathcal{C}}_0)$ is then given by

$$(\boldsymbol{U}_0, \boldsymbol{V}_0, \boldsymbol{W}_0) \cdot \boldsymbol{\mathcal{C}}_0 = \mathcal{H}_{\boldsymbol{r}}(\mathcal{A}^*(\boldsymbol{y}_{\mathsf{trunc}})), \tag{5.11}$$

i.e. the top-$\boldsymbol{r}$ higher-order SVD (HOSVD) of $\mathcal{A}^*(\boldsymbol{y}_{\mathsf{trunc}})$. More specifically, $\boldsymbol{U}_0$ (resp. $\boldsymbol{V}_0$ and $\boldsymbol{W}_0$) is the top-r_1 (resp. r_2 and r_3) left singular vectors of $\mathcal{M}_1(\mathcal{A}^*(\boldsymbol{y}_{\mathsf{trunc}}))$ (resp. $\mathcal{M}_2(\mathcal{A}^*(\boldsymbol{y}_{\mathsf{trunc}}))$ and $\mathcal{M}_3(\mathcal{A}^*(\boldsymbol{y}_{\mathsf{trunc}}))$), and $\boldsymbol{\mathcal{C}}_0 = (\boldsymbol{U}_0^\top, \boldsymbol{V}_0^\top, \boldsymbol{W}_0^\top) \cdot \mathcal{A}^*(\boldsymbol{y}_{\mathsf{trunc}})$ is the core tensor. The full algorithm is stated in Algorithm 6.

Algorithm 6 `ScaledSM` for low-rank tensor recovery

Input parameters: parameters λ, q, multilinear rank $\boldsymbol{r} = (r_1, r_2, r_3)$, fraction of outlier p_s.
Truncated spectral initialization: Let $(\boldsymbol{U}_0, \boldsymbol{V}_0, \boldsymbol{W}_0) \cdot \boldsymbol{\mathcal{C}}_0 = \mathcal{H}_{\boldsymbol{r}}(\mathcal{A}^*(\boldsymbol{y}_{\mathsf{trunc}}))$, with $\boldsymbol{y}_{\mathsf{trunc}}$ defined in (5.10).
Scaled subgradient updates: for $t = 0, 1, 2, \ldots, T-1$ **do**

$$\begin{aligned}
\boldsymbol{U}_{t+1} &\coloneqq \boldsymbol{U}_t - \eta_t \mathcal{M}_1(\boldsymbol{\mathcal{G}}_t)\breve{\boldsymbol{U}}_t(\breve{\boldsymbol{U}}_t^\top \breve{\boldsymbol{U}}_t)^{-1}, \\
\boldsymbol{V}_{t+1} &\coloneqq \boldsymbol{V}_t - \eta_t \mathcal{M}_2(\boldsymbol{\mathcal{G}}_t)\breve{\boldsymbol{V}}_t(\breve{\boldsymbol{V}}_t^\top \breve{\boldsymbol{V}}_t)^{-1}, \\
\boldsymbol{W}_{t+1} &\coloneqq \boldsymbol{W}_t - \eta_t \mathcal{M}_3(\boldsymbol{\mathcal{G}}_t)\breve{\boldsymbol{W}}_t(\breve{\boldsymbol{W}}_t^\top \breve{\boldsymbol{W}}_t)^{-1}, \\
\boldsymbol{\mathcal{C}}_{t+1} &\coloneqq \boldsymbol{\mathcal{C}}_t - \eta_t \left((\boldsymbol{U}_t^\top \boldsymbol{U}_t)^{-1}\boldsymbol{U}_t^\top, (\boldsymbol{V}_t^\top \boldsymbol{V}_t)^{-1}\boldsymbol{V}_t^\top, (\boldsymbol{W}_t^\top \boldsymbol{W}_t)^{-1}\boldsymbol{W}_t^\top\right) \cdot \boldsymbol{\mathcal{G}}_t,
\end{aligned} \tag{5.12}$$

where $\boldsymbol{\mathcal{G}}_t \coloneqq \mathcal{A}^*(\mathrm{sgn}(\mathcal{A}((\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t) \cdot \boldsymbol{\mathcal{C}}_t) - \boldsymbol{y}))$, $\breve{\boldsymbol{U}}_t$, $\breve{\boldsymbol{V}}_t$, $\breve{\boldsymbol{W}}_t$ are defined in (5.6), and $\eta_t = \lambda q^t / N_t$ is defined in (5.8).

5.3 Theoretical Guarantees

We focus on presenting the local linear convergence of the proposed scaled subgradient method while leaving a complete account of global convergence to the future work.

5.3.1 A general theory of local linear convergence

Our convergence guarantees are built on standard geometric assumptions [DR19,CCD+21,TMC21b] on the loss function $f(\cdot)$ for the analysis of subgradient-type algorithms, which are defined as follows.

Definition 13 (Restricted Lipschitz continuity)**.** A function $f : \mathbb{R}^{n_1 \times n_2 \times n_3} \mapsto \mathbb{R}$ is said to be rank-$\boldsymbol{r}$ restricted L-Lipschitz continuous for some quantity $L > 0$ if

$$|f(\boldsymbol{\mathcal{X}}_1) - f(\boldsymbol{\mathcal{X}}_2)| \le L\|\boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_2\|_{\mathsf{F}}$$

holds for any $\boldsymbol{\mathcal{X}}_1, \boldsymbol{\mathcal{X}}_2 \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ such that $\boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_2$ has multilinear rank at most $2\boldsymbol{r}$.

Definition 14 (Restricted sharpness)**.** A function $f : \mathbb{R}^{n_1 \times n_2 \times n_3} \mapsto \mathbb{R}$ is said to be rank-$\boldsymbol{r}$ restricted μ-sharp w.r.t. $\boldsymbol{\mathcal{X}}_\star$ for some $\mu > 0$ if

$$f(\boldsymbol{\mathcal{X}}) - f(\boldsymbol{\mathcal{X}}_\star) \ge \mu\|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}$$

holds for any $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ with multilinear rank at most $\boldsymbol{r}$.

The condition number of a function $f(\cdot)$ that is both restricted L-Lipschitz continuous and μ-sharp is then denoted by

$$\chi_f \coloneqq L/\mu. \tag{5.13}$$

To fully capture the performance progress of `ScaledSM`, we measure the performance of factor quadruple $\boldsymbol{F} = (\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}, \boldsymbol{\mathcal{C}})$ using the following error metric (4.24), which takes into consideration both the representation ambiguity of the factorization up to invertible transforms and the scaling of different factors due to the presence of preconditioners. With this metric in place, we state the linear convergence of the scaled subgradient method when $f(\cdot)$ satisfies both the rank-$\boldsymbol{r}$ restricted L-Lipschitz continuity and μ-sharpness, as follows.

Theorem 11 (Scaled subgradient method with exact convergence)**.** *Suppose that $f(\boldsymbol{\mathcal{X}}) : \mathbb{R}^{n_1 \times n_2 \times n_3} \mapsto \mathbb{R}$ is convex in $\boldsymbol{\mathcal{X}}$, and satisfies rank-$\boldsymbol{r}$ restricted L-Lipschitz continuity and μ-sharpness (cf. Definitions 13 and 14). In addition, suppose that the initialization $\boldsymbol{F}_0$ satisfies*

$$\operatorname{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \leq 10^{-3} \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)/\chi_f, \tag{5.14}$$

and the scaled subgradient method adopts the geometrically decaying step sizes in (5.8) with $\lambda = \frac{(\sqrt{2}-1)^{3/2}}{2} 10^{-3} \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)/\chi_f^2$ and $q = (1 - 0.016/\chi_f^2)^{1/2}$. Then for all $t \geq 0$, the iterates satisfy

$$\operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \leq (1 - 0.016/\chi_f^2)^{t/2} 10^{-3} \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)/\chi_f, \quad \text{and} \quad \|\boldsymbol{\mathcal{X}}_t - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \leq 3 \operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star).$$

Theorem 11 shows that the iterates of the scaled subgradient method converges at a linear rate; to reach ϵ-accuracy, i.e. $\|\boldsymbol{\mathcal{X}}_t - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \leq \epsilon \sigma_r(\boldsymbol{\mathcal{X}}_\star)$, it takes at most $O(\chi_f^2 \log \frac{1}{\epsilon})$ iterations, which, importantly, is *independent of the condition number* κ of $\boldsymbol{\mathcal{X}}_\star$. Finally, it is worth noting that the choices of constants in Theorem 11 are pessimistic to simplify analysis.

5.3.2 Case study: Gaussian design

It turns out that under the Gaussian design, where all the sensing tensors are composed of i.i.d. Gaussian entries, the resulting loss function obeys the rank-$\boldsymbol{r}$ restricted L-Lipschitz continuity and μ-sharpness with high probability.

Proposition 5 (Gaussian designs). *Let $n \coloneqq \max\{n_1, n_2, n_3\}$ and $r \coloneqq \max\{r_1, r_2, r_3\}$. Suppose that $[\mathcal{A}(\boldsymbol{\mathcal{X}})]_i = \frac{1}{m}\langle \boldsymbol{\mathcal{A}}_i, \boldsymbol{\mathcal{X}}\rangle$ with tensors $\boldsymbol{\mathcal{A}}_1, \ldots, \boldsymbol{\mathcal{A}}_m$ composed of i.i.d. standard Gaussian entries. Then with probability exceeding $1 - c_1 n^{-c_2}$, the loss function $f(\boldsymbol{\mathcal{X}}) = \|\mathcal{A}(\boldsymbol{\mathcal{X}}) - \boldsymbol{y}\|_1$ in (5.3) satisfies the rank-$\boldsymbol{r}$ restricted L-Lipschitz continuity and μ-sharpness with*

$$L = 0.8, \quad \mu = 0.79(1 - 2p_s), \tag{5.15}$$

as long as $m \geq \frac{C(nr+r^3)}{(1-2p_s)^2} \log\left(\frac{1}{1-2p_s}\right)$. Here, C, c_1, c_2 are some universal constants.

Combining Theorem 11 and Proposition 5, it is guaranteed that `ScaledSM` reaches ϵ-accuracy in at most $O\left(\frac{1}{(1-2p_s)^2}\log\frac{1}{\epsilon}\right)$ iterations, as long as the sample size is sufficiently large. This amounts to a near-optimal sample complexity $O(nr + r^3)$ and dimension-free iteration complexity $O(\log\frac{1}{\epsilon})$ even with a constant fraction of outliers.

Beyond the Gaussian design, similar guarantees can be established when the observation operator satisfies the mixed-norm restricted isometry property; see Chapter 3.

5.4 Numerical Experiments

In this section, we provide numerical experiments to illustrate the performance of `ScaledSM` for robust tensor regression, and highlight its advantage compared to the vanilla subgradient method (SM). For simplicity, we set $n_1 = n_2 = n_3 = 30$, and $r_1 = r_2 = r_3 = 3$, and collect $m = 5000$ measurements according to (5.2). The ground truth tensor $\boldsymbol{\mathcal{X}}_\star$ is generated as described in Section 4.4. Each outlier is independently generated as $s_i = \bar{s}_i\Omega_i$, with Ω_i drawn from a Bernoulli distribution with parameter p_s, and $\bar{s}_i$ drawn from a uniform distribution in $[-10\|\mathcal{A}(\boldsymbol{\mathcal{X}}_\star)\|_\infty, 10\|\mathcal{A}(\boldsymbol{\mathcal{X}}_\star)\|_\infty]$. Both `ScaledSM` and SM start from the same truncated spectral initialization (5.11), and for sim-

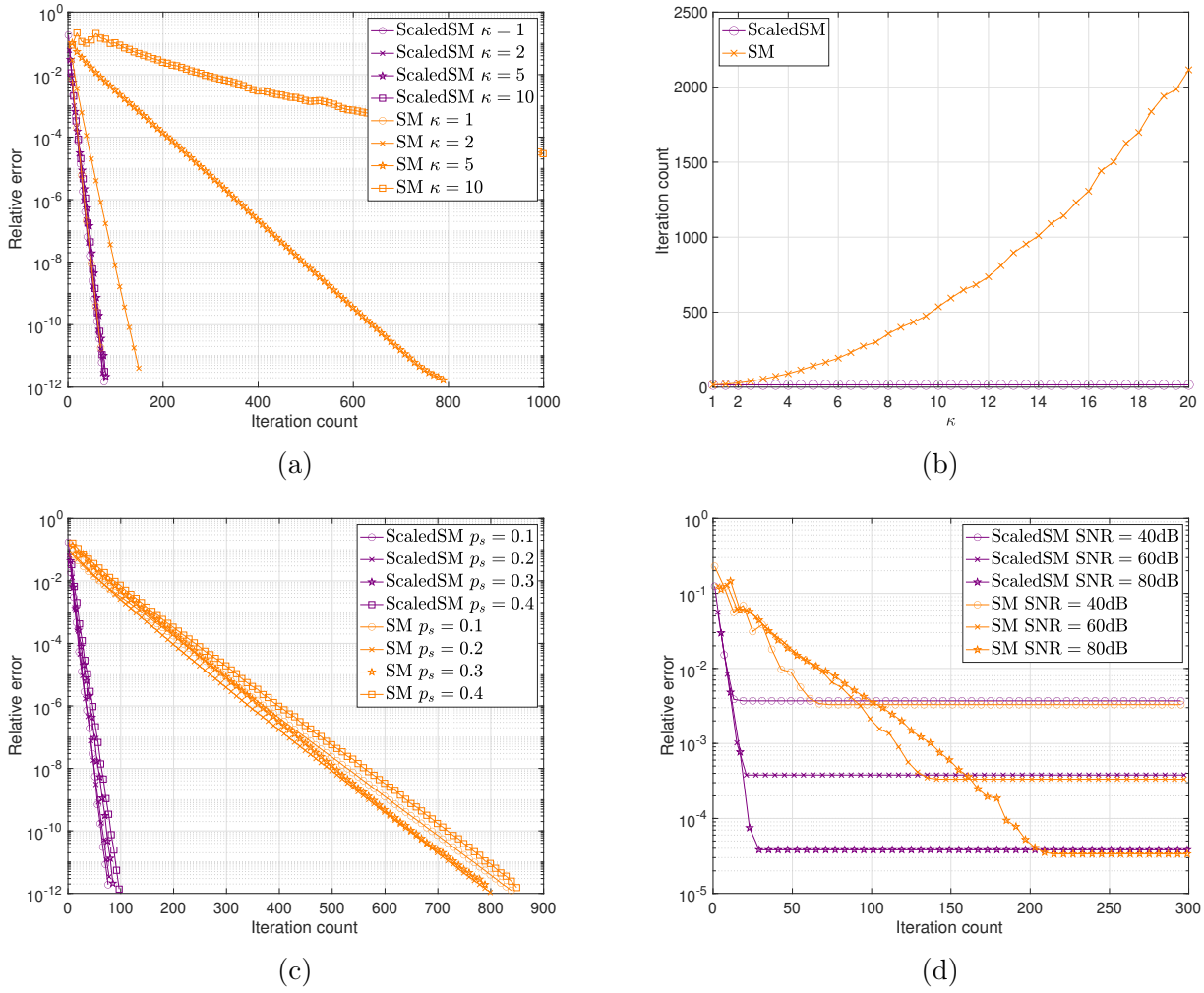


Figure 5.1: Performance comparisons of `ScaledSM` and the vanilla subgradient method (SM). (a) The reconstruction errors $\|\boldsymbol{\mathcal{X}}_t - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}/\|\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}$ w.r.t. the iteration count under different condition numbers $\kappa = 1, 2, 5, 10$ with $p_s = 0.2$. (b) The iteration complexities w.r.t. the condition number for achieving $\|\boldsymbol{\mathcal{X}}_t - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \leq 10^{-3}\|\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}$ with $p_s = 0.2$. (c) The reconstruction errors w.r.t. the iteration count under different amounts of outliers $p_s = 0.1, 0.2, 0.3, 0.4$ with $\kappa = 5$. (d) The reconstruction errors w.r.t. the iteration count under different signal-to-noise ratios SNR $= 40, 60, 80$dB with $p_s = 0.2$.

plicity use the Polyak's step size (which amounts to using optimally tuned geometrically decaying step sizes).

Fig. 5.1 shows the detailed performance comparison of `ScaledSM` and SM under various settings. Thanks to the robustness of the least absolute deviation loss, both algorithms converge linearly in the presence of outliers. Noteworthily, `ScaledSM` converges as a fast rate that is indepen-

dent with κ, while SM slows down dramatically as κ increases. Indeed, the iteration complexity of SM grows super linearly with respect to condition number κ, while `ScaledSM` takes a much smaller number of iterations and therefore accelerates the convergence for ill-conditioned instances.

5.5 Conclusions

This chapter develops a scaled subgradient method for robust low-rank tensor regression from corrupted measurements, by minimizing the a natural nonsmooth and nonconvex loss function based on least absolute deviation. In addition, it is of interest to examine if it is possible to develop provably efficient algorithms for the related problem called robust low-rank tensor completion [LFC+16].

Appendix A

Proofs for Low-rank Matrix Estimation

A.1 Technical Lemmas

This section gathers several useful lemmas that will be used in the appendix. Throughout all lemmas, we use $\boldsymbol{X}_\star$ to denote the ground truth low-rank matrix, with its compact SVD as $\boldsymbol{X}_\star = \boldsymbol{U}_\star \boldsymbol{\Sigma}_\star \boldsymbol{V}_\star^\top$, and the stacked factor matrix is defined as $\boldsymbol{F}_\star = \begin{bmatrix} \boldsymbol{L}_\star \\ \boldsymbol{R}_\star \end{bmatrix} = \begin{bmatrix} \boldsymbol{U}_\star \boldsymbol{\Sigma}_\star^{1/2} \\ \boldsymbol{V}_\star \boldsymbol{\Sigma}_\star^{1/2} \end{bmatrix}$.

A.1.1 New distance metric

We begin with the investigation of the new distance metric (2.8), where the matrix $\boldsymbol{Q}$ that attains the infimum, if exists, is called the optimal alignment matrix between $\boldsymbol{F}$ and $\boldsymbol{F}_\star$; see (2.9). Notice that (2.8) involves a minimization problem over an open set (the set of invertible matrices). Hence the minimizer, i.e. the optimal alignment matrix between $\boldsymbol{F}$ and $\boldsymbol{F}_\star$ is not guaranteed to be attained. Fortunately, a simple sufficient condition guarantees the existence of the minimizer; see the lemma below.

Lemma 14. *Fix any factor matrix* $\boldsymbol{F} = \begin{bmatrix} \boldsymbol{L} \\ \boldsymbol{R} \end{bmatrix} \in \mathbb{R}^{(n_1+n_2)\times r}$. *Suppose that*

$$\operatorname{dist}(\boldsymbol{F}, \boldsymbol{F}_\star) = \sqrt{\inf_{\boldsymbol{Q}\in \mathrm{GL}(r)} \left\| (\boldsymbol{L}\boldsymbol{Q} - \boldsymbol{L}_\star)\, \boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2 + \left\| (\boldsymbol{R}\boldsymbol{Q}^{-\top} - \boldsymbol{R}_\star)\, \boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2} < \sigma_r(\boldsymbol{X}_\star), \tag{A.1}$$

then the minimizer of the above minimization problem is attained at some $\boldsymbol{Q} \in \mathrm{GL}(r)$, *i.e. the optimal alignment matrix* $\boldsymbol{Q}$ *between* $\boldsymbol{F}$ *and* $\boldsymbol{F}_\star$ *exists.*

Proof. In view of the condition (A.1) and the definition of infimum, one knows that there must exist

a matrix $\bar{\boldsymbol{Q}} \in \mathrm{GL}(r)$ such that

$$\sqrt{\left\|\left(\boldsymbol{L}\bar{\boldsymbol{Q}}-\boldsymbol{L}_\star\right)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2+\left\|\left(\boldsymbol{R}\bar{\boldsymbol{Q}}^{-\top}-\boldsymbol{R}_\star\right)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2}\le\epsilon\sigma_r(\boldsymbol{X}_\star),$$

for some ϵ obeying $0<\epsilon<1$. It further implies that

$$\left\|\left(\boldsymbol{L}\bar{\boldsymbol{Q}}-\boldsymbol{L}_\star\right)\boldsymbol{\Sigma}_\star^{-1/2}\right\|\vee\left\|\left(\boldsymbol{R}\bar{\boldsymbol{Q}}^{-\top}-\boldsymbol{R}_\star\right)\boldsymbol{\Sigma}_\star^{-1/2}\right\|\le\epsilon.$$

Invoke Weyl's inequality $|\sigma_r(\boldsymbol{A})-\sigma_r(\boldsymbol{B})|\le\|\boldsymbol{A}-\boldsymbol{B}\|$, and use that $\sigma_r(\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2})=\sigma_r(\boldsymbol{U}_\star)=1$ to obtain

$$\sigma_r(\boldsymbol{L}\bar{\boldsymbol{Q}}\boldsymbol{\Sigma}_\star^{-1/2})\ge\sigma_r(\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2})-\left\|\left(\boldsymbol{L}\bar{\boldsymbol{Q}}-\boldsymbol{L}_\star\right)\boldsymbol{\Sigma}_\star^{-1/2}\right\|\ge1-\epsilon. \tag{A.2}$$

In addition, it is straightforward to verify that

$$\inf_{\boldsymbol{Q}\in\mathrm{GL}(r)}\ \left\|\left(\boldsymbol{L}\boldsymbol{Q}-\boldsymbol{L}_\star\right)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2+\left\|\left(\boldsymbol{R}\boldsymbol{Q}^{-\top}-\boldsymbol{R}_\star\right)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 \tag{A.3}$$

$$=\inf_{\boldsymbol{H}\in\mathrm{GL}(r)}\ \left\|\left(\boldsymbol{L}\bar{\boldsymbol{Q}}\boldsymbol{H}-\boldsymbol{L}_\star\right)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2+\left\|\left(\boldsymbol{R}\bar{\boldsymbol{Q}}^{-\top}\boldsymbol{H}^{-\top}-\boldsymbol{R}_\star\right)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2. \tag{A.4}$$

Indeed, if the minimizer of the second optimization problem (cf. (A.4)) is attained at some $\boldsymbol{H}$, then $\bar{\boldsymbol{Q}}\boldsymbol{H}$ must be the minimizer of the first problem (A.3). Therefore, from now on, we focus on proving that the minimizer of the second problem (A.4) is attained at some $\boldsymbol{H}$. In view of (A.3) and (A.4), one has

$$\begin{aligned}&\inf_{\boldsymbol{H}\in\mathrm{GL}(r)}\ \left\|\left(\boldsymbol{L}\bar{\boldsymbol{Q}}\boldsymbol{H}-\boldsymbol{L}_\star\right)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2+\left\|\left(\boldsymbol{R}\bar{\boldsymbol{Q}}^{-\top}\boldsymbol{H}^{-\top}-\boldsymbol{R}_\star\right)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2\\&\le\left\|\left(\boldsymbol{L}\bar{\boldsymbol{Q}}-\boldsymbol{L}_\star\right)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2+\left\|\left(\boldsymbol{R}\bar{\boldsymbol{Q}}^{-\top}-\boldsymbol{R}_\star\right)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2,\end{aligned}$$

Clearly, for any $\bar{\boldsymbol{Q}}\boldsymbol{H}$ to yield a smaller distance than $\bar{\boldsymbol{Q}}$, $\boldsymbol{H}$ must obey

$$\sqrt{\left\|\left(\boldsymbol{L}\bar{\boldsymbol{Q}}\boldsymbol{H}-\boldsymbol{L}_\star\right)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2+\left\|\left(\boldsymbol{R}\bar{\boldsymbol{Q}}^{-\top}\boldsymbol{H}^{-\top}-\boldsymbol{R}_\star\right)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2}\le\epsilon\sigma_r(\boldsymbol{X}_\star).$$

It further implies that

$$\left\|\left(\boldsymbol{L}\bar{\boldsymbol{Q}}\boldsymbol{H}-\boldsymbol{L}_\star\right)\boldsymbol{\Sigma}_\star^{-1/2}\right\| \vee \left\|\left(\boldsymbol{R}\bar{\boldsymbol{Q}}^{-\top}\boldsymbol{H}^{-\top}-\boldsymbol{R}_\star\right)\boldsymbol{\Sigma}_\star^{-1/2}\right\| \le \epsilon.$$

Invoke Weyl's inequality $|\sigma_1(\boldsymbol{A})-\sigma_1(\boldsymbol{B})| \le \|\boldsymbol{A}-\boldsymbol{B}\|$, and use that $\sigma_1(\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}) = \sigma_1(\boldsymbol{U}_\star) = 1$ to obtain

$$\sigma_1(\boldsymbol{L}\bar{\boldsymbol{Q}}\boldsymbol{H}\boldsymbol{\Sigma}_\star^{-1/2}) \le \sigma_1(\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}) + \left\|\left(\boldsymbol{L}\bar{\boldsymbol{Q}}\boldsymbol{H}-\boldsymbol{L}_\star\right)\boldsymbol{\Sigma}_\star^{-1/2}\right\| \le 1+\epsilon. \tag{A.5}$$

Combine (A.2) and (A.5), and use the relation $\sigma_r(\boldsymbol{A})\sigma_1(\boldsymbol{B}) \le \sigma_1(\boldsymbol{AB})$ to obtain

$$\sigma_r(\boldsymbol{L}\bar{\boldsymbol{Q}}\boldsymbol{\Sigma}_\star^{-1/2})\sigma_1(\boldsymbol{\Sigma}_\star^{1/2}\boldsymbol{H}\boldsymbol{\Sigma}_\star^{-1/2}) \le \sigma_1(\boldsymbol{L}\bar{\boldsymbol{Q}}\boldsymbol{H}\boldsymbol{\Sigma}_\star^{-1/2}) \le \frac{1+\epsilon}{1-\epsilon}\sigma_r(\boldsymbol{L}\bar{\boldsymbol{Q}}\boldsymbol{\Sigma}_\star^{-1/2}).$$

As a result, one has $\sigma_1(\boldsymbol{\Sigma}_\star^{1/2}\boldsymbol{H}\boldsymbol{\Sigma}_\star^{-1/2}) \le \frac{1+\epsilon}{1-\epsilon}$.

Similarly, one can show that $\sigma_1(\boldsymbol{\Sigma}_\star^{1/2}\boldsymbol{H}^{-\top}\boldsymbol{\Sigma}_\star^{-1/2}) \le \frac{1+\epsilon}{1-\epsilon}$, equivalently, $\sigma_r(\boldsymbol{\Sigma}_\star^{1/2}\boldsymbol{H}\boldsymbol{\Sigma}_\star^{-1/2}) \ge \frac{1-\epsilon}{1+\epsilon}$. Combining the above two arguments reveals that the minimization problem (A.4) is equivalent to the constrained problem:

$$\begin{aligned} \underset{\boldsymbol{H}\in\mathrm{GL}(r)}{\text{minimize}} \quad & \left\|\left(\boldsymbol{L}\bar{\boldsymbol{Q}}\boldsymbol{H}-\boldsymbol{L}_\star\right)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 + \left\|\left(\boldsymbol{R}\bar{\boldsymbol{Q}}^{-\top}\boldsymbol{H}^{-\top}-\boldsymbol{R}_\star\right)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 \\ \text{s.t.} \quad & \frac{1-\epsilon}{1+\epsilon} \le \sigma_r(\boldsymbol{\Sigma}_\star^{1/2}\boldsymbol{H}\boldsymbol{\Sigma}_\star^{-1/2}) \le \sigma_1(\boldsymbol{\Sigma}_\star^{1/2}\boldsymbol{H}\boldsymbol{\Sigma}_\star^{-1/2}) \le \frac{1+\epsilon}{1-\epsilon}. \end{aligned}$$

Notice that this is a continuous optimization problem over a compact set. Apply the Weierstrass extreme value theorem to finish the proof. □

With the existence of the optimal alignment matrix in place, the following lemma provides the first-order necessary condition for the minimizer.

Lemma 15. *For any factor matrix* $\boldsymbol{F} = \begin{bmatrix}\boldsymbol{L}\\ \boldsymbol{R}\end{bmatrix} \in \mathbb{R}^{(n_1+n_2)\times r}$, *suppose that the optimal alignment*

matrix

$$\boldsymbol{Q} = \operatorname*{argmin}_{\boldsymbol{Q}\in \mathrm{GL}(r)} \left\|(\boldsymbol{L}\boldsymbol{Q} - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{R}\boldsymbol{Q}^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2$$

between $\boldsymbol{F}$ *and* $\boldsymbol{F}_\star$ *exists, then* $\boldsymbol{Q}$ *obeys*

$$(\boldsymbol{L}\boldsymbol{Q})^\top(\boldsymbol{L}\boldsymbol{Q} - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star = \boldsymbol{\Sigma}_\star(\boldsymbol{R}\boldsymbol{Q}^{-\top} - \boldsymbol{R}_\star)^\top\boldsymbol{R}\boldsymbol{Q}^{-\top}. \tag{A.6}$$

Proof. Expand the squares in the definition of $\boldsymbol{Q}$ to obtain

$$\boldsymbol{Q} = \operatorname*{argmin}_{\boldsymbol{Q}\in \mathrm{GL}(r)} \operatorname{tr}\left((\boldsymbol{L}\boldsymbol{Q} - \boldsymbol{L}_\star)^\top(\boldsymbol{L}\boldsymbol{Q} - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star\right) + \operatorname{tr}\left((\boldsymbol{R}\boldsymbol{Q}^{-\top} - \boldsymbol{R}_\star)^\top(\boldsymbol{R}\boldsymbol{Q}^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star\right).$$

Clearly, the first order necessary condition (i.e. the gradient is zero) yields

$$2\boldsymbol{L}^\top(\boldsymbol{L}\boldsymbol{Q} - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star - 2\boldsymbol{Q}^{-\top}\boldsymbol{\Sigma}_\star(\boldsymbol{R}\boldsymbol{Q}^{-\top} - \boldsymbol{R}_\star)^\top\boldsymbol{R}\boldsymbol{Q}^{-\top} = \boldsymbol{0},$$

which implies the optimal alignment criterion (A.6). □

Last but not least, we connect the newly proposed distance to the usual Frobenius norm in Lemma 16, the proof of which is a slight modification to [TBS+16, Lemma 5.4] and [GJZ17, Lemma 41].

Lemma 16. *For any factor matrix* $\boldsymbol{F} = \begin{bmatrix}\boldsymbol{L}\\ \boldsymbol{R}\end{bmatrix} \in \mathbb{R}^{(n_1+n_2)\times r}$*, the distance between* $\boldsymbol{F}$ *and* $\boldsymbol{F}_\star$ *satisfies*

$$\operatorname{dist}(\boldsymbol{F}, \boldsymbol{F}_\star) \le \left(\sqrt{2}+1\right)^{1/2} \|\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star\|_{\mathsf{F}}.$$

Proof. Suppose that $\boldsymbol{X} \coloneqq \boldsymbol{L}\boldsymbol{R}^\top$ has compact SVD as $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$. Without loss of generality, we can assume that $\boldsymbol{F} = \begin{bmatrix}\boldsymbol{U}\boldsymbol{\Sigma}^{1/2}\\ \boldsymbol{V}\boldsymbol{\Sigma}^{1/2}\end{bmatrix}$, since any factorization of $\boldsymbol{L}\boldsymbol{R}^\top$ yields the same distance. Introduce

two auxiliary matrices $\bar{\boldsymbol{F}} \coloneqq \begin{bmatrix} \boldsymbol{U}\boldsymbol{\Sigma}^{1/2} \\ -\boldsymbol{V}\boldsymbol{\Sigma}^{1/2} \end{bmatrix}$ and $\bar{\boldsymbol{F}}_\star \coloneqq \begin{bmatrix} \boldsymbol{U}_\star\boldsymbol{\Sigma}_\star^{1/2} \\ -\boldsymbol{V}_\star\boldsymbol{\Sigma}_\star^{1/2} \end{bmatrix}$. Apply the dilation trick to obtain

$$2\begin{bmatrix} \boldsymbol{0} & \boldsymbol{X} \\ \boldsymbol{X}^\top & \boldsymbol{0} \end{bmatrix} = \boldsymbol{F}\boldsymbol{F}^\top - \bar{\boldsymbol{F}}\bar{\boldsymbol{F}}^\top, \quad 2\begin{bmatrix} \boldsymbol{0} & \boldsymbol{X}_\star \\ \boldsymbol{X}_\star^\top & \boldsymbol{0} \end{bmatrix} = \boldsymbol{F}_\star\boldsymbol{F}_\star^\top - \bar{\boldsymbol{F}}_\star\bar{\boldsymbol{F}}_\star^\top.$$

As a result, the squared Frobenius norm of $\boldsymbol{X} - \boldsymbol{X}_\star$ is given by

$$\begin{aligned} 8\|\boldsymbol{X} - \boldsymbol{X}_\star\|_{\mathsf{F}}^2 &= \left\|\boldsymbol{F}\boldsymbol{F}^\top - \bar{\boldsymbol{F}}\bar{\boldsymbol{F}}^\top - \boldsymbol{F}_\star\boldsymbol{F}_\star^\top + \bar{\boldsymbol{F}}_\star\bar{\boldsymbol{F}}_\star^\top\right\|_{\mathsf{F}}^2 \\ &= \left\|\boldsymbol{F}\boldsymbol{F}^\top - \boldsymbol{F}_\star\boldsymbol{F}_\star^\top\right\|_{\mathsf{F}}^2 + \left\|\bar{\boldsymbol{F}}\bar{\boldsymbol{F}}^\top - \bar{\boldsymbol{F}}_\star\bar{\boldsymbol{F}}_\star^\top\right\|_{\mathsf{F}}^2 - 2\operatorname{tr}\left((\boldsymbol{F}\boldsymbol{F}^\top - \boldsymbol{F}_\star\boldsymbol{F}_\star^\top)(\bar{\boldsymbol{F}}\bar{\boldsymbol{F}}^\top - \bar{\boldsymbol{F}}_\star\bar{\boldsymbol{F}}_\star^\top)\right) \\ &= 2\left\|\boldsymbol{F}\boldsymbol{F}^\top - \boldsymbol{F}_\star\boldsymbol{F}_\star^\top\right\|_{\mathsf{F}}^2 + 2\|\boldsymbol{F}^\top\bar{\boldsymbol{F}}_\star\|_{\mathsf{F}}^2 + 2\|\boldsymbol{F}_\star^\top\bar{\boldsymbol{F}}\|_{\mathsf{F}}^2 \\ &\geq 2\left\|\boldsymbol{F}\boldsymbol{F}^\top - \boldsymbol{F}_\star\boldsymbol{F}_\star^\top\right\|_{\mathsf{F}}^2, \end{aligned}$$

where we use the facts that $\left\|\boldsymbol{F}\boldsymbol{F}^\top - \boldsymbol{F}_\star\boldsymbol{F}_\star^\top\right\|_{\mathsf{F}}^2 = \left\|\bar{\boldsymbol{F}}\bar{\boldsymbol{F}}^\top - \bar{\boldsymbol{F}}_\star\bar{\boldsymbol{F}}_\star^\top\right\|_{\mathsf{F}}^2$ and $\boldsymbol{F}^\top\bar{\boldsymbol{F}} = \boldsymbol{F}_\star^\top\bar{\boldsymbol{F}}_\star = \boldsymbol{0}$.

Let $\boldsymbol{O} \coloneqq \operatorname{sgn}(\boldsymbol{F}^\top\boldsymbol{F}_\star)$[1] be the optimal orthonormal alignment matrix between $\boldsymbol{F}$ and $\boldsymbol{F}_\star$. Denote $\boldsymbol{\Delta} \coloneqq \boldsymbol{F}\boldsymbol{O} - \boldsymbol{F}_\star$. Follow the same argument as [TBS+16, Lemma 5.14] and [GJZ17, Lemma 41] to obtain

$$\begin{aligned} 4\|\boldsymbol{X} - \boldsymbol{X}_\star\|_{\mathsf{F}}^2 &\geq \left\|\boldsymbol{F}_\star\boldsymbol{\Delta}^\top + \boldsymbol{\Delta}\boldsymbol{F}_\star^\top + \boldsymbol{\Delta}\boldsymbol{\Delta}^\top\right\|_{\mathsf{F}}^2 \\ &= \operatorname{tr}\left(2\boldsymbol{F}_\star^\top\boldsymbol{F}_\star\boldsymbol{\Delta}^\top\boldsymbol{\Delta} + (\boldsymbol{\Delta}^\top\boldsymbol{\Delta})^2 + 2(\boldsymbol{F}_\star^\top\boldsymbol{\Delta})^2 + 4\boldsymbol{F}_\star^\top\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\boldsymbol{\Delta}\right) \\ &= \operatorname{tr}\left(2\boldsymbol{F}_\star^\top\boldsymbol{F}_\star\boldsymbol{\Delta}^\top\boldsymbol{\Delta} + (\boldsymbol{\Delta}^\top\boldsymbol{\Delta} + \sqrt{2}\boldsymbol{F}_\star^\top\boldsymbol{\Delta})^2 + (4 - 2\sqrt{2})\boldsymbol{F}_\star^\top\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\boldsymbol{\Delta}\right) \\ &= \operatorname{tr}\left(2(\sqrt{2}-1)\boldsymbol{F}_\star^\top\boldsymbol{F}_\star\boldsymbol{\Delta}^\top\boldsymbol{\Delta} + (\boldsymbol{\Delta}^\top\boldsymbol{\Delta} + \sqrt{2}\boldsymbol{F}_\star^\top\boldsymbol{\Delta})^2 + (4 - 2\sqrt{2})\boldsymbol{F}_\star^\top\boldsymbol{F}\boldsymbol{O}\boldsymbol{\Delta}^\top\boldsymbol{\Delta}\right) \\ &\geq \operatorname{tr}\left(4(\sqrt{2}-1)\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}^\top\boldsymbol{\Delta}\right) = 4(\sqrt{2}-1)\left\|(\boldsymbol{F}\boldsymbol{O} - \boldsymbol{F}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2, \end{aligned}$$

where the last inequality follows from the facts that $\boldsymbol{F}_\star^\top\boldsymbol{F}_\star = 2\boldsymbol{\Sigma}_\star$ and that $\boldsymbol{F}_\star^\top\boldsymbol{F}\boldsymbol{O}$ is positive

[1] Let $\boldsymbol{A}\boldsymbol{S}\boldsymbol{B}^\top$ be the SVD of $\boldsymbol{F}^\top\boldsymbol{F}_\star$, then the matrix sign is $\operatorname{sgn}(\boldsymbol{F}^\top\boldsymbol{F}_\star) \coloneqq \boldsymbol{A}\boldsymbol{B}^\top$.

semi-definite. Therefore we obtain

$$\left\|(\boldsymbol{F}\boldsymbol{O}-\boldsymbol{F}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}} \le \left(\sqrt{2}+1\right)^{1/2}\|\boldsymbol{X}-\boldsymbol{X}_\star\|_{\mathsf{F}}.$$

This in conjunction with $\mathrm{dist}(\boldsymbol{F},\boldsymbol{F}_\star) \le \|(\boldsymbol{F}\boldsymbol{O}-\boldsymbol{F}_\star)\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}$ yields the claimed result. □

A.1.2 Matrix perturbation bounds

Lemma 17. *For any $\boldsymbol{L}\in\mathbb{R}^{n_1\times r},\boldsymbol{R}\in\mathbb{R}^{n_2\times r}$, denote $\boldsymbol{\Delta}_L \coloneqq \boldsymbol{L}-\boldsymbol{L}_\star$ and $\boldsymbol{\Delta}_R \coloneqq \boldsymbol{R}-\boldsymbol{R}_\star$. Suppose that $\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\| \vee \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\| < 1$, then one has*

$$\left\|\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\| \le \frac{1}{1-\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|}; \tag{A.7a}$$

$$\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\| \le \frac{1}{1-\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\|}; \tag{A.7b}$$

$$\left\|\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{\Sigma}_\star^{1/2}-\boldsymbol{U}_\star\right\| \le \frac{\sqrt{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|}{1-\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|}; \tag{A.7c}$$

$$\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}-\boldsymbol{V}_\star\right\| \le \frac{\sqrt{2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\|}{1-\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\|}. \tag{A.7d}$$

Proof. We only prove claims (A.7a) and (A.7c) on the factor $\boldsymbol{L}$, while the claims on the factor $\boldsymbol{R}$ follow from a similar argument. We start to prove (A.7a). Notice that

$$\left\|\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\| = \frac{1}{\sigma_r(\boldsymbol{L}\boldsymbol{\Sigma}_\star^{-1/2})}.$$

In addition, invoke Weyl's inequality to obtain

$$\sigma_r(\boldsymbol{L}\boldsymbol{\Sigma}_\star^{-1/2}) \ge \sigma_r(\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}) - \|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\| = 1-\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|,$$

where we have used the fact that $\boldsymbol{U}_\star = \boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}$ satisfies $\sigma_r(\boldsymbol{U}_\star)=1$. Combine the preceding two relations to prove (A.7a).

We proceed to prove (A.7c). Combine $\boldsymbol{L}_\star^\top\boldsymbol{U}_\star = \boldsymbol{\Sigma}_\star^{1/2}$ and $(\boldsymbol{I}_{n_1}-\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{L}^\top)\boldsymbol{L}=\boldsymbol{0}$ to

obtain the decomposition

$$\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{\Sigma}_\star^{1/2}-\boldsymbol{U}_\star=-\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{\Delta}_L^\top\boldsymbol{U}_\star+(\boldsymbol{I}_{n_1}-\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{L}^\top)\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}.$$

The fact that $\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{\Delta}_L^\top\boldsymbol{U}_\star$ and $(\boldsymbol{I}_{n_1}-\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{L}^\top)\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}$ are orthogonal implies

$$\begin{aligned}\left\|\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{\Sigma}_\star^{1/2}-\boldsymbol{U}_\star\right\|^2&\le\left\|\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{\Delta}_L^\top\boldsymbol{U}_\star\right\|^2+\left\|(\boldsymbol{I}_{n_1}-\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{L}^\top)\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\right\|^2\\&\le\left\|\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|^2\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|^2+\left\|\boldsymbol{I}_{n_1}-\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{L}^\top\right\|^2\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|^2\\&\le\frac{\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|^2}{(1-\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|)^2}+\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|^2\\&\le\frac{2\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|^2}{(1-\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|)^2},\end{aligned}$$

where we have used (A.7a) and the fact that $\|\boldsymbol{I}_{n_1}-\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{L}^\top\|\le 1$ in the third line. □

Lemma 18. *For any $\boldsymbol{L}\in\mathbb{R}^{n_1\times r},\boldsymbol{R}\in\mathbb{R}^{n_2\times r}$, denote $\boldsymbol{\Delta}_L\coloneqq\boldsymbol{L}-\boldsymbol{L}_\star$ and $\boldsymbol{\Delta}_R\coloneqq\boldsymbol{R}-\boldsymbol{R}_\star$, then one has*

$$\begin{aligned}\|\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star\|_{\mathsf{F}}&\le\|\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top\|_{\mathsf{F}}+\|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}}\\&\le\left(1+\frac{1}{2}(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|\vee\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\|)\right)\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right).\end{aligned}$$

Proof. In light of the decomposition $\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star=\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top+\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top+\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top$ and the triangle inequality, one has

$$\begin{aligned}\|\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star\|_{\mathsf{F}}&\le\|\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top\|_{\mathsf{F}}+\|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}}\\&=\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}},\end{aligned}$$

where we have used the facts that

$$\|\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top\|_{\mathsf{F}}=\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\boldsymbol{V}_\star^\top\|_{\mathsf{F}}=\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}},\quad\text{and}\quad\|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}}=\|\boldsymbol{U}_\star\boldsymbol{\Sigma}_\star^{1/2}\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}}=\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}.$$

This together with the simple upper bound

$$\begin{aligned}\|\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}} &= \frac{1}{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}(\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2})^\top\|_{\mathsf{F}} + \frac{1}{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}(\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2})^\top\|_{\mathsf{F}} \\ &\le \frac{1}{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\| + \frac{1}{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} \\ &\le \frac{1}{2}(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\| \vee \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\|)\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)\end{aligned}$$

finishes the proof. □

Lemma 19. *For any $\boldsymbol{L} \in \mathbb{R}^{n_1\times r}, \boldsymbol{R} \in \mathbb{R}^{n_2\times r}$ and any invertible matrices $\boldsymbol{Q}, \bar{\boldsymbol{Q}} \in \mathrm{GL}(r)$, suppose that $\|(\boldsymbol{L}\boldsymbol{Q} - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{-1/2}\| \vee \|(\boldsymbol{R}\boldsymbol{Q}^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{-1/2}\| < 1$, then one has*

$$\begin{aligned}\left\|\boldsymbol{\Sigma}_\star^{1/2}\bar{\boldsymbol{Q}}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}_\star^{1/2} - \boldsymbol{\Sigma}_\star\right\| &\le \frac{\|\boldsymbol{R}(\bar{\boldsymbol{Q}}^{-\top} - \boldsymbol{Q}^{-\top})\boldsymbol{\Sigma}_\star^{1/2}\|}{1 - \|(\boldsymbol{R}\boldsymbol{Q}^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{-1/2}\|}; \\ \left\|\boldsymbol{\Sigma}_\star^{1/2}\bar{\boldsymbol{Q}}^{\top}\boldsymbol{Q}^{-\top}\boldsymbol{\Sigma}_\star^{1/2} - \boldsymbol{\Sigma}_\star\right\| &\le \frac{\|\boldsymbol{L}(\bar{\boldsymbol{Q}} - \boldsymbol{Q})\boldsymbol{\Sigma}_\star^{1/2}\|}{1 - \|(\boldsymbol{L}\boldsymbol{Q} - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{-1/2}\|}.\end{aligned}$$

Proof. Insert $\boldsymbol{R}^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}$, and use the relation $\|\boldsymbol{A}\boldsymbol{B}\| \le \|\boldsymbol{A}\|\|\boldsymbol{B}\|$ to obtain

$$\begin{aligned}\left\|\boldsymbol{\Sigma}_\star^{1/2}\bar{\boldsymbol{Q}}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}_\star^{1/2} - \boldsymbol{\Sigma}_\star\right\| &= \left\|\boldsymbol{\Sigma}_\star^{1/2}(\bar{\boldsymbol{Q}}^{-1} - \boldsymbol{Q}^{-1})\boldsymbol{R}^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}_\star^{1/2}\right\| \\ &\le \left\|\boldsymbol{R}(\bar{\boldsymbol{Q}}^{-\top} - \boldsymbol{Q}^{-\top})\boldsymbol{\Sigma}_\star^{1/2}\right\| \left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}_\star^{1/2}\right\| \\ &= \left\|\boldsymbol{R}(\bar{\boldsymbol{Q}}^{-\top} - \boldsymbol{Q}^{-\top})\boldsymbol{\Sigma}_\star^{1/2}\right\| \left\|\boldsymbol{R}\boldsymbol{Q}^{-\top}((\boldsymbol{R}\boldsymbol{Q}^{-\top})^\top\boldsymbol{R}\boldsymbol{Q}^{-\top})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\| \\ &\le \frac{\|\boldsymbol{R}(\bar{\boldsymbol{Q}}^{-\top} - \boldsymbol{Q}^{-\top})\boldsymbol{\Sigma}_\star^{1/2}\|}{1 - \|(\boldsymbol{R}\boldsymbol{Q}^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{-1/2}\|},\end{aligned}$$

where the last line uses Lemma 17.

Similarly, insert $\boldsymbol{L}^\top\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}$, and use the relation $\|\boldsymbol{A}\boldsymbol{B}\| \le \|\boldsymbol{A}\|\|\boldsymbol{B}\|$ to obtain

$$\begin{aligned}\left\|\boldsymbol{\Sigma}_\star^{1/2}\bar{\boldsymbol{Q}}^{\top}\boldsymbol{Q}^{-\top}\boldsymbol{\Sigma}_\star^{1/2} - \boldsymbol{\Sigma}_\star\right\| &= \left\|\boldsymbol{\Sigma}_\star^{1/2}(\bar{\boldsymbol{Q}}^{\top} - \boldsymbol{Q}^{\top})\boldsymbol{L}^\top\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{Q}^{-\top}\boldsymbol{\Sigma}_\star^{1/2}\right\| \\ &\le \left\|\boldsymbol{L}(\bar{\boldsymbol{Q}} - \boldsymbol{Q})\boldsymbol{\Sigma}_\star^{1/2}\right\| \left\|\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{Q}^{-\top}\boldsymbol{\Sigma}_\star^{1/2}\right\| \\ &= \left\|\boldsymbol{L}(\bar{\boldsymbol{Q}} - \boldsymbol{Q})\boldsymbol{\Sigma}_\star^{1/2}\right\| \left\|\boldsymbol{L}\boldsymbol{Q}((\boldsymbol{L}\boldsymbol{Q})^\top\boldsymbol{L}\boldsymbol{Q})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|\end{aligned}$$

$$\leq \frac{\|\boldsymbol{L}(\bar{\boldsymbol{Q}} - \boldsymbol{Q})\boldsymbol{\Sigma}_\star^{1/2}\|}{1 - \|(\boldsymbol{L}\boldsymbol{Q} - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{-1/2}\|},$$

where the last line uses Lemma 17. □

A.1.3 Partial Frobenius norm

We introduce the partial Frobenius norm

$$\|\boldsymbol{X}\|_{\mathsf{F},r} := \sqrt{\sum_{i=1}^{r} \sigma_i^2(\boldsymbol{X})} = \|\mathcal{P}_r(\boldsymbol{X})\|_{\mathsf{F}} \tag{A.8}$$

as the ℓ_2 norm of the vector composed of the top-r singular values of the matrix $\boldsymbol{X}$, or equivalently as the Frobenius norm of the rank-r approximation $\mathcal{P}_r(\boldsymbol{X})$ defined in (1.3). It is straightforward to verify that $\|\cdot\|_{\mathsf{F},r}$ is a norm; see also [Maz16]. The following lemma provides several equivalent and useful characterizations of this partial Frobenius norm.

Lemma 20. *For any $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$, one has*

$$\|\boldsymbol{X}\|_{\mathsf{F},r} = \max_{\widetilde{\boldsymbol{V}} \in \mathbb{R}^{n_2 \times r} : \widetilde{\boldsymbol{V}}^\top \widetilde{\boldsymbol{V}} = \boldsymbol{I}_r} \|\boldsymbol{X}\widetilde{\boldsymbol{V}}\|_{\mathsf{F}} \tag{A.9a}$$

$$= \max_{\widetilde{\boldsymbol{X}} \in \mathbb{R}^{n_1 \times n_2} : \|\widetilde{\boldsymbol{X}}\|_{\mathsf{F}} \leq 1, \mathrm{rank}(\widetilde{\boldsymbol{X}}) \leq r} |\langle \boldsymbol{X}, \widetilde{\boldsymbol{X}} \rangle| \tag{A.9b}$$

$$= \max_{\widetilde{\boldsymbol{R}} \in \mathbb{R}^{n_2 \times r} : \|\widetilde{\boldsymbol{R}}\| \leq 1} \|\boldsymbol{X}\widetilde{\boldsymbol{R}}\|_{\mathsf{F}}. \tag{A.9c}$$

Proof. The first representation (A.9a) follows immediately from the extremal partial trace identity; see [Maz16, Proposition 4.4], by noticing the following relation

$$\sum_{i=1}^{r} \sigma_i^2(\boldsymbol{X}) = \max_{\mathbb{V} \subseteq \mathbb{R}^{n_2} : \dim(\mathbb{V}) = r} \mathrm{tr}\left(\boldsymbol{X}^\top \boldsymbol{X} \mid \mathbb{V}\right) = \max_{\widetilde{\boldsymbol{V}} \in \mathbb{R}^{n_2 \times r} : \widetilde{\boldsymbol{V}}^\top \widetilde{\boldsymbol{V}} = \boldsymbol{I}_r} \|\boldsymbol{X}\widetilde{\boldsymbol{V}}\|_{\mathsf{F}}^2.$$

Here the partial trace over a vector space $\mathbb{V}$ is defined as

$$\mathrm{tr}(\boldsymbol{X}^\top \boldsymbol{X} \mid \mathbb{V}) := \sum_{i=1}^{r} \widetilde{\boldsymbol{v}}_i^\top \boldsymbol{X}^\top \boldsymbol{X} \widetilde{\boldsymbol{v}}_i,$$

where $\{\widetilde{\boldsymbol{v}}_i\}_{1\le i\le r}$ is any orthonormal basis of $\mathbb{V}$. The partial trace is invariant to the choice of orthonormal basis and therefore well-defined.

To prove the second representation (A.9b), for any $\widetilde{\boldsymbol{X}} \in \mathbb{R}^{n_1\times n_2}$ obeying $\mathrm{rank}(\widetilde{\boldsymbol{X}}) \le r$ and $\|\widetilde{\boldsymbol{X}}\|_{\mathsf{F}} \le 1$, denoting $\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{V}}^\top$ as its compact SVD, one has

$$|\langle \boldsymbol{X}, \widetilde{\boldsymbol{X}}\rangle| = |\langle \boldsymbol{X}, \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{V}}^\top\rangle| = |\langle \boldsymbol{X}\widetilde{\boldsymbol{V}}, \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Sigma}}\rangle| \le \|\boldsymbol{X}\widetilde{\boldsymbol{V}}\|_{\mathsf{F}}\|\widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Sigma}}\|_{\mathsf{F}} \le \|\boldsymbol{X}\|_{\mathsf{F},r},$$

where the last inequality follows from (A.9a). In addition, the maximum in (A.9b) is attained at $\widetilde{\boldsymbol{X}} = \mathcal{P}_r(\boldsymbol{X})/\|\mathcal{P}_r(\boldsymbol{X})\|_{\mathsf{F}}$.

To prove the third representation (A.9c), for any $\widetilde{\boldsymbol{R}} \in \mathbb{R}^{n_2\times r}$ obeying $\|\widetilde{\boldsymbol{R}}\| \le 1$, combine the variational representation of the Frobenius norm and (A.9b) to obtain

$$\begin{aligned}\|\boldsymbol{X}\widetilde{\boldsymbol{R}}\|_{\mathsf{F}} &= \max_{\widetilde{\boldsymbol{L}}\in\mathbb{R}^{n_1\times n_2}:\|\widetilde{\boldsymbol{L}}\|_{\mathsf{F}}\le 1} |\langle \boldsymbol{X}\widetilde{\boldsymbol{R}}, \widetilde{\boldsymbol{L}}\rangle| \\ &= \max_{\widetilde{\boldsymbol{L}}\in\mathbb{R}^{n_1\times n_2}:\|\widetilde{\boldsymbol{L}}\|_{\mathsf{F}}\le 1} |\langle \boldsymbol{X}, \widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{R}}^\top\rangle| \le \|\boldsymbol{X}\|_{\mathsf{F},r},\end{aligned}$$

where the last inequality follows from (A.9b). In addition, the maximum in (A.9c) is attained at $\widetilde{\boldsymbol{R}} = \boldsymbol{V}$, where $\boldsymbol{V}$ denotes the top-r right singular vectors of $\boldsymbol{X}$. □

Remark 8. For self-completeness, we also provide a detailed proof of the first representation (A.9a). This proof is inductive on r. When $r = 1$, we have

$$\sigma_1(\boldsymbol{X}) = \|\boldsymbol{X}\boldsymbol{v}_1\|_2 = \max_{\widetilde{\boldsymbol{v}}\in\mathbb{R}^{n_2}:\|\widetilde{\boldsymbol{v}}\|_2=1} \|\boldsymbol{X}\widetilde{\boldsymbol{v}}\|_2,$$

where $\boldsymbol{v}_1$ denotes the top right singular vector of $\boldsymbol{X}$. Assume that the statement holds for $\|\cdot\|_{\mathsf{F},r-1}$. Now consider $\|\cdot\|_{\mathsf{F},r}$. For any $\widetilde{\boldsymbol{V}} \in \mathbb{R}^{n_2\times r}$ such that $\widetilde{\boldsymbol{V}}^\top\widetilde{\boldsymbol{V}} = \boldsymbol{I}_r$, we can first pick $\widetilde{\boldsymbol{v}}_2, \ldots, \widetilde{\boldsymbol{v}}_r$ as a set of orthonormal vectors in the column space of $\widetilde{\boldsymbol{V}}$ that are orthogonal to $\boldsymbol{v}_1$, and then pick $\widetilde{\boldsymbol{v}}_1$ via the Gram-Schmidt process, so that $\{\widetilde{\boldsymbol{v}}_i\}_{i=1}^r$ provides an orthonormal basis of the column space of $\widetilde{\boldsymbol{V}}$. Further, by the orthogonality of $\widetilde{\boldsymbol{V}}$, there exists an orthonormal matrix $\boldsymbol{O}$ such that

$$\widetilde{\boldsymbol{V}} = [\widetilde{\boldsymbol{v}}_1, \ldots, \widetilde{\boldsymbol{v}}_r]\boldsymbol{O}.$$

Combining this formula with the induction hypothesis yields

$$\begin{aligned}\|\boldsymbol{X}\widetilde{\boldsymbol{V}}\|_{\mathsf{F}}^2 &= \|\boldsymbol{X}[\widetilde{\boldsymbol{v}}_1,\ldots,\widetilde{\boldsymbol{v}}_r]\|_{\mathsf{F}}^2 \\ &= \|\boldsymbol{X}\widetilde{\boldsymbol{v}}_1\|_2^2 + \|\boldsymbol{X}[\widetilde{\boldsymbol{v}}_2,\ldots,\widetilde{\boldsymbol{v}}_r]\|_{\mathsf{F}}^2 \\ &= \|\boldsymbol{X}\widetilde{\boldsymbol{v}}_1\|_2^2 + \|(\boldsymbol{X}-\mathcal{P}_1(\boldsymbol{X}))[\widetilde{\boldsymbol{v}}_2,\ldots,\widetilde{\boldsymbol{v}}_r]\|_{\mathsf{F}}^2 \\ &\le \sigma_1^2(\boldsymbol{X}) + \|\boldsymbol{X}-\mathcal{P}_1(\boldsymbol{X})\|_{\mathsf{F},r-1}^2 \\ &= \sum_{i=1}^{r}\sigma_i^2(\boldsymbol{X}) = \|\boldsymbol{X}\|_{\mathsf{F},r}^2,\end{aligned}$$

where the first line holds since $\boldsymbol{O}$ is orthonormal, the third line holds since $\mathcal{P}_1(\boldsymbol{X})[\widetilde{\boldsymbol{v}}_2,\ldots,\widetilde{\boldsymbol{v}}_r]=\boldsymbol{0}$, the fourth line follows from the induction hypothesis, and the last line follows from the definition (A.8). In addition, the maximum in (A.9a) is attained at $\widetilde{\boldsymbol{V}}=\boldsymbol{V}$, where $\boldsymbol{V}$ denotes the top-r right singular vectors of $\boldsymbol{X}$. This finishes the proof.

Recall that $\mathcal{P}_r(\boldsymbol{X})$ denotes the best rank-r approximation of $\boldsymbol{X}$ under the Frobenius norm. It turns out that $\mathcal{P}_r(\boldsymbol{X})$ is also the best rank-r approximation of $\boldsymbol{X}$ under the partial Frobenius norm $\|\cdot\|_{\mathsf{F},r}$. This claim is formally stated below; see also [Maz16, Theorem 4.21].

Lemma 21. *Fix any $\boldsymbol{X}\in\mathbb{R}^{n_1\times n_2}$ and recall the definition of $\mathcal{P}_r(\boldsymbol{X})$ in (1.3). One has*

$$\mathcal{P}_r(\boldsymbol{X}) = \operatorname*{argmin}_{\widetilde{\boldsymbol{X}}\in\mathbb{R}^{n_1\times n_2}:\operatorname{rank}(\widetilde{\boldsymbol{X}})\le r} \|\boldsymbol{X}-\widetilde{\boldsymbol{X}}\|_{\mathsf{F},r}.$$

Proof. For any $\widetilde{\boldsymbol{X}}$ of rank at most r, invoke Weyl's inequality to obtain $\sigma_{r+i}(\boldsymbol{X})\le\sigma_i(\boldsymbol{X}-\widetilde{\boldsymbol{X}})+\sigma_{r+1}(\widetilde{\boldsymbol{X}})=\sigma_i(\boldsymbol{X}-\widetilde{\boldsymbol{X}})$, for $i=1,\ldots,r$. Thus one has

$$\|\boldsymbol{X}-\mathcal{P}_r(\boldsymbol{X})\|_{\mathsf{F},r}^2 = \sum_{i=1}^{r}\sigma_{r+i}^2(\boldsymbol{X}) \le \sum_{i=1}^{r}\sigma_i^2(\boldsymbol{X}-\widetilde{\boldsymbol{X}}) = \|\boldsymbol{X}-\widetilde{\boldsymbol{X}}\|_{\mathsf{F},r}^2.$$

The proof is finished by observing that the rank of $\mathcal{P}_r(\boldsymbol{X})$ is at most r. □

A.2 Proof for Low-Rank Matrix Factorization

A.2.1 Proof of Proposition 2

The gradients of $\mathcal{L}(\boldsymbol{F})$ in (2.27) with respect to $\boldsymbol{L}$ and $\boldsymbol{R}$ are given as

$$\nabla_{\boldsymbol{L}}\mathcal{L}(\boldsymbol{F}) = (\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star)\boldsymbol{R}, \quad \nabla_{\boldsymbol{R}}\mathcal{L}(\boldsymbol{F}) = (\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star)^\top\boldsymbol{L},$$

which can be used to compute the Hessian with respect to $\boldsymbol{L}$ and $\boldsymbol{R}$. Writing for the vectorized variables, the Hessians are given as

$$\nabla^2_{\boldsymbol{L},\boldsymbol{L}}\mathcal{L}(\boldsymbol{F}) = (\boldsymbol{R}^\top\boldsymbol{R}) \otimes \boldsymbol{I}_{n_1}, \quad \nabla^2_{\boldsymbol{R},\boldsymbol{R}}\mathcal{L}(\boldsymbol{F}) = (\boldsymbol{L}^\top\boldsymbol{L}) \otimes \boldsymbol{I}_{n_2}.$$

Viewed in the vectorized form, the `ScaledGD` update in (2.2) can be rewritten as

$$\begin{aligned}
\operatorname{vec}(\boldsymbol{L}_{t+1}) &= \operatorname{vec}(\boldsymbol{L}_t) - \eta((\boldsymbol{R}_t^\top\boldsymbol{R}_t)^{-1} \otimes \boldsymbol{I}_{n_1})\operatorname{vec}((\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star)\boldsymbol{R}_t) \\
&= \operatorname{vec}(\boldsymbol{L}_t) - \eta(\nabla^2_{\boldsymbol{L},\boldsymbol{L}}\mathcal{L}(\boldsymbol{F}_t))^{-1}\operatorname{vec}(\nabla_{\boldsymbol{L}}\mathcal{L}(\boldsymbol{F}_t)), \\
\operatorname{vec}(\boldsymbol{R}_{t+1}) &= \operatorname{vec}(\boldsymbol{R}_t) - \eta((\boldsymbol{L}_t^\top\boldsymbol{L}_t)^{-1} \otimes \boldsymbol{I}_{n_2})\operatorname{vec}((\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star)^\top\boldsymbol{L}_t) \\
&= \operatorname{vec}(\boldsymbol{R}_t) - \eta(\nabla^2_{\boldsymbol{R},\boldsymbol{R}}\mathcal{L}(\boldsymbol{F}_t))^{-1}\operatorname{vec}(\nabla_{\boldsymbol{R}}\mathcal{L}(\boldsymbol{F}_t)).
\end{aligned}$$

A.2.2 Proof of Theorem 5

The proof is inductive in nature. More specifically, we intend to show that for all $t \geq 0$,

1. $\operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \leq (1 - 0.7\eta)^t \operatorname{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \leq 0.1(1 - 0.7\eta)^t\sigma_r(\boldsymbol{X}_\star)$, and
2. the optimal alignment matrix $\boldsymbol{Q}_t$ between $\boldsymbol{F}_t$ and $\boldsymbol{F}_\star$ exists.

For the base case, i.e. $t = 0$, the first induction hypothesis trivially holds, while the second also holds true in view of Lemma 14 and the assumption that $\operatorname{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \leq 0.1\sigma_r(\boldsymbol{X}_\star)$. We therefore concentrate on the induction step. Suppose that the t-th iterate $\boldsymbol{F}_t$ obeys the aforementioned induction hypotheses. Our goal is to show that $\boldsymbol{F}_{t+1}$ continues to satisfy those.

For notational convenience, denote $\boldsymbol{L} \coloneqq \boldsymbol{L}_t\boldsymbol{Q}_t$, $\boldsymbol{R} \coloneqq \boldsymbol{R}_t\boldsymbol{Q}_t^{-\top}$, $\boldsymbol{\Delta}_L \coloneqq \boldsymbol{L} - \boldsymbol{L}_\star$, $\boldsymbol{\Delta}_R \coloneqq \boldsymbol{R} - \boldsymbol{R}_\star$, and $\epsilon \coloneqq 0.1$. By the definition of $\mathrm{dist}(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star)$, one has

$$\mathrm{dist}^2(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) \le \left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2, \tag{A.10}$$

where we recall that $\boldsymbol{Q}_t$ is the optimal alignment matrix between $\boldsymbol{F}_t$ and $\boldsymbol{F}_\star$. Utilize the `ScaledGD` update rule (2.28) and the decomposition $\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star = \boldsymbol{\Delta}_L\boldsymbol{R}^\top + \boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top$ to obtain

$$\begin{aligned}
(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2} &= \left(\boldsymbol{L} - \eta(\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1} - \boldsymbol{L}_\star\right)\boldsymbol{\Sigma}_\star^{1/2} \\
&= \left(\boldsymbol{\Delta}_L - \eta(\boldsymbol{\Delta}_L\boldsymbol{R}^\top + \boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\right)\boldsymbol{\Sigma}_\star^{1/2} \\
&= (1-\eta)\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2} - \eta\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}.
\end{aligned}$$

As a result, one can expand the first square in (A.10) as

$$\begin{aligned}
\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 = (1-\eta)^2\,\mathrm{tr}\left(\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\right) - 2\eta(1-\eta)\underbrace{\mathrm{tr}\left(\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\right)}_{\mathfrak{M}_1} \\
+ \eta^2\underbrace{\left\|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2}_{\mathfrak{M}_2}.
\end{aligned} \tag{A.11}$$

The first term $\mathrm{tr}(\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top)$ is closely related to $\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$, and hence our focus will be on relating $\mathfrak{M}_1$ and $\mathfrak{M}_2$ to $\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$. We start with the term $\mathfrak{M}_1$. Since $\boldsymbol{L}$ and $\boldsymbol{R}$ are aligned with $\boldsymbol{L}_\star$ and $\boldsymbol{R}_\star$, Lemma 15 tells that $\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\boldsymbol{L} = \boldsymbol{R}^\top\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star$. This together with $\boldsymbol{L}_\star = \boldsymbol{L} - \boldsymbol{\Delta}_L$ allows us to rewrite $\mathfrak{M}_1$ as

$$\begin{aligned}
\mathfrak{M}_1 &= \mathrm{tr}\left(\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\right) \\
&= \mathrm{tr}\left(\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\boldsymbol{L}\boldsymbol{\Delta}_R^\top\right) - \mathrm{tr}\left(\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\right) \\
&= \mathrm{tr}\left(\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_R^\top\right) - \mathrm{tr}\left(\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\right).
\end{aligned}$$

Moving on to $\mathfrak{M}_2$, we can utilize the fact $\boldsymbol{L}_\star^\top\boldsymbol{L}_\star = \boldsymbol{\Sigma}_\star$ and the decomposition $\boldsymbol{\Sigma}_\star = \boldsymbol{R}^\top\boldsymbol{R} - (\boldsymbol{R}^\top\boldsymbol{R} -$

$\boldsymbol{\Sigma}_\star$) to obtain

$$\begin{aligned}\mathfrak{M}_2 &= \mathrm{tr}\left(\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_R^\top\right)\\ &= \mathrm{tr}\left(\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_R^\top\right) - \mathrm{tr}\left(\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}(\boldsymbol{R}^\top\boldsymbol{R}-\boldsymbol{\Sigma}_\star)(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_R^\top\right).\end{aligned}$$

Putting $\mathfrak{M}_1$ and $\mathfrak{M}_2$ back to (A.11) yields

$$\begin{aligned}\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 = (1-\eta)^2\,\mathrm{tr}\left(\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\right) - \eta(2-3\eta)\underbrace{\mathrm{tr}\left(\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_R^\top\right)}_{\mathfrak{F}_1}\\ + 2\eta(1-\eta)\underbrace{\mathrm{tr}\left(\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\right)}_{\mathfrak{F}_2}\\ - \eta^2\underbrace{\mathrm{tr}\left(\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}(\boldsymbol{R}^\top\boldsymbol{R}-\boldsymbol{\Sigma}_\star)(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_R^\top\right)}_{\mathfrak{F}_3}.\end{aligned}$$

In what follows, we will control the three terms $\mathfrak{F}_1, \mathfrak{F}_2$ and $\mathfrak{F}_3$ separately.

1. Notice that $\mathfrak{F}_1$ is the inner product of two positive semi-definite matrices $\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top$ and $\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_R^\top$. Consequently we have $\mathfrak{F}_1 \geq 0$.

2. To control $\mathfrak{F}_2$, we need certain control on $\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|$ and $\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\|$. The first induction hypothesis

$$\mathrm{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star) = \sqrt{\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\boldsymbol{\Sigma}_\star\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\boldsymbol{\Sigma}_\star\|_{\mathsf{F}}^2} \leq \epsilon\sigma_r(\boldsymbol{X}_\star)$$

together with the relation $\|\boldsymbol{A}\boldsymbol{B}\|_{\mathsf{F}} \geq \|\boldsymbol{A}\|_{\mathsf{F}}\sigma_r(\boldsymbol{B})$ tells that

$$\sqrt{\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{F}}^2}\,\sigma_r(\boldsymbol{X}_\star) \leq \epsilon\sigma_r(\boldsymbol{X}_\star).$$

In light of the relation $\|\boldsymbol{A}\| \leq \|\boldsymbol{A}\|_{\mathsf{F}}$, this further implies

$$\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\| \vee \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\| \leq \epsilon. \tag{A.12}$$

Invoke Lemma 17 to see

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

With these consequences, one can bound $|\mathfrak{F}_2|$ by

$$\begin{aligned}
|\mathfrak{F}_2| &= \left|\operatorname{tr}\left(\boldsymbol{\Sigma}_\star^{-1/2}\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\right)\right| \\
&\le \left\|\boldsymbol{\Sigma}_\star^{-1/2}\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\| \operatorname{tr}\left(\boldsymbol{\Sigma}_\star^{1/2}\boldsymbol{\Delta}_L^\top\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\right) \\
&\le \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\| \left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\| \operatorname{tr}\left(\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\right) \\
&\le \frac{\epsilon}{1-\epsilon}\operatorname{tr}\left(\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\right).
\end{aligned}$$

3. Similarly, one can bound $|\mathfrak{F}_3|$ by

$$\begin{aligned}
|\mathfrak{F}_3| &\le \left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}(\boldsymbol{R}^\top\boldsymbol{R}-\boldsymbol{\Sigma}_\star)(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\right\| \operatorname{tr}\left(\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_R^\top\right) \\
&\le \left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|^2 \left\|\boldsymbol{\Sigma}_\star^{-1/2}(\boldsymbol{R}^\top\boldsymbol{R}-\boldsymbol{\Sigma}_\star)\boldsymbol{\Sigma}_\star^{-1/2}\right\| \operatorname{tr}\left(\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_R^\top\right) \\
&\le \frac{1}{(1-\epsilon)^2}\left\|\boldsymbol{\Sigma}_\star^{-1/2}(\boldsymbol{R}^\top\boldsymbol{R}-\boldsymbol{\Sigma}_\star)\boldsymbol{\Sigma}_\star^{-1/2}\right\| \operatorname{tr}\left(\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_R^\top\right).
\end{aligned}$$

Further notice that

$$\begin{aligned}
\left\|\boldsymbol{\Sigma}_\star^{-1/2}(\boldsymbol{R}^\top\boldsymbol{R}-\boldsymbol{\Sigma}_\star)\boldsymbol{\Sigma}_\star^{-1/2}\right\| &= \left\|\boldsymbol{\Sigma}_\star^{-1/2}(\boldsymbol{R}_\star^\top\boldsymbol{\Delta}_R+\boldsymbol{\Delta}_R^\top\boldsymbol{R}_\star+\boldsymbol{\Delta}_R^\top\boldsymbol{\Delta}_R)\boldsymbol{\Sigma}_\star^{-1/2}\right\| \\
&\le 2\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\| + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\|^2 \\
&\le 2\epsilon+\epsilon^2.
\end{aligned}$$

Take the preceding two bounds together to arrive at

$$|\mathfrak{F}_3| \le \frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}\operatorname{tr}\left(\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_R^\top\right).$$

Combining the bounds for $\mathfrak{F}_1, \mathfrak{F}_2, \mathfrak{F}_3$, one has

$$\begin{aligned}\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 &= \left\|(1-\eta)\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2} - \eta\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 \\ &\le \left((1-\eta)^2 + \frac{2\epsilon}{1-\epsilon}\eta(1-\eta)\right)\operatorname{tr}\left(\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\right) + \frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}\eta^2\operatorname{tr}\left(\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_R^\top\right). \end{aligned} \tag{A.13}$$

A similarly bound holds for the second square $\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2$ in (A.10). Therefore we obtain

$$\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 \le \rho^2(\eta;\epsilon)\operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star),$$

where we identify

$$\operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star) = \operatorname{tr}(\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top) + \operatorname{tr}(\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_R^\top) \tag{A.14}$$

and the contraction rate $\rho^2(\eta;\epsilon)$ is given by

$$\rho^2(\eta;\epsilon) \coloneqq (1-\eta)^2 + \frac{2\epsilon}{1-\epsilon}\eta(1-\eta) + \frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}\eta^2.$$

With $\epsilon = 0.1$ and $0 < \eta \le 2/3$, one has $\rho(\eta;\epsilon) \le 1 - 0.7\eta$. Thus we conclude that

$$\begin{aligned}\operatorname{dist}(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) &\le \sqrt{\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2} \\ &\le (1-0.7\eta)\operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \\ &\le (1-0.7\eta)^{t+1}\operatorname{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \le (1-0.7\eta)^{t+1}0.1\sigma_r(\boldsymbol{X}_\star).\end{aligned}$$

This proves the first induction hypothesis. The existence of the optimal alignment matrix $\boldsymbol{Q}_{t+1}$ between $\boldsymbol{F}_{t+1}$ and $\boldsymbol{F}_\star$ is assured by Lemma 14, which finishes the proof for the second hypothesis.

So far, we have demonstrated the first conclusion in the theorem. The second conclusion is

an easy consequence of Lemma 18 as

$$\begin{aligned}\left\|\boldsymbol{L}_t\boldsymbol{R}_t^\top-\boldsymbol{X}_\star\right\|_{\mathsf{F}} &\le \left(1+\frac{\epsilon}{2}\right)\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)\\ &\le \left(1+\frac{\epsilon}{2}\right)\sqrt{2}\,\mathrm{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star)\\ &\le 1.5\,\mathrm{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star).\end{aligned} \tag{A.15}$$

Here, the second line follows from the elementary inequality $a+b\le\sqrt{2(a^2+b^2)}$ and the expression of $\mathrm{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star)$ in (A.14). The proof is now completed.

A.3 Proof for Low-Rank Matrix Sensing

We start by recording a useful lemma.

Lemma 22 ([CP11]). *Suppose that $\mathcal{A}(\cdot)$ obeys the $2r$-RIP with a constant δ_{2r}. Then for any $\boldsymbol{X}_1,\boldsymbol{X}_2\in\mathbb{R}^{n_1\times n_2}$ of rank at most r, one has*

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

which can be stated equivalently as

$$\left|\mathrm{tr}\left((\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{X}_1)\boldsymbol{X}_2^\top\right)\right|\le\delta_{2r}\|\boldsymbol{X}_1\|_{\mathsf{F}}\|\boldsymbol{X}_2\|_{\mathsf{F}}. \tag{A.16}$$

As a simple corollary, one has that for any matrix $\boldsymbol{R}\in\mathbb{R}^{n_2\times r}$:

$$\|(\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{X}_1)\boldsymbol{R}\|_{\mathsf{F}}\le\delta_{2r}\|\boldsymbol{X}_1\|_{\mathsf{F}}\|\boldsymbol{R}\|. \tag{A.17}$$

This is due to the fact that

$$\begin{aligned}\|(\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{X}_1)\boldsymbol{R}\|_{\mathsf{F}} &= \max_{\widetilde{\boldsymbol{L}}:\|\widetilde{\boldsymbol{L}}\|_{\mathsf{F}}\le1}\ \mathrm{tr}\left((\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{X}_1)\boldsymbol{R}\widetilde{\boldsymbol{L}}^\top\right)\\ &\le \max_{\widetilde{\boldsymbol{L}}:\|\widetilde{\boldsymbol{L}}\|_{\mathsf{F}}\le1}\ \delta_{2r}\|\boldsymbol{X}_1\|_{\mathsf{F}}\|\widetilde{\boldsymbol{L}}\boldsymbol{R}^\top\|_{\mathsf{F}}\end{aligned}$$

$$\leq \delta_{2r}\|\boldsymbol{X}_1\|_{\mathsf{F}}\|\boldsymbol{R}\|.$$

Here, the first line follows from the variational representation of the Frobenius norm, the second line follows from (A.16), and the last line follows from the relation $\|\boldsymbol{AB}\|_{\mathsf{F}} \leq \|\boldsymbol{A}\|_{\mathsf{F}}\|\boldsymbol{B}\|$.

A.3.1 Proof of Lemma 1

The proof mostly mirrors that in Section A.2.2. First, in view of the condition $\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \leq 0.1\sigma_r(\boldsymbol{X}_\star)$ and Lemma 14, one knows that $\boldsymbol{Q}_t$, the optimal alignment matrix between $\boldsymbol{F}_t$ and $\boldsymbol{F}_\star$ exists. Therefore, for notational convenience, denote $\boldsymbol{L} \coloneqq \boldsymbol{L}_t\boldsymbol{Q}_t$, $\boldsymbol{R} \coloneqq \boldsymbol{R}_t\boldsymbol{Q}_t^{-\top}$, $\boldsymbol{\Delta}_L \coloneqq \boldsymbol{L} - \boldsymbol{L}_\star$, $\boldsymbol{\Delta}_R \coloneqq \boldsymbol{R} - \boldsymbol{R}_\star$, and $\epsilon \coloneqq 0.1$. Similar to the derivation in (A.12), we have

$$\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\| \vee \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\| \leq \epsilon. \tag{A.18}$$

The conclusion $\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\|_{\mathsf{F}} \leq 1.5\,\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$ is a simple consequence of Lemma 18; see (A.15) for a detailed argument. From now on, we focus on proving the distance contraction.

With these notations in place, we have by the definition of $\mathrm{dist}(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star)$ that

$$\mathrm{dist}^2(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) \leq \left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2. \tag{A.19}$$

Apply the update rule (2.13) and the decomposition $\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star = \boldsymbol{\Delta}_L\boldsymbol{R}^\top + \boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top$ to obtain

$$\begin{aligned}
(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2} &= \left(\boldsymbol{L} - \eta\mathcal{A}^*\mathcal{A}(\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1} - \boldsymbol{L}_\star\right)\boldsymbol{\Sigma}_\star^{1/2} \\
&= \left(\boldsymbol{\Delta}_L - \eta(\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1} - \eta(\mathcal{A}^*\mathcal{A} - \mathcal{I})(\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\right)\boldsymbol{\Sigma}_\star^{1/2} \\
&= (1-\eta)\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2} - \eta\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2} - \eta(\mathcal{A}^*\mathcal{A} - \mathcal{I})(\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}.
\end{aligned}$$

This allows us to expand the first square in (A.19) as

$$\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 = \underbrace{\left\|(1-\eta)\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2} - \eta\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2}_{\mathfrak{S}_1}$$

$$
\begin{aligned}
&- 2\eta(1-\eta)\underbrace{\operatorname{tr}\Big((\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\Big)}_{\mathfrak{S}_2}\\
&+ 2\eta^2\underbrace{\operatorname{tr}\Big((\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\boldsymbol{\Delta}_R\boldsymbol{L}_\star^\top\Big)}_{\mathfrak{S}_3}\\
&+ \eta^2\underbrace{\Big\|(\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\Big\|_{\mathsf{F}}^2}_{\mathfrak{S}_4}.
\end{aligned}
$$

In what follows, we shall control the four terms separately, of which $\mathfrak{S}_1$ is the main term, and $\mathfrak{S}_2, \mathfrak{S}_3$ and $\mathfrak{S}_4$ are perturbation terms.

1. Notice that the main term $\mathfrak{S}_1$ has already been controlled in (A.13) under the condition (A.18). It obeys

$$
\mathfrak{S}_1 \le \left((1-\eta)^2+\frac{2\epsilon}{1-\epsilon}\eta(1-\eta)\right)\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+\frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}\eta^2\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2.
$$

2. For the second term $\mathfrak{S}_2$, decompose $\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star=\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top+\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top+\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top$ and apply the triangle inequality to obtain

$$
\begin{aligned}
|\mathfrak{S}_2| &= \Big|\operatorname{tr}\Big((\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top+\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top+\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\Big)\Big|\\
&\le \Big|\operatorname{tr}\Big((\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\Big)\Big|\\
&\quad+\Big|\operatorname{tr}\Big((\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\Big)\Big|\\
&\quad+\Big|\operatorname{tr}\Big((\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\Big)\Big|.
\end{aligned}
$$

Invoke Lemma 22 to further obtain

$$
\begin{aligned}
|\mathfrak{S}_2| &\le \delta_{2r}\left(\|\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top\|_{\mathsf{F}}+\|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}}\right)\Big\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\Big\|_{\mathsf{F}}\\
&\le \delta_{2r}\left(\|\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top\|_{\mathsf{F}}+\|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}}\right)\Big\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\Big\|\,\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}},
\end{aligned}
$$

where the second line follows from the relation $\|\boldsymbol{A}\boldsymbol{B}\|_{\mathsf{F}}\le\|\boldsymbol{A}\|\|\boldsymbol{B}\|_{\mathsf{F}}$. Take the condition (A.18)

and Lemmas 17 and 18 together to obtain

$$\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\| \leq \frac{1}{1-\epsilon};$$
$$\|\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top\|_{\mathsf{F}} + \|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}} \leq (1+\frac{\epsilon}{2})\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right).$$

These consequences further imply that

$$\begin{aligned}|\mathfrak{S}_2| &\leq \frac{\delta_{2r}(2+\epsilon)}{2(1-\epsilon)}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} \\ &= \frac{\delta_{2r}(2+\epsilon)}{2(1-\epsilon)}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right).\end{aligned}$$

For the term $\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}$, we can apply the elementary inequality $2ab \leq a^2+b^2$ to see

$$\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} \leq \frac{1}{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \frac{1}{2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2.$$

The preceding two bounds taken collectively yield

$$|\mathfrak{S}_2| \leq \frac{\delta_{2r}(2+\epsilon)}{2\,(1-\epsilon)}\left(\frac{3}{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \frac{1}{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right).$$

3. The third term $\mathfrak{S}_3$ can be similarly bounded as

$$\begin{aligned}|\mathfrak{S}_3| &\leq \delta_{2r}\left(\|\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top\|_{\mathsf{F}} + \|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}}\right)\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\boldsymbol{\Delta}_R\boldsymbol{L}_\star^\top\right\|_{\mathsf{F}} \\ &\leq \delta_{2r}\left(\|\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top\|_{\mathsf{F}} + \|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}}\right)\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|^2\|\boldsymbol{\Delta}_R\boldsymbol{L}_\star^\top\|_{\mathsf{F}} \\ &\leq \frac{\delta_{2r}(2+\epsilon)}{2(1-\epsilon)^2}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} \\ &\leq \frac{\delta_{2r}(2+\epsilon)}{2(1-\epsilon)^2}\left(\frac{1}{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \frac{3}{2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right).\end{aligned}$$

4. We are then left with the last term $\mathfrak{S}_4$, for which we have

$$\begin{aligned}\sqrt{\mathfrak{S}_4} &= \left\|(\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}\\ &\le \left\|(\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}\\ &\quad+\left\|(\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}\\ &\quad+\left\|(\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}},\end{aligned}$$

where once again we use the decomposition $\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star=\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top+\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top+\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top$. Use (A.17) to see that

$$\sqrt{\mathfrak{S}_4}\le\delta_{2r}\left(\|\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top\|_{\mathsf{F}}+\|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}}\right)\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|.$$

Repeating the same argument in bounding $\mathfrak{S}_2$ yields

$$\sqrt{\mathfrak{S}_4}\le\frac{\delta_{2r}\,(2+\epsilon)}{2\,(1-\epsilon)}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right).$$

We can then take the squares of both sides and use $(a+b)^2\le 2a^2+2b^2$ to reach

$$\mathfrak{S}_4\le\frac{\delta_{2r}^2(2+\epsilon)^2}{2(1-\epsilon)^2}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right).$$

Taking the bounds for $\mathfrak{S}_1,\mathfrak{S}_2,\mathfrak{S}_3,\mathfrak{S}_4$ collectively yields

$$\begin{aligned}\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 &\le\left((1-\eta)^2+\frac{2\epsilon}{1-\epsilon}\eta(1-\eta)\right)\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+\frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}\eta^2\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\\ &\quad+\frac{\delta_{2r}(2+\epsilon)}{1-\epsilon}\eta(1-\eta)\left(\frac{3}{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+\frac{1}{2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right)\\ &\quad+\frac{\delta_{2r}(2+\epsilon)}{(1-\epsilon)^2}\eta^2\left(\frac{1}{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+\frac{3}{2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right)\\ &\quad+\frac{\delta_{2r}^2(2+\epsilon)^2}{2(1-\epsilon)^2}\eta^2\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right).\end{aligned}$$

Similarly, we can expand the second square in (A.19) and obtain a similar bound. Combine both

to obtain

$$\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2+\left\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top}-\boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2\le\rho^2(\eta;\epsilon,\delta_{2r})\,\mathrm{dist}^2(\boldsymbol{F}_t,\boldsymbol{F}_\star),$$

where the contraction rate is given by

$$\rho^2(\eta;\epsilon,\delta_{2r}):=(1-\eta)^2+\frac{2\epsilon+\delta_{2r}(4+2\epsilon)}{1-\epsilon}\eta(1-\eta)+\frac{2\epsilon+\epsilon^2+\delta_{2r}(4+2\epsilon)+\delta_{2r}^2(2+\epsilon)^2}{(1-\epsilon)^2}\eta^2.$$

With $\epsilon=0.1$, $\delta_{2r}\le 0.02$, and $0<\eta\le 2/3$, one has $\rho(\eta;\epsilon,\delta_{2r})\le 1-0.6\eta$. Thus we conclude that

$$\begin{aligned}\mathrm{dist}(\boldsymbol{F}_{t+1},\boldsymbol{F}_\star)&\le\sqrt{\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2+\left\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top}-\boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2}\\&\le(1-0.6\eta)\,\mathrm{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star).\end{aligned}$$

A.3.2 Proof of Lemma 2

With the knowledge of partial Frobenius norm $\|\cdot\|_{\mathsf{F},r}$, we are ready to establish the claimed result. Invoke Lemma 16 to relate $\mathrm{dist}(\boldsymbol{F}_0,\boldsymbol{F}_\star)$ to $\|\boldsymbol{L}_0\boldsymbol{R}_0^\top-\boldsymbol{X}_\star\|_{\mathsf{F}}$, and use that $\boldsymbol{L}_0\boldsymbol{R}_0^\top-\boldsymbol{X}_\star$ has rank at most $2r$ to obtain

$$\mathrm{dist}(\boldsymbol{F}_0,\boldsymbol{F}_\star)\le\sqrt{\sqrt{2}+1}\left\|\boldsymbol{L}_0\boldsymbol{R}_0^\top-\boldsymbol{X}_\star\right\|_{\mathsf{F}}\le\sqrt{2(\sqrt{2}+1)}\left\|\boldsymbol{L}_0\boldsymbol{R}_0^\top-\boldsymbol{X}_\star\right\|_{\mathsf{F},r}.$$

Note that $\boldsymbol{L}_0\boldsymbol{R}_0^\top$ is the best rank-r approximation of $\mathcal{A}^*\mathcal{A}(\boldsymbol{X}_\star)$, and apply the triangle inequality combined with Lemma 21 to obtain

$$\begin{aligned}\left\|\boldsymbol{L}_0\boldsymbol{R}_0^\top-\boldsymbol{X}_\star\right\|_{\mathsf{F},r}&\le\left\|\mathcal{A}^*\mathcal{A}(\boldsymbol{X}_\star)-\boldsymbol{L}_0\boldsymbol{R}_0^\top\right\|_{\mathsf{F},r}+\|\mathcal{A}^*\mathcal{A}(\boldsymbol{X}_\star)-\boldsymbol{X}_\star\|_{\mathsf{F},r}\\&\le 2\,\|(\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{X}_\star)\|_{\mathsf{F},r}\le 2\delta_{2r}\|\boldsymbol{X}_\star\|_{\mathsf{F}}.\end{aligned}$$

Here, the last inequality follows from combining Lemma 20 and (A.17) as

$$\|(\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{X}_\star)\|_{\mathsf{F},r}=\max_{\widetilde{\boldsymbol{R}}\in\mathbb{R}^{n_2\times r}:\|\widetilde{\boldsymbol{R}}\|\le 1}\left\|(\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{X}_\star)\widetilde{\boldsymbol{R}}\right\|_{\mathsf{F}}\le\delta_{2r}\|\boldsymbol{X}_\star\|_{\mathsf{F}}.$$

As a result, one has

$$\operatorname{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \le 2\sqrt{2(\sqrt{2}+1)}\delta_{2r}\|\boldsymbol{X}_\star\|_{\mathsf{F}} \le 5\delta_{2r}\sqrt{r}\kappa\sigma_r(\boldsymbol{X}_\star).$$

A.4 Proof for Robust PCA

We first establish a useful property regarding the truncation operator $\mathcal{T}_{2\alpha}[\cdot]$.

Lemma 23. *Given $\boldsymbol{S}_\star \in \mathcal{S}_\alpha$ and $\boldsymbol{S} = \mathcal{T}_{2\alpha}[\boldsymbol{X}_\star + \boldsymbol{S}_\star - \boldsymbol{L}\boldsymbol{R}^\top]$, one has*

$$\|\boldsymbol{S} - \boldsymbol{S}_\star\|_\infty \le 2\|\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star\|_\infty. \tag{A.20}$$

In addition, for any low-rank matrix $\boldsymbol{M} = \boldsymbol{L}_M\boldsymbol{R}_M^\top \in \mathbb{R}^{n_1\times n_2}$ with $\boldsymbol{L}_M \in \mathbb{R}^{n_1\times r}, \boldsymbol{R}_M \in \mathbb{R}^{n_2\times r}$, one has

$$\begin{aligned}|\langle \boldsymbol{S} - \boldsymbol{S}_\star, \boldsymbol{M}\rangle| \le{}& \sqrt{3\alpha}\nu\left(\|(\boldsymbol{L}-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|(\boldsymbol{R}-\boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)\|\boldsymbol{M}\|_{\mathsf{F}} \\ &+ 2\sqrt{\alpha}\left(\sqrt{n_1}\|\boldsymbol{L}_M\|_{2,\infty}\|\boldsymbol{R}_M\|_{\mathsf{F}} \wedge \sqrt{n_2}\|\boldsymbol{L}_M\|_{\mathsf{F}}\|\boldsymbol{R}_M\|_{2,\infty}\right)\|\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star\|_{\mathsf{F}},\end{aligned} \tag{A.21}$$

where ν obeys

$$\nu \ge \frac{\sqrt{n_1}}{2}\left(\|\boldsymbol{L}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} + \|\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\right) \vee \frac{\sqrt{n_2}}{2}\left(\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} + \|\boldsymbol{R}_\star\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\right).$$

Proof. Denote $\boldsymbol{\Delta}_L \coloneqq \boldsymbol{L} - \boldsymbol{L}_\star$, $\boldsymbol{\Delta}_R \coloneqq \boldsymbol{R} - \boldsymbol{R}_\star$, and $\boldsymbol{\Delta}_X \coloneqq \boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star$. Let $\Omega, \Omega_\star$ be the support of $\boldsymbol{S}$ and $\boldsymbol{S}_\star$, respectively. As a result, $\boldsymbol{S} - \boldsymbol{S}_\star$ is supported on $\Omega \cup \Omega_\star$.

We start with proving the first claim, i.e. (A.20). For $(i,j) \in \Omega$, by the definition of $\mathcal{T}_{2\alpha}[\cdot]$, we have $(\boldsymbol{S} - \boldsymbol{S}_\star)_{i,j} = (-\boldsymbol{\Delta}_X)_{i,j}$. For $(i,j) \in \Omega_\star \setminus \Omega$, one necessarily has $\boldsymbol{S}_{i,j} = 0$ and therefore $(\boldsymbol{S} - \boldsymbol{S}_\star)_{i,j} = (-\boldsymbol{S}_\star)_{i,j}$. Again by the definition of the operator $\mathcal{T}_{2\alpha}[\cdot]$, we know $|\boldsymbol{S}_\star - \boldsymbol{\Delta}_X|_{i,j}$ is either smaller than $|\boldsymbol{S}_\star - \boldsymbol{\Delta}_X|_{i,(2\alpha n_2)}$ or $|\boldsymbol{S}_\star - \boldsymbol{\Delta}_X|_{(2\alpha n_1),j}$. Furthermore, we know that $\boldsymbol{S}_\star$ contains at most α-fraction nonzero entries per row and column. Consequently, one has $|\boldsymbol{S}_\star - \boldsymbol{\Delta}_X|_{i,j} \le$

$|\boldsymbol{\Delta}_X|_{i,(\alpha n_2)} \vee |\boldsymbol{\Delta}_X|_{(\alpha n_1),j}$. Combining the two cases above, we conclude that

$$|\boldsymbol{S}-\boldsymbol{S}_\star|_{i,j} \le \begin{cases} |\boldsymbol{\Delta}_X|_{i,j}, & (i,j)\in\Omega \\ |\boldsymbol{\Delta}_X|_{i,j} + \left(|\boldsymbol{\Delta}_X|_{i,(\alpha n_2)} \vee |\boldsymbol{\Delta}_X|_{(\alpha n_1),j}\right), & (i,j)\in\Omega_\star\setminus\Omega \end{cases}. \tag{A.22}$$

This immediately implies the ℓ_∞ norm bound (A.20).

Next, we prove the second claim (A.21). Recall that $\boldsymbol{S}-\boldsymbol{S}_\star$ is supported on $\Omega\cup\Omega_\star$. We then have

$$\begin{aligned} |\langle \boldsymbol{S}-\boldsymbol{S}_\star, \boldsymbol{M}\rangle| &\le \sum_{(i,j)\in\Omega} |\boldsymbol{S}-\boldsymbol{S}_\star|_{i,j}|\boldsymbol{M}|_{i,j} + \sum_{(i,j)\in\Omega_\star\setminus\Omega} |\boldsymbol{S}-\boldsymbol{S}_\star|_{i,j}|\boldsymbol{M}|_{i,j} \\ &\le \sum_{(i,j)\in\Omega\cup\Omega_\star} |\boldsymbol{\Delta}_X|_{i,j}|\boldsymbol{M}|_{i,j} + \sum_{(i,j)\in\Omega_\star\setminus\Omega} \left(|\boldsymbol{\Delta}_X|_{i,(\alpha n_2)} + |\boldsymbol{\Delta}_X|_{(\alpha n_1),j}\right)|\boldsymbol{M}|_{i,j}, \end{aligned}$$

where the second line follows from (A.22). Let $\beta>0$ be some positive number, whose value will be determined later. Use $2ab \le \beta^{-1}a^2+\beta b^2$ to further obtain

$$|\langle \boldsymbol{S}-\boldsymbol{S}_\star, \boldsymbol{M}\rangle| \le \underbrace{\sum_{(i,j)\in\Omega\cup\Omega_\star} |\boldsymbol{\Delta}_X|_{i,j}|\boldsymbol{M}|_{i,j}}_{\mathfrak{A}_1} + \frac{1}{2\beta}\underbrace{\sum_{(i,j)\in\Omega_\star\setminus\Omega}\left(|\boldsymbol{\Delta}_X|^2_{i,(\alpha n_2)} + |\boldsymbol{\Delta}_X|^2_{(\alpha n_1),j}\right)}_{\mathfrak{A}_2} + \beta\underbrace{\sum_{(i,j)\in\Omega_\star\setminus\Omega}|\boldsymbol{M}|^2_{i,j}}_{\mathfrak{A}_3}.$$

In regard to the three terms $\mathfrak{A}_1, \mathfrak{A}_2$ and $\mathfrak{A}_3$, we have the following claims, whose proofs are deferred to the end.

Claim 1. *The first term $\mathfrak{A}_1$ satisfies*

$$\mathfrak{A}_1 \le \sqrt{3\alpha}\nu\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)\|\boldsymbol{M}\|_{\mathsf{F}}.$$

Claim 2. *The second term $\mathfrak{A}_2$ satisfies*

$$\mathfrak{A}_2 \le 2\|\boldsymbol{\Delta}_X\|_{\mathsf{F}}^2.$$

Claim 3. *The third term* $\mathfrak{A}_3$ *satisfies*

$$\mathfrak{A}_3 \le \alpha \left(n_1 \|\boldsymbol{L}_M\|_{2,\infty}^2 \|\boldsymbol{R}_M\|_{\mathsf{F}}^2 \wedge n_2 \|\boldsymbol{L}_M\|_{\mathsf{F}}^2 \|\boldsymbol{R}_M\|_{2,\infty}^2 \right).$$

Combine the pieces to reach

$$\begin{aligned} |\langle \boldsymbol{S} - \boldsymbol{S}_\star, \boldsymbol{M} \rangle| &\le \sqrt{3\alpha}\nu \left(\|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} \right) \|\boldsymbol{M}\|_{\mathsf{F}} \\ &\quad + \frac{\|\boldsymbol{\Delta}_X\|_{\mathsf{F}}^2}{\beta} + \beta\alpha \left(n_1 \|\boldsymbol{L}_M\|_{2,\infty}^2 \|\boldsymbol{R}_M\|_{\mathsf{F}}^2 \wedge n_2 \|\boldsymbol{L}_M\|_{\mathsf{F}}^2 \|\boldsymbol{R}_M\|_{2,\infty}^2 \right). \end{aligned}$$

One can then choose β optimally to yield

$$\begin{aligned} |\langle \boldsymbol{S} - \boldsymbol{S}_\star, \boldsymbol{M} \rangle| &\le \sqrt{3\alpha}\nu \left(\|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} \right) \|\boldsymbol{M}\|_{\mathsf{F}} \\ &\quad + 2\sqrt{\alpha} \left(\sqrt{n_1} \|\boldsymbol{L}_M\|_{2,\infty} \|\boldsymbol{R}_M\|_{\mathsf{F}} \wedge \sqrt{n_2} \|\boldsymbol{L}_M\|_{\mathsf{F}} \|\boldsymbol{R}_M\|_{2,\infty} \right) \|\boldsymbol{\Delta}_X\|_{\mathsf{F}}. \end{aligned}$$

This finishes the proof. □

Proof of Claim 1. Use the decomposition $\boldsymbol{\Delta}_X = \boldsymbol{\Delta}_L \boldsymbol{R}^\top + \boldsymbol{L}_\star \boldsymbol{\Delta}_R^\top = \boldsymbol{\Delta}_L \boldsymbol{R}_\star^\top + \boldsymbol{L} \boldsymbol{\Delta}_R^\top$ to obtain

$$\begin{aligned} |\boldsymbol{\Delta}_X|_{i,j} &\le \|(\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2})_{i,\cdot}\|_2 \|\boldsymbol{R} \boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} + \|\boldsymbol{L}_\star \boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} \|(\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{1/2})_{j,\cdot}\|_2, \quad \text{and} \\ |\boldsymbol{\Delta}_X|_{i,j} &\le \|(\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2})_{i,\cdot}\|_2 \|\boldsymbol{R}_\star \boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} + \|\boldsymbol{L} \boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} \|(\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{1/2})_{j,\cdot}\|_2. \end{aligned}$$

Take the average to yield

$$|\boldsymbol{\Delta}_X|_{i,j} \le \frac{\nu}{\sqrt{n_2}} \|(\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2})_{i,\cdot}\|_2 + \frac{\nu}{\sqrt{n_1}} \|(\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{1/2})_{j,\cdot}\|_2,$$

where we have used the assumption on ν. With this upper bound on $|\boldsymbol{\Delta}_X|_{i,j}$ in place, we can further control $\mathfrak{A}_1$ as

$$\mathfrak{A}_1 \le \sum_{(i,j)\in\Omega\cup\Omega_\star} \frac{\nu}{\sqrt{n_2}} \|(\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2})_{i,\cdot}\|_2 |\boldsymbol{M}|_{i,j} + \sum_{(i,j)\in\Omega\cup\Omega_\star} \frac{\nu}{\sqrt{n_1}} \|(\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{1/2})_{j,\cdot}\|_2 |\boldsymbol{M}|_{i,j}$$

$$\leq \left(\sqrt{\sum_{(i,j)\in\Omega\cup\Omega_\star} \|(\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2})_{i,\cdot}\|_2^2/n_2} + \sqrt{\sum_{(i,j)\in\Omega\cup\Omega_\star} \|(\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{1/2})_{j,\cdot}\|_2^2/n_1} \right) \nu \|\boldsymbol{M}\|_{\mathsf{F}}.$$

Regarding the first term, one has

$$\begin{aligned}
\sum_{(i,j)\in\Omega\cup\Omega_\star} \|(\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2})_{i,\cdot}\|_2^2 &= \sum_{i=1}^{n_1} \sum_{j:(i,j)\in\Omega\cup\Omega_\star} \|(\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2})_{i,\cdot}\|_2^2 \\
&\leq 3\alpha n_2 \sum_{i=1}^{n_1} \|(\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2})_{i,\cdot}\|_2^2 \\
&= 3\alpha n_2 \|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2,
\end{aligned}$$

where the second line follows from the fact that $\Omega \cup \Omega_\star$ contains at most $3\alpha n_2$ non-zero entries in each row. Similarly, we can show that

$$\sum_{(i,j)\in\Omega\cup\Omega_\star} \|(\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{1/2})_{j,\cdot}\|_2^2 \leq 3\alpha n_1 \|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2.$$

In all, we arrive at

$$\mathfrak{A}_1 \leq \sqrt{3\alpha}\nu \left(\|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} \right) \|\boldsymbol{M}\|_{\mathsf{F}},$$

which is the desired claim. □

Proof of Claim 2. Recall that $(\boldsymbol{\Delta}_X)_{i,(\alpha n_2)}$ denotes the (αn_2)-th largest entry in the i-th row of $\boldsymbol{\Delta}_X$. One necessarily has

$$\alpha n_2 |\boldsymbol{\Delta}_X|_{i,(\alpha n_2)}^2 \leq \|(\boldsymbol{\Delta}_X)_{i,\cdot}\|_2^2.$$

As a result, we obtain

$$\sum_{(i,j)\in\Omega_\star\setminus\Omega} |\boldsymbol{\Delta}_X|_{i,(\alpha n_2)}^2 \leq \sum_{(i,j)\in\Omega_\star} |\boldsymbol{\Delta}_X|_{i,(\alpha n_2)}^2$$

$$\leq \sum_{i=1}^{n_1} \sum_{j:(i,j)\in\Omega_\star} \frac{\|(\boldsymbol{\Delta}_X)_{i,\cdot}\|_2^2}{\alpha n_2}$$
$$\leq \sum_{i=1}^{n_1} \|(\boldsymbol{\Delta}_X)_{i,\cdot}\|_2^2 = \|\boldsymbol{\Delta}_X\|_{\mathsf{F}}^2,$$

where the last line follows from the fact that $\Omega_\star$ contains at most αn_2 nonzero entries in each row. Similarly one can show that

$$\sum_{(i,j)\in\Omega_\star\setminus\Omega} |\boldsymbol{\Delta}_X|_{(\alpha n_1),j}^2 \leq \|\boldsymbol{\Delta}_X\|_{\mathsf{F}}^2.$$

Combining the above two bounds with the definition of $\mathfrak{A}_2$ completes the proof. □

Proof of Claim 3. By definition, $\boldsymbol{M} = \boldsymbol{L}_M \boldsymbol{R}_M^\top$, and hence one has

$$\mathfrak{A}_3 = \sum_{(i,j)\in\Omega_\star\setminus\Omega} |(\boldsymbol{L}_M)_{i,\cdot}(\boldsymbol{R}_M)_{j,\cdot}^\top|^2 \leq \sum_{(i,j)\in\Omega_\star} |(\boldsymbol{L}_M)_{i,\cdot}(\boldsymbol{R}_M)_{j,\cdot}^\top|^2.$$

We can further upper bound $\mathfrak{A}_3$ as

$$\mathfrak{A}_3 \leq \sum_{(i,j)\in\Omega_\star} \|(\boldsymbol{L}_M)_{i,\cdot}\|_2^2 \|(\boldsymbol{R}_M)_{j,\cdot}\|_2^2$$
$$\leq \sum_{i=1}^{n_1} \sum_{j:(i,j)\in\Omega_\star} \|(\boldsymbol{L}_M)_{i,\cdot}\|_2^2 \|\boldsymbol{R}_M\|_{2,\infty}^2$$
$$\leq \sum_{i=1}^{n_1} \alpha n_2 \|(\boldsymbol{L}_M)_{i,\cdot}\|_2^2 \|\boldsymbol{R}_M\|_{2,\infty}^2 = \alpha n_2 \|\boldsymbol{L}_M\|_{\mathsf{F}}^2 \|\boldsymbol{R}_M\|_{2,\infty}^2,$$

where the last line follows from the fact that $\Omega_\star$ contains at most αn_2 non-zero entries in each row. Similarly, one can obtain

$$\mathfrak{A}_3 \leq \alpha n_1 \|\boldsymbol{L}_M\|_{2,\infty}^2 \|\boldsymbol{R}_M\|_{\mathsf{F}}^2,$$

which completes the proof. □

A.4.1 Proof of Lemma 3

We begin with introducing several useful notations and facts. In view of the condition $\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \le 0.02\sigma_r(\boldsymbol{X}_\star)$ and Lemma 14, one knows that $\boldsymbol{Q}_t$, the optimal alignment matrix between $\boldsymbol{F}_t$ and $\boldsymbol{F}_\star$ exists. Therefore, for notational convenience, denote $\boldsymbol{L} \coloneqq \boldsymbol{L}_t\boldsymbol{Q}_t$, $\boldsymbol{R} \coloneqq \boldsymbol{R}_t\boldsymbol{Q}_t^{-\top}$, $\boldsymbol{\Delta}_L \coloneqq \boldsymbol{L} - \boldsymbol{L}_\star$, $\boldsymbol{\Delta}_R \coloneqq \boldsymbol{R} - \boldsymbol{R}_\star$, $\boldsymbol{S} \coloneqq \boldsymbol{S}_t = \mathcal{T}_{2\alpha}[\boldsymbol{X}_\star + \boldsymbol{S}_\star - \boldsymbol{L}\boldsymbol{R}^\top]$, and $\epsilon \coloneqq 0.02$. Similar to the derivation in (A.12), we have

$$\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\| \vee \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\| \le \epsilon. \tag{A.23}$$

Moreover, the incoherence condition

$$\sqrt{n_1}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \vee \sqrt{n_2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \le \sqrt{\mu r}\sigma_r(\boldsymbol{X}_\star) \tag{A.24}$$

implies

$$\sqrt{n_1}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} \vee \sqrt{n_2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} \le \sqrt{\mu r}, \tag{A.25}$$

which combined with the triangle inequality further implies

$$\sqrt{n_1}\|\boldsymbol{L}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} \vee \sqrt{n_2}\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} \le 2\sqrt{\mu r}. \tag{A.26}$$

The conclusion $\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\|_{\mathsf{F}} \le 1.5\,\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$ is a simple consequence of Lemma 18; see (A.15) for a detailed argument. In what follows, we shall prove the distance contraction and the incoherence condition separately.

Distance contraction

By the definition of $\mathrm{dist}^2(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star)$, one has

$$\mathrm{dist}^2(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) \le \left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2. \tag{A.27}$$

From now on, we focus on controlling the first square $\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2$. In view of the update rule (2.18), one has

$$\begin{aligned}(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2} &= \Big(\boldsymbol{L} - \eta(\boldsymbol{L}\boldsymbol{R}^\top + \boldsymbol{S} - \boldsymbol{X}_\star - \boldsymbol{S}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1} - \boldsymbol{L}_\star\Big)\boldsymbol{\Sigma}_\star^{1/2} \\ &= \Big(\boldsymbol{\Delta}_L - \eta(\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1} - \eta(\boldsymbol{S} - \boldsymbol{S}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\Big)\boldsymbol{\Sigma}_\star^{1/2} \\ &= (1-\eta)\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2} - \eta\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2} - \eta(\boldsymbol{S} - \boldsymbol{S}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}.\end{aligned} \tag{A.28}$$

Here, we use the notation introduced above and the decomposition $\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star = \boldsymbol{\Delta}_L\boldsymbol{R}^\top + \boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top$. Take the squared Frobenius norm of both sides of (A.28) to obtain

$$\begin{aligned}\Big\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\Big\|_{\mathsf{F}}^2 = &\underbrace{\Big\|(1-\eta)\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2} - \eta\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\Big\|_{\mathsf{F}}^2}_{\mathfrak{R}_1} \\ &- 2\eta(1-\eta)\underbrace{\operatorname{tr}\Big((\boldsymbol{S} - \boldsymbol{S}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\Big)}_{\mathfrak{R}_2} \\ &+ 2\eta^2\underbrace{\operatorname{tr}\Big((\boldsymbol{S} - \boldsymbol{S}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\boldsymbol{\Delta}_R\boldsymbol{L}_\star^\top\Big)}_{\mathfrak{R}_3} \\ &+ \eta^2\underbrace{\Big\|(\boldsymbol{S} - \boldsymbol{S}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\Big\|_{\mathsf{F}}^2}_{\mathfrak{R}_4}.\end{aligned}$$

In the sequel, we shall bound the four terms separately, of which $\mathfrak{R}_1$ is the main term, and $\mathfrak{R}_2, \mathfrak{R}_3$ and $\mathfrak{R}_4$ are perturbation terms.

1. Notice that the main term $\mathfrak{R}_1$ has already been controlled in (A.13) under the condition (A.23). It obeys

$$\mathfrak{R}_1 \le \left((1-\eta)^2 + \frac{2\epsilon}{1-\epsilon}\eta(1-\eta)\right)\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}\eta^2\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2.$$

2. For the second term $\mathfrak{R}_2$, set $\boldsymbol{M} \coloneqq \boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top$ with $\boldsymbol{L}_M \coloneqq \boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}$, $\boldsymbol{R}_M \coloneqq$

$\boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2}$, and then invoke Lemma 23 with $\nu \coloneqq 3\sqrt{\mu r}/2$ to see

$$\begin{aligned}
|\mathfrak{R}_2| &\le \frac{3}{2}\sqrt{3\alpha\mu r}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)\left\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\right\|_{\mathsf{F}} \\
&\quad + 2\sqrt{\alpha n_2}\left\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\|\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star\|_{\mathsf{F}} \\
&\le \frac{3}{2}\sqrt{3\alpha\mu r}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\| \\
&\quad + 2\sqrt{\alpha n_2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\left\|\boldsymbol{\Sigma}_\star^{1/2}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\|\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star\|_{\mathsf{F}}.
\end{aligned}$$

Take the condition (A.23) and Lemmas 17 and 18 together to obtain

$$\begin{aligned}
\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\| &\le \frac{1}{1-\epsilon}; \\
\left\|\boldsymbol{\Sigma}_\star^{1/2}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\| = \left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|^2 &\le \frac{1}{(1-\epsilon)^2}; \\
\|\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star\|_{\mathsf{F}} &\le (1+\frac{\epsilon}{2})\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right).
\end{aligned} \tag{A.29}$$

These consequences combined with the condition (A.26) yield

$$\begin{aligned}
|\mathfrak{R}_2| &\le \frac{3\sqrt{3\alpha\mu r}}{2(1-\epsilon)}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} \\
&\quad + \frac{4\sqrt{\alpha\mu r}}{(1-\epsilon)^2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}(1+\frac{\epsilon}{2})\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right) \\
&\le \sqrt{\alpha\mu r}\frac{3\sqrt{3} + \frac{4(2+\epsilon)}{1-\epsilon}}{2(1-\epsilon)}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right) \\
&\le \sqrt{\alpha\mu r}\frac{3\sqrt{3} + \frac{4(2+\epsilon)}{1-\epsilon}}{2(1-\epsilon)}\left(\frac{3}{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \frac{1}{2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right),
\end{aligned}$$

where the last inequality holds since $2ab \le a^2 + b^2$.

3. The third term $\mathfrak{R}_3$ can be controlled similarly. Set $\boldsymbol{M} \coloneqq \boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top$ with $\boldsymbol{L}_M \coloneqq \boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}$ and $\boldsymbol{R}_M \coloneqq \boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}$, and invoke Lemma 23 with $\nu \coloneqq 3\sqrt{\mu r}/2$ to arrive at

$$|\mathfrak{R}_3| \le \frac{3}{2}\sqrt{3\alpha\mu r}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)\left\|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\right\|_{\mathsf{F}}$$

$$
\begin{aligned}
&+ 2\sqrt{\alpha n_1}\|\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}\|\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star\|_{\mathsf{F}}\\
&\le \frac{3}{2}\sqrt{3\alpha\mu r}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|^2\\
&\quad+ 2\sqrt{\alpha n_1}\|\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|^2\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star\|_{\mathsf{F}}.
\end{aligned}
$$

Use the consequences (A.29) again to obtain

$$
\begin{aligned}
|\mathfrak{R}_3| &\le \frac{3\sqrt{3\alpha\mu r}}{2(1-\epsilon)^2}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\\
&\quad+\frac{2\sqrt{\alpha\mu r}}{(1-\epsilon)^2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}(1+\frac{\epsilon}{2})\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)\\
&\le \sqrt{\alpha\mu r}\frac{3\sqrt{3}+2(2+\epsilon)}{2(1-\epsilon)^2}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right)\\
&\le \sqrt{\alpha\mu r}\frac{3\sqrt{3}+2(2+\epsilon)}{2(1-\epsilon)^2}\left(\frac{1}{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+\frac{3}{2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right).
\end{aligned}
$$

4. For the last term $\mathfrak{R}_4$, utilize the variational representation of the Frobenius norm to see

$$
\sqrt{\mathfrak{R}_4}=\mathrm{tr}\left((\boldsymbol{S}-\boldsymbol{S}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\widetilde{\boldsymbol{L}}^\top\right)
$$

for some $\widetilde{\boldsymbol{L}}\in\mathbb{R}^{n_1\times r}$ obeying $\|\widetilde{\boldsymbol{L}}\|_{\mathsf{F}}=1$. Setting $\boldsymbol{M}:=\widetilde{\boldsymbol{L}}\boldsymbol{\Sigma}_\star^{1/2}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top=\boldsymbol{L}_M\boldsymbol{R}_M^\top$ with $\boldsymbol{L}_M:=\widetilde{\boldsymbol{L}}\boldsymbol{\Sigma}_\star^{1/2}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}$ and $\boldsymbol{R}_M:=\boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2}$, we are ready to apply Lemma 23 again with $\nu:=3\sqrt{\mu r}/2$ to see

$$
\begin{aligned}
\sqrt{\mathfrak{R}_4} &\le \frac{3}{2}\sqrt{3\alpha\mu r}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)\left\|\widetilde{\boldsymbol{L}}\boldsymbol{\Sigma}_\star^{1/2}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\right\|_{\mathsf{F}}\\
&\quad+2\sqrt{\alpha n_2}\left\|\widetilde{\boldsymbol{L}}\boldsymbol{\Sigma}_\star^{1/2}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\|\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star\|_{\mathsf{F}}\\
&\le \frac{3}{2}\sqrt{3\alpha\mu r}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|\\
&\quad+2\sqrt{\alpha n_2}\left\|\boldsymbol{\Sigma}_\star^{1/2}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\|\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star\|_{\mathsf{F}}.
\end{aligned}
$$

This combined with the consequences (A.29) and condition (A.26) yields

$$
\sqrt{\mathfrak{R}_4}\le\sqrt{\alpha\mu r}\frac{3\sqrt{3}+\frac{4(2+\epsilon)}{1-\epsilon}}{2(1-\epsilon)}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right).
$$

Take the square, and use the elementary inequality $(a+b)^2 \le 2a^2 + 2b^2$ to reach

$$\mathfrak{R}_4 \le \alpha\mu r \frac{(3\sqrt{3} + \frac{4(2+\epsilon)}{1-\epsilon})^2}{2(1-\epsilon)^2} \left(\|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 \right).$$

Taking collectively the bounds for $\mathfrak{R}_1, \mathfrak{R}_2, \mathfrak{R}_3$ and $\mathfrak{R}_4$ yields the control of $\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2$ as

$$\begin{aligned}
\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 \le{}& \left((1-\eta)^2 + \frac{2\epsilon}{1-\epsilon}\eta(1-\eta)\right) \|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}\eta^2 \|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 \\
&+ \sqrt{\alpha\mu r}\frac{3\sqrt{3} + \frac{4(2+\epsilon)}{1-\epsilon}}{1-\epsilon}\eta(1-\eta)\left(\frac{3}{2}\|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \frac{1}{2}\|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right) \\
&+ \sqrt{\alpha\mu r}\frac{3\sqrt{3} + 2(2+\epsilon)}{(1-\epsilon)^2}\eta^2\left(\frac{1}{2}\|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \frac{3}{2}\|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right) \\
&+ \alpha\mu r \frac{(3\sqrt{3} + \frac{4(2+\epsilon)}{1-\epsilon})^2}{2(1-\epsilon)^2}\eta^2\left(\|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right).
\end{aligned}$$

Similarly, we can obtain the control of $\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2$. Combine them together and identify $\mathrm{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star) = \|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2$ to reach

$$\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 \le \rho^2(\eta; \epsilon, \alpha\mu r)\,\mathrm{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star),$$

where the contraction rate $\rho^2(\eta; \epsilon, \alpha\mu r)$ is given by

$$\begin{aligned}
\rho^2(\eta; \epsilon, \alpha\mu r) \coloneqq{}& (1-\eta)^2 + \frac{2\epsilon + \sqrt{\alpha\mu r}(6\sqrt{3} + \frac{8(2+\epsilon)}{1-\epsilon})}{1-\epsilon}\eta(1-\eta) \\
&+ \frac{2\epsilon + \epsilon^2 + \sqrt{\alpha\mu r}(6\sqrt{3} + 4(2+\epsilon)) + \alpha\mu r(3\sqrt{3} + \frac{4(2+\epsilon)}{1-\epsilon})^2}{(1-\epsilon)^2}\eta^2.
\end{aligned}$$

With $\epsilon = 0.02$, $\alpha\mu r \le 10^{-4}$, and $0 < \eta \le 2/3$, one has $\rho(\eta; \epsilon, \alpha\mu r) \le 1 - 0.6\eta$. Thus we conclude that

$$\begin{aligned}
\mathrm{dist}(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) &\le \sqrt{\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2} \\
&\le (1 - 0.6\eta)\,\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star). \qquad \text{(A.30)}
\end{aligned}$$

Incoherence condition

We start by controlling the term $\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty}$. We know from (A.28) that

$$(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2} = (1-\eta)\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2} - \eta\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2} - \eta(\boldsymbol{S}-\boldsymbol{S}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}.$$

Apply the triangle inequality to obtain

$$\begin{aligned}\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty} \le (1-\eta)\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} + \eta\underbrace{\left\|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty}}_{\mathfrak{T}_1} \\ + \eta\underbrace{\left\|(\boldsymbol{S}-\boldsymbol{S}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty}}_{\mathfrak{T}_2}.\end{aligned}$$

The first term $\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty}$ follows from the incoherence condition (A.24) as

$$\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \le \sqrt{\frac{\mu r}{n_1}}\sigma_r(\boldsymbol{X}_\star).$$

In the sequel, we shall bound the terms $\mathfrak{T}_1$ and $\mathfrak{T}_2$.

1. For the term $\mathfrak{T}_1$, use the relation $\|\boldsymbol{AB}\|_{2,\infty} \le \|\boldsymbol{A}\|_{2,\infty}\|\boldsymbol{B}\|$, and combine the condition (A.23) with the consequences (A.29) to obtain

$$\begin{aligned}\mathfrak{T}_1 &\le \|\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\left\|\boldsymbol{\Sigma}_\star^{1/2}\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\| \\ &\le \|\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\| \\ &\le \frac{\epsilon}{1-\epsilon}\sqrt{\frac{\mu r}{n_1}}\sigma_r(\boldsymbol{X}_\star),\end{aligned}$$

2. For the term $\mathfrak{T}_2$, use the relation $\|\boldsymbol{AB}\|_{2,\infty} \le \|\boldsymbol{A}\|_{2,\infty}\|\boldsymbol{B}\|$ to obtain

$$\mathfrak{T}_2 \le \|\boldsymbol{S}-\boldsymbol{S}_\star\|_{2,\infty}\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|.$$

We know from Lemma 23 that $\boldsymbol{S}-\boldsymbol{S}_\star$ has at most $3\alpha n_2$ non-zero entries in each row, and

$\|\boldsymbol{S}-\boldsymbol{S}_\star\|_\infty \le 2\|\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star\|_\infty$. Upper bound the $\ell_{2,\infty}$ norm by the ℓ_∞ norm as

$$\|\boldsymbol{S}-\boldsymbol{S}_\star\|_{2,\infty} \le \sqrt{3\alpha n_2}\|\boldsymbol{S}-\boldsymbol{S}_\star\|_\infty \le 2\sqrt{3\alpha n_2}\|\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star\|_\infty.$$

Split $\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star = \boldsymbol{\Delta}_L\boldsymbol{R}^\top + \boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top$, and take the conditions (A.24) and (A.26) to obtain

$$\begin{aligned}
\|\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star\|_\infty &\le \|\boldsymbol{\Delta}_L\boldsymbol{R}^\top\|_\infty + \|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\|_\infty \\
&\le \|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty}\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} + \|\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \\
&\le \sqrt{\frac{\mu r}{n_1}}\sigma_r(\boldsymbol{X}_\star)2\sqrt{\frac{\mu r}{n_2}} + \sqrt{\frac{\mu r}{n_1}}\sqrt{\frac{\mu r}{n_2}}\sigma_r(\boldsymbol{X}_\star) \\
&= \frac{3\mu r}{\sqrt{n_1 n_2}}\sigma_r(\boldsymbol{X}_\star).
\end{aligned}$$

This combined with the consequences (A.29) yields

$$\mathfrak{T}_2 \le \frac{6\sqrt{3\alpha\mu r}}{1-\epsilon}\sqrt{\frac{\mu r}{n_1}}\sigma_r(\boldsymbol{X}_\star).$$

Taking collectively the bounds for $\mathfrak{T}_1, \mathfrak{T}_2$ yields the control

$$\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty} \le \left(1-\eta+\frac{\epsilon+6\sqrt{3\alpha\mu r}}{1-\epsilon}\eta\right)\sqrt{\frac{\mu r}{n_1}}\sigma_r(\boldsymbol{X}_\star). \tag{A.31}$$

The last step is to switch the alignment matrix from $\boldsymbol{Q}_t$ to $\boldsymbol{Q}_{t+1}$. (A.30) together with Lemma 14 demonstrates the existence of $\boldsymbol{Q}_{t+1}$. Apply the triangle inequality to obtain

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

We deduct from (A.31) that

$$\|\boldsymbol{L}_{t+1}\boldsymbol{Q}_t\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} \le \|\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} + \left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{-1/2}\right\|_{2,\infty} \le \left(2-\eta+\frac{\epsilon+6\sqrt{3\alpha\mu r}}{1-\epsilon}\eta\right)\sqrt{\frac{\mu r}{n_1}}.$$

Regarding the alignment matrix term, invoke Lemma 19 to obtain

$$\begin{aligned}\left\|\boldsymbol{\Sigma}_\star^{1/2}\boldsymbol{Q}_t^{-1}\boldsymbol{Q}_{t+1}\boldsymbol{\Sigma}_\star^{1/2}-\boldsymbol{\Sigma}_\star\right\| &\le \frac{\|(\boldsymbol{R}_{t+1}(\boldsymbol{Q}_t^{-\top}-\boldsymbol{Q}_{t+1}^{-\top})\boldsymbol{\Sigma}_\star^{1/2}\|}{1-\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_{t+1}^{-\top}-\boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{-1/2}\|} \\ &\le \frac{\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top}-\boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\|+\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_{t+1}^{-\top}-\boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\|}{1-\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_{t+1}^{-\top}-\boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{-1/2}\|} \\ &\le \frac{2\epsilon}{1-\epsilon}\sigma_r(\boldsymbol{X}_\star),\end{aligned}$$

where we deduct from (A.30) that the distances using either $\boldsymbol{Q}_t$ or $\boldsymbol{Q}_{t+1}$ are bounded by

$$\begin{aligned}\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top}-\boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\| &\le \epsilon\sigma_r(\boldsymbol{X}_\star); \\ \|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_{t+1}^{-\top}-\boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\| &\le \epsilon\sigma_r(\boldsymbol{X}_\star); \\ \|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_{t+1}^{-\top}-\boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{-1/2}\| &\le \epsilon.\end{aligned}$$

Combine all pieces to reach

$$\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_{t+1}-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty} \le \left(\frac{1+\epsilon}{1-\epsilon}\left(1-\eta+\frac{\epsilon+6\sqrt{3\alpha\mu r}}{1-\epsilon}\eta\right)+\frac{2\epsilon}{1-\epsilon}\right)\sqrt{\frac{\mu r}{n_1}}\sigma_r(\boldsymbol{X}_\star).$$

With $\epsilon = 0.02$, $\alpha\mu r \le 10^{-4}$, and $0.1 \le \eta \le 2/3$, we get the desired incoherence condition

$$\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_{t+1}-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty} \le \sqrt{\frac{\mu r}{n_1}}\sigma_r(\boldsymbol{X}_\star).$$

Similarly, we can prove the other part

$$\left\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_{t+1}^{-\top}-\boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty} \le \sqrt{\frac{\mu r}{n_2}}\sigma_r(\boldsymbol{X}_\star).$$

A.4.2 Proof of Lemma 4

We first record two lemmas from [YPCC16], which are useful for studying the properties of the initialization.

Lemma 24 ([YPCC16, Section 6.1]). *Given $\boldsymbol{S}_\star \in \mathcal{S}_\alpha$, one has $\|\boldsymbol{S}_\star - \mathcal{T}_\alpha[\boldsymbol{X}_\star+\boldsymbol{S}_\star]\|_\infty \le 2\|\boldsymbol{X}_\star\|_\infty$.*

Lemma 25 ([YPCC16, Lemma 1]). *For any matrix $\boldsymbol{M} \in \mathcal{S}_\alpha$, one has $\|\boldsymbol{M}\| \le \alpha\sqrt{n_1 n_2}\|\boldsymbol{M}\|_\infty$.*

With these two lemmas in place, we are ready to establish the claimed result. Invoke Lemma 16 to obtain

$$\mathrm{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \le \sqrt{\sqrt{2}+1}\left\|\boldsymbol{L}_0\boldsymbol{R}_0^\top - \boldsymbol{X}_\star\right\|_{\mathsf{F}} \le \sqrt{(\sqrt{2}+1)2r}\left\|\boldsymbol{L}_0\boldsymbol{R}_0^\top - \boldsymbol{X}_\star\right\|,$$

where the last relation uses the fact that $\boldsymbol{L}_0\boldsymbol{R}_0^\top - \boldsymbol{X}_\star$ has rank at most $2r$. We can further apply the triangle inequality to see

$$\begin{aligned}\left\|\boldsymbol{L}_0\boldsymbol{R}_0^\top - \boldsymbol{X}_\star\right\| &\le \left\|\boldsymbol{Y} - \mathcal{T}_\alpha[\boldsymbol{Y}] - \boldsymbol{L}_0\boldsymbol{R}_0^\top\right\| + \|\boldsymbol{Y} - \mathcal{T}_\alpha[\boldsymbol{Y}] - \boldsymbol{X}_\star\| \\ &\le 2\|\boldsymbol{Y} - \mathcal{T}_\alpha[\boldsymbol{Y}] - \boldsymbol{X}_\star\| = 2\|\boldsymbol{S}_\star - \mathcal{T}_\alpha[\boldsymbol{X}_\star + \boldsymbol{S}_\star]\|.\end{aligned}$$

Here the second inequality hinges on the fact that $\boldsymbol{L}_0\boldsymbol{R}_0^\top$ is the best rank-r approximation of $\boldsymbol{Y} - \mathcal{T}_\alpha[\boldsymbol{Y}]$, and the last identity arises from $\boldsymbol{Y} = \boldsymbol{X}_\star + \boldsymbol{S}_\star$. Follow the same argument as [YPCC16, Section 6.1], combining Lemmas 24 and 25 to reach

$$\begin{aligned}\|\boldsymbol{S}_\star - \mathcal{T}_\alpha[\boldsymbol{X}_\star + \boldsymbol{S}_\star]\| &\le 2\alpha\sqrt{n_1 n_2}\,\|\boldsymbol{S}_\star - \mathcal{T}_\alpha[\boldsymbol{X}_\star + \boldsymbol{S}_\star]\|_\infty \\ &\le 4\alpha\sqrt{n_1 n_2}\|\boldsymbol{X}_\star\|_\infty \le 4\alpha\mu r\kappa\sigma_r(\boldsymbol{X}_\star),\end{aligned}$$

where the last inequality follows from the incoherence assumption

$$\|\boldsymbol{X}_\star\|_\infty \le \|\boldsymbol{U}_\star\|_{2,\infty}\|\boldsymbol{\Sigma}_\star\|\,\|\boldsymbol{V}_\star\|_{2,\infty} \le \frac{\mu r}{\sqrt{n_1 n_2}}\kappa\sigma_r(\boldsymbol{X}_\star). \tag{A.32}$$

Take the above inequalities together to arrive at

$$\mathrm{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \le 8\sqrt{2(\sqrt{2}+1)}\alpha\mu r^{3/2}\kappa\sigma_r(\boldsymbol{X}_\star) \le 20\alpha\mu r^{3/2}\kappa\sigma_r(\boldsymbol{X}_\star).$$

A.4.3 Proof of Lemma 5

In view of the condition $\mathrm{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \le 0.02\sigma_r(\boldsymbol{X}_\star)$ and Lemma 14, one knows that $\boldsymbol{Q}_0$, the optimal alignment matrix between $\boldsymbol{F}_0$ and $\boldsymbol{F}_\star$ exists. Therefore, for notational convenience, denote $\boldsymbol{L} \coloneqq \boldsymbol{L}_0\boldsymbol{Q}_0$, $\boldsymbol{R} \coloneqq \boldsymbol{R}_0\boldsymbol{Q}_0^{-\top}$, $\boldsymbol{\Delta}_L \coloneqq \boldsymbol{L} - \boldsymbol{L}_\star$, $\boldsymbol{\Delta}_R \coloneqq \boldsymbol{R} - \boldsymbol{R}_\star$, and $\epsilon \coloneqq 0.02$. Our objective is then translated to demonstrate

$$\sqrt{n_1}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \vee \sqrt{n_2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \le \sqrt{\mu r}\sigma_r(\boldsymbol{X}_\star).$$

From now on, we focus on bounding $\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty}$. Since $\boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^\top$ is the top-r SVD of $\boldsymbol{Y} - \mathcal{T}_\alpha[\boldsymbol{Y}]$, and recall that $\boldsymbol{Y} = \boldsymbol{X}_\star + \boldsymbol{S}_\star$, we have the relation

$$(\boldsymbol{X}_\star + \boldsymbol{S}_\star - \mathcal{T}_\alpha[\boldsymbol{X}_\star + \boldsymbol{S}_\star])\boldsymbol{V}_0 = \boldsymbol{U}_0\boldsymbol{\Sigma}_0,$$

which further implies the following decomposition of $\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}$.

Claim 4. *One has*

$$\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2} = (\boldsymbol{S}_\star - \mathcal{T}_\alpha[\boldsymbol{X}_\star + \boldsymbol{S}_\star])\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2} - \boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}.$$

Combining Claim 4 with the triangle inequality yields

$$\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \le \underbrace{\left\|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty}}_{\mathfrak{I}_1} + \underbrace{\left\|(\boldsymbol{S}_\star - \mathcal{T}_\alpha[\boldsymbol{X}_\star + \boldsymbol{S}_\star])\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty}}_{\mathfrak{I}_2}.$$

In what follows, we shall control $\mathfrak{I}_1$ and $\mathfrak{I}_2$ in turn.

1. For the term $\mathfrak{I}_1$, use the relation $\|\boldsymbol{A}\boldsymbol{B}\|_{2,\infty} \le \|\boldsymbol{A}\|_{2,\infty}\|\boldsymbol{B}\|$ to obtain

$$\mathfrak{I}_1 \le \|\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\| \left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|.$$

The incoherence assumption tells $\|\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} = \|\boldsymbol{U}_\star\|_{2,\infty} \le \sqrt{\mu r/n_1}$. In addition, the assumption $\mathrm{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \le \epsilon\sigma_r(\boldsymbol{X}_\star)$ entails the bound $\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\| \le \epsilon\sigma_r(\boldsymbol{X}_\star)$. Finally, repeating the

argument for obtaining (A.23) yields $\|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{-1/2}\| \le \epsilon$, which together with Lemma 17 reveals

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

In all, we arrive at

$$\mathfrak{I}_1 \le \frac{\epsilon}{1-\epsilon}\sqrt{\frac{\mu r}{n_1}}\sigma_r(\boldsymbol{X}_\star).$$

2. Proceeding to the term $\mathfrak{I}_2$, use the relations $\|\boldsymbol{AB}\|_{2,\infty} \le \|\boldsymbol{A}\|_{1,\infty}\|\boldsymbol{B}\|_{2,\infty}$ and $\|\boldsymbol{AB}\|_{2,\infty} \le \|\boldsymbol{A}\|_{2,\infty}\|\boldsymbol{B}\|$ to obtain

$$\begin{aligned}
\mathfrak{I}_2 &\le \|\boldsymbol{S}_\star - \mathcal{T}_\alpha[\boldsymbol{X}_\star + \boldsymbol{S}_\star]\|_{1,\infty} \left\|\boldsymbol{R}(\boldsymbol{R}^\top \boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty} \\
&\le \|\boldsymbol{S}_\star - \mathcal{T}_\alpha[\boldsymbol{X}_\star + \boldsymbol{S}_\star]\|_{1,\infty} \|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} \left\|\boldsymbol{\Sigma}_\star^{1/2}(\boldsymbol{R}^\top \boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|.
\end{aligned}$$

Regarding $\boldsymbol{S}_\star - \mathcal{T}_\alpha[\boldsymbol{X}_\star + \boldsymbol{S}_\star]$, Lemma 24 tells that $\boldsymbol{S}_\star - \mathcal{T}_\alpha[\boldsymbol{X}_\star + \boldsymbol{S}_\star]$ has at most $2\alpha n_2$ non-zero entries in each row, and $\|\boldsymbol{S}_\star - \mathcal{T}_\alpha[\boldsymbol{X}_\star + \boldsymbol{S}_\star]\|_\infty \le 2\|\boldsymbol{X}_\star\|_\infty$. Consequently, we can upper bound the $\ell_{1,\infty}$ norm by the ℓ_∞ norm as

$$\begin{aligned}
\|\boldsymbol{S}_\star - \mathcal{T}_\alpha[\boldsymbol{X}_\star + \boldsymbol{S}_\star]\|_{1,\infty} &\le 2\alpha n_2 \|\boldsymbol{S}_\star - \mathcal{T}_\alpha[\boldsymbol{X}_\star + \boldsymbol{S}_\star]\|_\infty \\
&\le 4\alpha n_2 \|\boldsymbol{X}_\star\|_\infty \\
&\le 4\alpha n_2 \frac{\mu r}{\sqrt{n_1 n_2}}\kappa\sigma_r(\boldsymbol{X}_\star).
\end{aligned}$$

Here the last inequality follows from the incoherence assumption (A.32). For the term $\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}$, one can apply the triangle inequality to see

$$\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} \le \|\boldsymbol{R}_\star\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} \le \sqrt{\frac{\mu r}{n_2}} + \frac{\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty}}{\sigma_r(\boldsymbol{X}_\star)}.$$

Last but not least, repeat the argument for (A.29) to obtain

$$\left\|\boldsymbol{\Sigma}_\star^{1/2}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\| = \left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|^2 \le \frac{1}{(1-\epsilon)^2}.$$

Taking together the above bounds yields

$$\mathfrak{I}_2 \le \frac{4\alpha\mu r\kappa}{(1-\epsilon)^2}\sqrt{\frac{\mu r}{n_1}}\sigma_r(\boldsymbol{X}_\star) + \frac{4\alpha\mu r\kappa}{(1-\epsilon)^2}\sqrt{\frac{n_2}{n_1}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty}.$$

Combine the bounds on $\mathfrak{I}_1$ and $\mathfrak{I}_2$ to reach

$$\sqrt{n_1}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \le \left(\frac{\epsilon}{1-\epsilon} + \frac{4\alpha\mu r\kappa}{(1-\epsilon)^2}\right)\sqrt{\mu r}\sigma_r(\boldsymbol{X}_\star) + \frac{4\alpha\mu r\kappa}{(1-\epsilon)^2}\sqrt{n_2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty}.$$

Similarly, we have

$$\sqrt{n_2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \le \left(\frac{\epsilon}{1-\epsilon} + \frac{4\alpha\mu r\kappa}{(1-\epsilon)^2}\right)\sqrt{\mu r}\sigma_r(\boldsymbol{X}_\star) + \frac{4\alpha\mu r\kappa}{(1-\epsilon)^2}\sqrt{n_1}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty}.$$

Taking the maximum and solving for $\sqrt{n_1}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \vee \sqrt{n_2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty}$ yield the relation

$$\sqrt{n_1}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \vee \sqrt{n_2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \le \frac{\epsilon(1-\epsilon)+4\alpha\mu r\kappa}{(1-\epsilon)^2-4\alpha\mu r\kappa}\sqrt{\mu r}\sigma_r(\boldsymbol{X}_\star).$$

With $\epsilon = 0.02$ and $\alpha\mu r\kappa \le 0.1$, we get the desired conclusion

$$\sqrt{n_1}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \vee \sqrt{n_2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \le \sqrt{\mu r}\sigma_r(\boldsymbol{X}_\star).$$

Proof of Claim 4. Identify $\boldsymbol{U}_0$ (resp. $\boldsymbol{V}_0$) with $\boldsymbol{L}_0\boldsymbol{\Sigma}_0^{-1/2}$ (resp. $\boldsymbol{R}_0\boldsymbol{\Sigma}_0^{-1/2}$) to yield

$$(\boldsymbol{X}_\star + \boldsymbol{S}_\star - \mathcal{T}_\alpha[\boldsymbol{X}_\star + \boldsymbol{S}_\star])\boldsymbol{R}_0\boldsymbol{\Sigma}_0^{-1} = \boldsymbol{L}_0,$$

which is equivalent to $(\boldsymbol{X}_\star + \boldsymbol{S}_\star - \mathcal{T}_\alpha[\boldsymbol{X}_\star + \boldsymbol{S}_\star])\boldsymbol{R}_0(\boldsymbol{R}_0^\top\boldsymbol{R}_0)^{-1} = \boldsymbol{L}_0$ since $\boldsymbol{\Sigma}_0 = \boldsymbol{R}_0^\top\boldsymbol{R}_0$. Multiply

both sides by $\boldsymbol{Q}_0\boldsymbol{\Sigma}_\star^{1/2}$ to obtain

$$(\boldsymbol{X}_\star + \boldsymbol{S}_\star - \mathcal{T}_\alpha[\boldsymbol{X}_\star + \boldsymbol{S}_\star])\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2} = \boldsymbol{L}\boldsymbol{\Sigma}_\star^{1/2},$$

where we recall that $\boldsymbol{L} = \boldsymbol{L}_0\boldsymbol{Q}_0$ and $\boldsymbol{R} = \boldsymbol{R}_0\boldsymbol{Q}_0^{-\top}$. In the end, subtract $\boldsymbol{X}_\star\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}$ from both sides to reach

$$\begin{aligned}(\boldsymbol{S}_\star - \mathcal{T}_\alpha[\boldsymbol{X}_\star + \boldsymbol{S}_\star])\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2} &= \boldsymbol{L}\boldsymbol{\Sigma}_\star^{1/2} - \boldsymbol{L}_\star\boldsymbol{R}_\star^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2} \\ &= (\boldsymbol{L} - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2} + \boldsymbol{L}_\star(\boldsymbol{R} - \boldsymbol{R}_\star)^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2} \\ &= \boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2} + \boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}.\end{aligned}$$

This finishes the proof. □

A.5 Proof for Matrix Completion

A.5.1 New projection operator

Proof of Proposition 1

First, notice that the optimization of $\boldsymbol{L}$ and $\boldsymbol{R}$ in (2.21) can be decomposed and done in parallel, hence we focus on the optimization of $\boldsymbol{L}$ below:

$$\boldsymbol{L} = \underset{\boldsymbol{L}\in\mathbb{R}^{n_1\times r}}{\operatorname{argmin}} \left\|(\boldsymbol{L} - \widetilde{\boldsymbol{L}})(\widetilde{\boldsymbol{R}}^\top\widetilde{\boldsymbol{R}})^{1/2}\right\|_{\mathsf{F}}^2 \qquad \text{s.t.} \quad \sqrt{n_1}\left\|\boldsymbol{L}(\widetilde{\boldsymbol{R}}^\top\widetilde{\boldsymbol{R}})^{1/2}\right\|_{2,\infty} \le B.$$

By a change of variables as $\boldsymbol{G} \coloneqq \boldsymbol{L}(\widetilde{\boldsymbol{R}}^\top\widetilde{\boldsymbol{R}})^{1/2}$ and $\widetilde{\boldsymbol{G}} \coloneqq \widetilde{\boldsymbol{L}}(\widetilde{\boldsymbol{R}}^\top\widetilde{\boldsymbol{R}})^{1/2}$, we rewrite the above problem equivalently as

$$\boldsymbol{G} = \underset{\boldsymbol{G}\in\mathbb{R}^{n_1\times r}}{\operatorname{argmin}} \|\boldsymbol{G} - \widetilde{\boldsymbol{G}}\|_{\mathsf{F}}^2 \qquad \text{s.t.} \quad \sqrt{n_1}\|\boldsymbol{G}\|_{2,\infty} \le B,$$

whose solution is given as [CW15]

$$\boldsymbol{G}_{i,\cdot} = \left(1 \wedge \frac{B}{\sqrt{n_1}\|\widetilde{\boldsymbol{G}}_{i,\cdot}\|_2}\right)\widetilde{\boldsymbol{G}}_{i,\cdot}, \quad 1 \le i \le n_1.$$

By applying again the change of variable $\boldsymbol{L} = \boldsymbol{G}(\widetilde{\boldsymbol{R}}^\top \widetilde{\boldsymbol{R}})^{-1/2}$ and $\widetilde{\boldsymbol{L}} = \widetilde{\boldsymbol{G}}(\widetilde{\boldsymbol{R}}^\top \widetilde{\boldsymbol{R}})^{-1/2}$, we obtain the claimed solution.

Proof of Lemma 6

We begin with proving the non-expansiveness property. Denote the optimal alignment matrix between $\widetilde{\boldsymbol{F}}$ and $\boldsymbol{F}_\star$ as $\widetilde{\boldsymbol{Q}}$, whose existence is guaranteed by Lemma 14. Denoting $\mathcal{P}_B(\widetilde{\boldsymbol{F}}) = [\boldsymbol{L}^\top, \boldsymbol{R}^\top]^\top$, by the definition of $\mathrm{dist}(\mathcal{P}_B(\widetilde{\boldsymbol{F}}), \boldsymbol{F}_\star)$, we know that

$$\mathrm{dist}^2(\mathcal{P}_B(\widetilde{\boldsymbol{F}}), \boldsymbol{F}_\star) \le \sum_{i=1}^{n_1} \left\|\boldsymbol{L}_{i,\cdot}\widetilde{\boldsymbol{Q}}\boldsymbol{\Sigma}_\star^{1/2} - (\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{1/2})_{i,\cdot}\right\|_2^2 + \sum_{j=1}^{n_2} \left\|\boldsymbol{R}_{j,\cdot}\widetilde{\boldsymbol{Q}}^{-\top}\boldsymbol{\Sigma}_\star^{1/2} - (\boldsymbol{R}_\star\boldsymbol{\Sigma}_\star^{1/2})_{j,\cdot}\right\|_2^2. \quad \text{(A.33)}$$

Recall that the condition $\mathrm{dist}(\widetilde{\boldsymbol{F}}, \boldsymbol{F}_\star) \le \epsilon\sigma_r(\boldsymbol{X}_\star)$ implies

$$\left\|(\widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{Q}} - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{-1/2}\right\| \vee \left\|(\widetilde{\boldsymbol{R}}\widetilde{\boldsymbol{Q}}^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{-1/2}\right\| \le \epsilon,$$

which, together with $\boldsymbol{R}_\star\boldsymbol{\Sigma}_\star^{-1/2} = \boldsymbol{V}_\star$, further implies that

$$\begin{aligned}\left\|\widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{R}}^\top\right\|_2 &\le \left\|\widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{Q}}\boldsymbol{\Sigma}_\star^{1/2}\right\|_2 \left\|\widetilde{\boldsymbol{R}}\widetilde{\boldsymbol{Q}}^{-\top}\boldsymbol{\Sigma}_\star^{-1/2}\right\| \\ &\le \left\|\widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{Q}}\boldsymbol{\Sigma}_\star^{1/2}\right\|_2 \left(\|\boldsymbol{V}_\star\| + \left\|(\widetilde{\boldsymbol{R}}\widetilde{\boldsymbol{Q}}^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{-1/2}\right\|\right) \le (1+\epsilon)\left\|\widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{Q}}\boldsymbol{\Sigma}_\star^{1/2}\right\|_2.\end{aligned}$$

In addition, the μ-incoherence of $\boldsymbol{X}_\star$ yields

$$\sqrt{n_1}\left\|(\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{1/2})_{i,\cdot}\right\|_2 \le \sqrt{n_1}\|\boldsymbol{U}_\star\|_{2,\infty}\|\boldsymbol{\Sigma}_\star\| \le \sqrt{\mu r}\sigma_1(\boldsymbol{X}_\star) \le \frac{B}{1+\epsilon},$$

where the last inequality follows from the choice of B. Take the above two relations collectively to reach

$$\frac{B}{\sqrt{n_1}\|\widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{R}}^\top\|_2} \geq \frac{\left\|(\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{1/2})_{i,\cdot}\right\|_2}{\left\|\widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{Q}}\boldsymbol{\Sigma}_\star^{1/2}\right\|_2}.$$

We claim that performing the following projection yields a contraction on each row; see also [ZL16, Lemma 11].

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*

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

Apply Claim 5 with $\boldsymbol{u} \coloneqq \widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{Q}}\boldsymbol{\Sigma}_\star^{1/2}$, $\boldsymbol{u}_\star \coloneqq (\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{1/2})_{i,\cdot}$, and $\lambda \coloneqq B/(\sqrt{n_1}\|\widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{R}}^\top\|_2)$ to obtain

$$\begin{aligned}\left\|\boldsymbol{L}_{i,\cdot}\widetilde{\boldsymbol{Q}}\boldsymbol{\Sigma}_\star^{1/2} - (\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{1/2})_{i,\cdot}\right\|_2^2 &= \left\|\left(1\wedge\frac{B}{\sqrt{n_1}\|\widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{R}}^\top\|_2}\right)\widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{Q}}\boldsymbol{\Sigma}_\star^{1/2} - (\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{1/2})_{i,\cdot}\right\|_2^2 \\ &\leq \left\|\widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{Q}}\boldsymbol{\Sigma}_\star^{1/2} - (\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{1/2})_{i,\cdot}\right\|_2^2.\end{aligned}$$

Following a similar argument for $\boldsymbol{R}$, and plugging them back to (A.33), we conclude that

$$\mathrm{dist}^2(\mathcal{P}_B(\widetilde{\boldsymbol{F}}), \boldsymbol{F}_\star) \leq \sum_{i=1}^{n_1}\left\|\widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{Q}}\boldsymbol{\Sigma}_\star^{1/2} - (\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{1/2})_{i,\cdot}\right\|_2^2 + \sum_{j=1}^{n_2}\left\|\widetilde{\boldsymbol{R}}_{j,\cdot}\widetilde{\boldsymbol{Q}}^{-\top}\boldsymbol{\Sigma}_\star^{1/2} - (\boldsymbol{R}_\star\boldsymbol{\Sigma}_\star^{1/2})_{j,\cdot}\right\|_2^2 = \mathrm{dist}^2(\widetilde{\boldsymbol{F}}, \boldsymbol{F}_\star).$$

We move on to the incoherence condition. For any $1 \leq i \leq n_1$, one has

$$\begin{aligned}\|\boldsymbol{L}_{i,\cdot}\boldsymbol{R}^\top\|_2^2 = \sum_{j=1}^{n_2}\langle\boldsymbol{L}_{i,\cdot}, \boldsymbol{R}_{j,\cdot}\rangle^2 &= \sum_{j=1}^{n_2}\left(1\wedge\frac{B}{\sqrt{n_1}\|\widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{R}}^\top\|_2}\right)^2\langle\widetilde{\boldsymbol{L}}_{i,\cdot}, \widetilde{\boldsymbol{R}}_{j,\cdot}\rangle^2\left(1\wedge\frac{B}{\sqrt{n_2}\|\widetilde{\boldsymbol{R}}_{j,\cdot}\widetilde{\boldsymbol{L}}^\top\|_2}\right)^2 \\ &\overset{\text{(i)}}{\leq}\left(1\wedge\frac{B}{\sqrt{n_1}\|\widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{R}}^\top\|_2}\right)^2\sum_{j=1}^{n_2}\langle\widetilde{\boldsymbol{L}}_{i,\cdot}, \widetilde{\boldsymbol{R}}_{j,\cdot}\rangle^2 = \left(1\wedge\frac{B}{\sqrt{n_1}\|\widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{R}}^\top\|_2}\right)^2\|\widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{R}}^\top\|_2^2 \\ &\overset{\text{(ii)}}{\leq}\frac{B^2}{n_1}.\end{aligned}$$

where (i) follows from $1 \wedge \frac{B}{\sqrt{n_2}\|\widetilde{\boldsymbol{R}}_{j,\cdot}\widetilde{\boldsymbol{L}}^\top\|_2} \le 1$, and (ii) follows from $1 \wedge \frac{B}{\sqrt{n_1}\|\widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{R}}^\top\|_2} \le \frac{B}{\sqrt{n_1}\|\widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{R}}^\top\|_2}$. Similarly, one has $\|\boldsymbol{R}_{j,\cdot}\boldsymbol{L}^\top\|_2^2 \le B^2/n_2$. Combining these two bounds completes the proof.

Proof of Claim 5. When $\lambda > 1$, the claim holds as an identity. Otherwise $\lambda \le 1$. Denote $h(\bar{\lambda}) \coloneqq \|\bar{\lambda}\boldsymbol{u} - \boldsymbol{u}_\star\|_2^2$. Calculate its derivative to conclude that $h(\bar{\lambda})$ is monotonically increasing when $\bar{\lambda} \ge \lambda_\star \coloneqq \langle \boldsymbol{u}, \boldsymbol{u}_\star\rangle/\|\boldsymbol{u}\|_2^2$. Note that $\lambda \ge \|\boldsymbol{u}_\star\|_2/\|\boldsymbol{u}\|_2 \ge \lambda_\star$, thus $h(\lambda) \le h(1)$, i.e. the claim holds. □

A.5.2 Proof of Lemma 7

We first record two useful lemmas regarding the projector $\mathcal{P}_\Omega(\cdot)$.

Lemma 26 ([ZL16, Lemma 10]). *Suppose that $\boldsymbol{X}_\star$ is μ-incoherent, and $p \gtrsim \mu r \log(n_1 \vee n_2)/(n_1 \wedge n_2)$. With overwhelming probability, one has*

$$\begin{aligned}&\left|\left\langle (p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{L}_\star\boldsymbol{R}_A^\top + \boldsymbol{L}_A\boldsymbol{R}_\star^\top), \boldsymbol{L}_\star\boldsymbol{R}_B^\top + \boldsymbol{L}_B\boldsymbol{R}_\star^\top\right\rangle\right| \\ &\qquad \le C_1\sqrt{\frac{\mu r\log(n_1\vee n_2)}{p(n_1\wedge n_2)}}\|\boldsymbol{L}_\star\boldsymbol{R}_A^\top + \boldsymbol{L}_A\boldsymbol{R}_\star^\top\|_{\mathsf{F}}\|\boldsymbol{L}_\star\boldsymbol{R}_B^\top + \boldsymbol{L}_B\boldsymbol{R}_\star^\top\|_{\mathsf{F}},\end{aligned}$$

simultaneously for all $\boldsymbol{L}_A, \boldsymbol{L}_B \in \mathbb{R}^{n_1\times r}$ and $\boldsymbol{R}_A, \boldsymbol{R}_B \in \mathbb{R}^{n_2\times r}$, where $C_1 > 0$ is some universal constant.

Lemma 27 ([CL19, Lemma 8], [CLL20, Lemma 12]). *Suppose that $p \gtrsim \log(n_1 \vee n_2)/(n_1 \wedge n_2)$. With overwhelming probability, one has*

$$\begin{aligned}&\left|\left\langle (p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{L}_A\boldsymbol{R}_A^\top), \boldsymbol{L}_B\boldsymbol{R}_B^\top\right\rangle\right| \\ &\qquad \le C_2\sqrt{\frac{n_1\vee n_2}{p}}\left(\|\boldsymbol{L}_A\|_{\mathsf{F}}\|\boldsymbol{L}_B\|_{2,\infty} \wedge \|\boldsymbol{L}_A\|_{2,\infty}\|\boldsymbol{L}_B\|_{\mathsf{F}}\right)\left(\|\boldsymbol{R}_A\|_{\mathsf{F}}\|\boldsymbol{R}_B\|_{2,\infty} \wedge \|\boldsymbol{R}_A\|_{2,\infty}\|\boldsymbol{R}_B\|_{\mathsf{F}}\right),\end{aligned}$$

simultaneously for all $\boldsymbol{L}_A, \boldsymbol{L}_B \in \mathbb{R}^{n_1\times r}$ and $\boldsymbol{R}_A, \boldsymbol{R}_B \in \mathbb{R}^{n_2\times r}$, where $C_2 > 0$ is some universal constant.

In view of the above two lemmas, define the event $\mathcal{E}$ as the intersection of the events that the bounds in Lemmas 26 and 27 hold, which happens with overwhelming probability. The rest of the proof is then performed under the event that $\mathcal{E}$ holds.

By the condition $\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \le 0.02\sigma_r(\boldsymbol{X}_\star)$ and Lemma 14, one knows that $\boldsymbol{Q}_t$, the optimal alignment matrix between $\boldsymbol{F}_t$ and $\boldsymbol{F}_\star$ exists. Therefore, for notational convenience, we denote $\boldsymbol{L} \coloneqq \boldsymbol{L}_t\boldsymbol{Q}_t$, $\boldsymbol{R} \coloneqq \boldsymbol{R}_t\boldsymbol{Q}_t^{-\top}$, $\boldsymbol{\Delta}_L \coloneqq \boldsymbol{L} - \boldsymbol{L}_\star$, $\boldsymbol{\Delta}_R \coloneqq \boldsymbol{R} - \boldsymbol{R}_\star$, and $\epsilon \coloneqq 0.02$. In addition, denote $\widetilde{\boldsymbol{F}}_{t+1}$ as the update before projection as

$$\widetilde{\boldsymbol{F}}_{t+1} \coloneqq \begin{bmatrix} \widetilde{\boldsymbol{L}}_{t+1} \\ \widetilde{\boldsymbol{R}}_{t+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{L}_t - \eta p^{-1}\mathcal{P}_\Omega(\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star)\boldsymbol{R}_t(\boldsymbol{R}_t^\top\boldsymbol{R}_t)^{-1} \\ \boldsymbol{R}_t - \eta p^{-1}\mathcal{P}_\Omega(\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star)^\top\boldsymbol{L}_t(\boldsymbol{L}_t^\top\boldsymbol{L}_t)^{-1} \end{bmatrix},$$

and therefore $\boldsymbol{F}_{t+1} = \mathcal{P}_B(\widetilde{\boldsymbol{F}}_{t+1})$. Note that in view of Lemma 6, it suffices to prove the following relation

$$\mathrm{dist}(\widetilde{\boldsymbol{F}}_{t+1}, \boldsymbol{F}_\star) \le (1 - 0.6\eta)\,\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star). \tag{A.34}$$

The conclusion $\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\|_{\mathsf{F}} \le 1.5\,\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$ is a simple consequence of Lemma 18; see (A.15) for a detailed argument. In what follows, we concentrate on proving (A.34).

To begin with, we list a few easy consequences under the assumed conditions.

Claim 6. *Under conditions* $\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \le \epsilon\sigma_r(\boldsymbol{X}_\star)$ *and* $\sqrt{n_1}\|\boldsymbol{L}\boldsymbol{R}^\top\|_{2,\infty} \vee \sqrt{n_2}\|\boldsymbol{R}\boldsymbol{L}^\top\|_{2,\infty} \le C_B\sqrt{\mu r}\sigma_1(\boldsymbol{X}_\star)$, *one has*

$$\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\| \vee \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\| \le \epsilon; \tag{A.35a}$$

$$\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\| \le \frac{1}{1-\epsilon}; \tag{A.35b}$$

$$\left\|\boldsymbol{\Sigma}_\star^{1/2}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\| \le \frac{1}{(1-\epsilon)^2}; \tag{A.35c}$$

$$\sqrt{n_1}\|\boldsymbol{L}\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \vee \sqrt{n_2}\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \le \frac{C_B}{1-\epsilon}\sqrt{\mu r}\sigma_1(\boldsymbol{X}_\star); \tag{A.35d}$$

$$\sqrt{n_1}\|\boldsymbol{L}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} \vee \sqrt{n_2}\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} \le \frac{C_B\kappa}{1-\epsilon}\sqrt{\mu r}; \tag{A.35e}$$

$$\sqrt{n_1}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \vee \sqrt{n_2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \le \left(1 + \frac{C_B}{1-\epsilon}\right)\sqrt{\mu r}\sigma_1(\boldsymbol{X}_\star). \tag{A.35f}$$

Now we are ready to embark on the proof of (A.34). By the definition of $\mathrm{dist}(\widetilde{\boldsymbol{F}}_{t+1}, \boldsymbol{F}_\star)$, one

has

$$\mathrm{dist}^2(\widetilde{\boldsymbol{F}}_{t+1}, \boldsymbol{F}_\star) \le \left\|(\widetilde{\boldsymbol{L}}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 + \left\|(\widetilde{\boldsymbol{R}}_{t+1}\boldsymbol{Q}_t^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2, \tag{A.36}$$

where we recall that $\boldsymbol{Q}_t$ is the optimal alignment matrix between $\boldsymbol{F}_t$ and $\boldsymbol{F}_\star$. Plug in the update rule (2.24) and the decomposition $\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star = \boldsymbol{\Delta}_L\boldsymbol{R}^\top + \boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top$ to obtain

$$\begin{aligned}
(\widetilde{\boldsymbol{L}}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2} &= \left(\boldsymbol{L} - \eta p^{-1}\mathcal{P}_\Omega(\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1} - \boldsymbol{L}_\star\right)\boldsymbol{\Sigma}_\star^{1/2} \\
&= \boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2} - \eta(\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2} - \eta(p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2} \\
&= (1-\eta)\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2} - \eta\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2} - \eta(p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}.
\end{aligned}$$

This allows us to expand the first square in (A.36) as

$$\begin{aligned}
\left\|(\widetilde{\boldsymbol{L}}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 = {} & \underbrace{\left\|(1-\eta)\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2} - \eta\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2}_{\mathfrak{P}_1} \\
& - 2\eta(1-\eta)\underbrace{\mathrm{tr}\left((p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\right)}_{\mathfrak{P}_2} \\
& + 2\eta^2\underbrace{\mathrm{tr}\left((p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\boldsymbol{\Delta}_R\boldsymbol{L}_\star^\top\right)}_{\mathfrak{P}_3} \\
& + \eta^2\underbrace{\left\|(p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2}_{\mathfrak{P}_4}.
\end{aligned}$$

In the sequel, we shall control the four terms separately, of which $\mathfrak{P}_1$ is the main term, and $\mathfrak{P}_2, \mathfrak{P}_3$ and $\mathfrak{P}_4$ are perturbation terms.

1. Notice that the main term $\mathfrak{P}_1$ has already been controlled in (A.13) under the condition (A.35a). It obeys

$$\mathfrak{P}_1 \le \left((1-\eta)^2 + \frac{2\epsilon}{1-\epsilon}\eta(1-\eta)\right)\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}\eta^2\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2.$$

2. For the second term $\mathfrak{P}_2$, decompose $\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star = \boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top + \boldsymbol{L}\boldsymbol{\Delta}_R^\top$ and apply the triangle inequality

to obtain

$$
\begin{aligned}
|\mathfrak{P}_2| &= \left| \operatorname{tr}\left((p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{\Delta}_L \boldsymbol{R}_\star^\top + \boldsymbol{L}\boldsymbol{\Delta}_R^\top)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star \boldsymbol{\Delta}_L^\top \right) \right| \\
&\le \underbrace{\left| \operatorname{tr}\left((p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{\Delta}_L \boldsymbol{R}_\star^\top)\boldsymbol{R}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star \boldsymbol{\Delta}_L^\top \right) \right|}_{\mathfrak{P}_{2,1}} \\
&\quad + \underbrace{\left| \operatorname{tr}\left((p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{\Delta}_L \boldsymbol{R}_\star^\top)\boldsymbol{\Delta}_R(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star \boldsymbol{\Delta}_L^\top \right) \right|}_{\mathfrak{P}_{2,2}} \\
&\quad + \underbrace{\left| \operatorname{tr}\left((p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{L}\boldsymbol{\Delta}_R^\top)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star \boldsymbol{\Delta}_L^\top \right) \right|}_{\mathfrak{P}_{2,3}}.
\end{aligned}
$$

For the first term $\mathfrak{P}_{2,1}$, under the event $\mathcal{E}$, we can invoke Lemma 26 to obtain

$$
\begin{aligned}
\mathfrak{P}_{2,1} &\le C_1 \sqrt{\frac{\mu r \log(n_1 \vee n_2)}{p(n_1 \wedge n_2)}} \|\boldsymbol{\Delta}_L \boldsymbol{R}_\star^\top\|_{\mathsf{F}} \left\| \boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star (\boldsymbol{R}^\top\boldsymbol{R})^{-1} \boldsymbol{R}_\star^\top \right\|_{\mathsf{F}} \\
&\le C_1 \sqrt{\frac{\mu r \log(n_1 \vee n_2)}{p(n_1 \wedge n_2)}} \|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 \left\| \boldsymbol{\Sigma}_\star^{1/2} (\boldsymbol{R}^\top\boldsymbol{R})^{-1} \boldsymbol{\Sigma}_\star^{1/2} \right\|,
\end{aligned}
$$

where the second line follows from the relation $\|\boldsymbol{A}\boldsymbol{B}\|_{\mathsf{F}} \le \|\boldsymbol{A}\|\|\boldsymbol{B}\|_{\mathsf{F}}$. Use the condition (A.35c) to obtain

$$
\mathfrak{P}_{2,1} \le \frac{C_1}{(1-\epsilon)^2} \sqrt{\frac{\mu r \log(n_1 \vee n_2)}{p(n_1 \wedge n_2)}} \|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2.
$$

Regarding the remaining terms $\mathfrak{P}_{2,2}$ and $\mathfrak{P}_{2,3}$, our main hammer is Lemma 27. Invoking Lemma 27 under the event $\mathcal{E}$ with $\boldsymbol{L}_A \coloneqq \boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2}$, $\boldsymbol{R}_A \coloneqq \boldsymbol{R}_\star \boldsymbol{\Sigma}_\star^{-1/2}$, $\boldsymbol{L}_B \coloneqq \boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2}$, and $\boldsymbol{R}_B \coloneqq \boldsymbol{\Delta}_R (\boldsymbol{R}^\top\boldsymbol{R})^{-1} \boldsymbol{\Sigma}_\star^{1/2}$, we arrive at

$$
\begin{aligned}
\mathfrak{P}_{2,2} &\le C_2 \sqrt{\frac{n_1 \vee n_2}{p}} \|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} \|\boldsymbol{R}_\star \boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} \left\| \boldsymbol{\Delta}_R (\boldsymbol{R}^\top\boldsymbol{R})^{-1} \boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}} \\
&\le C_2 \sqrt{\frac{n_1 \vee n_2}{p}} \|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} \|\boldsymbol{R}_\star \boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} \|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{F}} \left\| \boldsymbol{\Sigma}_\star^{1/2} (\boldsymbol{R}^\top\boldsymbol{R})^{-1} \boldsymbol{\Sigma}_\star^{1/2} \right\|.
\end{aligned}
$$

Similarly, with the help of Lemma 27, one has

$$\mathfrak{P}_{2,3} \le C_2\sqrt{\frac{n_1 \vee n_2}{p}}\|\boldsymbol{L}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\left\|\boldsymbol{\Sigma}_\star^{1/2}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|.$$

Utilizing the consequences in Claim 6, we arrive at

$$\begin{aligned}
\mathfrak{P}_{2,2} &\le \frac{C_2\kappa}{(1-\epsilon)^2}\left(1+\frac{C_B}{1-\epsilon}\right)\frac{\mu r}{\sqrt{p(n_1\wedge n_2)}}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}};\\
\mathfrak{P}_{2,3} &\le \frac{C_2C_B^2\kappa^2}{(1-\epsilon)^4}\frac{\mu r}{\sqrt{p(n_1\wedge n_2)}}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}.
\end{aligned}$$

We then combine the bounds for $\mathfrak{P}_{2,1}, \mathfrak{P}_{2,2}$ and $\mathfrak{P}_{2,3}$ to see

$$\begin{aligned}
\mathfrak{P}_2 &\le \frac{C_1}{(1-\epsilon)^2}\sqrt{\frac{\mu r\log(n_1\vee n_2)}{p(n_1\wedge n_2)}}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\\
&\quad+\frac{C_2\kappa}{(1-\epsilon)^2}\left(1+\frac{C_B}{1-\epsilon}+\frac{C_B^2\kappa}{(1-\epsilon)^2}\right)\frac{\mu r}{\sqrt{p(n_1\wedge n_2)}}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\\
&= \delta_1\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+\delta_2\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\\
&\le (\delta_1+\frac{\delta_2}{2})\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+\frac{\delta_2}{2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2,
\end{aligned}$$

where we denote

$$\delta_1 \coloneqq \frac{C_1}{(1-\epsilon)^2}\sqrt{\frac{\mu r\log(n_1\vee n_2)}{p(n_1\wedge n_2)}},\quad\text{and}\quad \delta_2 \coloneqq \frac{C_2\kappa}{(1-\epsilon)^2}\left(1+\frac{C_B}{1-\epsilon}+\frac{C_B^2\kappa}{(1-\epsilon)^2}\right)\frac{\mu r}{\sqrt{p(n_1\wedge n_2)}}.$$

3. Following a similar argument for controlling $\mathfrak{P}_2$ (i.e. repeatedly using Lemmas 26 and 27), we can obtain the following bounds for $\mathfrak{P}_3$ and $\mathfrak{P}_4$, whose proof are deferred to the end of this section.

Claim 7. *Under the event $\mathcal{E}$, one has*

$$\begin{aligned}
\mathfrak{P}_3 &\le \frac{\delta_2}{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+(\delta_1+\frac{\delta_2}{2})\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2;\\
\mathfrak{P}_4 &\le \delta_1(\delta_1+\delta_2)\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+\delta_2(\delta_1+\delta_2)\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2.
\end{aligned}$$

Taking the bounds for $\mathfrak{P}_1, \mathfrak{P}_2, \mathfrak{P}_3$ and $\mathfrak{P}_4$ collectively yields

$$\begin{aligned}\left\|(\widetilde{\boldsymbol{L}}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 \le &\left((1-\eta)^2 + \frac{2\epsilon}{1-\epsilon}\eta(1-\eta)\right)\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}\eta^2\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 \\ &+ \eta(1-\eta)\left((2\delta_1+\delta_2)\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \delta_2\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right) \\ &+ \eta^2\left(\delta_2\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + (2\delta_1+\delta_2)\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right) \\ &+ \eta^2\left(\delta_1(\delta_1+\delta_2)\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \delta_2(\delta_1+\delta_2)\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right).\end{aligned}$$

A similar upper bound holds for the second square in (A.36). As a result, we reach the conclusion that

$$\left\|(\widetilde{\boldsymbol{L}}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 + \left\|(\widetilde{\boldsymbol{R}}_{t+1}\boldsymbol{Q}_t^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 \le \rho^2(\eta;\epsilon,\delta_1,\delta_2)\,\mathrm{dist}^2(\boldsymbol{F}_t,\boldsymbol{F}_\star),$$

where the contraction rate $\rho^2(\eta;\epsilon,\delta_1,\delta_2)$ is given by

$$\rho^2(\eta;\epsilon,\delta_1,\delta_2) \coloneqq (1-\eta)^2 + \left(\frac{2\epsilon}{1-\epsilon} + 2(\delta_1+\delta_2)\right)\eta(1-\eta) + \left(\frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2} + 2(\delta_1+\delta_2) + (\delta_1+\delta_2)^2\right)\eta^2.$$

As long as $p \ge C(\mu r\kappa^4 \vee \log(n_1\vee n_2))\mu r/(n_1\wedge n_2)$ for some sufficiently large constant C, one has $\delta_1+\delta_2 \le 0.1$ under the setting $\epsilon = 0.02$. When $0 < \eta \le 2/3$, one further has $\rho(\eta;\epsilon,\delta_1,\delta_2) \le 1-0.6\eta$. Thus we conclude that

$$\begin{aligned}\mathrm{dist}(\widetilde{\boldsymbol{F}}_{t+1},\boldsymbol{F}_\star) &\le \sqrt{\left\|(\widetilde{\boldsymbol{L}}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 + \left\|(\widetilde{\boldsymbol{R}}_{t+1}\boldsymbol{Q}_t^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2} \\ &\le (1-0.6\eta)\,\mathrm{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star),\end{aligned}$$

which is exactly the upper bound we are after; see (A.34). This finishes the proof.

Proof of Claim 6. First, repeating the derivation for (A.12) obtains (A.35a). Second, take the condition (A.35a) and Lemma 17 together to obtain (A.35b) and (A.35c). Third, take the incoherence

condition $\sqrt{n_1}\|\boldsymbol{L}\boldsymbol{R}^\top\|_{2,\infty} \vee \sqrt{n_2}\|\boldsymbol{R}\boldsymbol{L}^\top\|_{2,\infty} \leq C_B\sqrt{\mu r}\sigma_1(\boldsymbol{X}_\star)$ together with the relations

$$\begin{aligned}
\|\boldsymbol{L}\boldsymbol{R}^\top\|_{2,\infty} &\geq \sigma_r(\boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2})\|\boldsymbol{L}\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \\
&\geq \left(\sigma_r(\boldsymbol{R}_\star\boldsymbol{\Sigma}_\star^{-1/2}) - \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\|\right)\|\boldsymbol{L}\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \\
&\geq (1-\epsilon)\|\boldsymbol{L}\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty}; \\
\|\boldsymbol{R}\boldsymbol{L}^\top\|_{2,\infty} &\geq \sigma_r(\boldsymbol{L}\boldsymbol{\Sigma}_\star^{-1/2})\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \\
&\geq \left(\sigma_r(\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}) - \|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|\right)\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \\
&\geq (1-\epsilon)\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty}
\end{aligned}$$

to obtain (A.35d) and (A.35e). Finally, apply the triangle inequality together with incoherence assumption to obtain (A.35f). □

Proof of Claim 7. We start with the term $\mathfrak{P}_3$, for which we have

$$\begin{aligned}
|\mathfrak{P}_3| \leq &\underbrace{\left|\operatorname{tr}\left((p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\boldsymbol{\Delta}_R\boldsymbol{L}_\star^\top\right)\right|}_{\mathfrak{P}_{3,1}} \\
&+ \underbrace{\left|\operatorname{tr}\left((p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{\Delta}_L\boldsymbol{R}^\top)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\boldsymbol{\Delta}_R\boldsymbol{L}_\star^\top\right)\right|}_{\mathfrak{P}_{3,2}}.
\end{aligned}$$

Invoke Lemma 26 to bound $\mathfrak{P}_{3,1}$ as

$$\begin{aligned}
\mathfrak{P}_{3,1} &\leq C_1\sqrt{\frac{\mu r\log(n_1\vee n_2)}{p(n_1\wedge n_2)}}\|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}}\left\|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\right\|_{\mathsf{F}} \\
&\leq C_1\sqrt{\frac{\mu r\log(n_1\vee n_2)}{p(n_1\wedge n_2)}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|^2.
\end{aligned}$$

The condition (A.35b) allows us to obtain a simplified bound

$$\mathfrak{P}_{3,1} \leq \frac{C_1}{(1-\epsilon)^2}\sqrt{\frac{\mu r\log(n_1\vee n_2)}{p(n_1\wedge n_2)}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2.$$

In regard to $\mathfrak{P}_{3,2}$, we apply Lemma 27 with $\boldsymbol{L}_A \coloneqq \boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}$, $\boldsymbol{R}_A \coloneqq \boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2}$, $\boldsymbol{L}_B \coloneqq \boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}$, and

$\boldsymbol{R}_B \coloneqq \boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}$ to see

$$\begin{aligned}\mathfrak{P}_{3,2} &\le C_2\sqrt{\frac{n_1\vee n_2}{p}}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}\\ &\le C_2\sqrt{\frac{n_1\vee n_2}{p}}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|^2\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}.\end{aligned}$$

Again, use the consequences in Claim 6 to reach

$$\begin{aligned}\mathfrak{P}_{3,2} &\le C_2\sqrt{\frac{n_1\vee n_2}{p}}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\sqrt{\frac{\mu r}{n_1}}\frac{C_B\kappa}{1-\epsilon}\sqrt{\frac{\mu r}{n_2}}\frac{1}{(1-\epsilon)^2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\\ &= \frac{C_2C_B\kappa}{(1-\epsilon)^3}\frac{\mu r}{\sqrt{p(n_1\wedge n_2)}}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}.\end{aligned}$$

Combine the bounds of $\mathfrak{P}_{3,1}$ and $\mathfrak{P}_{3,2}$ to reach

$$\begin{aligned}\mathfrak{P}_3 &\le \frac{C_1}{(1-\epsilon)^2}\sqrt{\frac{\mu r\log(n_1\vee n_2)}{p(n_1\wedge n_2)}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\\ &\quad+\frac{C_2C_B\kappa}{(1-\epsilon)^3}\frac{\mu r}{\sqrt{p(n_1\wedge n_2)}}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\\ &\le \delta_1\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+\delta_2\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\\ &\le \frac{\delta_2}{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+(\delta_1+\frac{\delta_2}{2})\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2.\end{aligned}$$

Moving on to the term $\mathfrak{P}_4$, we have

$$\begin{aligned}\sqrt{\mathfrak{P}_4} &= \left\|(p^{-1}\mathcal{P}_\Omega-\mathcal{I})(\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}\\ &\le \underbrace{\left|\operatorname{tr}\left((p^{-1}\mathcal{P}_\Omega-\mathcal{I})(\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top)\boldsymbol{R}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\widetilde{\boldsymbol{L}}^\top\right)\right|}_{\mathfrak{P}_{4,1}}\\ &\quad+\underbrace{\left|\operatorname{tr}\left((p^{-1}\mathcal{P}_\Omega-\mathcal{I})(\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top)\boldsymbol{\Delta}_R(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\widetilde{\boldsymbol{L}}^\top\right)\right|}_{\mathfrak{P}_{4,2}}\\ &\quad+\underbrace{\left|\operatorname{tr}\left((p^{-1}\mathcal{P}_\Omega-\mathcal{I})(\boldsymbol{L}\boldsymbol{\Delta}_R^\top)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\widetilde{\boldsymbol{L}}^\top\right)\right|}_{\mathfrak{P}_{4,3}},\end{aligned}$$

where we have used the variational representation of the Frobenius norm for some $\widetilde{\boldsymbol{L}}\in\mathbb{R}^{n_1\times r}$ obeying

$\|\widetilde{\boldsymbol{L}}\|_{\mathsf{F}} = 1$. Note that the decomposition of $\sqrt{\mathfrak{P}_4}$ is extremely similar to that of $\mathfrak{P}_2$. Therefore we can follow a similar argument (i.e. applying Lemmas 26 and 27) to control these terms as

$$\begin{aligned}
\mathfrak{P}_{4,1} &\leq \frac{C_1}{(1-\epsilon)^2}\sqrt{\frac{\mu r \log(n_1 \vee n_2)}{p(n_1 \wedge n_2)}}\|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}; \\
\mathfrak{P}_{4,2} &\leq \frac{C_2\kappa}{(1-\epsilon)^2}\left(1+\frac{C_B}{1-\epsilon}\right)\frac{\mu r}{\sqrt{p(n_1 \wedge n_2)}}\|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}; \\
\mathfrak{P}_{4,3} &\leq \frac{C_2 C_B^2 \kappa^2}{(1-\epsilon)^4}\frac{\mu r}{\sqrt{p(n_1 \wedge n_2)}}\|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}.
\end{aligned}$$

For conciseness, we omit the details for bounding each term. Combine them to reach

$$\sqrt{\mathfrak{P}_4} \leq \delta_1 \|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \delta_2 \|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}.$$

Finally take the square on both sides and use $2ab \leq a^2 + b^2$ to obtain the upper bound

$$\mathfrak{P}_4 \leq \delta_1(\delta_1+\delta_2)\|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \delta_2(\delta_1+\delta_2)\|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2.$$

□

A.5.3 Proof of Lemma 8

We start by recording a useful lemma below.

Lemma 28 ([Che15, Lemma 2], [CLL20, Lemma 4]). *For any fixed $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$, with overwhelming probability, one has*

$$\left\|(p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{X})\right\| \leq C_0 \frac{\log(n_1 \vee n_2)}{p}\|\boldsymbol{X}\|_\infty + C_0\sqrt{\frac{\log(n_1 \vee n_2)}{p}}(\|\boldsymbol{X}\|_{2,\infty} \vee \|\boldsymbol{X}^\top\|_{2,\infty}),$$

where $C_0 > 0$ is some universal constant that does not depend on $\boldsymbol{X}$.

In view of Lemma 16, one has

$$\mathrm{dist}(\widetilde{\boldsymbol{F}}_0, \boldsymbol{F}_\star) \leq \sqrt{\sqrt{2}+1}\left\|\boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^\top - \boldsymbol{X}_\star\right\|_{\mathsf{F}} \leq \sqrt{(\sqrt{2}+1)2r}\left\|\boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^\top - \boldsymbol{X}_\star\right\|, \tag{A.37}$$

where the last relation uses the fact that $\boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^\top - \boldsymbol{X}_\star$ has rank at most $2r$. Applying the triangle inequality, we obtain

$$\begin{aligned}\left\|\boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^\top - \boldsymbol{X}_\star\right\| &\le \left\|p^{-1}\mathcal{P}_\Omega(\boldsymbol{X}_\star) - \boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^\top\right\| + \left\|p^{-1}\mathcal{P}_\Omega(\boldsymbol{X}_\star) - \boldsymbol{X}_\star\right\| \\ &\le 2\left\|(p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{X}_\star)\right\|. \end{aligned} \tag{A.38}$$

Here the second inequality hinges on the fact that $\boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^\top$ is the best rank-r approximation to $p^{-1}\mathcal{P}_\Omega(\boldsymbol{X}_\star)$, i.e.

$$\left\|p^{-1}\mathcal{P}_\Omega(\boldsymbol{X}_\star) - \boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^\top\right\| \le \left\|p^{-1}\mathcal{P}_\Omega(\boldsymbol{X}_\star) - \boldsymbol{X}_\star\right\|.$$

Combining (A.37) and (A.38) yields

$$\mathrm{dist}(\widetilde{\boldsymbol{F}}_0, \boldsymbol{F}_\star) \le 2\sqrt{(\sqrt{2}+1)2r}\left\|(p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{X}_\star)\right\| \le 5\sqrt{r}\left\|(p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{X}_\star)\right\|.$$

It then boils down to controlling $\left\|p^{-1}\mathcal{P}_\Omega(\boldsymbol{X}_\star) - \boldsymbol{X}_\star\right\|$, which is readily supplied by Lemma 28 as

$$\left\|(p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{X}_\star)\right\| \le C_0\frac{\log(n_1\vee n_2)}{p}\|\boldsymbol{X}_\star\|_\infty + C_0\sqrt{\frac{\log(n_1\vee n_2)}{p}}(\|\boldsymbol{X}_\star\|_{2,\infty}\vee\|\boldsymbol{X}_\star^\top\|_{2,\infty}),$$

which holds with overwhelming probability. The proof is finished by plugging the following bounds from incoherence assumption of $\boldsymbol{X}_\star$:

$$\begin{aligned}\|\boldsymbol{X}_\star\|_\infty &\le \|\boldsymbol{U}_\star\|_{2,\infty}\|\boldsymbol{\Sigma}_\star\|\|\boldsymbol{V}_\star\|_{2,\infty} \le \frac{\mu r}{\sqrt{n_1 n_2}}\kappa\sigma_r(\boldsymbol{X}_\star); \\ \|\boldsymbol{X}_\star\|_{2,\infty} &\le \|\boldsymbol{U}_\star\|_{2,\infty}\|\boldsymbol{\Sigma}_\star\|\|\boldsymbol{V}_\star\| \le \sqrt{\frac{\mu r}{n_1}}\kappa\sigma_r(\boldsymbol{X}_\star); \\ \|\boldsymbol{X}_\star^\top\|_{2,\infty} &\le \|\boldsymbol{U}_\star\|\|\boldsymbol{\Sigma}_\star\|\|\boldsymbol{V}_\star\|_{2,\infty} \le \sqrt{\frac{\mu r}{n_2}}\kappa\sigma_r(\boldsymbol{X}_\star).\end{aligned}$$

A.6 Proof for General Loss Functions

We first present a useful property of restricted smooth and convex functions.

Lemma 29. *Suppose that $f : \mathbb{R}^{n_1 \times n_2} \mapsto \mathbb{R}$ is rank-$2r$ restricted L-smooth and rank-$2r$ restricted convex. Then for any $\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathbb{R}^{n_1 \times n_2}$ of rank at most r, one has*

$$\langle \nabla f(\boldsymbol{X}_1) - \nabla f(\boldsymbol{X}_2), \boldsymbol{X}_1 - \boldsymbol{X}_2 \rangle \geq \frac{1}{L} \|\nabla f(\boldsymbol{X}_1) - \nabla f(\boldsymbol{X}_2)\|_{\mathsf{F},r}^2.$$

Proof. Since $f(\cdot)$ is rank-$2r$ restricted L-smooth and convex, it holds for any $\bar{\boldsymbol{X}} \in \mathbb{R}^{n_1 \times n_2}$ with rank at most $2r$ that

$$f(\boldsymbol{X}_1) + \langle \nabla f(\boldsymbol{X}_1), \bar{\boldsymbol{X}} - \boldsymbol{X}_1 \rangle \leq f(\bar{\boldsymbol{X}}) \leq f(\boldsymbol{X}_2) + \langle \nabla f(\boldsymbol{X}_2), \bar{\boldsymbol{X}} - \boldsymbol{X}_2 \rangle + \frac{L}{2} \|\bar{\boldsymbol{X}} - \boldsymbol{X}_2\|_{\mathsf{F}}^2.$$

Reorganize the terms to yield

$$f(\boldsymbol{X}_1) + \langle \nabla f(\boldsymbol{X}_1), \boldsymbol{X}_2 - \boldsymbol{X}_1 \rangle \leq f(\boldsymbol{X}_2) + \langle \nabla f(\boldsymbol{X}_2) - \nabla f(\boldsymbol{X}_1), \bar{\boldsymbol{X}} - \boldsymbol{X}_2 \rangle + \frac{L}{2} \|\bar{\boldsymbol{X}} - \boldsymbol{X}_2\|_{\mathsf{F}}^2.$$

Take $\bar{\boldsymbol{X}} = \boldsymbol{X}_2 - \frac{1}{L} \mathcal{P}_r(\nabla f(\boldsymbol{X}_2) - \nabla f(\boldsymbol{X}_1))$, whose rank is at most $2r$, to see

$$f(\boldsymbol{X}_1) + \langle \nabla f(\boldsymbol{X}_1), \boldsymbol{X}_2 - \boldsymbol{X}_1 \rangle + \frac{1}{2L} \|\nabla f(\boldsymbol{X}_2) - \nabla f(\boldsymbol{X}_1)\|_{\mathsf{F},r}^2 \leq f(\boldsymbol{X}_2).$$

We can further switch the roles of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ to obtain

$$f(\boldsymbol{X}_2) + \langle \nabla f(\boldsymbol{X}_2), \boldsymbol{X}_1 - \boldsymbol{X}_2 \rangle + \frac{1}{2L} \|\nabla f(\boldsymbol{X}_2) - \nabla f(\boldsymbol{X}_1)\|_{\mathsf{F},r}^2 \leq f(\boldsymbol{X}_1).$$

Adding the above two inequalities yields the desired bound. □

A.6.1 Proof of Theorem 4

Suppose that the t-th iterate $\boldsymbol{F}_t$ obeys the condition $\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \leq 0.1\sigma_r(\boldsymbol{X}_\star)/\sqrt{\kappa_f}$. In view of Lemma 14, one knows that $\boldsymbol{Q}_t$, the optimal alignment matrix between $\boldsymbol{F}_t$ and $\boldsymbol{F}_\star$ exists. Therefore, for notational convenience, denote $\boldsymbol{L} \coloneqq \boldsymbol{L}_t \boldsymbol{Q}_t$, $\boldsymbol{R} \coloneqq \boldsymbol{R}_t \boldsymbol{Q}_t^{-\top}$, $\boldsymbol{\Delta}_L \coloneqq \boldsymbol{L} - \boldsymbol{L}_\star$, $\boldsymbol{\Delta}_R \coloneqq \boldsymbol{R} - \boldsymbol{R}_\star$, and

$\epsilon \coloneqq 0.1/\sqrt{\kappa_f}$. Similar to the derivation in (A.12), we have

$$\|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{-1/2}\| \vee \|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{-1/2}\| \le \epsilon. \tag{A.39}$$

The conclusion $\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\|_{\mathsf{F}} \le 1.5\,\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$ is a simple consequence of Lemma 18; see (A.15) for a detailed argument. From now on, we focus on proving the distance contraction.

By the definition of $\mathrm{dist}(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star)$, one has

$$\mathrm{dist}^2(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) \le \left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2. \tag{A.40}$$

Introduce an auxiliary function

$$f_\mu(\boldsymbol{X}) = f(\boldsymbol{X}) - \frac{\mu}{2}\|\boldsymbol{X} - \boldsymbol{X}_\star\|_{\mathsf{F}}^2,$$

which is rank-$2r$ restricted $(L-\mu)$-smooth and rank-$2r$ restricted convex. Using the `ScaledGD` update rule (2.25) and the decomposition $\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star = \boldsymbol{\Delta}_L\boldsymbol{R}^\top + \boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top$, we obtain

$$\begin{aligned}
(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2} &= \left(\boldsymbol{L} - \eta\nabla f(\boldsymbol{L}\boldsymbol{R}^\top)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1} - \boldsymbol{L}_\star\right)\boldsymbol{\Sigma}_\star^{1/2} \\
&= \left(\boldsymbol{L} - \eta\mu(\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1} - \eta\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1} - \boldsymbol{L}_\star\right)\boldsymbol{\Sigma}_\star^{1/2} \\
&= (1-\eta\mu)\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2} - \eta\mu\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2} - \eta\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}.
\end{aligned}$$

As a result, one can expand the first square in (A.40) as

$$\begin{aligned}
\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 = {} & \underbrace{\left\|(1-\eta\mu)\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2} - \eta\mu\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2}_{\mathfrak{G}_1} \\
& - 2\eta(1-\eta\mu)\underbrace{\left\langle \nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top), \boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top - \boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top - \frac{1}{2}\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top \right\rangle}_{\mathfrak{G}_2} \\
& - 2\eta(1-\eta\mu)\left\langle \nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top), \boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top + \frac{1}{2}\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top \right\rangle
\end{aligned}$$

$$+ 2\eta^2\mu \underbrace{\left\langle \nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top), \boldsymbol{L}_\star \boldsymbol{\Delta}_R^\top \boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top \right\rangle}_{\mathfrak{G}_3}$$
$$+ \eta^2 \underbrace{\left\| \nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2}_{\mathfrak{G}_4}.$$

In the sequel, we shall bound the four terms separately.

1. Notice that the main term $\mathfrak{G}_1$ has already been controlled in (A.13) under the condition (A.39). It obeys

$$\mathfrak{G}_1 \le \left((1-\eta\mu)^2 + \frac{2\epsilon}{1-\epsilon}\eta\mu(1-\eta\mu)\right)\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}\eta^2\mu^2\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2,$$

as long as $\eta\mu \le 2/3$.

2. For the second term $\mathfrak{G}_2$, note that $\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top - \boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top - \frac{1}{2}\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top$ has rank at most r. Hence we can invoke Lemma 20 to obtain

$$\begin{aligned}
|\mathfrak{G}_2| &\le \|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}\left\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top - \boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top - \frac{1}{2}\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\right\|_{\mathsf{F}} \\
&\le \|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\left(\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2} - \boldsymbol{V}_\star\right\| + \frac{1}{2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\|\right),
\end{aligned}$$

where the second line uses $\boldsymbol{R}_\star = \boldsymbol{V}_\star\boldsymbol{\Sigma}_\star^{1/2}$. Take the condition (A.39) and Lemma 17 together to obtain

$$\begin{aligned}
\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\| &\le \frac{1}{1-\epsilon}; \\
\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2} - \boldsymbol{V}_\star\right\| &\le \frac{\sqrt{2}\epsilon}{1-\epsilon}.
\end{aligned}$$

These consequences further imply that

$$|\mathfrak{G}_2| \le \left(\frac{\sqrt{2}\epsilon}{1-\epsilon} + \frac{\epsilon}{2}\right)\|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}.$$

3. As above, the third term $\mathfrak{G}_3$ can be similarly bounded as

$$\begin{aligned}|\mathfrak{G}_3| &\le \|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}\left\|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\right\|_{\mathsf{F}}\\ &\le \|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|^2\\ &\le \frac{1}{(1-\epsilon)^2}\|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}.\end{aligned}$$

4. For the last term $\mathfrak{G}_4$, invoke Lemma 20 to obtain

$$\mathfrak{G}_4 \le \|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}^2\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|^2 \le \frac{1}{(1-\epsilon)^2}\|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}^2.$$

Taking collectively the bounds for $\mathfrak{G}_1,\mathfrak{G}_2,\mathfrak{G}_3$ and $\mathfrak{G}_4$ yields

$$\begin{aligned}\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 \le{}& \left((1-\eta\mu)^2+\frac{2\epsilon}{1-\epsilon}\eta\mu(1-\eta\mu)\right)\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+\frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}\eta^2\mu^2\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\\ &+2\eta(\frac{\sqrt{2}\epsilon}{1-\epsilon}+\frac{\epsilon}{2})(1-\eta\mu)\|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\\ &-2\eta(1-\eta\mu)\left\langle\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top),\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top+\frac{1}{2}\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\right\rangle\\ &+\frac{2\eta^2\mu}{(1-\epsilon)^2}\|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}+\frac{\eta^2}{(1-\epsilon)^2}\|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}^2.\end{aligned}$$

Similarly, we can obtain the control of $\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top}-\boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2$. Combine them together to reach

$$\begin{aligned}&\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2+\left\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top}-\boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2\\ &\quad\le\left((1-\eta\mu)^2+\frac{2\epsilon}{1-\epsilon}\eta\mu(1-\eta\mu)+\frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}\eta^2\mu^2\right)\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right)\\ &\qquad+2\eta\left((\frac{\sqrt{2}\epsilon}{1-\epsilon}+\frac{\epsilon}{2})(1-\eta\mu)+\frac{\eta\mu}{(1-\epsilon)^2}\right)\|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)\\ &\qquad-2\eta(1-\eta\mu)\left\langle\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top),\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top+\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top+\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\right\rangle+\frac{2\eta^2}{(1-\epsilon)^2}\|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}^2\\ &\quad\le\left((1-\eta\mu)^2+\frac{2\epsilon}{1-\epsilon}\eta\mu(1-\eta\mu)+\frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}\eta^2\mu^2\right)\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right)\end{aligned}$$

$$+ 2\eta \underbrace{\left((\frac{\sqrt{2}\epsilon}{1-\epsilon} + \frac{\epsilon}{2})(1-\eta\mu) + \frac{\eta\mu}{(1-\epsilon)^2} \right)}_{\mathfrak{C}_1} \|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r} \left(\|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} \right)$$
$$- 2\eta \underbrace{\left(\frac{1-\eta\mu}{L-\mu} - \frac{\eta}{(1-\epsilon)^2} \right)}_{\mathfrak{C}_2} \|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}^2,$$

where the last line follows from Lemma 29 (notice that $\nabla f_\mu(\boldsymbol{X}_\star) = \boldsymbol{0}$) as

$$\langle \nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top), \boldsymbol{\Delta}_L \boldsymbol{R}_\star^\top + \boldsymbol{L}_\star \boldsymbol{\Delta}_R^\top + \boldsymbol{\Delta}_L \boldsymbol{\Delta}_R^\top \rangle = \langle \nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top), \boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star \rangle \geq \frac{1}{L-\mu} \|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}^2.$$

Notice that $\mathfrak{C}_2 > 0$ as long as $\eta \leq (1-\epsilon)^2/L$. Maximizing the quadratic function of $\|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}$ yields

$$\begin{aligned}\mathfrak{C}_1 \|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r} \left(\|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} \right) - \mathfrak{C}_2 \|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}^2 &\leq \frac{\mathfrak{C}_1^2}{4\mathfrak{C}_2} \left(\|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} \right)^2 \\ &\leq \frac{\mathfrak{C}_1^2}{2\mathfrak{C}_2} \left(\|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 \right),\end{aligned}$$

where the last inequality holds since $(a+b)^2 \leq 2(a^2+b^2)$. Identify $\mathrm{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star) = \|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2$ to obtain

$$\left\| (\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2 + \left\| (\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2 \leq \rho^2(\eta;\epsilon,\mu,L)\, \mathrm{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star),$$

where the contraction rate is given by

$$\rho^2(\eta;\epsilon,\mu,L) \coloneqq (1-\eta\mu)^2 + \frac{2\epsilon}{1-\epsilon}\eta\mu(1-\eta\mu) + \frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}\eta^2\mu^2 + \frac{\left((\frac{\sqrt{2}\epsilon}{1-\epsilon} + \frac{\epsilon}{2})(1-\eta\mu) + \frac{\eta\mu}{(1-\epsilon)^2} \right)^2}{1-\eta\mu-\frac{\eta(L-\mu)}{(1-\epsilon)^2}} \eta(L-\mu).$$

With $\epsilon = 0.1/\sqrt{\kappa_f}$ and $0 < \eta \leq 0.4/L$, one has $\rho(\eta;\epsilon,\mu,L) \leq 1 - 0.7\eta\mu$. Thus we conclude that

$$\begin{aligned}\mathrm{dist}(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) &\leq \sqrt{\left\| (\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2 + \left\| (\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2} \\ &\leq (1-0.7\eta\mu)\, \mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star),\end{aligned}$$

which is the desired claim.

Remark 9. We provide numerical details for the contraction rate. For simplicity, we shall prove $\rho(\eta;\epsilon,\mu,L) \le 1-0.7\eta\mu$ under a stricter condition $\epsilon = 0.02/\sqrt{\kappa_f}$. The stronger result under the condition $\epsilon = 0.1/\sqrt{\kappa_f}$ can be verified through a subtler analysis.

With $\epsilon = 0.02/\sqrt{\kappa_f}$ and $0 < \eta \le 0.4/L$, one can bound the terms in $\rho^2(\eta;\epsilon,\mu,L)$ as

$$(1-\eta\mu)^2 + \frac{2\epsilon}{1-\epsilon}\eta\mu(1-\eta\mu) + \frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}\eta^2\mu^2 \le 1 - 1.959\eta\mu + 1.002\eta^2\mu^2; \tag{A.41}$$

$$\begin{aligned}\frac{\left((\frac{\sqrt{2}\epsilon}{1-\epsilon}+\frac{\epsilon}{2})(1-\eta\mu)+\frac{\eta\mu}{(1-\epsilon)^2}\right)^2}{1-\eta\mu-\frac{\eta(L-\mu)}{(1-\epsilon)^2}}\eta(L-\mu) &\le \frac{\frac{0.0016}{\kappa_f}+0.078\eta\mu+1.005\eta^2\mu^2}{1-1.042\eta L}\eta L \\ &\le \frac{0.0016\eta\frac{L}{\kappa_f}+0.4\times(0.078\eta\mu+1.005\eta^2\mu^2)}{1-0.4\times 1.042} \\ &\le 0.057\eta\mu + 0.69\eta^2\mu^2, \end{aligned} \tag{A.42}$$

where the last line uses the definition (2.26) of κ_f. Putting (A.41) and (A.42) together further implies

$$\rho^2(\eta;\epsilon,\mu,L) \le 1 - 1.9\eta\mu + 1.7\eta^2\mu^2 \le (1-0.7\eta\mu)^2,$$

as long as $0 < \eta\mu \le 0.4$.

Appendix B

Proofs for Robust Low-rank Matrix Estimation

Lemma 30. *Suppose that $f(\cdot) : \mathbb{R}^{n_1 \times n_2} \mapsto \mathbb{R}$ is convex and rank-r restricted L-Lipschitz continuous (cf. Definition 5). Then for any subgradient $\boldsymbol{S} \in \partial f(\boldsymbol{X})$, one has $\|\boldsymbol{S}\|_{\mathsf{F},r} \leq L$.*

Proof. Fix any subgradient $\boldsymbol{S} \in \partial f(\boldsymbol{X})$. By the definition of a subgradient, for any $\widetilde{\boldsymbol{X}} \in \mathbb{R}^{n_1 \times n_2}$, one has

$$f(\widetilde{\boldsymbol{X}}) \geq f(\boldsymbol{X}) + \langle \boldsymbol{S}, \widetilde{\boldsymbol{X}} - \boldsymbol{X} \rangle.$$

In particular, taking $\widetilde{\boldsymbol{X}} = \boldsymbol{X} + \mathcal{P}_r(\boldsymbol{S})$ arrives at

$$f(\boldsymbol{X} + \mathcal{P}_r(\boldsymbol{S})) \geq f(\boldsymbol{X}) + \langle \boldsymbol{S}, \mathcal{P}_r(\boldsymbol{S}) \rangle = f(\boldsymbol{X}) + \|\boldsymbol{S}\|_{\mathsf{F},r}^2, \tag{B.1}$$

where the last equality follows from the definition (A.8). Note that $\mathcal{P}_r(\boldsymbol{S})$ has rank at most r. By the rank-r restricted L-Lipschitz continuity of $f(\cdot)$, we have

$$f(\boldsymbol{X} + \mathcal{P}_r(\boldsymbol{S})) - f(\boldsymbol{X}) \leq L\|\mathcal{P}_r(\boldsymbol{S})\|_{\mathsf{F}} = L\|\boldsymbol{S}\|_{\mathsf{F},r}.$$

Combining the above inequality with (B.1), we conclude $\|\boldsymbol{S}\|_{\mathsf{F},r} \leq L$. □

B.1 Proof of Theorem 6

Suppose that the t-th iterate $\boldsymbol{F}_t$ obeys the condition

$$\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \le 0.02\sigma_r(\boldsymbol{X}_\star)/\chi_f. \tag{B.2}$$

Lemma 14 ensures that $\boldsymbol{Q}_t$, the optimal alignment matrix between $\boldsymbol{F}_t$ and $\boldsymbol{F}_\star$ exists. For notational convenience, we denote $\boldsymbol{L} \coloneqq \boldsymbol{L}_t\boldsymbol{Q}_t$, $\boldsymbol{R} \coloneqq \boldsymbol{R}_t\boldsymbol{Q}_t^{-\top}$, $\boldsymbol{\Delta}_L \coloneqq \boldsymbol{L} - \boldsymbol{L}_\star$, $\boldsymbol{\Delta}_R \coloneqq \boldsymbol{R} - \boldsymbol{R}_\star$, $\boldsymbol{S} \coloneqq \boldsymbol{S}_t$, and $\epsilon \coloneqq 0.02/\chi_f$. By the definition

$$\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) = \sqrt{\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2} \tag{B.3}$$

and the relation $\|\boldsymbol{AB}\|_{\mathsf{F}} \ge \|\boldsymbol{A}\|_{\mathsf{F}}\sigma_r(\boldsymbol{B}) \ge \|\boldsymbol{A}\|\sigma_r(\boldsymbol{B})$, we have

$$\max\{\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|, \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\|\} \le \epsilon. \tag{B.4}$$

We start by relating $\|\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star\|_{\mathsf{F}}$ to $\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$ given (B.4). Applying the triangle inequality to the basic relation $\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star = \boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star = \boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top + \boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top + \boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top$, we have

$$\begin{aligned}
\|\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star\|_{\mathsf{F}} &\le \|\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top\|_{\mathsf{F}} + \|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}} \\
&\le \|\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top\|_{\mathsf{F}} + \|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}} + \frac{1}{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \frac{1}{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\| \\
&\le \left(1 + \frac{1}{2}\max\{\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|, \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\|\}\right)\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right) \\
&\le \left(1 + \frac{\epsilon}{2}\right)\sqrt{2}\sqrt{\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2} \le 1.5\,\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star),
\end{aligned} \tag{B.5}$$

where the last line uses the basic inequality $\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} \le \sqrt{2}\,\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$ and (B.4).

From now on, we focus on proving the distance contraction. By the definition of $\mathrm{dist}(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star)$, one has

$$\mathrm{dist}^2(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) \le \left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2. \tag{B.6}$$

We expand the first square in (B.6) as

$$\begin{aligned}\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 &= \left\|\left(\boldsymbol{L}-\eta_t\boldsymbol{S}\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}-\boldsymbol{L}_\star\right)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 \\ &= \|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 - 2\eta_t\left\langle \boldsymbol{S}, \boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\right\rangle + \eta_t^2\left\|\boldsymbol{S}\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 \\ &= \|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 - 2\eta_t\left\langle \boldsymbol{S}, \boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top + \frac{1}{2}\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\right\rangle + \eta_t^2\underbrace{\left\|\boldsymbol{S}\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2}_{\mathfrak{S}_1} \\ &\quad - 2\eta_t\underbrace{\left\langle \boldsymbol{S}, \boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top - \boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top - \frac{1}{2}\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\right\rangle}_{\mathfrak{S}_2}, \end{aligned} \tag{B.7}$$

where in the first line, we used the fact that the update rule (3.13) is covariant with respect to $\boldsymbol{Q}_t$, implying that

$$\boldsymbol{L}_{t+1}\boldsymbol{Q}_t = \boldsymbol{L} - \eta_t\boldsymbol{S}\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}.$$

We proceed to bound $\mathfrak{S}_1$ and $\mathfrak{S}_2$. The term $\mathfrak{S}_1$ can be bounded by

$$\begin{aligned}\mathfrak{S}_1 &\le \left\|\boldsymbol{S}\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1/2}\right\|_{\mathsf{F}}^2\left\|(\boldsymbol{R}^\top\boldsymbol{R})^{-1/2}\boldsymbol{\Sigma}_\star^{1/2}\right\|^2 \\ &\le \left\|\boldsymbol{S}\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1/2}\right\|_{\mathsf{F}}^2\frac{1}{(1-\epsilon)^2},\end{aligned}$$

where the second line follows from the condition (B.4) and Lemma 17 (cf. (A.7b)):

$$\left\|(\boldsymbol{R}^\top\boldsymbol{R})^{-1/2}\boldsymbol{\Sigma}_\star^{1/2}\right\| = \left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\| \le \frac{1}{1-\epsilon}.$$

For the term $\mathfrak{S}_2$, note that

$$\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top - \boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top - \frac{1}{2}\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top = \boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\left(\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2} - \boldsymbol{V}_\star - \frac{1}{2}\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\right)^\top$$

has rank at most r. Hence we can invoke Lemma 20 (cf. (A.9b)) to obtain

$$|\mathfrak{S}_2| \le \|\boldsymbol{S}\|_{\mathsf{F},r}\left\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top - \boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top - \frac{1}{2}\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\right\|_{\mathsf{F}}$$

$$\le \|\boldsymbol{S}\|_{\mathsf{F},r}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\left(\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}-\boldsymbol{V}_\star\right\|+\frac{1}{2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\|\right)$$
$$\le L\left(\frac{\sqrt{2}\epsilon}{1-\epsilon}+\frac{\epsilon}{2}\right)\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}},$$

where the second line follows from the triangle inequality, and the third line follows from $\|\boldsymbol{S}\|_{\mathsf{F},r}\le L$ (cf. Lemma 30), (B.4), and

$$\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}-\boldsymbol{V}_\star\right\|\le\frac{\sqrt{2}\epsilon}{1-\epsilon}$$

from Lemma 17 (cf. (A.7d)).

Plugging collectively the bounds for $\mathfrak{S}_1$ and $\mathfrak{S}_2$ into (B.7) yields

$$\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2\le\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2-2\eta_t\left\langle\boldsymbol{S},\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top+\frac{1}{2}\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\right\rangle+\frac{\eta_t^2}{(1-\epsilon)^2}\left\|\boldsymbol{S}\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1/2}\right\|_{\mathsf{F}}^2$$
$$+\eta_tL\left(\frac{2\sqrt{2}\epsilon}{1-\epsilon}+\epsilon\right)\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}.$$

Similarly, we can obtain the control of $\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top}-\boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2$. Combine them together to reach

$$\mathrm{dist}^2(\boldsymbol{F}_{t+1},\boldsymbol{F}_\star)\le\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2-2\eta_t\left\langle\boldsymbol{S},\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top+\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top+\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\right\rangle$$
$$+\frac{\eta_t^2}{(1-\epsilon)^2}\left(\left\|\boldsymbol{S}\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1/2}\right\|_{\mathsf{F}}^2+\left\|\boldsymbol{S}^\top\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1/2}\right\|_{\mathsf{F}}^2\right)+\eta_tL\left(\frac{2\sqrt{2}\epsilon}{1-\epsilon}+\epsilon\right)\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right).$$

Using the subgradient optimality of $\boldsymbol{S}$, we obtain

$$\langle\boldsymbol{S},\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top+\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top+\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\rangle=\langle\boldsymbol{S},\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star\rangle\ge f(\boldsymbol{L}\boldsymbol{R}^\top)-f(\boldsymbol{X}_\star),$$

together with (B.3), which further implies that

$$\mathrm{dist}^2(\boldsymbol{F}_{t+1},\boldsymbol{F}_\star)\le\mathrm{dist}^2(\boldsymbol{F}_t,\boldsymbol{F}_\star)-2\eta_t\left(f(\boldsymbol{L}\boldsymbol{R}^\top)-f(\boldsymbol{X}_\star)\right)$$
$$+\frac{\eta_t^2}{(1-\epsilon)^2}\left(\left\|\boldsymbol{S}\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1/2}\right\|_{\mathsf{F}}^2+\left\|\boldsymbol{S}^\top\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1/2}\right\|_{\mathsf{F}}^2\right)+\eta_tL\left(\frac{4\epsilon}{1-\epsilon}+\sqrt{2}\epsilon\right)\mathrm{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star),\quad\text{(B.8)}$$

where the last term uses the basic inequality $\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} \le \sqrt{2}\,\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$.

Before proceeding to different cases of stepsize schedules, we record two useful properties. First, by the restricted μ-sharpness of $f(\cdot)$ together with Lemma 16, we have

$$f(\boldsymbol{L}\boldsymbol{R}^\top) - f(\boldsymbol{X}_\star) \ge \mu\|\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star\|_{\mathsf{F}} \ge \mu\sqrt{\sqrt{2}-1}\,\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star). \tag{B.9}$$

On the other end, by Lemma 20 (cf. (A.9c)), we have

$$\begin{aligned}\|\boldsymbol{S}\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1/2}\|_{\mathsf{F}}^2 + \|\boldsymbol{S}^\top\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1/2}\|_{\mathsf{F}}^2 &\le \|\boldsymbol{S}\|_{\mathsf{F},r}^2\left(\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1/2}\|^2 + \|\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1/2}\|^2\right)\\ &\le 2L^2,\end{aligned} \tag{B.10}$$

where the second line follows from $\|\boldsymbol{S}\|_{\mathsf{F},r} \le L$ (cf. Lemma 30) and

$$\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1/2}\|^2 = \|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\| = 1, \quad \|\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1/2}\|^2 = \|\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{L}^\top\| = 1.$$

B.1.1 Convergence with Polyak's stepsizes

Let $\eta_t = \eta_t^{\mathsf{P}}$ be the Polyak's stepsize in (3.14), which is

$$\begin{aligned}\eta_t &= \frac{f(\boldsymbol{L}_t\boldsymbol{R}_t^\top) - f(\boldsymbol{X}_\star)}{\|\boldsymbol{S}_t\boldsymbol{R}_t(\boldsymbol{R}_t^\top\boldsymbol{R}_t)^{-1/2}\|_{\mathsf{F}}^2 + \|\boldsymbol{S}_t^\top\boldsymbol{L}_t(\boldsymbol{L}_t^\top\boldsymbol{L}_t)^{-1/2}\|_{\mathsf{F}}^2}\\ &= \frac{f(\boldsymbol{L}\boldsymbol{R}^\top) - f(\boldsymbol{X}_\star)}{\|\boldsymbol{S}\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1/2}\|_{\mathsf{F}}^2 + \|\boldsymbol{S}^\top\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1/2}\|_{\mathsf{F}}^2},\end{aligned} \tag{B.11}$$

where the second line follows since $\boldsymbol{L}_t\boldsymbol{R}_t^\top = \boldsymbol{L}\boldsymbol{R}^\top$, $\boldsymbol{L}_t(\boldsymbol{L}_t^\top\boldsymbol{L}_t)^{-1}\boldsymbol{L}_t^\top = \boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{L}^\top$ and $\boldsymbol{R}_t(\boldsymbol{R}_t^\top\boldsymbol{R}_t)^{-1}\boldsymbol{R}_t^\top = \boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top$. Plugging (B.11) into (B.8), we have

$$\begin{aligned}\mathrm{dist}^2(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) &\le \mathrm{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star) - \eta_t\left(2 - \frac{1}{(1-\epsilon)^2}\right)\left(f(\boldsymbol{L}\boldsymbol{R}^\top) - f(\boldsymbol{X}_\star)\right) + \eta_t L\left(\frac{4\epsilon}{1-\epsilon} + \sqrt{2}\epsilon\right)\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)\\ &\le \mathrm{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star) - \eta_t\mu\left(\sqrt{\sqrt{2}-1}\left(2 - \frac{1}{(1-\epsilon)^2}\right) - \chi_f\left(\frac{4\epsilon}{1-\epsilon} + \sqrt{2}\epsilon\right)\right)\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star),\end{aligned} \tag{B.12}$$

where the second line follows from (B.9) and $\chi_f = L/\mu$.

To continue, combining (B.9) and (B.10), we can lower bound the Polyak's stepsize (B.11) as

$$\eta_t \geq \frac{\sqrt{\sqrt{2}-1}\mu \operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)}{2L^2}.$$

This, combined with (B.12), leads to

$$\operatorname{dist}^2(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) \leq \rho(\epsilon, \chi_f) \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star),$$

where the contraction rate $\rho(\epsilon, \chi_f)$ is

$$\rho(\epsilon, \chi_f) \coloneqq 1 - \frac{\sqrt{\sqrt{2}-1}}{2\chi_f^2}\left(\sqrt{\sqrt{2}-1}\left(2 - \frac{1}{(1-\epsilon)^2}\right) - \chi_f\left(\frac{4\epsilon}{1-\epsilon} + \sqrt{2}\epsilon\right)\right). \tag{B.13}$$

Under the condition $\epsilon = 0.02/\chi_f$, we calculate $(1-\rho(\epsilon, \chi_f))\chi_f^2$ as

$$\begin{aligned}&\frac{\sqrt{\sqrt{2}-1}}{2}\left(\sqrt{\sqrt{2}-1}\left(2 - \frac{1}{(1-\epsilon)^2}\right) - \chi_f\left(\frac{4\epsilon}{1-\epsilon} + \sqrt{2}\epsilon\right)\right)\\ &\quad\geq 0.32\left(0.64 \times \left(2 - \frac{1}{0.98^2}\right) - 0.02\left(\frac{4}{0.98} + \sqrt{2}\right)\right) \geq 0.16,\end{aligned}$$

thus $\rho(\epsilon, \chi_f) \leq 1 - 0.16/\chi_f^2$. We conclude that

$$\operatorname{dist}^2(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) \leq (1 - 0.16/\chi_f^2) \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star),$$

which is the desired claim.

B.1.2 Convergence with geometrically decaying stepsizes

Let $\eta_t = \eta_t^{\mathsf{G}}$ be the geometrically decaying stepsize in (3.15), which is

$$\eta_t = \frac{\lambda q^t}{\sqrt{\left\|\boldsymbol{S}\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1/2}\right\|_{\mathsf{F}}^2 + \left\|\boldsymbol{S}^\top\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1/2}\right\|_{\mathsf{F}}^2}}.$$

Plugging the above into (B.8), we have

$$\begin{aligned}\operatorname{dist}^2(\boldsymbol{F}_{t+1},\boldsymbol{F}_\star) &\le \operatorname{dist}^2(\boldsymbol{F}_t,\boldsymbol{F}_\star) - \eta_t\mu\left(2\sqrt{\sqrt{2}-1}-\chi_f\left(\frac{4\epsilon}{1-\epsilon}+\sqrt{2}\epsilon\right)\right)\operatorname{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star)+\frac{\lambda^2q^{2t}}{(1-\epsilon)^2}\\ &\le \operatorname{dist}^2(\boldsymbol{F}_t,\boldsymbol{F}_\star) - \frac{\lambda q^t}{\sqrt{2}\chi_f}\left(2\sqrt{\sqrt{2}-1}-\chi_f\left(\frac{4\epsilon}{1-\epsilon}+\sqrt{2}\epsilon\right)\right)\operatorname{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star)+\frac{\lambda^2q^{2t}}{(1-\epsilon)^2},\end{aligned}$$

where the first line follows from (B.9) and $\chi_f = L/\mu$, and the second line follows from $\eta_t \ge \frac{\lambda q^t}{\sqrt{2}L}$ due to (B.10). We now aim to show that

$$\operatorname{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star) \le (1-0.16/\chi_f^2)^{t/2}0.02\sigma_r(\boldsymbol{X}_\star)/\chi_f$$

in an inductive manner. Assume the above induction hypothesis holds at the t-iteration. By the setting of parameters, i.e.

$$\lambda q^t = \sqrt{\frac{\sqrt{2}-1}{2}}(1-0.16/\chi_f^2)^{t/2}0.02\sigma_r(\boldsymbol{X}_\star)/\chi_f^2,$$

we have

$$\operatorname{dist}^2(\boldsymbol{F}_{t+1},\boldsymbol{F}_\star) \le \rho(\epsilon,\chi_f)(1-0.16/\chi_f^2)^t(0.02\sigma_r(\boldsymbol{X}_\star)/\chi_f)^2,$$

where the contraction rate $\rho(\epsilon,\chi_f)$ matches exactly (B.13). Therefore, under the condition $\epsilon = 0.02/\chi_f$, we have $\rho(\epsilon,\chi_f) \le 1-0.16/\chi_f^2$, thus we conclude that

$$\operatorname{dist}(\boldsymbol{F}_{t+1},\boldsymbol{F}_\star) \le (1-0.16/\chi_f^2)^{\frac{t+1}{2}}0.02\sigma_r(\boldsymbol{X}_\star)/\chi_f,$$

which is the desired claim.

B.2 Proof of Theorem 7

We start by introducing the short-hand notation $d_t := (1-0.13/\chi_f^2)^{t/2}0.02\sigma_r(\boldsymbol{X}_\star)/\chi_f$. The parameters are set as

$$\lambda q^t = \sqrt{\frac{\sqrt{2}-1}{2}}(1-0.13/\chi_f^2)^{t/2}0.02\sigma_r(\boldsymbol{X}_\star)/\chi_f^2 = \sqrt{\frac{\sqrt{2}-1}{2}}\frac{d_t}{\chi_f}.$$

Therefore, the geometric stepsize

$$\eta_t = \frac{\lambda q^t}{\sqrt{\left\|\boldsymbol{S}\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1/2}\right\|_{\mathsf{F}}^2 + \left\|\boldsymbol{S}^\top\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1/2}\right\|_{\mathsf{F}}^2}},$$

in view of (B.10), satisfies

$$\eta_t \geq \frac{\lambda q^t}{\sqrt{2}L} = \frac{\sqrt{\sqrt{2}-1}}{2}\frac{d_t}{\chi_f^2\mu}. \tag{B.14}$$

Follow the same derivations as the proof of Theorem 6 until (B.8). Plugging the stepsize (B.14) into (B.8), together with the approximate restricted sharpness property

$$f(\boldsymbol{L}\boldsymbol{R}^\top) - f(\boldsymbol{X}_\star) \geq \mu\|\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star\|_{\mathsf{F}} - \xi \geq \sqrt{\sqrt{2}-1}\mu\,\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) - \xi,$$

we have

$$\mathrm{dist}^2(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) \leq \mathrm{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star) - \eta_t\mu\left(2\sqrt{\sqrt{2}-1} - \left(\frac{4}{1-\epsilon} + \sqrt{2}\right)\epsilon\chi_f\right)\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) + \frac{\lambda^2 q^{2t}}{(1-\epsilon)^2} + 2\eta_t\xi.$$

Under the conditions $\chi_f \geq 1$ and $\epsilon = 0.02/\chi_f \leq 0.02$, the above relation can be simplified to

$$\mathrm{dist}^2(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) \leq \mathrm{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star) - 1.177\eta_t\mu\,\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) + \frac{0.216}{\chi_f^2}d_t^2 + 2\eta_t\xi. \tag{B.15}$$

We next prove the theorem by induction, where the base case is established trivially by the initial

condition. By the induction hypothesis, the distance at the t-th iterate is bounded by

$$\operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \le \max\{d_t, 20\xi/\mu\}.$$

To obtain the control of $\operatorname{dist}(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star)$, we split the discussion in two cases.

1. If $d_t \ge 20\xi/\mu$, or equivalently, $\xi \le 0.05\mu d_t$, in view of (B.15), we have

$$\begin{aligned}
\operatorname{dist}^2(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) &\overset{\text{(i)}}{\le} d_t^2 - 1.177\eta_t\mu d_t + \frac{0.216}{\chi_f^2} d_t^2 + 0.1\eta_t\mu d_t \\
&= d_t^2 - 1.077\eta_t\mu d_t + \frac{0.216}{\chi_f^2} d_t^2 \\
&\overset{\text{(ii)}}{\le} d_t^2 - \frac{0.346}{\chi_f^2} d_t^2 + \frac{0.216}{\chi_f^2} d_t^2 \\
&= (1 - 0.13/\chi_f^2) d_t^2,
\end{aligned}$$

   where (i) uses $\xi \le 0.05\mu d_t$, and (ii) uses the condition (B.14). We conclude that $\operatorname{dist}(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) \le (1 - 0.13/\chi_f^2)^{1/2} d_t$.

2. If $0 \le d_t < 20\xi/\mu$, we have

$$\begin{aligned}
\operatorname{dist}^2(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) &\le \left(\frac{20\xi}{\mu}\right)^2 - 1.177\eta_t\mu \frac{20\xi}{\mu} + \frac{0.216}{\chi_f^2} d_t^2 + 2\eta_t\xi \\
&= \left(\frac{20\xi}{\mu}\right)^2 - 1.077\eta_t\mu \frac{20\xi}{\mu} + \frac{0.216}{\chi_f^2} d_t^2 \\
&\le \left(\frac{20\xi}{\mu}\right)^2 - \frac{1.077\sqrt{\sqrt{2}-1}}{2\chi_f^2} d_t \frac{20\xi}{\mu} + \frac{0.216}{\chi_f^2} d_t^2 \\
&\le \left(\frac{20\xi}{\mu}\right)^2 - 0.13 d_t \frac{20\xi}{\mu} \\
&\le \left(\frac{20\xi}{\mu}\right)^2,
\end{aligned}$$

   where the third line uses the condition (B.14), and the last line holds since $d_t \ge 0$.

In sum, we conclude

$$\operatorname{dist}(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) \le \max\left\{(1 - 0.13/\chi_f^2)^{\frac{t+1}{2}} 0.02\sigma_r(\boldsymbol{X}_\star)/\chi_f, 20\xi/\mu\right\},$$

which is the desired claim.

B.3 Proof of Proposition 3

For $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ where $\boldsymbol{X}_1 - \boldsymbol{X}_2$ has rank at most $2r$, we have

$$\begin{aligned} |f(\boldsymbol{X}_1) - f(\boldsymbol{X}_2)| &= \left| \|\mathcal{A}(\boldsymbol{X}_1 - \boldsymbol{X}_\star)\|_1 - \|\mathcal{A}(\boldsymbol{X}_2 - \boldsymbol{X}_\star)\|_1 \right| \\ &\le \|\mathcal{A}(\boldsymbol{X}_1 - \boldsymbol{X}_2)\|_1 \le \delta_2 \|\boldsymbol{X}_1 - \boldsymbol{X}_2\|_{\mathsf{F}}, \end{aligned}$$

where the second line follows from the inverse triangle inequality and the assumed rank-$2r$ mixed-norm RIP (cf. Definition 8) of $\mathcal{A}(\cdot)$. As a result, we have $L = \delta_2$. On the other end, we note

$$f(\boldsymbol{X}) - f(\boldsymbol{X}_\star) = \|\mathcal{A}(\boldsymbol{X} - \boldsymbol{X}_\star)\|_1 \ge \delta_1 \|\boldsymbol{X} - \boldsymbol{X}_\star\|_{\mathsf{F}},$$

where the first equality uses $f(\boldsymbol{X}_\star) = 0$ and the second inequality follows from the rank-$2r$ mixed-norm RIP; thus $\mu = \delta_1$.

B.4 Proof of Proposition 4

For $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ with $\operatorname{rank}(\boldsymbol{X}_1 - \boldsymbol{X}_2) \le 2r$, we have

$$\begin{aligned} |f(\boldsymbol{X}_1) - f(\boldsymbol{X}_2)| &= \left| \|\mathcal{A}(\boldsymbol{X}_1 - \boldsymbol{X}_\star) - \boldsymbol{w} - \boldsymbol{s}\|_1 - \|\mathcal{A}(\boldsymbol{X}_2 - \boldsymbol{X}_\star) - \boldsymbol{w} - \boldsymbol{s}\|_1 \right| \\ &\le \|\mathcal{A}(\boldsymbol{X}_1 - \boldsymbol{X}_2)\|_1 \le \delta_2 \|\boldsymbol{X}_1 - \boldsymbol{X}_2\|_{\mathsf{F}}, \end{aligned}$$

where the second line follows from the inverse triangle inequality and the rank-$2r$ mixed-norm RIP; hence $L = \delta_2$. For approximate restricted sharpness, note that

$$\begin{aligned}
f(\boldsymbol{X}) - f(\boldsymbol{X}_\star) &= \|\mathcal{A}(\boldsymbol{X} - \boldsymbol{X}_\star) - \boldsymbol{w} - \boldsymbol{s}\|_1 - \|\boldsymbol{w} + \boldsymbol{s}\|_1 \\
&\geq \|\mathcal{A}(\boldsymbol{X} - \boldsymbol{X}_\star) - \boldsymbol{s}\|_1 - \|\boldsymbol{w}\|_1 - \|\boldsymbol{s}\|_1 - \|\boldsymbol{w}\|_1 \\
&= \|\mathcal{A}_{\mathcal{S}^c}(\boldsymbol{X} - \boldsymbol{X}_\star)\|_1 + \|\mathcal{A}_{\mathcal{S}}(\boldsymbol{X} - \boldsymbol{X}_\star) - \boldsymbol{s}\|_1 - \|\boldsymbol{s}\|_1 - 2\|\boldsymbol{w}\|_1 \\
&\geq \|\mathcal{A}_{\mathcal{S}^c}(\boldsymbol{X} - \boldsymbol{X}_\star)\|_1 - \|\mathcal{A}_{\mathcal{S}}(\boldsymbol{X} - \boldsymbol{X}_\star)\|_1 - 2\|\boldsymbol{w}\|_1 \\
&\geq \delta_3 \|\boldsymbol{X} - \boldsymbol{X}_\star\|_{\mathsf{F}} - 2\|\boldsymbol{w}\|_1 \\
&\geq \delta_3 \|\boldsymbol{X} - \boldsymbol{X}_\star\|_{\mathsf{F}} - 2\sigma_w,
\end{aligned}$$

where the second and the fourth lines follow from the triangle inequality, the third line follows from the definition of $\mathcal{S}$, and the last line follows from the definition of the $\mathcal{S}$-outlier bound and the noise upper bound $\|\boldsymbol{w}\|_1 \leq \sigma_w$. Therefore, we have $\mu = \delta_3$ and $\xi = 2\sigma_w$.

Appendix C

Proofs for Low-rank Tensor Estimation

C.1 Preliminaries

This section gathers several technical lemmas that will be used later in the proof. More specifically, Section C.1.1 is devoted to understanding the scaled distance defined in the equation (4.24), and in Section C.1.2, 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.

C.1.1 Understanding the scaled distance

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

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

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 (C.1) as long as $\operatorname{dist}(\boldsymbol{F}, \boldsymbol{F}_\star)$ is sufficiently small.

Lemma 31. *Fix any factor quadruple $\boldsymbol{F} = (\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}, \boldsymbol{\mathcal{S}})$. Suppose that $\operatorname{dist}(\boldsymbol{F}, \boldsymbol{F}_\star) < \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$, then the infimum of (C.1) 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.*

Proof. This proof mimics that of Lemma 14 in Chapter 2. The high level idea is to translate the optimization problem (C.1) 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

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

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

$$\big\|\boldsymbol{U}\bar{\boldsymbol{Q}}_1 - \boldsymbol{U}_\star\big\| \vee \big\|\boldsymbol{V}\bar{\boldsymbol{Q}}_2 - \boldsymbol{V}_\star\big\| \vee \big\|\boldsymbol{W}\bar{\boldsymbol{Q}}_3 - \boldsymbol{W}_\star\big\| \vee \Big\|(\bar{\boldsymbol{Q}}_3^{-1} \otimes \bar{\boldsymbol{Q}}_2^{-1})\mathcal{M}_1(\boldsymbol{\mathcal{S}})^\top \bar{\boldsymbol{Q}}_1^{-\top}\boldsymbol{\Sigma}_{\star,1}^{-1} - \mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)^\top \boldsymbol{\Sigma}_{\star,1}^{-1}\Big\| \le \epsilon.$$

Invoke Weyl's inequality, and use the fact that $\boldsymbol{U}_\star, \boldsymbol{V}_\star, \boldsymbol{W}_\star, \mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)^\top \boldsymbol{\Sigma}_{\star,1}^{-1}$ have orthonormal columns to obtain

$$\sigma_{\min}(\boldsymbol{U}\bar{\boldsymbol{Q}}_1) \wedge \sigma_{\min}(\boldsymbol{V}\bar{\boldsymbol{Q}}_2) \wedge \sigma_{\min}(\boldsymbol{W}\bar{\boldsymbol{Q}}_3) \wedge \sigma_{\min}\Big((\bar{\boldsymbol{Q}}_3^{-1} \otimes \bar{\boldsymbol{Q}}_2^{-1})\mathcal{M}_1(\boldsymbol{\mathcal{S}})^\top \bar{\boldsymbol{Q}}_1^{-\top}\boldsymbol{\Sigma}_{\star,1}^{-1}\Big) \ge 1 - \epsilon. \quad \text{(C.2)}$$

In addition, it is straightforward to see that the minimization problem on the right hand side of (C.1) is equivalent to

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

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

$$\Big(\big\|(\boldsymbol{U}\bar{\boldsymbol{Q}}_1\boldsymbol{H}_1 - \boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1}\big\|_{\mathsf{F}}^2 + \big\|(\boldsymbol{V}\bar{\boldsymbol{Q}}_2\boldsymbol{H}_2 - \boldsymbol{V}_\star)\boldsymbol{\Sigma}_{\star,2}\big\|_{\mathsf{F}}^2 + \big\|(\boldsymbol{W}\bar{\boldsymbol{Q}}_3\boldsymbol{H}_3 - \boldsymbol{W}_\star)\boldsymbol{\Sigma}_{\star,3}\big\|_{\mathsf{F}}^2$$

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

Follow similar reasoning and invoke Weyl's inequality again to obtain

$$\begin{aligned}
&\sigma_{\max}(\boldsymbol{U}\bar{\boldsymbol{Q}}_1\boldsymbol{H}_1) \vee \sigma_{\max}(\boldsymbol{V}\bar{\boldsymbol{Q}}_2\boldsymbol{H}_2) \vee \sigma_{\max}(\boldsymbol{W}\bar{\boldsymbol{Q}}_3\boldsymbol{H}_3) \\
&\qquad \vee \sigma_{\max}\left((\boldsymbol{H}_3^{-1} \otimes \boldsymbol{H}_2^{-1})(\bar{\boldsymbol{Q}}_3^{-1} \otimes \bar{\boldsymbol{Q}}_2^{-1}) \mathcal{M}_1(\boldsymbol{\mathcal{S}})^\top \bar{\boldsymbol{Q}}_1^{-\top} \boldsymbol{H}_1^{-\top} \boldsymbol{\Sigma}_{\star,1}^{-1} \right) \leq 1 + \epsilon.
\end{aligned}$$

Use the relation $\sigma_{\min}(\boldsymbol{A})\sigma_{\max}(\boldsymbol{B}) \leq \sigma_{\max}(\boldsymbol{AB})$, combined with (C.2), to further obtain

$$\begin{aligned}
&\sigma_{\max}(\boldsymbol{H}_k) \leq \frac{1+\epsilon}{1-\epsilon}, \quad k = 1, 2, 3, \\
&\sigma_{\max}\left(\boldsymbol{\Sigma}_{\star,1} \boldsymbol{H}_1^{-\top} \boldsymbol{\Sigma}_{\star,1}^{-1} \right) \sigma_{\max}(\boldsymbol{H}_2^{-1}) \sigma_{\max}(\boldsymbol{H}_3^{-1}) \leq \frac{1+\epsilon}{1-\epsilon} \\
&\implies \sigma_{\min}\left(\boldsymbol{\Sigma}_{\star,1} \boldsymbol{H}_1 \boldsymbol{\Sigma}_{\star,1}^{-1} \right) \sigma_{\min}(\boldsymbol{H}_2) \sigma_{\min}(\boldsymbol{H}_3) \geq \frac{1-\epsilon}{1+\epsilon}.
\end{aligned}$$

As a result, the minimization problem (C.3) is equivalent to the constrained problem:

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

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

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

Lemma 32. *The optimal alignment matrices $\{\boldsymbol{Q}_k\}_{k=1,2,3}$ between $\boldsymbol{F}$ and $\boldsymbol{F}_\star$, if exist, must satisfy*

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

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

Proof. Set the gradient of the expression on the right hand side of (C.1) with respect to $\boldsymbol{Q}_1$ as zero to see

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

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

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

Lemma 33. *For any factor quadruple $\boldsymbol{F}=(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W},\boldsymbol{\mathcal{S}})$, the scaled distance (C.1) satisfies*

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

Proof. We begin by applying the mode-1 matricization (see (4.12)), and invoking Lemma 16 with $\boldsymbol{L}\coloneqq\boldsymbol{U}$, $\boldsymbol{R}\coloneqq(\boldsymbol{W}\otimes\boldsymbol{V})\mathcal{M}_1(\boldsymbol{\mathcal{S}})^\top$, $\boldsymbol{X}_\star\coloneqq\boldsymbol{U}_\star\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{W}_\star\otimes\boldsymbol{V}_\star)^\top$ to arrive at

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

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 (4.12)), and invoke Lemma 16 with $\boldsymbol{L}\coloneqq\boldsymbol{V}$, $\boldsymbol{R}\coloneqq(\boldsymbol{W}\otimes\boldsymbol{Q}_1^{-1})\mathcal{M}_2(\boldsymbol{\mathcal{S}})^\top$, $\boldsymbol{X}_\star\coloneqq\boldsymbol{V}_\star\mathcal{M}_2(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{W}_\star\otimes\boldsymbol{I}_{r_1})^\top$ to arrive at

$$\left\|(\boldsymbol{Q}_1^{-1},\boldsymbol{V},\boldsymbol{W})\cdot\boldsymbol{\mathcal{S}}-(\boldsymbol{I}_{r_1},\boldsymbol{V}_\star,\boldsymbol{W}_\star)\cdot\boldsymbol{\mathcal{S}}_\star\right\|_{\mathsf{F}}^2=\left\|\boldsymbol{V}\mathcal{M}_2(\boldsymbol{\mathcal{S}})(\boldsymbol{W}\otimes\boldsymbol{Q}_1^{-1})^\top-\boldsymbol{V}_\star\mathcal{M}_2(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{W}_\star\otimes\boldsymbol{I}_{r_1})^\top\right\|_{\mathsf{F}}^2$$

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

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

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

Finally, combine these results to conclude

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

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

C.1.2 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:

$$
\begin{aligned}
&\boldsymbol{\Delta}_U \coloneqq \boldsymbol{U} - \boldsymbol{U}_\star, \qquad \boldsymbol{\Delta}_V \coloneqq \boldsymbol{V} - \boldsymbol{V}_\star, \qquad \boldsymbol{\Delta}_W \coloneqq \boldsymbol{W} - \boldsymbol{W}_\star, \qquad \boldsymbol{\Delta}_{\mathcal{S}} \coloneqq \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{S}}_\star, \\
&\breve{\boldsymbol{U}} \coloneqq (\boldsymbol{W}\otimes\boldsymbol{V})\mathcal{M}_1(\boldsymbol{\mathcal{S}})^\top, \qquad \breve{\boldsymbol{V}} \coloneqq (\boldsymbol{W}\otimes\boldsymbol{U})\mathcal{M}_2(\boldsymbol{\mathcal{S}})^\top, \qquad \breve{\boldsymbol{W}} \coloneqq (\boldsymbol{V}\otimes\boldsymbol{U})\mathcal{M}_3(\boldsymbol{\mathcal{S}})^\top, \\
&\breve{\boldsymbol{U}}_\star \coloneqq (\boldsymbol{W}_\star\otimes\boldsymbol{V}_\star)\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)^\top, \qquad \breve{\boldsymbol{V}}_\star \coloneqq (\boldsymbol{W}_\star\otimes\boldsymbol{U}_\star)\mathcal{M}_2(\boldsymbol{\mathcal{S}}_\star)^\top, \qquad \breve{\boldsymbol{W}}_\star \coloneqq (\boldsymbol{V}_\star\otimes\boldsymbol{U}_\star)\mathcal{M}_3(\boldsymbol{\mathcal{S}}_\star)^\top,
\end{aligned} \tag{C.4}
$$

$$
\boldsymbol{\mathcal{T}}_U \coloneqq (\boldsymbol{U}_\star^\top \boldsymbol{\Delta}_U, \boldsymbol{I}_{r_2}, \boldsymbol{I}_{r_3})\cdot\boldsymbol{\mathcal{S}}_\star, \qquad \boldsymbol{\mathcal{T}}_V \coloneqq (\boldsymbol{I}_{r_1}, \boldsymbol{V}_\star^\top \boldsymbol{\Delta}_V, \boldsymbol{I}_{r_3})\cdot\boldsymbol{\mathcal{S}}_\star, \qquad \boldsymbol{\mathcal{T}}_W \coloneqq (\boldsymbol{I}_{r_1}, \boldsymbol{I}_{r_2}, \boldsymbol{W}_\star^\top \boldsymbol{\Delta}_W)\cdot\boldsymbol{\mathcal{S}}_\star,
$$

$$\boldsymbol{D}_U \coloneqq (\boldsymbol{U}^\top \boldsymbol{U})^{-1/2} \boldsymbol{U}^\top \boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}, \qquad \boldsymbol{D}_V \coloneqq (\boldsymbol{V}^\top \boldsymbol{V})^{-1/2} \boldsymbol{V}^\top \boldsymbol{\Delta}_V \boldsymbol{\Sigma}_{\star,2}, \quad \boldsymbol{D}_W \coloneqq (\boldsymbol{W}^\top \boldsymbol{W})^{-1/2} \boldsymbol{W}^\top \boldsymbol{\Delta}_W \boldsymbol{\Sigma}_{\star,3}.$$

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

Lemma 34. *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:*

$$\|\boldsymbol{\Delta}_U\| \vee \|\boldsymbol{\Delta}_V\| \vee \|\boldsymbol{\Delta}_W\| \vee \|\mathcal{M}_k(\boldsymbol{\Delta}_{\mathcal{S}})^\top \boldsymbol{\Sigma}_{\star,k}^{-1}\| \le \epsilon, \qquad k = 1, 2, 3; \tag{C.5a}$$

$$\|\boldsymbol{U}(\boldsymbol{U}^\top \boldsymbol{U})^{-1}\| \le (1-\epsilon)^{-1}; \tag{C.5b}$$

$$\left\|\boldsymbol{U}(\boldsymbol{U}^\top \boldsymbol{U})^{-1} - \boldsymbol{U}_\star\right\| \le \frac{\sqrt{2}\epsilon}{1-\epsilon}; \tag{C.5c}$$

$$\left\|(\boldsymbol{U}^\top \boldsymbol{U})^{-1}\right\| \le (1-\epsilon)^{-2}; \tag{C.5d}$$

$$\left\|(\breve{\boldsymbol{U}} - \breve{\boldsymbol{U}}_\star)\boldsymbol{\Sigma}_{\star,1}^{-1}\right\| \le 3\epsilon + 3\epsilon^2 + \epsilon^3; \tag{C.5e}$$

$$\left\|\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}\right\| \le (1-\epsilon)^{-3}; \tag{C.5f}$$

$$\left\|\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1} - \breve{\boldsymbol{U}}_\star \boldsymbol{\Sigma}_{\star,1}^{-1}\right\| \le \frac{\sqrt{2}(3\epsilon + 3\epsilon^2 + \epsilon^3)}{(1-\epsilon)^3}; \tag{C.5g}$$

$$\left\|\boldsymbol{\Sigma}_{\star,1}(\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}\right\| \le (1-\epsilon)^{-6}; \tag{C.5h}$$

$$\left\|\boldsymbol{\Sigma}_{\star,1}(\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1}\mathcal{M}_1(\boldsymbol{\mathcal{S}})\right\| \le (1-\epsilon)^{-5}. \tag{C.5i}$$

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:

$$\|(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \cdot \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \le (1 + \frac{3}{2}\epsilon + \epsilon^2 + \frac{\epsilon^3}{4}) \left(\|\boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_V \boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_W \boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\right); \tag{C.6a}$$

$$\|(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \cdot \boldsymbol{\mathcal{S}}_\star - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \le (1 + \epsilon + \frac{\epsilon^2}{3}) \left(\|\boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_V \boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_W \boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}\right); \tag{C.6b}$$

$$\left\|\breve{\boldsymbol{U}} - \breve{\boldsymbol{U}}_\star\right\|_{\mathsf{F}} \le (1 + \epsilon + \frac{\epsilon^2}{3}) \left(\|\boldsymbol{\Delta}_V \boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_W \boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\right). \tag{C.6c}$$

As a straightforward consequence of (C.6a)*, the following important relation holds when* $\epsilon \le 0.2$*:*

$$\|(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\cdot\boldsymbol{\mathcal{S}}-\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \le 2(1+\frac{3}{2}\epsilon+\epsilon^2+\frac{\epsilon^3}{4})\,\mathrm{dist}(\boldsymbol{F},\boldsymbol{F}_\star) \le 3\,\mathrm{dist}(\boldsymbol{F},\boldsymbol{F}_\star). \tag{C.7}$$

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

Proof of spectral norm perturbation bounds. To begin, recalling the notation in (C.4), (C.5a) follows directly from the definition

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

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

For (C.5b), 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 (C.5c), 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

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

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 (C.5d), 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 (C.5b).

For (C.5e), we first expand the expression as

$$\begin{aligned}
\breve{\boldsymbol{U}} - \breve{\boldsymbol{U}}_\star &= (\boldsymbol{W}\otimes\boldsymbol{V})\mathcal{M}_1(\boldsymbol{\mathcal{S}})^\top - (\boldsymbol{W}_\star\otimes\boldsymbol{V}_\star)\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)^\top \\
&= (\boldsymbol{W}\otimes\boldsymbol{V} - \boldsymbol{W}_\star\otimes\boldsymbol{V}_\star)\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)^\top + (\boldsymbol{W}\otimes\boldsymbol{V})\mathcal{M}_1(\boldsymbol{\mathcal{S}})^\top - (\boldsymbol{W}\otimes\boldsymbol{V})\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)^\top \\
&= (\boldsymbol{W}\otimes\boldsymbol{\Delta}_V + \boldsymbol{\Delta}_W\otimes\boldsymbol{V}_\star)\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)^\top + (\boldsymbol{W}\otimes\boldsymbol{V})\mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})^\top. \qquad \text{(C.8)}
\end{aligned}$$

Apply the triangle inequality to obtain

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

where we have used (C.5a) and the fact $\|\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)^\top\boldsymbol{\Sigma}_{\star,1}^{-1}\| = 1$ (see (4.13)) in the last line.

(C.5f) follows from combining

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

With regard to (C.5g), repeat the same proof as (C.5c), 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

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

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 (C.5h) follows from (C.5f) and the relation:

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

With regard to (C.5i), we have

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

where the first line follows from

$$\breve{\boldsymbol{U}}^\top=\mathcal{M}_1(\boldsymbol{\mathcal{S}})(\boldsymbol{W}\otimes\boldsymbol{V})^\top\qquad\Longrightarrow\qquad\mathcal{M}_1(\boldsymbol{\mathcal{S}})=\breve{\boldsymbol{U}}^\top\left(\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\otimes\boldsymbol{V}(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\right),\quad\text{(C.9)}$$

and the last inequality uses (C.5c) and (C.5f).

Proof of Frobenius norm perturbation bounds. We proceed to prove the perturbation bounds regarding the Frobenius norm. For (C.6a), we begin with the following decomposition

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

$$= (\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \cdot \boldsymbol{\Delta}_{\mathcal{S}} + (\boldsymbol{\Delta}_U, \boldsymbol{V}, \boldsymbol{W}) \cdot \boldsymbol{\mathcal{S}}_\star + (\boldsymbol{U}_\star, \boldsymbol{\Delta}_V, \boldsymbol{W}) \cdot \boldsymbol{\mathcal{S}}_\star + (\boldsymbol{U}_\star, \boldsymbol{V}_\star, \boldsymbol{\Delta}_W) \cdot \boldsymbol{\mathcal{S}}_\star. \tag{C.10}$$

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

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

where the second inequality follows from (4.6e), and the last inequality follows from (4.13) and (C.5a). By symmetry, one can permute the occurrence of $\boldsymbol{\Delta}_U, \boldsymbol{\Delta}_V, \boldsymbol{\Delta}_W, \boldsymbol{\Delta}_{\mathcal{S}}$ in the decomposition (C.10). 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

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

To complete the proof of (C.6a), 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 (C.6b) can be proved in a similar fashion; for the sake of brevity, we omit its proof.

Turning to (C.6c), apply the triangle inequality to (C.8) to obtain

$$\|\breve{\boldsymbol{U}}-\breve{\boldsymbol{U}}_\star\|_{\mathsf{F}}\le\Big\|(\boldsymbol{W}\otimes\boldsymbol{\Delta}_V)\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)^\top\Big\|_{\mathsf{F}}+\Big\|(\boldsymbol{\Delta}_W\otimes\boldsymbol{V}_\star)\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)^\top\Big\|_{\mathsf{F}}+\|(\boldsymbol{W}\otimes\boldsymbol{V})\mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})\|_{\mathsf{F}}\,. \tag{C.11}$$

To bound the first term, change the mode of matricization (see (4.12)) to arrive at

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

where the last inequality uses (C.5a). Similarly, the last two terms in (C.11) can be bounded as

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

Plugging the above bounds back to (C.11), 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 (C.8), we arrive at the final advertised bound (C.6c).

C.2 Proof for Tensor Factorization (Theorem 10)

We prove Theorem 10 via induction. Suppose that for some $t\ge0$, 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 $\operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \le \epsilon \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$, Lemma 31 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 $\operatorname{dist}(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star)$, one has

$$\begin{aligned}\operatorname{dist}^2(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) \le \|(\boldsymbol{U}_{t+1}\boldsymbol{Q}_{t,1} - \boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 + \|(\boldsymbol{V}_{t+1}\boldsymbol{Q}_{t,2} - \boldsymbol{V}_\star)\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}^2 + \|(\boldsymbol{W}_{t+1}\boldsymbol{Q}_{t,3} - \boldsymbol{W}_\star)\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}^2 \\ + \left\|(\boldsymbol{Q}_{t,1}^{-1}, \boldsymbol{Q}_{t,2}^{-1}, \boldsymbol{Q}_{t,3}^{-1}) \cdot \boldsymbol{\mathcal{S}}_{t+1} - \boldsymbol{\mathcal{S}}_\star\right\|_{\mathsf{F}}^2. \qquad \text{(C.12)}\end{aligned}$$

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

$$\boldsymbol{U} \coloneqq \boldsymbol{U}_t\boldsymbol{Q}_{t,1}, \qquad \boldsymbol{V} \coloneqq \boldsymbol{V}_t\boldsymbol{Q}_{t,2}, \qquad \boldsymbol{W} \coloneqq \boldsymbol{W}_t\boldsymbol{Q}_{t,3}, \qquad \boldsymbol{\mathcal{S}} \coloneqq (\boldsymbol{Q}_{t,1}^{-1}, \boldsymbol{Q}_{t,2}^{-1}, \boldsymbol{Q}_{t,3}^{-1}) \cdot \boldsymbol{\mathcal{S}}_t, \qquad \text{(C.13)}$$

and adopt the set of notation defined in (C.4) for the rest of the proof. Clearly, $\boldsymbol{F}$ is aligned with $\boldsymbol{F}_\star$. With these notation, we can rephrase the consequences of Lemma 32 as:

$$\begin{aligned}\boldsymbol{U}^\top \boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}^2 &= \mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})\mathcal{M}_1(\boldsymbol{\mathcal{S}})^\top, \\ \boldsymbol{V}^\top \boldsymbol{\Delta}_V \boldsymbol{\Sigma}_{\star,2}^2 &= \mathcal{M}_2(\boldsymbol{\Delta}_{\mathcal{S}})\mathcal{M}_2(\boldsymbol{\mathcal{S}})^\top, \\ \boldsymbol{W}^\top \boldsymbol{\Delta}_W \boldsymbol{\Sigma}_{\star,3}^2 &= \mathcal{M}_3(\boldsymbol{\Delta}_{\mathcal{S}})\mathcal{M}_3(\boldsymbol{\mathcal{S}})^\top.\end{aligned} \qquad \text{(C.14)}$$

We aim to establish the following bounds for the four terms in (C.12) as long as $\eta < 1$:

$$\begin{aligned}\|(\boldsymbol{U}_{t+1}\boldsymbol{Q}_{t,1} - \boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 \le (1-\eta)^2 \|\boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 \\ - 2\eta(1-\eta)\left\langle \boldsymbol{\mathcal{T}}_U, \boldsymbol{\mathcal{T}}_U + \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W \right\rangle + \eta^2 \left\|\boldsymbol{\mathcal{T}}_U + \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W\right\|_{\mathsf{F}}^2 \\ + 2\eta(1-\eta)C_1\epsilon \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star) + \eta^2 C_2 \epsilon \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star); \qquad \text{(C.15a)} \\ \|(\boldsymbol{V}_{t+1}\boldsymbol{Q}_{t,2} - \boldsymbol{V}_\star)\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}^2 \le (1-\eta)^2 \|\boldsymbol{\Delta}_V \boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}^2 \\ - 2\eta(1-\eta)\left\langle \boldsymbol{\mathcal{T}}_V, \boldsymbol{\mathcal{T}}_U + \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W \right\rangle + \eta^2 \left\|\boldsymbol{\mathcal{T}}_U + \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W\right\|_{\mathsf{F}}^2 \\ + 2\eta(1-\eta)C_1\epsilon \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star) + \eta^2 C_2 \epsilon \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star); \qquad \text{(C.15b)} \\ \|(\boldsymbol{W}_{t+1}\boldsymbol{Q}_{t,3} - \boldsymbol{W}_\star)\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}^2 \le (1-\eta)^2 \|\boldsymbol{\Delta}_W \boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}^2 \\ - 2\eta(1-\eta)\left\langle \boldsymbol{\mathcal{T}}_W, \boldsymbol{\mathcal{T}}_U + \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W \right\rangle + \eta^2 \left\|\boldsymbol{\mathcal{T}}_U + \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W\right\|_{\mathsf{F}}^2\end{aligned}$$

$$+ 2\eta(1-\eta)C_1\epsilon \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star) + \eta^2 C_2 \epsilon \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star); \quad \text{(C.15c)}$$

$$\left\|(\boldsymbol{Q}_{t,1}^{-1}, \boldsymbol{Q}_{t,2}^{-1}, \boldsymbol{Q}_{t,3}^{-1}) \cdot \boldsymbol{\mathcal{S}}_{t+1} - \boldsymbol{\mathcal{S}}_\star\right\|_{\mathsf{F}}^2 \le (1-\eta)^2 \|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}^2 - \eta(2-5\eta)\left(\|\boldsymbol{D}_U\|_{\mathsf{F}}^2 + \|\boldsymbol{D}_V\|_{\mathsf{F}}^2 + \|\boldsymbol{D}_W\|_{\mathsf{F}}^2\right)$$
$$+ 2\eta(1-\eta)C_1\epsilon \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star) + \eta^2 C_2 \epsilon \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star), \quad \text{(C.15d)}$$

where $C_1, C_2 > 1$ are two universal constants. Suppose for the moment that the four bounds (C.15) hold. We can then combine them all to deduce

$$\begin{aligned}\operatorname{dist}^2(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) \le (1-\eta)^2\Big(&\|\boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_V \boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_W \boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}^2\Big) \\ &- \eta(2-5\eta)\,\|\boldsymbol{\mathcal{T}}_U + \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W\|_{\mathsf{F}}^2 - \eta(2-5\eta)\Big(\|\boldsymbol{D}_U\|_{\mathsf{F}}^2 + \|\boldsymbol{D}_V\|_{\mathsf{F}}^2 + \|\boldsymbol{D}_W\|_{\mathsf{F}}^2\Big) \\ &+ 2\eta(1-\eta)C\epsilon \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star) + \eta^2 C\epsilon \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star). \end{aligned} \quad \text{(C.16)}$$

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 (C.7) in Lemma 34 guarantees that $\|(\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t) \cdot \boldsymbol{\mathcal{S}}_t - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \le 3 \operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$.

It then boils down to demonstrating the four bounds (C.15). Due to the symmetry among $\boldsymbol{U}, \boldsymbol{V}$ and $\boldsymbol{W}$, we will focus on proving the bounds (C.15a) and (C.15d), omitting the proofs for the other two.

Proof of (C.15a). Utilize the `ScaledGD` update rule (4.26) to write

$$\begin{aligned}(\boldsymbol{U}_{t+1}\boldsymbol{Q}_{t,1} - \boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1} &= \Big(\boldsymbol{U} - \eta \mathcal{M}_1\left((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \cdot \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_\star\right) \breve{\boldsymbol{U}} (\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1} - \boldsymbol{U}_\star\Big) \boldsymbol{\Sigma}_{\star,1} \\ &= (1-\eta)\boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1} - \eta \boldsymbol{U}_\star (\breve{\boldsymbol{U}} - \breve{\boldsymbol{U}}_\star)^\top \breve{\boldsymbol{U}} (\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1} \boldsymbol{\Sigma}_{\star,1}, \end{aligned} \quad \text{(C.17)}$$

where we use the decomposition of the mode-1 matricization

$$\mathcal{M}_1\left((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \cdot \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_\star\right) = \boldsymbol{U}\mathcal{M}_1(\boldsymbol{\mathcal{S}})(\boldsymbol{W} \otimes \boldsymbol{V})^\top - \boldsymbol{U}_\star \mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{W}_\star \otimes \boldsymbol{V}_\star)^\top$$

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

Take the squared norm of both sides of the identity (C.17) to obtain

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

The following two claims bound the two terms $\mathfrak{U}_1$ and $\mathfrak{U}_2$, whose proofs can be found in Appendix C.2.1 and Appendix C.2.2, respectively.

Claim 8. $\mathfrak{U}_1 \geq \langle \boldsymbol{\mathcal{T}}_U, \boldsymbol{\mathcal{T}}_U + \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W \rangle - C_1 \epsilon \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star)$.

Claim 9. $\mathfrak{U}_2 \leq \|\boldsymbol{\mathcal{T}}_U + \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W\|_{\mathsf{F}}^2 + C_2 \epsilon \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star)$.

We can combine the above two claims to obtain that

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

as long as $\eta < 1$. This proves the bound (C.15a).

Proof of (C.15d). Again, we use the `ScaledGD` update rule (4.26) and the decomposition $\boldsymbol{\mathcal{S}} = \boldsymbol{\Delta}_{\mathcal{S}} + \boldsymbol{\mathcal{S}}_\star$ to obtain

$$\begin{aligned}
&(\boldsymbol{Q}_{t,1}^{-1}, \boldsymbol{Q}_{t,2}^{-1}, \boldsymbol{Q}_{t,3}^{-1}) \cdot \boldsymbol{\mathcal{S}}_{t+1} - \boldsymbol{\mathcal{S}}_\star \\
&\quad = \boldsymbol{\mathcal{S}} - \eta \left((\boldsymbol{U}^\top \boldsymbol{U})^{-1} \boldsymbol{U}^\top, (\boldsymbol{V}^\top \boldsymbol{V})^{-1} \boldsymbol{V}^\top, (\boldsymbol{W}^\top \boldsymbol{W})^{-1} \boldsymbol{W}^\top \right) \cdot \left((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \cdot \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{X}}_\star \right) - \boldsymbol{\mathcal{S}}_\star \\
&\quad = (1-\eta) \boldsymbol{\Delta}_{\mathcal{S}} - \eta \left((\boldsymbol{U}^\top \boldsymbol{U})^{-1} \boldsymbol{U}^\top, (\boldsymbol{V}^\top \boldsymbol{V})^{-1} \boldsymbol{V}^\top, (\boldsymbol{W}^\top \boldsymbol{W})^{-1} \boldsymbol{W}^\top \right) \cdot \left((\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \cdot \boldsymbol{\mathcal{S}}_\star - \boldsymbol{\mathcal{X}}_\star \right),
\end{aligned} \tag{C.18}$$

where we used (4.6c) in the last line. Expand the squared norm of both sides to reach

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

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 C.2.3 and Appendix C.2.4, respectively.

Claim 10. $\mathfrak{S}_1 \geq \|\boldsymbol{D}_U\|_{\mathsf{F}}^2+\|\boldsymbol{D}_V\|_{\mathsf{F}}^2+\|\boldsymbol{D}_W\|_{\mathsf{F}}^2-C_1\epsilon\,\mathrm{dist}^2(\boldsymbol{F}_t,\boldsymbol{F}_\star)$.

Claim 11. $\mathfrak{S}_2 \leq 3\left(\|\boldsymbol{D}_U\|_{\mathsf{F}}^2+\|\boldsymbol{D}_V\|_{\mathsf{F}}^2+\|\boldsymbol{D}_W\|_{\mathsf{F}}^2\right)+C_2\epsilon\,\mathrm{dist}^2(\boldsymbol{F}_t,\boldsymbol{F}_\star)$.

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

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

as long as $\eta<1$. This recovers the bound (C.15d).

C.2.1 Proof of Claim 8

Use the relation (C.8) to decompose $\mathfrak{U}_1$ as

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

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

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

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

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 (C.4) and the relation (4.12), the main term $\mathfrak{U}_{1,1}^{\mathrm{m}}$ can be rewritten as an inner product in the tensor space:

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

To control the other three perturbation terms, Lemma 34 turns out to be extremely useful. For instance, the perturbation term $\mathfrak{U}_{1,1}^{\mathrm{p},1}$ is bounded by

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

Here in the last inequality, we used the upper bound (C.5g) and changed the matricization mode to obtain

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

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

$$
\begin{aligned}
\mathfrak{U}_{1,2} &= \Big\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}\Big\rangle\\
&\quad+\underbrace{\Big\langle \boldsymbol{U}_\star^\top\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}, \mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})(\boldsymbol{W}_\star\otimes\boldsymbol{V}_\star)^\top\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)\Big\rangle}_{=:\mathfrak{U}_{1,2}^{\mathrm{p},1}}\\
&\quad+\underbrace{\Big\langle \boldsymbol{U}_\star^\top\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}, \mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})(\boldsymbol{W}\otimes\boldsymbol{V}-\boldsymbol{W}_\star\otimes\boldsymbol{V}_\star)^\top\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}\Big\rangle}_{=:\mathfrak{U}_{1,2}^{\mathrm{p},2}}\\
&= \Big\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}\Big\rangle\underbrace{-\Big\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}\Big\rangle}_{=:\mathfrak{U}_{1,2}^{\mathrm{p},3}}+\mathfrak{U}_{1,2}^{\mathrm{p},1}+\mathfrak{U}_{1,2}^{\mathrm{p},2}\\
&= \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{U}_{1,2}^{\mathrm{p},1}+\mathfrak{U}_{1,2}^{\mathrm{p},2}+\mathfrak{U}_{1,2}^{\mathrm{p},3}\\
&= \|\boldsymbol{\mathcal{T}}_U\|_{\mathsf{F}}^2+\mathfrak{U}_{1,2}^{\mathrm{p},1}+\mathfrak{U}_{1,2}^{\mathrm{p},2}+\mathfrak{U}_{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{U}_{1,2}^{\mathrm{p},4}},
\end{aligned}
$$

where in the penultimate identity we have applied the identity (C.14) to replace $\mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})\mathcal{M}_1(\boldsymbol{\mathcal{S}})^\top$. Again, by Lemma 34, the perturbation term $\mathfrak{U}_{1,2}^{\mathrm{p},1}$ is bounded by

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

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

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

where we have used

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

Following similar arguments (i.e. repeatedly using Lemma 34), we can bound $\mathfrak{U}_{1,2}^{\mathrm{p},3}$ and $\mathfrak{U}_{1,2}^{\mathrm{p},4}$ as

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

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

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

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

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

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

C.2.2 Proof of Claim 9

Note that

$$\begin{aligned}\mathfrak{U}_2 &= \left\|(\breve{\boldsymbol{U}}-\breve{\boldsymbol{U}}_\star)^\top\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}\right\|_{\mathsf{F}}^2 \\ &\le \left\|(\breve{\boldsymbol{U}}-\breve{\boldsymbol{U}}_\star)^\top\breve{\boldsymbol{U}}\boldsymbol{\Sigma}_{\star,1}^{-1}\right\|_{\mathsf{F}}^2\left\|\boldsymbol{\Sigma}_{\star,1}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}\right\|^2 \\ &\le \left\|(\breve{\boldsymbol{U}}-\breve{\boldsymbol{U}}_\star)^\top\breve{\boldsymbol{U}}\boldsymbol{\Sigma}_{\star,1}^{-1}\right\|_{\mathsf{F}}^2(1-\epsilon)^{-12}, \qquad \text{(C.19)}\end{aligned}$$

where the last relation arises from the bound (C.5h) in Lemma 34. We can then use the decomposition (C.8) to obtain

$$\begin{aligned}
\|(\breve{\boldsymbol{U}}-\breve{\boldsymbol{U}}_\star)^\top\breve{\boldsymbol{U}}\boldsymbol{\Sigma}_{\star,1}^{-1}\|_{\mathsf{F}} &= \left\|\Big(\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{W}\otimes\boldsymbol{\Delta}_V+\boldsymbol{\Delta}_W\otimes\boldsymbol{V}_\star)^\top+\mathcal{M}_1(\boldsymbol{\Delta}_{\mathcal{S}})(\boldsymbol{W}\otimes\boldsymbol{V})^\top\Big)(\boldsymbol{W}\otimes\boldsymbol{V})\mathcal{M}_1(\boldsymbol{\mathcal{S}})^\top\boldsymbol{\Sigma}_{\star,1}^{-1}\right\|_{\mathsf{F}} \\
&\le \underbrace{\left\|\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)\left(\boldsymbol{I}_{r_3}\otimes\boldsymbol{\Delta}_V^\top\boldsymbol{V}_\star+\boldsymbol{\Delta}_W^\top\boldsymbol{W}_\star\otimes\boldsymbol{I}_{r_2}\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{W}^\top\boldsymbol{W}\otimes\boldsymbol{\Delta}_V^\top\boldsymbol{V}-\boldsymbol{I}_{r_3}\otimes\boldsymbol{\Delta}_V^\top\boldsymbol{V}_\star\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{\Delta}_W^\top\boldsymbol{W}\otimes\boldsymbol{V}_\star^\top\boldsymbol{V}-\boldsymbol{\Delta}_W^\top\boldsymbol{W}_\star\otimes\boldsymbol{I}_{r_2}\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{W}^\top\boldsymbol{W}\otimes\boldsymbol{\Delta}_V^\top\boldsymbol{V}+\boldsymbol{\Delta}_W^\top\boldsymbol{W}\otimes\boldsymbol{V}_\star^\top\boldsymbol{V}\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{W}^\top\boldsymbol{W}\otimes\boldsymbol{V}^\top\boldsymbol{V}-\boldsymbol{I}_{r_3}\otimes\boldsymbol{I}_{r_2}\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 (C.14) 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

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

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

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

They follow from similar calculations as those in bounding $\mathfrak{U}_1$ with the aid of Lemma 34; hence we omit the details for brevity. Combine these results to see

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

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

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

Next take the square to obtain

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

Finally plug this back into (C.19) to conclude

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

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

C.2.3 Proof of Claim 10

Use the decomposition

$$(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\cdot\boldsymbol{\mathcal{S}}_\star-\boldsymbol{\mathcal{X}}_\star=(\boldsymbol{\Delta}_U,\boldsymbol{V},\boldsymbol{W})\cdot\boldsymbol{\mathcal{S}}_\star+(\boldsymbol{U}_\star,\boldsymbol{\Delta}_V,\boldsymbol{W})\cdot\boldsymbol{\mathcal{S}}_\star+(\boldsymbol{U}_\star,\boldsymbol{V}_\star,\boldsymbol{\Delta}_W)\cdot\boldsymbol{\mathcal{S}}_\star \tag{C.20}$$

to rewrite $\mathfrak{S}_1$ as

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

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

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

Again, the identity (C.14) is helpful in characterizing the main term $\mathfrak{S}_{1,1}^{\mathrm{m}}$:

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

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

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

which follows directly from Lemma 34.

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

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

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

In view of the relation (C.14), we can rewrite the main term $\mathfrak{S}_{1,2}^{\mathrm{m}}$ as

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

In addition, for the perturbation terms, Lemma 34 allows us to obtain

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

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

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

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

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

The main term obeys (thanks again to the identity (C.14))

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

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

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

where we use the decomposition

$$\boldsymbol{V}_\star^\top\boldsymbol{V}(\boldsymbol{V}^\top\boldsymbol{V})^{-1} \otimes \boldsymbol{U}_\star^\top\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1} - \boldsymbol{I}_{r_2}\otimes\boldsymbol{I}_{r_1} = (\boldsymbol{V}_\star\otimes\boldsymbol{U}_\star - \boldsymbol{V}\otimes\boldsymbol{U})^\top \left(\boldsymbol{V}(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\otimes\boldsymbol{U}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\right)$$

and its immediate consequence

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

Step 4: putting all pieces together. 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 aggregated perturbation term $\mathfrak{S}_1^{\mathrm{p}}$ obeys

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

It is straightforward to check that $|\mathfrak{S}_1^{\mathrm{p}}| \le C_1\epsilon\,\mathrm{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star)$ for some absolute constant $C_1 > 1$.

C.2.4 Proof of Claim 11

Reuse the decomposition (C.20) 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 34 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

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

and

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

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

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

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

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

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.3 Proof for Tensor Completion

This section is devoted to the proofs of claims related to tensor completion. To begin with, we state several bounds regarding the $\ell_{2,\infty}$ norm that will be repeatedly used throughout this section.

Lemma 35. *Suppose that $\boldsymbol{\mathcal{X}}_\star$ is μ-incoherent, and that $\boldsymbol{F} = (\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}, \boldsymbol{\mathcal{S}})$ satisfies $\mathrm{dist}(\boldsymbol{F}, \boldsymbol{F}_\star) \le \epsilon\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$ for $\epsilon < 1$ and the incoherence condition (4.29). Then one has the following bounds regarding the $\ell_{2,\infty}$ norm:*

$$\sqrt{n_1}\|\boldsymbol{U}\mathcal{M}_1(\boldsymbol{\mathcal{S}})\|_{2,\infty} \le (1-\epsilon)^{-2}C_B\sqrt{\mu r}\sigma_{\max}(\boldsymbol{\mathcal{X}}_\star); \tag{C.21a}$$

$$\sqrt{n_1}\|\boldsymbol{U}\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)\|_{2,\infty} = \sqrt{n_1}\|\boldsymbol{U}\boldsymbol{\Sigma}_{\star,1}\|_{2,\infty} \le (1-\epsilon)^{-3}C_B\sqrt{\mu r}\sigma_{\max}(\boldsymbol{\mathcal{X}}_\star); \tag{C.21b}$$

$$\sqrt{n_1}\|\boldsymbol{U}\|_{2,\infty} \le (1-\epsilon)^{-3}C_B\kappa\sqrt{\mu r}. \tag{C.21c}$$

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

Proof. For (C.21a), we have

$$\begin{aligned}\|\boldsymbol{U}\mathcal{M}_1(\boldsymbol{\mathcal{S}})\|_{2,\infty} &= \big\|\boldsymbol{U}\breve{\boldsymbol{U}}^\top\big(\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\otimes\boldsymbol{V}(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\big)\big\|_{2,\infty}\\ &\le \|\boldsymbol{U}\breve{\boldsymbol{U}}^\top\|_{2,\infty}\Big\|\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\Big\|\,\Big\|\boldsymbol{V}(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\Big\|\\ &\le \|\boldsymbol{U}\breve{\boldsymbol{U}}^\top\|_{2,\infty}(1-\epsilon)^{-2},\end{aligned}$$

where the first line uses (C.9), the second line follows from $\|\boldsymbol{AB}\|_{2,\infty}\le\|\boldsymbol{A}\|_{2,\infty}\|\boldsymbol{B}\|$, and the last inequality uses (C.5c). This combined with condition (4.29) leads to the declared bound.

Similarly for (C.21b), we have

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

where the last line follows from (C.5f).

Finally, observe that

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

Combining the above inequality with (C.21b), we reach the bound (C.21c). □

C.3.1 Proof of Lemma 10

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 (4.19). For the purpose of

understanding, it is instructive to view $\boldsymbol{F}$ as the solution to the following optimization problems:

$$\begin{aligned}
\boldsymbol{U} &= \operatorname*{argmin}_{\boldsymbol{U}} \left\|(\boldsymbol{U}-\boldsymbol{U}_+)\check{\boldsymbol{U}}_+^\top\right\|_{\mathsf{F}}^2 \qquad \text{s.t.} \quad \sqrt{n_1}\|\boldsymbol{U}\check{\boldsymbol{U}}_+^\top\|_{2,\infty} \le B, \\
\boldsymbol{V} &= \operatorname*{argmin}_{\boldsymbol{V}} \left\|(\boldsymbol{V}-\boldsymbol{V}_+)\check{\boldsymbol{V}}_+^\top\right\|_{\mathsf{F}}^2 \qquad \text{s.t.} \quad \sqrt{n_2}\|\boldsymbol{V}\check{\boldsymbol{V}}_+^\top\|_{2,\infty} \le B, \\
\boldsymbol{W} &= \operatorname*{argmin}_{\boldsymbol{W}} \left\|(\boldsymbol{W}-\boldsymbol{W}_+)\check{\boldsymbol{W}}_+^\top\right\|_{\mathsf{F}}^2 \qquad \text{s.t.} \quad \sqrt{n_3}\|\boldsymbol{W}\check{\boldsymbol{W}}_+^\top\|_{2,\infty} \le B.
\end{aligned} \tag{C.22}$$

The remaining proof follows similar arguments as Chapter 2. To begin, we collect a useful claim as follows.

Claim 12 (Claim 5 in Chapter 2)**.** *For vectors $\boldsymbol{u}, \boldsymbol{u}_\star \in \mathbb{R}^n$ and $\lambda \ge \|\boldsymbol{u}_\star\|_2/\|\boldsymbol{u}\|_2$, it holds that*

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

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 31. Assume for now (which shall be established at the end of the proof) that for any $1 \le i_1 \le n_1$, we have

$$\frac{B}{\sqrt{n_1}\left\|\boldsymbol{U}_+(i_1,:)\check{\boldsymbol{U}}_+^\top\right\|_2} \ge \frac{\left\|\boldsymbol{U}_\star(i_1,:)\boldsymbol{\Sigma}_{\star,1}\right\|_2}{\left\|\boldsymbol{U}_+(i_1,:)\boldsymbol{Q}_{+,1}\boldsymbol{\Sigma}_{\star,1}\right\|_2}. \tag{C.23}$$

This taken together with Claim 12 immediately implies

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

Repeating similar arguments for the other two factors, we obtain

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

Combining the above bounds, we have

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

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

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

Here, (i) and (iii) follow from the definition of the scaled projection (4.19), (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 (C.23).

Proof of (C.23). Under the condition $\operatorname{dist}(\boldsymbol{F}_+, \boldsymbol{F}_\star) \le \epsilon\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$, invoke (C.5a) in Lemma 34 on the factor quadruple $\left(\boldsymbol{U}_+\boldsymbol{Q}_{+,1}, \boldsymbol{V}_+\boldsymbol{Q}_{+,2}, \boldsymbol{W}_+\boldsymbol{Q}_{+,3}, (\boldsymbol{Q}_{+,1}^{-1}, \boldsymbol{Q}_{+,2}^{-1}, \boldsymbol{Q}_{+,3}^{-1})\cdot\boldsymbol{\mathcal{S}}_+\right)$ to see

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

which further implies that

$$\left\|\breve{\boldsymbol{U}}_{+}\boldsymbol{Q}_{+,1}^{-\top}\boldsymbol{\Sigma}_{\star,1}^{-1}\right\| \leq \|\boldsymbol{V}_{+}\boldsymbol{Q}_{+,2}\|\|\boldsymbol{W}_{+}\boldsymbol{Q}_{+,3}\| \left\|\mathcal{M}_1\left((\boldsymbol{Q}_{+,1}^{-1},\boldsymbol{Q}_{+,2}^{-1},\boldsymbol{Q}_{+,3}^{-1})\cdot\boldsymbol{\mathcal{S}}_{+}\right)^{\top}\boldsymbol{\Sigma}_{\star,1}^{-1}\right\| \leq (1+\epsilon)^3. \tag{C.24}$$

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

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

where the second line follows from the bound (C.24). In addition, the incoherence assumption of $\boldsymbol{\mathcal{X}}_\star$ (4.15) implies that

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

where the last inequality follows from the choice of B. Take the above two relations collectively to reach the advertised bound (C.23).

C.3.2 Concentration inequalities

We gather several useful concentration inequalities regarding the partial observation operator $\mathcal{P}_\Omega(\cdot)$ for the Bernoulli observation model (4.17).

Lemma 36. *Suppose that $\boldsymbol{\mathcal{X}}_\star$ is μ-incoherent, and that $pn_1n_2n_3 \gtrsim n\mu^2r^2\log n$. With overwhelming probability, one has*

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

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

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

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

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 37 ([CLPC19, Lemma D.2]). *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 38. *With overwhelming probability, one has*

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

simultaneously for all tensors $(\boldsymbol{U}_A, \boldsymbol{V}_A, \boldsymbol{W}_A)\cdot\boldsymbol{\mathcal{S}}_A$ *and* $(\boldsymbol{U}_B, \boldsymbol{V}_B, \boldsymbol{W}_B)\cdot\boldsymbol{\mathcal{S}}_B$, *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 permuting the occurrences of $\boldsymbol{U}$, $\boldsymbol{V}$, *and* $\boldsymbol{W}$.

Lemma 39 ([CCFM21, Lemma 3.24], [CLC+21, Lemma 1]). *For any fixed* $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1\times n_2\times n_3}$, $k = 1, 2, 3$, *with overwhelming probability, one has*

$$\begin{aligned}&\left\|\mathcal{P}_{\mathsf{off\text{-}diag}}\left(p^{-2}\mathcal{M}_k(\mathcal{P}_\Omega(\boldsymbol{\mathcal{X}}))\mathcal{M}_k(\mathcal{P}_\Omega(\boldsymbol{\mathcal{X}}))^\top\right) - \mathcal{M}_k(\boldsymbol{\mathcal{X}})\mathcal{M}_k(\boldsymbol{\mathcal{X}})^\top\right\|\\ &\quad\le C_M\left(p^{-1}\sqrt{\log n}\|\mathcal{M}_k(\boldsymbol{\mathcal{X}})\|_{2,\infty}\|\mathcal{M}_k(\boldsymbol{\mathcal{X}})^\top\|_{2,\infty} + \sqrt{p^{-1}\log n}\,\sigma_{\max}(\mathcal{M}_k(\boldsymbol{\mathcal{X}}))\|\mathcal{M}_k(\boldsymbol{\mathcal{X}})^\top\|_{2,\infty}\right)\\ &\qquad + C_M\left(p^{-1}\log n\|\boldsymbol{\mathcal{X}}\|_\infty + \sqrt{p^{-1}\log n}\|\mathcal{M}_k(\boldsymbol{\mathcal{X}})^\top\|_{2,\infty}\right)^2\log n + \|\mathcal{M}_k(\boldsymbol{\mathcal{X}})\|_{2,\infty}^2,\end{aligned}$$

where $C_M > 0$ *is some universal constant.*

Proof of Lemma 36

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 $\delta_{i_1,i_2,i_3} \sim \text{Bernoulli}(p)$ 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

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

where $\boldsymbol{V}_{\star\perp}, \boldsymbol{W}_{\star\perp}$ denote the orthogonal complements of $\boldsymbol{V}_\star, \boldsymbol{W}_\star$. It is straightforward to verify that $\mathcal{P}_T(\cdot)$ defines a projection, and that

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

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

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

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

$$\left\| \sum_{i_1,i_2,i_3} (p^{-1}\delta_{i_1,i_2,i_3} - 1) \operatorname{vec}\left(\mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3})\right) \operatorname{vec}\left(\mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3})\right)^\top \right\| \lesssim \left(\frac{n\mu^2 r^2 \log n}{p n_1 n_2 n_3} + \sqrt{\frac{n\mu^2 r^2 \log n}{p n_1 n_2 n_3}} \right) \tag{C.25}$$

$$\lesssim \sqrt{\frac{n\mu^2 r^2 \log n}{p n_1 n_2 n_3}},$$

where the second line holds as long as $p n_1 n_2 n_3 \gtrsim n\mu^2 r^2 \log n$. Plugging the above bound (which will be proved at the end) in the previous one, we immediately arrive at the desired result:

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

Proof of (C.25)**.** By standard matrix Bernstein inequality, we have

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

where

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

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

- Here, L obeys

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

  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

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

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

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

Finally use the definition of incoherence (cf. Definition 11) to conclude

$$L \le p^{-1}\left(\|\boldsymbol{V}_\star\|_{2,\infty}^2\|\boldsymbol{W}_\star\|_{2,\infty}^2 + \|\boldsymbol{U}_\star\|_{2,\infty}^2\|\boldsymbol{W}_\star\|_{2,\infty}^2 + \|\boldsymbol{U}_\star\|_{2,\infty}^2\|\boldsymbol{V}_\star\|_{2,\infty}^2\right) \le \frac{3n\mu^2r^2}{pn_1n_2n_3}.$$

- In addition, σ^2 obeys

$$\sigma^2 \le p^{-1}\max_{i_1,i_2,i_3}\|\mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3})\|_{\mathsf{F}}^2\left\|\sum_{i_1,i_2,i_3}\mathrm{vec}\left(\mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3})\right)\mathrm{vec}\left(\mathcal{P}_T(\boldsymbol{\mathcal{E}}_{i_1,i_2,i_3})\right)^\top\right\| \le \frac{3n\mu^2r^2}{pn_1n_2n_3},$$

where we have used the variational representation to conclude

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

Plugging the expressions of L and σ leads to the advertised bound (C.25).

Proof of Lemma 38

This lemma generalizes [CL19, Lemma 8] to the tensor setting, which is a powerful tool in the analysis of matrix completion [CLL20, TMC21a]. We begin by decomposing $(\boldsymbol{U}_A, \boldsymbol{V}_A, \boldsymbol{W}_A) \cdot \boldsymbol{\mathcal{S}}_A$ into a sum of $r_2 r_3$ rank-$\mathbf{1}$ tensors:

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

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 $(\boldsymbol{U}_B, \boldsymbol{V}_B, \boldsymbol{W}_B) \cdot \boldsymbol{\mathcal{S}}_B$ as

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

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 $1 \le i_k \le n_k$, $k = 1, 2, 3$. With these preparation in hand, by the triangle inequality we have

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

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

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

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 37 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.3.3 Proof of spectral initialization (Lemma 9)

In view of Lemma 33, we start by relating $\mathrm{dist}(\boldsymbol{F}_+, \boldsymbol{F}_\star)$ to $\|(\boldsymbol{U}_+, \boldsymbol{V}_+, \boldsymbol{W}_+) \cdot \boldsymbol{\mathcal{S}}_+ - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}$ as

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

With this bound in mind, it suffices to control $\|(\boldsymbol{U}_+, \boldsymbol{V}_+, \boldsymbol{W}_+) \cdot \boldsymbol{\mathcal{S}}_+ - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}$. 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

$$\boldsymbol{\mathcal{X}}_\star = (\boldsymbol{P}_U, \boldsymbol{P}_V, \boldsymbol{P}_W) \cdot \boldsymbol{\mathcal{X}}_\star + (\boldsymbol{P}_{U_\perp}, \boldsymbol{P}_V, \boldsymbol{P}_W) \cdot \boldsymbol{\mathcal{X}}_\star + (\boldsymbol{I}_{n_1}, \boldsymbol{P}_{V_\perp}, \boldsymbol{P}_W) \cdot \boldsymbol{\mathcal{X}}_\star + (\boldsymbol{I}_{n_1}, \boldsymbol{I}_{n_2}, \boldsymbol{P}_{W_\perp}) \cdot \boldsymbol{\mathcal{X}}_\star.$$

Expand the following squared norm and use that the four terms are mutually orthogonal to see

$$\begin{aligned}
&\|(\boldsymbol{U}_+, \boldsymbol{V}_+, \boldsymbol{W}_+) \cdot \boldsymbol{\mathcal{S}}_+ - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}^2 = \left\|(\boldsymbol{P}_U, \boldsymbol{P}_V, \boldsymbol{P}_W) \cdot (p^{-1}\boldsymbol{\mathcal{Y}}) - \boldsymbol{\mathcal{X}}_\star\right\|_{\mathsf{F}}^2 \\
&\quad = \left\|(\boldsymbol{P}_U, \boldsymbol{P}_V, \boldsymbol{P}_W) \cdot (p^{-1}\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}_\star) - (\boldsymbol{P}_{U_\perp}, \boldsymbol{P}_V, \boldsymbol{P}_W) \cdot \boldsymbol{\mathcal{X}}_\star - (\boldsymbol{I}_{n_1}, \boldsymbol{P}_{V_\perp}, \boldsymbol{P}_W) \cdot \boldsymbol{\mathcal{X}}_\star - (\boldsymbol{I}_{n_1}, \boldsymbol{I}_{n_2}, \boldsymbol{P}_{W_\perp}) \cdot \boldsymbol{\mathcal{X}}_\star\right\|_{\mathsf{F}}^2 \\
&\quad = \left\|(\boldsymbol{P}_U, \boldsymbol{P}_V, \boldsymbol{P}_W) \cdot (p^{-1}\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}_\star)\right\|_{\mathsf{F}}^2 + \|(\boldsymbol{P}_{U_\perp}, \boldsymbol{P}_V, \boldsymbol{P}_W) \cdot \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}^2 + \|(\boldsymbol{I}_{n_1}, \boldsymbol{P}_{V_\perp}, \boldsymbol{P}_W) \cdot \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}^2 \\
&\qquad + \|(\boldsymbol{I}_{n_1}, \boldsymbol{I}_{n_2}, \boldsymbol{P}_{W_\perp}) \cdot \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}^2 \\
&\quad \le \left\|(\boldsymbol{P}_U, \boldsymbol{P}_V, \boldsymbol{P}_W) \cdot (p^{-1}\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}_\star)\right\|_{\mathsf{F}}^2 + \|\boldsymbol{P}_{U_\perp}\mathcal{M}_1(\boldsymbol{\mathcal{X}}_\star)\|_{\mathsf{F}}^2 + \|\boldsymbol{P}_{V_\perp}\mathcal{M}_2(\boldsymbol{\mathcal{X}}_\star)\|_{\mathsf{F}}^2 + \|\boldsymbol{P}_{W_\perp}\mathcal{M}_3(\boldsymbol{\mathcal{X}}_\star)\|_{\mathsf{F}}^2 .
\end{aligned} \tag{C.26}$$

We next control the terms in (C.26) one by one.

Bounding $\|(\boldsymbol{P}_U, \boldsymbol{P}_V, \boldsymbol{P}_W) \cdot (\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}_\star)\|_{\mathsf{F}}$. For the first term in (C.26), since $(\boldsymbol{P}_U, \boldsymbol{P}_V, \boldsymbol{P}_W) \cdot (p^{-1}\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}_\star)$ has a multilinear rank of at most $\boldsymbol{r}$, applying the relation (4.7) leads to

$$\left\|(\boldsymbol{P}_U, \boldsymbol{P}_V, \boldsymbol{P}_W) \cdot (p^{-1}\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}_\star)\right\|_{\mathsf{F}} \le r \left\|(\boldsymbol{P}_U, \boldsymbol{P}_V, \boldsymbol{P}_W) \cdot (p^{-1}\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}_\star)\right\| \le r \left\|(p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{\mathcal{X}}_\star)\right\| .$$

Therefore, it comes down to control $\left\|(p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{\mathcal{X}}_\star)\right\|$. Lemma 37 tells with overwhelming probability that

$$\begin{aligned}
\left\|(p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{\mathcal{X}}_\star)\right\| &\lesssim \left(p^{-1}\log^3 n\|\boldsymbol{\mathcal{X}}_\star\|_\infty + \sqrt{p^{-1}\log^5 n}\max_{k=1,2,3}\|\mathcal{M}_k(\boldsymbol{\mathcal{X}}_\star)^\top\|_{2,\infty}\right) \\
&\lesssim \left(\frac{\mu^{3/2}r^{3/2}\log^3 n}{p\sqrt{n_1n_2n_3}} + \sqrt{\frac{n\mu^2r^2\log^5 n}{pn_1n_2n_3}}\right)\sigma_{\max}(\boldsymbol{X}_\star),
\end{aligned}$$

where the second line follows from the following relations in view of the incoherence property of $\boldsymbol{\mathcal{X}}_\star$ (cf. Definition 11):

$$\begin{aligned}
\|\boldsymbol{\mathcal{X}}_\star\|_\infty &\le \sigma_{\max}(\boldsymbol{\mathcal{X}}_\star)\|\boldsymbol{U}_\star\|_{2,\infty}\|\boldsymbol{V}_\star\|_{2,\infty}\|\boldsymbol{W}_\star\|_{2,\infty} \le \sigma_{\max}(\boldsymbol{\mathcal{X}}_\star)\sqrt{\frac{\mu^3 r^3}{n_1 n_2 n_3}}; \\
\|\mathcal{M}_1(\boldsymbol{\mathcal{X}}_\star)^\top\|_{2,\infty} &\le \|\boldsymbol{U}_\star\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)\|\,\|\boldsymbol{W}_\star\|_{2,\infty}\|\boldsymbol{V}_\star\|_{2,\infty} \le \sigma_{\max}(\boldsymbol{\mathcal{X}}_\star)\sqrt{\frac{\mu^2 r^2}{n_2 n_3}}; \\
\|\mathcal{M}_2(\boldsymbol{\mathcal{X}}_\star)^\top\|_{2,\infty} &\le \|\boldsymbol{V}_\star\mathcal{M}_2(\boldsymbol{\mathcal{S}}_\star)\|\,\|\boldsymbol{W}_\star\|_{2,\infty}\|\boldsymbol{U}_\star\|_{2,\infty} \le \sigma_{\max}(\boldsymbol{\mathcal{X}}_\star)\sqrt{\frac{\mu^2 r^2}{n_1 n_3}}; \\
\|\mathcal{M}_3(\boldsymbol{\mathcal{X}}_\star)^\top\|_{2,\infty} &\le \|\boldsymbol{W}_\star\mathcal{M}_3(\boldsymbol{\mathcal{S}}_\star)\|\,\|\boldsymbol{V}_\star\|_{2,\infty}\|\boldsymbol{U}_\star\|_{2,\infty} \le \sigma_{\max}(\boldsymbol{\mathcal{X}}_\star)\sqrt{\frac{\mu^2 r^2}{n_1 n_2}}.
\end{aligned} \tag{C.27}$$

In total, the first term in (C.26) is bounded by

$$\left\|(\boldsymbol{P}_U, \boldsymbol{P}_V, \boldsymbol{P}_W)\cdot(p^{-1}\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}_\star)\right\|_{\mathsf{F}} \lesssim \left(\frac{\mu^{3/2} r^{3/2}\log^3 n}{p\sqrt{n_1 n_2 n_3}} + \sqrt{\frac{n\mu^2 r^2 \log^5 n}{p n_1 n_2 n_3}}\right) r\kappa\sigma_{\min}(\boldsymbol{X}_\star).$$

Bounding $\|\boldsymbol{P}_{U_\perp}\mathcal{M}_1(\boldsymbol{\mathcal{X}}_\star)\|_{\mathsf{F}}$. For the second term in (C.26), first bound it by

$$\|\boldsymbol{P}_{U_\perp}\mathcal{M}_1(\boldsymbol{\mathcal{X}}_\star)\|_{\mathsf{F}} \le \frac{\sqrt{r_1}}{\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)}\left\|\boldsymbol{P}_{U_\perp}\mathcal{M}_1(\boldsymbol{\mathcal{X}}_\star)\mathcal{M}_1(\boldsymbol{\mathcal{X}}_\star)^\top\right\|,$$

where we use the facts that $\boldsymbol{P}_{U_\perp}\mathcal{M}_1(\boldsymbol{\mathcal{X}}_\star)$ has rank at most r_1 and $\|\boldsymbol{AB}\| \ge \|\boldsymbol{A}\|\sigma_{\min}(\boldsymbol{B})$. For notation simplicity, we abbreviate

$$\boldsymbol{G} \coloneqq \mathcal{P}_{\mathsf{off\text{-}diag}}(p^{-2}\mathcal{M}_1(\boldsymbol{\mathcal{Y}})\mathcal{M}_1(\boldsymbol{\mathcal{Y}})^\top), \quad \text{and} \quad \boldsymbol{G}_\star \coloneqq \mathcal{M}_1(\boldsymbol{\mathcal{X}}_\star)\mathcal{M}_1(\boldsymbol{\mathcal{X}}_\star)^\top.$$

Invoke Lemma 39 together with incoherence conditions (C.27) as well as

$$\|\mathcal{M}_1(\boldsymbol{\mathcal{X}}_\star)\|_{2,\infty} \le \|\boldsymbol{U}_\star\|_{2,\infty}\left\|\mathcal{M}_1(\boldsymbol{\mathcal{S}}_\star)(\boldsymbol{W}_\star\otimes\boldsymbol{V}_\star)^\top\right\| \le \sigma_{\max}(\boldsymbol{\mathcal{X}}_\star)\sqrt{\frac{\mu r_1}{n_1}}$$

to conclude with overwhelming probability that

$$\|\boldsymbol{G}-\boldsymbol{G}_\star\| \lesssim \left(\frac{\mu^{3/2}r^{3/2}\sqrt{\log n}}{p\sqrt{n_1n_2n_3}}+\sqrt{\frac{n\mu^2r^2\log n}{pn_1n_2n_3}}+\frac{\mu^3r^3\log^3 n}{p^2n_1n_2n_3}+\frac{n\mu^2r^2\log^2 n}{pn_1n_2n_3}+\frac{\mu r_1}{n_1}\right)\sigma_{\max}^2(\boldsymbol{\mathcal{X}}_\star).$$

Under the conditions $n_1 \gtrsim \epsilon_0^{-1}\mu r_1^{3/2}\kappa^2$ and

$$pn_1n_2n_3 \gtrsim \epsilon_0^{-1}\sqrt{n_1n_2n_3}\mu^{3/2}r^{5/2}\kappa^2\log^3 n+\epsilon_0^{-2}n\mu^2r^4\kappa^4\log^5 n$$

for some small constant $\epsilon_0>0$, we have $\|\boldsymbol{G}-\boldsymbol{G}_\star\|\le\epsilon_0\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star)$, which implies that $\boldsymbol{G}$ is positive semi-definite, and therefore $\|\boldsymbol{P}_{U_\perp}\boldsymbol{G}\|=\sigma_{r_1+1}(\boldsymbol{G})$. By the triangle inequality, we obtain

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

where the second line follows from Weyl's inequality and that $\boldsymbol{G}_\star$ has rank r_1. In total, the second term of (C.26) is bounded by

$$\begin{aligned}\|\boldsymbol{P}_{U_\perp}\mathcal{M}_1(\boldsymbol{\mathcal{X}}_\star)\|_{\mathsf{F}} &\le \frac{2\sqrt{r_1}}{\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)}\|\boldsymbol{G}-\boldsymbol{G}_\star\|\\ &\lesssim \left(\frac{\mu^{3/2}r^2\sqrt{\log n}}{p\sqrt{n_1n_2n_3}}+\sqrt{\frac{n\mu^2r^3\log n}{pn_1n_2n_3}}+\frac{\mu^3r^{7/2}\log^3 n}{p^2n_1n_2n_3}+\frac{n\mu^2r^{5/2}\log^2 n}{pn_1n_2n_3}+\frac{\mu r_1^{3/2}}{n_1}\right)\kappa^2\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star).\end{aligned}$$

Completing the proof. The third and fourth terms in (C.26) can be bounded similarly. In all, we conclude that

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

C.3.4 Proof of local convergence (Lemma 11)

Define the event $\mathcal{E}$ as the intersection of the events that Lemmas 36 and 38 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 $\|(\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t)\cdot\boldsymbol{\mathcal{S}}_t - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \le 3\,\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$ follows from the relation (C.7) in Lemma 34. As in the proof of Theorem 10, we reuse the notations in (C.4) and (C.13). By the definition of $\mathrm{dist}(\boldsymbol{F}_{t+}, \boldsymbol{F}_\star)$, where $\boldsymbol{F}_{t+}$ is the update before projection, one has

$$\begin{aligned}\mathrm{dist}^2(\boldsymbol{F}_{t+}, \boldsymbol{F}_\star) \le \|(\boldsymbol{U}_{t+}\boldsymbol{Q}_{t,1} - \boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 + \|(\boldsymbol{V}_{t+}\boldsymbol{Q}_{t,2} - \boldsymbol{V}_\star)\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}^2 + \|(\boldsymbol{W}_{t+}\boldsymbol{Q}_{t,3} - \boldsymbol{W}_\star)\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}^2 \\ + \left\|(\boldsymbol{Q}_{t,1}^{-1}, \boldsymbol{Q}_{t,2}^{-1}, \boldsymbol{Q}_{t,3}^{-1})\cdot\boldsymbol{\mathcal{S}}_{t+} - \boldsymbol{\mathcal{S}}_\star\right\|_{\mathsf{F}}^2. \qquad \text{(C.28)}\end{aligned}$$

In the sequel, we shall bound each square on the right hand side of equation (C.28) separately. After a long journey of computation, the final result is

$$\begin{aligned}\mathrm{dist}^2(\boldsymbol{F}_{t+}, \boldsymbol{F}_\star) \le (1-\eta)^2\left(\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}^2\right) \\ - \eta(2-5\eta)\,\|\boldsymbol{\mathcal{T}}_U + \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W\|_{\mathsf{F}}^2 - \eta(2-5\eta)\left(\|\boldsymbol{D}_U\|_{\mathsf{F}}^2 + \|\boldsymbol{D}_V\|_{\mathsf{F}}^2 + \|\boldsymbol{D}_W\|_{\mathsf{F}}^2\right) \\ + 2\eta(1-\eta)C(\epsilon+\delta+\delta^2)\,\mathrm{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star) + \eta^2 C(\epsilon+\delta+\delta^2)\,\mathrm{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star), \qquad \text{(C.29)}\end{aligned}$$

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

$$\delta \coloneqq C_T\sqrt{\frac{n\mu^2 r^2\log n}{pn_1n_2n_3}} + C_Y\left(p^{-1}\log^3 n + \sqrt{p^{-1}n\log^5 n}\right)\sqrt{\frac{\mu^3 r^4}{n_1n_2n_3}}C_B^3\kappa^3. \qquad \text{(C.30)}$$

Under the condition

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

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 10 implies $\mathrm{dist}(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) \le \mathrm{dist}(\boldsymbol{F}_{t+}, \boldsymbol{F}_\star) \le (1-0.6\eta)\,\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$ and the incoherence condition.

It then boils down to expanding and bounding the four terms in (C.28). As before, we omit the control of the terms pertaining to $\boldsymbol{V}$ and $\boldsymbol{W}$.

Bounding the term related to $\boldsymbol{U}$

The first term in (C.28) is related to

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

Take the squared norm of both sides to reach

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

As before, the main term $\mathfrak{P}_U^{\mathrm{m}}$ has been handled in the tensor factorization problem in Section C.2; see (C.17) and the bound (C.15a). Hence we shall focus on the perturbation terms.

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

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

Apply the decomposition

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

$$= (\boldsymbol{U}, \boldsymbol{\Delta}_V, \boldsymbol{W}) \cdot \boldsymbol{\mathcal{S}} + (\boldsymbol{U}, \boldsymbol{V}_\star, \boldsymbol{\Delta}_W) \cdot \boldsymbol{\mathcal{S}} + (\boldsymbol{U}, \boldsymbol{V}_\star, \boldsymbol{W}_\star) \cdot \boldsymbol{\Delta}_{\mathcal{S}} + (\boldsymbol{\Delta}_U, \boldsymbol{V}_\star, \boldsymbol{W}_\star) \cdot \boldsymbol{\mathcal{S}}_\star \tag{C.31}$$

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

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

We shall bound each term in the sequel.

- For the first term $\mathfrak{P}_U^{\mathrm{p},1,1}$, we resort to Lemma 36, which leads to

$$|\mathfrak{P}_U^{\mathrm{p},1,1}| \le C_T \sqrt{\frac{n\mu^2 r^2 \log n}{p n_1 n_2 n_3}} \left\| \left(\boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}^2 (\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1}, \boldsymbol{V}_\star, \boldsymbol{W}_\star\right) \cdot \boldsymbol{\mathcal{S}} \right\|_{\mathsf{F}} \left\| (\boldsymbol{U}, \boldsymbol{V}_\star, \boldsymbol{W}_\star) \cdot \boldsymbol{\Delta}_{\mathcal{S}} + (\boldsymbol{\Delta}_U, \boldsymbol{V}_\star, \boldsymbol{W}_\star) \cdot \boldsymbol{\mathcal{S}}_\star \right\|_{\mathsf{F}}.$$

  Further use (C.5i) to bound that

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

  and that

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

Combine the preceding bounds to see

$$|\mathfrak{P}_U^{\mathrm{p},1,1}| \le C_T \sqrt{\frac{n\mu^2 r^2 \log n}{p n_1 n_2 n_3}} \frac{\|\boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}}{(1-\epsilon)^5} \left(\|\boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}} + (1+\epsilon)\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\right).$$

- For the second term $\mathfrak{P}_U^{\mathrm{p},1,2}$, our main hammer is Lemma 38, which implies

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

Use results in Lemma 35, 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

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

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

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

$$
\begin{aligned}
&\le C_Y\left(p^{-1}\log^3 n+\sqrt{p^{-1}n\log^5 n}\right)\frac{\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}}{(1-\epsilon)^5}\frac{C_B}{(1-\epsilon)^2}\sqrt{\frac{\mu r}{n_1}}\sigma_{\max}(\boldsymbol{\mathcal{X}}_\star)\\
&\quad\frac{C_B\kappa}{(1-\epsilon)^3}\sqrt{\frac{\mu r}{n_2}}\frac{C_B\kappa}{(1-\epsilon)^3}\sqrt{\frac{\mu r}{n_3}}\left(\frac{\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}}{\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)}\sqrt{r}(1+\epsilon)+\sqrt{r}\frac{\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}}{\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)}\right)\\
&\le C_Y\left(p^{-1}\log^3 n+\sqrt{p^{-1}n\log^5 n}\right)\sqrt{\frac{\mu^3r^4}{n_1n_2n_3}}\frac{C_B^3\kappa^3}{(1-\epsilon)^{13}}\\
&\quad\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}\left((1+\epsilon)\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}\right).
\end{aligned}
$$

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 (C.30).

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

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

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

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

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

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

$$
\begin{aligned}
&+ C_Y \left(p^{-1} \log^3 n + \sqrt{p^{-1} n \log^5 n} \right) \sqrt{\frac{\mu^3 r^4}{n_1 n_2 n_3}} \frac{(1+\epsilon+\frac{1}{3}\epsilon^2) C_B^3 \kappa^3}{(1-\epsilon)^{16}} \\
&\qquad (\|\boldsymbol{\Delta}_V \boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_W \boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}) \left((1+\epsilon) \|\boldsymbol{\Delta}_V \boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_W \boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}} \right) \\
&\lesssim \delta \, \mathrm{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star).
\end{aligned}
$$

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

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

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

$$
\begin{aligned}
\sqrt{\mathfrak{P}_U^{\mathrm{p},3}} &\le C_T \sqrt{\frac{n\mu^2 r^2 \log n}{p n_1 n_2 n_3}} (1-\epsilon)^{-5} \left(\|\boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}} + (1+\epsilon)\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}} \right) \\
&\quad + C_Y \left(p^{-1} \log^3 n + \sqrt{p^{-1} n \log^5 n} \right) \sqrt{\frac{\mu^3 r^4}{n_1 n_2 n_3}} \frac{(1-\epsilon)^{-2} C_B + 1}{(1-\epsilon)^5} \kappa \left((1+\epsilon)\|\boldsymbol{\Delta}_V \boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_W \boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}} \right) \\
&\quad + C_Y \left(p^{-1} \log^3 n + \sqrt{p^{-1} n \log^5 n} \right) \sqrt{\frac{\mu^3 r^4}{n_1 n_2 n_3}} \frac{C_B^3 \kappa^3}{(1-\epsilon)^{13}} \left((1+\epsilon)\|\boldsymbol{\Delta}_V \boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_W \boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}} \right) \\
&\lesssim \delta \, \mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star).
\end{aligned}
$$

Then take the square of both sides to see

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

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

The last term of (C.28) is related to

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

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

Expand its squared norm to obtain

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

Recall that the main term $\mathfrak{P}_{\mathcal{S}}^{\mathrm{m}}$ has been controlled in Section C.2; see (C.18) and the bound (C.15d). We therefore concentrate on the remaining perturbation terms.

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

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

Use the decomposition (C.31) to further obtain

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

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

We then bound each term in sequel.

- Regarding the first term $\mathfrak{P}_{\mathcal{S}}^{\mathrm{p},1,1}$, we can apply Lemma 36 to see

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

  In addition, notice that

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

  which further implies

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

- Now we turn to the second term $\mathfrak{P}_{\mathcal{S}}^{\mathrm{p},1,2}$, for which Lemma 38 yields

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

  The results in Lemma 35 together with the bounds

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

$$\left\|\boldsymbol{\Delta}_W(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\right\|_{\mathsf{F}} \le \|\boldsymbol{\Delta}_W\|_{\mathsf{F}}\left\|(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\right\| \le \|\boldsymbol{\Delta}_W\|_{\mathsf{F}}(1-\epsilon)^{-2} \le \frac{\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}}{(1-\epsilon)^2\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)},$$

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

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

Further combine (C.21c) and (C.5d) to see

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

These taken collectively with the results in Lemma 35 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 (C.30).

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

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

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

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

whose Frobenius norm can be bounded by

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

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

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

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.

C.4 Proof for Tensor Regression

Before embarking on the proof, we state a useful lemma regarding TRIP (cf. Definition 12).

Lemma 40 ([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*

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

or equivalently,

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

C.4.1 Proof of local convergence (Lemma 12)

Given that $\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \le \epsilon \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)$, the conclusion $\|(\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t)\cdot \boldsymbol{\mathcal{S}}_t - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \le 3\,\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$ directly follows from the relation (C.7) in Lemma 34. Hence we will focus on controlling $\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$.

As in the proof of Theorem 10, we reuse the notations in (C.4) and (C.13), and the definition of $\mathrm{dist}(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star)$ to obtain

$$\begin{aligned}\mathrm{dist}^2(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) \le{}& \|(\boldsymbol{U}_{t+1}\boldsymbol{Q}_{t,1} - \boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 + \|(\boldsymbol{V}_{t+1}\boldsymbol{Q}_{t,2} - \boldsymbol{V}_\star)\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}^2 + \|(\boldsymbol{W}_{t+1}\boldsymbol{Q}_{t,3} - \boldsymbol{W}_\star)\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}^2 \\ &+ \Big\|(\boldsymbol{Q}_{t,1}^{-1}, \boldsymbol{Q}_{t,2}^{-1}, \boldsymbol{Q}_{t,3}^{-1})\cdot \boldsymbol{\mathcal{S}}_{t+1} - \boldsymbol{\mathcal{S}}_\star\Big\|_{\mathsf{F}}^2. \end{aligned} \tag{C.32}$$

We shall bound each square in the right hand side of the bound (C.32) separately. The final result is

$$\begin{aligned}\mathrm{dist}^2(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) \le{}& (1-\eta)^2 \left(\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}^2\right) \\ &- \eta(2-5\eta)\,\|\boldsymbol{\mathcal{T}}_U + \boldsymbol{\mathcal{T}}_V + \boldsymbol{\mathcal{T}}_W\|_{\mathsf{F}}^2 - \eta(2-5\eta)\left(\|\boldsymbol{D}_U\|_{\mathsf{F}}^2 + \|\boldsymbol{D}_V\|_{\mathsf{F}}^2 + \|\boldsymbol{D}_W\|_{\mathsf{F}}^2\right) \\ &+ 2\eta(1-\eta)C(\epsilon + \delta_{2\boldsymbol{r}} + \delta_{2\boldsymbol{r}}^2)\,\mathrm{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star) + \eta^2 C(\epsilon + \delta_{2\boldsymbol{r}} + \delta_{2\boldsymbol{r}}^2)\,\mathrm{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star), \end{aligned} \tag{C.33}$$

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 (C.32). In a nutshell, the bounds that are sought after are reminiscent of those established in (C.15), with additional perturbation terms introduced due to incomplete measurements, manifested via the TRIP parameter $\delta_{2\boldsymbol{r}}$. Once established, the claimed bound (C.33) 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 $\boldsymbol{W}$.

Bounding the term pertaining to $\boldsymbol{U}$

The first term in (C.32) is given by

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

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

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

The main term $\mathfrak{R}_U^{\mathrm{m}}$ has been handled in Section C.2; see (C.17) and the bound (C.15a). In the sequel, we shall bound the three perturbation terms.

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

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

where the second relation uses the decomposition (C.10). Apply Lemma 40 to each of the four terms to obtain

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

For the prefactor, we have

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

where the last step arises from Lemma 34. In addition, the same argument as in (C.6a) yields

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

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

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

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

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

For the prefactor, we can use (C.5f) and (C.6c) 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

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

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

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

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 (C.5f) in Lemma 34. Then take the square on both sides to conclude

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

as long as ϵ is sufficiently small.

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

The last term of (C.32) can be rewritten as

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

which further gives

$$
\begin{aligned}
&\left\|(\boldsymbol{Q}_{t,1}^{-1},\boldsymbol{Q}_{t,2}^{-1},\boldsymbol{Q}_{t,3}^{-1})\cdot\boldsymbol{\mathcal{S}}_{t+1}-\boldsymbol{\mathcal{S}}_\star\right\|_{\mathsf{F}}^2 \\
&\quad=\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)\cdot((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\cdot\boldsymbol{\mathcal{S}}_\star-\boldsymbol{\mathcal{X}}_\star)\right\|_{\mathsf{F}}^2}_{=:\mathfrak{R}_{\mathcal{S}}^{\mathrm{m}}} \\
&\qquad-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)\cdot(\mathcal{A}^*\mathcal{A}-\mathcal{I})((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\cdot\boldsymbol{\mathcal{S}}-\boldsymbol{\mathcal{X}}_\star)\right\rangle}_{=:\mathfrak{R}_{\mathcal{S}}^{\mathrm{p},1}} \\
&\qquad+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)\cdot((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\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)\cdot(\mathcal{A}^*\mathcal{A}-\mathcal{I})((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\cdot\boldsymbol{\mathcal{S}}-\boldsymbol{\mathcal{X}}_\star)\end{array}\right\rangle}_{=:\mathfrak{R}_{\mathcal{S}}^{\mathrm{p},2}} \\
&\qquad+\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)\cdot(\mathcal{A}^*\mathcal{A}-\mathcal{I})((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\cdot\boldsymbol{\mathcal{S}}-\boldsymbol{\mathcal{X}}_\star)\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 C.2; see (C.18) and the bound (C.15d). Therefore we concentrate on the remaining perturbation terms.

Step 1: bounding $\mathfrak{R}_{\mathcal{S}}^{\mathrm{p},1}$. Use the property (4.6d) to write $\mathfrak{R}_{\mathcal{S}}^{\mathrm{p},1}$ as

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

We can use the decomposition (C.10) and Lemma 40 to derive

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

In addition, Lemma 34 tells us that

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

Combine the above two bounds to reach

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

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

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

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

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

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

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

Then take the square on both sides to conclude

$$
\mathfrak{R}_{\mathcal{S}}^{\mathrm{p},3} \leq \delta_{2\boldsymbol{r}}^2 \frac{(1+\frac{3}{2}\epsilon+\epsilon^2+\frac{1}{4}\epsilon^3)^2}{(1-\epsilon)^6} \left(\|\boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_V \boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_W \boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathsf{F}}\right)^2
$$

$$\lesssim \delta_{2\boldsymbol{r}}^2 \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star).$$

C.4.2 Proof of spectral initialization (Lemma 13)

In view of Lemma 33, we can relate $\operatorname{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star)$ to $\|(\boldsymbol{U}_0, \boldsymbol{V}_0, \boldsymbol{W}_0)\cdot\boldsymbol{\mathcal{S}}_0 - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}$ as

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

To proceed, we need to control $\|(\boldsymbol{U}_0, \boldsymbol{V}_0, \boldsymbol{W}_0)\cdot\boldsymbol{\mathcal{S}}_0 - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}$, where $(\boldsymbol{U}_0, \boldsymbol{V}_0, \boldsymbol{W}_0)\cdot\boldsymbol{\mathcal{S}}_0$ is the output of HOSVD. Similar results have been established in [LZ21, HWZ20, ZLRY20], which involve sophisticated subspace perturbation bounds. For conciseness and completeness, we provide an alternative proof directly tackling the distance.

Define $\boldsymbol{P}_U \coloneqq \boldsymbol{U}_0\boldsymbol{U}_0^\top$ as the projection matrix onto the column space of $\boldsymbol{U}_0$, $\boldsymbol{P}_{U_\perp} \coloneqq \boldsymbol{I}_{n_1} - \boldsymbol{P}_U$ as the projection onto its orthogonal complement, and define $\boldsymbol{P}_V, \boldsymbol{P}_{V_\perp}, \boldsymbol{P}_W, \boldsymbol{P}_{W_\perp}$ analogously. Similar to (C.26), we have the decomposition

$$\begin{aligned}
&\|(\boldsymbol{U}_0, \boldsymbol{V}_0, \boldsymbol{W}_0)\cdot\boldsymbol{\mathcal{S}}_0 - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}^2 \\
&\quad\le \|(\boldsymbol{P}_U, \boldsymbol{P}_V, \boldsymbol{P}_W)\cdot(\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}_\star)\|_{\mathsf{F}}^2 + \|\boldsymbol{P}_{U_\perp}\mathcal{M}_1(\boldsymbol{\mathcal{X}}_\star)\|_{\mathsf{F}}^2 + \|\boldsymbol{P}_{V_\perp}\mathcal{M}_2(\boldsymbol{\mathcal{X}}_\star)\|_{\mathsf{F}}^2 + \|\boldsymbol{P}_{W_\perp}\mathcal{M}_3(\boldsymbol{\mathcal{X}}_\star)\|_{\mathsf{F}}^2.
\end{aligned} \tag{C.34}$$

Below we bound the terms on the right hand side of (C.34) in order.

Bounding $\|(\boldsymbol{P}_U, \boldsymbol{P}_V, \boldsymbol{P}_W)\cdot(\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}_\star)\|_{\mathsf{F}}$. For the first term in the upper bound (C.34), apply the variational representation of the Frobenius norm to write

$$\|(\boldsymbol{P}_U, \boldsymbol{P}_V, \boldsymbol{P}_W)\cdot(\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}_\star)\|_{\mathsf{F}} = \left\langle (\boldsymbol{P}_U, \boldsymbol{P}_V, \boldsymbol{P}_W)\cdot(\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}_\star), \widetilde{\boldsymbol{\mathcal{T}}}\right\rangle = \left\langle (\mathcal{A}^*\mathcal{A} - \mathcal{I})\boldsymbol{\mathcal{X}}_\star, (\boldsymbol{P}_U, \boldsymbol{P}_V, \boldsymbol{P}_W)\cdot\widetilde{\boldsymbol{\mathcal{T}}}\right\rangle,$$

for some $\widetilde{\boldsymbol{\mathcal{T}}} \in \mathbb{R}^{n_1\times n_3\times n_3}$ obeying $\|\widetilde{\boldsymbol{\mathcal{T}}}\|_{\mathsf{F}} = 1$, where the last equality follows from (4.6d). Under the Gaussian design, we know from [RSS17, Theorem 2] that $\mathcal{A}(\cdot)$ obeys $2\boldsymbol{r}$-TRIP with a constant

$\delta_{2\boldsymbol{r}} \asymp \sqrt{\frac{nr+r^3}{m}}$. Therefore we can apply Lemma 40 to obtain

$$\begin{aligned}\|(\boldsymbol{P}_U, \boldsymbol{P}_V, \boldsymbol{P}_W)\cdot(\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}_\star)\|_{\mathsf{F}} &\le \delta_{2\boldsymbol{r}}\|\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}\big\|(\boldsymbol{P}_U, \boldsymbol{P}_V, \boldsymbol{P}_W)\cdot\widetilde{\boldsymbol{\mathcal{T}}}\big\|_{\mathsf{F}} \le \delta_{2\boldsymbol{r}}\|\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \\ &\lesssim \sqrt{\frac{nr+r^3}{m}}\|\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \le \sqrt{\frac{nr^2+r^4}{m}}\kappa\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star).\end{aligned}$$

Bounding $\|\boldsymbol{P}_{U_\perp}\mathcal{M}_1(\boldsymbol{\mathcal{X}}_\star)\|_{\mathsf{F}}$. For the second term in (C.34), first bound it by

$$\|\boldsymbol{P}_{U_\perp}\mathcal{M}_1(\boldsymbol{\mathcal{X}}_\star)\|_{\mathsf{F}} \le \frac{\sqrt{r_1}}{\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)}\left\|\boldsymbol{P}_{U_\perp}\mathcal{M}_1(\boldsymbol{\mathcal{X}}_\star)\mathcal{M}_1(\boldsymbol{\mathcal{X}}_\star)^\top\right\|,$$

where we use the facts that $\boldsymbol{P}_{U_\perp}\mathcal{M}_1(\boldsymbol{\mathcal{X}}_\star)$ has rank at most r_1 and $\|\boldsymbol{AB}\| \ge \|\boldsymbol{A}\|\sigma_{\min}(\boldsymbol{B})$. For notation simplicity, we abbreviate

$$\boldsymbol{G} \coloneqq \mathcal{M}_1(\mathcal{A}^*(\boldsymbol{y}))\mathcal{M}_1(\mathcal{A}^*(\boldsymbol{y}))^\top - \frac{\|\boldsymbol{y}\|_2^2}{m}(n_2n_3 - r_1)\boldsymbol{I}_{n_1}, \quad \text{and} \quad \boldsymbol{G}_\star \coloneqq \mathcal{M}_1(\boldsymbol{\mathcal{X}}_\star)\mathcal{M}_1(\boldsymbol{\mathcal{X}}_\star)^\top.$$

We claim for the moment that with overwhelming probability that

$$\|\boldsymbol{G} - \boldsymbol{G}_\star\| \lesssim \frac{\sqrt{n_1n_2n_3} + n\log n}{m}\|\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}^2 + \sqrt{\frac{n\log n}{m}}\|\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}\sigma_{\max}(\boldsymbol{\mathcal{X}}_\star), \tag{C.35}$$

whose proof is deferred to Appendix C.4.2. Under the sample size condition

$$m \gtrsim \epsilon_0^{-1}\sqrt{n_1n_2n_3}r^{3/2}\kappa^2 + \epsilon_0^{-2}(nr^2\kappa^4\log n + r^4\kappa^2)$$

for some small constant ϵ_0, we have $\|\boldsymbol{G} - \boldsymbol{G}_\star\| \le \epsilon_0\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_\star)$, which implies that $\boldsymbol{G}$ is positive semi-definite. Therefore, the top-r_1 eigenvectors of $\boldsymbol{G}$ coincide with $\boldsymbol{U}_0$, the top-r_1 left singular vectors of $\mathcal{M}_1(\mathcal{A}^*(\boldsymbol{y}))$, which implies $\|\boldsymbol{P}_{U_\perp}\boldsymbol{G}\| = \sigma_{r_1+1}(\boldsymbol{G})$. By the triangle inequality, we obtain

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

where the second line follows from Weyl's inequality and that $\boldsymbol{G}_\star$ has rank r_1. In total, the second term of (C.34) is bounded by

$$\|\boldsymbol{P}_{U_\perp}\mathcal{M}_1(\boldsymbol{\mathcal{X}}_\star)\|_{\mathsf{F}} \le \frac{2\sqrt{r_1}}{\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)}\|\boldsymbol{G}-\boldsymbol{G}_\star\| \lesssim \left(\frac{(\sqrt{n_1n_2n_3}+n\log n)r^{3/2}}{m}+\sqrt{\frac{nr^2\log n}{m}}\right)\kappa^2\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star).$$

Completing the proof. The third and fourth terms of (C.34) can be bounded similarly. In all, we conclude that

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

under the assumed sample size.

Proof of (C.35)

We start with stating a few useful concentration inequalities.

Lemma 41. *Suppose that $\boldsymbol{A}_i \in \mathbb{R}^{n_1\times n_2}$ has i.i.d. $\mathcal{N}(0,1/m)$ entries, and $\boldsymbol{y}_i = \langle \boldsymbol{A}_i, \boldsymbol{X}\rangle$ for a fixed $\boldsymbol{X}\in\mathbb{R}^{n_1\times n_2}$, $i=1,\dots,m$. Further suppose that $\boldsymbol{B}\in\mathbb{R}^{n_1\times n_2}$ has i.i.d. $\mathcal{N}(0,\sigma^2)$ entries. Then there exists a universal constant $C>0$ such that for any $t>0$, the following concentration inequalities hold:*

1. *Gaussian ensemble [ZLRY20, Lemma 4]:*

$$\mathbb{P}\left(\Big\|\sum_{i=1}^m y_i\boldsymbol{A}_i-\boldsymbol{X}\Big\| \ge C\|\boldsymbol{X}\|_{\mathsf{F}}\sqrt{n_1+n_2}\left(\sqrt{\frac{\log(n_1+n_2)+t}{m}}+\frac{\log(n_1+n_2)+t}{m}\right)\right) \le \exp(-t). \tag{C.36}$$

2. *Chi-square upper tail [LM00, Lemma 1]:*

$$\mathbb{P}\left(\|\boldsymbol{y}\|_2^2 \ge \|\boldsymbol{X}\|_{\mathsf{F}}^2\frac{m+2\sqrt{mt}+2t}{m}\right) \le \exp(-t). \tag{C.37}$$

3. Gaussian covariance [CHZ20, Theorem 5]:

$$\mathbb{P}\left(\left\|\boldsymbol{B}\boldsymbol{B}^\top - \mathbb{E}[\boldsymbol{B}\boldsymbol{B}^\top]\right\| \geq C\sigma^2\left((\sqrt{n_1}+\sqrt{n_2}+\sqrt{\log(n_1\wedge n_2)}+\sqrt{t})^2 - n_2\right)\right) \leq \exp(-t). \tag{C.38}$$

We now proceed to prove (C.35). In what follows, we take $t \asymp \log n$, and assume $m \gtrsim \log n$ to keep only the dominant terms when invoking the concentration inequalities in Lemma 41.

Let $\mathcal{M}_1(\boldsymbol{\mathcal{X}}_\star) = \boldsymbol{U}_\star\boldsymbol{\Sigma}_{\star,1}\boldsymbol{R}_\star^\top$ be its rank-r_1 SVD, with $\boldsymbol{R}_\star \in \mathbb{R}^{n_2n_3\times r_1}$ containing right singular vectors. Denote $\boldsymbol{R}_{\star\perp}$ as the orthogonal complement of $\boldsymbol{R}_\star$. We have the following decomposition

$$\mathcal{M}_1(\mathcal{A}^*(\boldsymbol{y}))\mathcal{M}_1(\mathcal{A}^*(\boldsymbol{y}))^\top = \mathcal{M}_1(\mathcal{A}^*(\boldsymbol{y}))\boldsymbol{R}_\star\boldsymbol{R}_\star^\top\mathcal{M}_1(\mathcal{A}^*(\boldsymbol{y}))^\top + \mathcal{M}_1(\mathcal{A}^*(\boldsymbol{y}))\boldsymbol{R}_{\star\perp}\boldsymbol{R}_{\star\perp}^\top\mathcal{M}_1(\mathcal{A}^*(\boldsymbol{y}))^\top.$$

By the triangle inequality, we bound

$$\begin{aligned}\|\boldsymbol{G}-\boldsymbol{G}_\star\| &\leq \left\|\mathcal{M}_1(\mathcal{A}^*(\boldsymbol{y}))\boldsymbol{R}_\star\boldsymbol{R}_\star^\top\mathcal{M}_1(\mathcal{A}^*(\boldsymbol{y}))^\top - \mathcal{M}_1(\boldsymbol{\mathcal{X}}_\star)\mathcal{M}_1(\boldsymbol{\mathcal{X}}_\star)^\top\right\| \\ &\quad + \underbrace{\left\|\mathcal{M}_1(\mathcal{A}^*(\boldsymbol{y}))\boldsymbol{R}_{\star\perp}\boldsymbol{R}_{\star\perp}^\top\mathcal{M}_1(\mathcal{A}^*(\boldsymbol{y}))^\top - \frac{\|\boldsymbol{y}\|_2^2}{m}(n_2n_3-r_1)\boldsymbol{I}_{n_1}\right\|}_{=:\mathfrak{A}_2} \\ &\leq \underbrace{\left\|\mathcal{M}_1(\mathcal{A}^*(\boldsymbol{y}))\boldsymbol{R}_\star - \boldsymbol{U}_\star\boldsymbol{\Sigma}_{\star,1}\right\|^2}_{=(\mathfrak{A}_1)^2} + 2\underbrace{\left\|\mathcal{M}_1(\mathcal{A}^*(\boldsymbol{y}))\boldsymbol{R}_\star - \boldsymbol{U}_\star\boldsymbol{\Sigma}_{\star,1}\right\|}_{=:\mathfrak{A}_1}\sigma_{\max}(\boldsymbol{\mathcal{X}}_\star) + \mathfrak{A}_2. \end{aligned}\tag{C.39}$$

Here, the second line follows by applying the triangle inequality to the relation

$$\begin{aligned}&\mathcal{M}_1(\mathcal{A}^*(\boldsymbol{y}))\boldsymbol{R}_\star\boldsymbol{R}_\star^\top\mathcal{M}_1(\mathcal{A}^*(\boldsymbol{y}))^\top - \mathcal{M}_1(\boldsymbol{\mathcal{X}}_\star)\mathcal{M}_1(\boldsymbol{\mathcal{X}}_\star)^\top = \mathcal{M}_1(\mathcal{A}^*(\boldsymbol{y}))\boldsymbol{R}_\star\boldsymbol{R}_\star^\top\mathcal{M}_1(\mathcal{A}^*(\boldsymbol{y}))^\top - \boldsymbol{U}_\star\boldsymbol{\Sigma}_{\star,1}^2\boldsymbol{U}_\star^\top \\ &= \left(\mathcal{M}_1(\mathcal{A}^*(\boldsymbol{y}))\boldsymbol{R}_\star - \boldsymbol{U}_\star\boldsymbol{\Sigma}_{\star,1}\right)\left(\mathcal{M}_1(\mathcal{A}^*(\boldsymbol{y}))\boldsymbol{R}_\star - \boldsymbol{U}_\star\boldsymbol{\Sigma}_{\star,1}\right)^\top + \boldsymbol{U}_\star\boldsymbol{\Sigma}_{\star,1}\left(\mathcal{M}_1(\mathcal{A}^*(\boldsymbol{y}))\boldsymbol{R}_\star - \boldsymbol{U}_\star\boldsymbol{\Sigma}_{\star,1}\right)^\top \\ &\quad + \left(\mathcal{M}_1(\mathcal{A}^*(\boldsymbol{y}))\boldsymbol{R}_\star - \boldsymbol{U}_\star\boldsymbol{\Sigma}_{\star,1}\right)\left(\boldsymbol{U}_\star\boldsymbol{\Sigma}_{\star,1}\right)^\top.\end{aligned}$$

We proceed to bound the terms in (C.39) separately.

- For the first term $\mathfrak{A}_1$, we can expand

$$\mathcal{M}_1(\mathcal{A}^*(\boldsymbol{y}))\boldsymbol{R}_\star = \sum_{i=1}^m y_i \mathcal{M}_1(\boldsymbol{\mathcal{A}}_i)\boldsymbol{R}_\star,$$

where $\mathcal{M}_1(\boldsymbol{\mathcal{A}}_i)\boldsymbol{R}_\star \in \mathbb{R}^{n_1 \times r_1}$ has i.i.d. $\mathcal{N}(0, 1/m)$ entries, and

$$y_i = \langle \mathcal{M}_1(\boldsymbol{\mathcal{A}}_i)\boldsymbol{R}_\star, \boldsymbol{U}_\star \boldsymbol{\Sigma}_{\star,1} \rangle \sim \mathcal{N}(0, \|\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}^2/m).$$

Apply inequality (C.36) in Lemma 41 to obtain with overwhelming probability that

$$\mathfrak{A}_1 = \left\| \sum_{i=1}^m y_i \mathcal{M}_1(\boldsymbol{\mathcal{A}}_i)\boldsymbol{R}_\star - \boldsymbol{U}_\star \boldsymbol{\Sigma}_{\star,1} \right\| \lesssim \sqrt{\frac{n \log n}{m}} \|\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}. \tag{C.40}$$

- Regarding the second term $\mathfrak{A}_2$, one has

$$\mathcal{M}_1(\mathcal{A}^*(\boldsymbol{y}))\boldsymbol{R}_{\star\perp} = \sum_{i=1}^m y_i \mathcal{M}_1(\boldsymbol{\mathcal{A}}_i)\boldsymbol{R}_{\star\perp}.$$

By construction, y_i is independent of $\mathcal{M}_1(\boldsymbol{\mathcal{A}}_i)\boldsymbol{R}_{\star\perp}$. Therefore, conditioned on $\boldsymbol{y}$, $\mathcal{M}_1(\mathcal{A}^*(\boldsymbol{y}))\boldsymbol{R}_{\star\perp} \in \mathbb{R}^{n_1 \times (n_2 n_3 - r_1)}$ is a random matrix with i.i.d. $\mathcal{N}(0, \|\boldsymbol{y}\|_2^2/m)$ entries. We can apply inequality (C.38) in Lemma 41 to obtain with overwhelming probability that

$$\begin{aligned}
\mathfrak{A}_2 &\lesssim \frac{\|\boldsymbol{y}\|_2^2}{m} \left((\sqrt{n_1} + \sqrt{n_2 n_3 - r_1} + c\sqrt{\log n})^2 - (n_2 n_3 - r_1) \right) \\
&\lesssim \frac{\|\boldsymbol{y}\|_2^2}{m} \left(\sqrt{n_1 n_2 n_3} + n\sqrt{\log n} \right).
\end{aligned}$$

Inequality (C.37) in Lemma 41 tells that $\|\boldsymbol{y}\|_2^2 \lesssim \|\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}^2$ with overwhelming probability, which implies

$$\mathfrak{A}_2 \lesssim \frac{\sqrt{n_1 n_2 n_3} + n\sqrt{\log n}}{m} \|\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}^2. \tag{C.41}$$

Finally, plug the bounds (C.40) and (C.41) into (C.39) to conclude

$$\|\boldsymbol{G} - \boldsymbol{G}_\star\| \lesssim \frac{\sqrt{n_1 n_2 n_3} + n \log n}{m} \|\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}^2 + \sqrt{\frac{n \log n}{m}} \|\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \sigma_{\max}(\boldsymbol{\mathcal{X}}_\star).$$

Appendix D

Proofs for Robust Low-rank Tensor Estimation

Lemma 42. *Suppose that $f : \mathbb{R}^{n_1 \times n_2 \times n_3} \mapsto \mathbb{R}$ is convex and rank-$\boldsymbol{r}$ restricted L-Lipschitz continuous (cf. Definition 13). Then for any subgradient $\boldsymbol{\mathcal{G}} \in \partial f(\boldsymbol{\mathcal{X}})$ and any $\widetilde{\boldsymbol{\mathcal{X}}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ with multilinear rank at most $2\boldsymbol{r}$, one has*

$$|\langle \boldsymbol{\mathcal{G}}, \widetilde{\boldsymbol{\mathcal{X}}} \rangle| \leq L \|\widetilde{\boldsymbol{\mathcal{X}}}\|_{\mathsf{F}}.$$

Remark 10. When $f(\cdot)$ satisfies the usual L-Lipschitz continuity, i.e. without the rank restriction, the statement degenerates into $\|\boldsymbol{\mathcal{G}}\|_{\mathsf{F}} \leq L$.

Proof. Fix any subgradient $\boldsymbol{\mathcal{G}} \in \partial f(\boldsymbol{\mathcal{X}})$. By the definition of a subgradient, for any $\widetilde{\boldsymbol{\mathcal{X}}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, one has

$$f(\boldsymbol{\mathcal{X}} + \widetilde{\boldsymbol{\mathcal{X}}}) \geq f(\boldsymbol{\mathcal{X}}) + \langle \boldsymbol{\mathcal{G}}, \widetilde{\boldsymbol{\mathcal{X}}} \rangle, \tag{D.1}$$

By the rank-$\boldsymbol{r}$ restricted L-Lipschitz continuity of $f(\cdot)$, when $\widetilde{\boldsymbol{\mathcal{X}}}$ has multilinear rank at most $2\boldsymbol{r}$, one has

$$f(\boldsymbol{\mathcal{X}} + \widetilde{\boldsymbol{\mathcal{X}}}) - f(\boldsymbol{\mathcal{X}}) \leq L \|\widetilde{\boldsymbol{\mathcal{X}}}\|_{\mathsf{F}}.$$

This proof is complete by combining the above inequality with (D.1). □

D.1 Proof of Theorem 11

We prove the theorem by induction, where the base case is established trivially by the initial condition. Suppose that the t-th iterate $\boldsymbol{F}_t$ obeys the condition

$$\operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \le (1 - 0.016/\chi_f^2)^{t/2} 10^{-3} \sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)/\chi_f \le 10^{-3}\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)/\chi_f. \tag{D.2}$$

Lemma 31 ensures that the optimal alignment matrices $\{\boldsymbol{Q}_{t,k}\}_{k=1,2,3}$ between $\boldsymbol{F}_t$ and $\boldsymbol{F}_\star$ exists. For notational convenience, we denote $\epsilon \coloneqq 10^{-3}/\chi_f$,

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

and adopt the notations in (C.4). The relation $\|\boldsymbol{\mathcal{X}}_t - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \le 3\operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$ follows from (C.7). From now on, we focus on proving the distance contraction. By the definition of $\operatorname{dist}(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star)$, one has

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

We expand the first square in (D.3) as

$$\begin{aligned}\|(\boldsymbol{U}_{t+1}\boldsymbol{Q}_{t,1} - \boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 &= \left\|\left(\boldsymbol{U} - \eta_t \mathcal{M}_1(\boldsymbol{\mathcal{G}})\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1} - \boldsymbol{U}_\star\right)\boldsymbol{\Sigma}_{\star,1}\right\|_{\mathsf{F}}^2 \\ &= \|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 - 2\eta_t\left\langle \boldsymbol{\Delta}_U, \mathcal{M}_1(\boldsymbol{\mathcal{G}})\breve{\boldsymbol{U}}_\star\right\rangle + \eta_t^2 \underbrace{\left\|\mathcal{M}_1(\boldsymbol{\mathcal{G}})\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}\right\|_{\mathsf{F}}^2}_{\mathfrak{U}_1} \\ &\quad - 2\eta_t \underbrace{\left\langle \boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}, \mathcal{M}_1(\boldsymbol{\mathcal{G}})\left(\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1} - \breve{\boldsymbol{U}}_\star\boldsymbol{\Sigma}_{\star,1}^{-1}\right)\right\rangle}_{\mathfrak{U}_2},\end{aligned} \tag{D.4}$$

where in the first line, we used the fact that the update rule (5.7) is covariant with respect to $\{\boldsymbol{Q}_{t,k}\}$, implying that

$$\boldsymbol{U}_{t+1}\boldsymbol{Q}_{t,1} = \boldsymbol{U} - \eta_t \mathcal{M}_1(\boldsymbol{\mathcal{G}})\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}.$$

We proceed to bound $\mathfrak{U}_1$ and $\mathfrak{U}_2$. For $\mathfrak{U}_1$, use Lemma 34 (C.5f) to obtain

$$\begin{aligned}
\mathfrak{U}_1 &\le \left\|\mathcal{M}_1(\boldsymbol{\mathcal{G}})\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1/2}\right\|_{\mathsf{F}}^2 \left\|(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1/2}\boldsymbol{\Sigma}_{\star,1}\right\|^2 \\
&= \left\|\mathcal{M}_1(\boldsymbol{\mathcal{G}})\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1/2}\right\|_{\mathsf{F}}^2 \left\|\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1}\right\|^2 \\
&\le \left\|\mathcal{M}_1(\boldsymbol{\mathcal{G}})\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1/2}\right\|_{\mathsf{F}}^2 \frac{1}{(1-\epsilon)^6}.
\end{aligned}$$

For $\mathfrak{U}_2$, note that

$$\mathfrak{U}_2 = \left\langle \boldsymbol{\mathcal{G}}, (\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}^2\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}, \boldsymbol{V}, \boldsymbol{W})\cdot\boldsymbol{\mathcal{C}} - (\boldsymbol{\Delta}_U, \boldsymbol{V}_\star, \boldsymbol{W}_\star)\cdot\boldsymbol{\mathcal{C}}_\star \right\rangle,$$

is the inner product of the subgradient $\boldsymbol{\mathcal{G}}$ and a tensor with multiliear rank at most $2\boldsymbol{r}$, thus combining Lemmas 34 (C.5g) and 42 one has

$$\begin{aligned}
|\mathfrak{U}_2| &\le L\left\|(\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}^2\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}, \boldsymbol{V}, \boldsymbol{W})\cdot\boldsymbol{\mathcal{C}} - (\boldsymbol{\Delta}_U, \boldsymbol{V}_\star, \boldsymbol{W}_\star)\cdot\boldsymbol{\mathcal{C}}_\star\right\|_{\mathsf{F}} \\
&\le L\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}\left\|\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1}\boldsymbol{\Sigma}_{\star,1} - \breve{\boldsymbol{U}}_\star\boldsymbol{\Sigma}_{\star,1}^{-1}\right\| \\
&\le L\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}\frac{\sqrt{2}(3\epsilon+3\epsilon^2+\epsilon^3)}{(1-\epsilon)^3}.
\end{aligned}$$

Plugging collectively the bounds for $\mathfrak{U}_1$ and $\mathfrak{U}_2$ into (D.4) yields

$$\begin{aligned}
\|(\boldsymbol{U}_{t+1}\boldsymbol{Q}_{t,1} - \boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 \le \|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 - 2\eta_t\left\langle \boldsymbol{\mathcal{G}}, (\boldsymbol{\Delta}_U, \boldsymbol{V}_\star, \boldsymbol{W}_\star)\cdot\boldsymbol{\mathcal{C}}_\star\right\rangle + \frac{\eta_t^2}{(1-\epsilon)^6}\left\|\mathcal{M}_1(\boldsymbol{\mathcal{G}})\breve{\boldsymbol{U}}(\breve{\boldsymbol{U}}^\top\breve{\boldsymbol{U}})^{-1/2}\right\|_{\mathsf{F}}^2 \\
+ 2\eta_t L\frac{\sqrt{2}(3\epsilon+3\epsilon^2+\epsilon^3)}{(1-\epsilon)^3}\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}.
\end{aligned}$$

Similarly, other terms in (D.3) can be expanded and bounded as

$$\begin{aligned}
\|(\boldsymbol{V}_{t+1}\boldsymbol{Q}_{t,2}-\boldsymbol{V}_\star)\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}^2 &\le \|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}^2 - 2\eta_t\left\langle \boldsymbol{\mathcal{G}}, (\boldsymbol{U}_\star, \boldsymbol{\Delta}_V, \boldsymbol{W}_\star)\cdot\boldsymbol{\mathcal{C}}_\star\right\rangle \\
&\quad + \frac{\eta_t^2}{(1-\epsilon)^6}\left\|\mathcal{M}_2(\boldsymbol{\mathcal{G}})\breve{\boldsymbol{V}}(\breve{\boldsymbol{V}}^\top\breve{\boldsymbol{V}})^{-1/2}\right\|_{\mathsf{F}}^2 + 2\eta_t L\frac{\sqrt{2}(3\epsilon+3\epsilon^2+\epsilon^3)}{(1-\epsilon)^3}\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}; \\
\|(\boldsymbol{W}_{t+1}\boldsymbol{Q}_{t,3}-\boldsymbol{W}_\star)\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}^2 &\le \|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}^2 - 2\eta_t\left\langle \boldsymbol{\mathcal{G}}, (\boldsymbol{U}_\star, \boldsymbol{V}_\star, \boldsymbol{\Delta}_W)\cdot\boldsymbol{\mathcal{C}}_\star\right\rangle \\
&\quad + \frac{\eta_t^2}{(1-\epsilon)^6}\left\|\mathcal{M}_3(\boldsymbol{\mathcal{G}})\breve{\boldsymbol{W}}(\breve{\boldsymbol{W}}^\top\breve{\boldsymbol{W}})^{-1/2}\right\|_{\mathsf{F}}^2 + 2\eta_t L\frac{\sqrt{2}(3\epsilon+3\epsilon^2+\epsilon^3)}{(1-\epsilon)^3}\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}; \\
\left\|(\boldsymbol{Q}_{t,1}^{-1},\boldsymbol{Q}_{t,2}^{-1},\boldsymbol{Q}_{t,3}^{-1})\cdot\boldsymbol{\mathcal{C}}_{t+1}-\boldsymbol{\mathcal{C}}_\star\right\|_{\mathsf{F}}^2 &\le \|\boldsymbol{\Delta}_{\mathcal{C}}\|_{\mathsf{F}}^2 - 2\eta_t\left\langle \boldsymbol{\mathcal{G}}, (\boldsymbol{U}_\star, \boldsymbol{V}_\star, \boldsymbol{W}_\star)\cdot\boldsymbol{\Delta}_{\mathcal{C}}\right\rangle \\
&\quad + \frac{\eta_t^2}{(1-\epsilon)^6}\left\|\left((\boldsymbol{U}_t^\top\boldsymbol{U}_t)^{-1/2}\boldsymbol{U}_t^\top, (\boldsymbol{V}_t^\top\boldsymbol{V}_t)^{-1/2}\boldsymbol{V}_t^\top, (\boldsymbol{W}_t^\top\boldsymbol{W}_t)^{-1/2}\boldsymbol{W}_t^\top\right)\cdot\boldsymbol{\mathcal{G}}_t\right\|_{\mathsf{F}}^2 \\
&\quad + 2\eta_t L\frac{\sqrt{2}(3\epsilon+3\epsilon^2+\epsilon^3)}{(1-\epsilon)^3}\|\boldsymbol{\Delta}_{\mathcal{C}}\|_{\mathsf{F}}.
\end{aligned}$$

In addition, we claim that

$$\begin{aligned}
&\left\langle \boldsymbol{\mathcal{G}}, (\boldsymbol{\Delta}_U, \boldsymbol{V}_\star, \boldsymbol{W}_\star)\cdot\boldsymbol{\mathcal{C}}_\star + (\boldsymbol{U}_\star, \boldsymbol{\Delta}_V, \boldsymbol{W}_\star)\cdot\boldsymbol{\mathcal{C}}_\star + (\boldsymbol{U}_\star, \boldsymbol{V}_\star, \boldsymbol{\Delta}_W)\cdot\boldsymbol{\mathcal{C}}_\star + (\boldsymbol{U}_\star, \boldsymbol{V}_\star, \boldsymbol{W}_\star)\cdot\boldsymbol{\Delta}_{\mathcal{C}}\right\rangle \\
&\ge \langle \boldsymbol{\mathcal{G}}, \boldsymbol{\mathcal{X}}_t - \boldsymbol{\mathcal{X}}_\star\rangle - L(\frac{3}{2}\epsilon+\epsilon^2+\frac{1}{4}\epsilon^3)\left(\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_{\mathcal{C}}\|_{\mathsf{F}}\right). \quad \text{(D.5)}
\end{aligned}$$

Combine them together to reach that

$$\begin{aligned}
\operatorname{dist}^2(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) &\le \|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_{\mathcal{C}}\|_{\mathsf{F}}^2 - 2\eta_t\langle \boldsymbol{\mathcal{G}}, \boldsymbol{\mathcal{X}}_t - \boldsymbol{\mathcal{X}}_\star\rangle + \frac{\eta_t^2 N_t^2}{(1-\epsilon)^6} \\
&\quad + \eta_t L\left(\frac{2\sqrt{2}(3\epsilon+3\epsilon^2+\epsilon^3)}{(1-\epsilon)^3} + 3\epsilon + 2\epsilon^2 + \frac{1}{2}\epsilon^3\right)\left(\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_{\mathcal{C}}\|_{\mathsf{F}}\right),
\end{aligned}$$

in which N_t^2 is defined in (5.9). Using the subgradient optimality of $\boldsymbol{\mathcal{G}}$, we obtain

$$\langle \boldsymbol{\mathcal{G}}, \boldsymbol{\mathcal{X}}_t - \boldsymbol{\mathcal{X}}_\star\rangle \ge f(\boldsymbol{\mathcal{X}}_t) - f(\boldsymbol{\mathcal{X}}_\star),$$

which further implies that

$$\operatorname{dist}^2(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) \le \operatorname{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star) - 2\eta_t\left(f(\boldsymbol{\mathcal{X}}_t) - f(\boldsymbol{\mathcal{X}}_\star)\right) + \frac{\eta_t^2 N_t^2}{(1-\epsilon)^6} + \eta_t L c(\epsilon)\operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star), \quad \text{(D.6)}$$

where the last term uses the basic inequality

$$\|\boldsymbol{\Delta}_U\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_C\|_{\mathsf{F}} \le 2\,\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star),$$

and for conciseness we abbreviate

$$c(\epsilon) \coloneqq \frac{4\sqrt{2}(3\epsilon + 3\epsilon^2 + \epsilon^3)}{(1-\epsilon)^3} + 6\epsilon + 4\epsilon^2 + \epsilon^3. \tag{D.7}$$

Also, we claim that

$$N_t \le 2L. \tag{D.8}$$

Convergence with Polyak's stepsizes. Let $\eta_t = (f(\boldsymbol{\mathcal{X}}_t) - f(\boldsymbol{\mathcal{X}}_\star))/N_t^2$ be the Polyak's stepsize. Plugging it into (D.6), we have

$$\begin{aligned}\mathrm{dist}^2(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) &\le \mathrm{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star) - \eta_t\left(2 - \frac{1}{(1-\epsilon)^6}\right)(f(\boldsymbol{\mathcal{X}}_t) - f(\boldsymbol{\mathcal{X}}_\star)) + \eta_t L c(\epsilon)\,\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \\ &\le \mathrm{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star) - \eta_t\mu\left((\sqrt{2}-1)^{3/2}\left(2 - \frac{1}{(1-\epsilon)^6}\right) - \chi_f c(\epsilon)\right)\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star),\end{aligned} \tag{D.9}$$

where the second inequality follows from (D.10) and $\chi_f = L/\mu$.

The restricted μ-sharpness of $f(\cdot)$ together with Lemma 33 yields

$$f(\boldsymbol{\mathcal{X}}_t) - f(\boldsymbol{\mathcal{X}}_\star) \ge \mu\|\boldsymbol{\mathcal{X}}_t - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}} \ge \mu(\sqrt{2}-1)^{3/2}\,\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star). \tag{D.10}$$

To continue, combining (D.10) and (D.8), we can lower bound the Polyak's stepsize as

$$\eta_t \ge \frac{(\sqrt{2}-1)^{3/2}\mu\,\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)}{4L^2}.$$

This, combined with (D.9), leads to

$$\mathrm{dist}^2(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) \le \left(1 - \frac{(\sqrt{2}-1)^{3/2}}{4\chi_f^2}\left((\sqrt{2}-1)^{3/2}\left(2 - \frac{1}{(1-\epsilon)^6}\right) - \chi_f c(\epsilon)\right)\right)\mathrm{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star),$$

Under the condition $\epsilon = 10^{-3}/\chi_f$, we calculate that

$$\begin{aligned}&\frac{(\sqrt{2}-1)^{3/2}}{4}\left((\sqrt{2}-1)^{3/2}\left(2-\frac{1}{(1-\epsilon)^6}\right)-\chi_f c(\epsilon)\right)\\&\quad\geq\frac{(\sqrt{2}-1)^{3/2}}{4}\left((\sqrt{2}-1)^{3/2}\left(2-\frac{1}{(1-10^{-3})^6}\right)-c(10^{-3})\right)\geq 0.016,\end{aligned}$$

thus we conclude that

$$\mathrm{dist}^2(\boldsymbol{F}_{t+1},\boldsymbol{F}_\star)\leq(1-0.016/\chi_f^2)\,\mathrm{dist}^2(\boldsymbol{F}_t,\boldsymbol{F}_\star),$$

which is the desired claim.

Convergence with geometrically decaying stepsizes. Let $\eta_t = \lambda q^t/N_t$ be the geometrically decaying stepsize in (5.8). Plugging it into (D.6), we have

$$\begin{aligned}\mathrm{dist}^2(\boldsymbol{F}_{t+1},\boldsymbol{F}_\star)&\leq\mathrm{dist}^2(\boldsymbol{F}_t,\boldsymbol{F}_\star)-\eta_t\mu\left(2(\sqrt{2}-1)^{3/2}-\chi_f c(\epsilon)\right)\mathrm{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star)+\frac{\lambda^2q^{2t}}{(1-\epsilon)^6}\\&\leq\mathrm{dist}^2(\boldsymbol{F}_t,\boldsymbol{F}_\star)-\frac{\lambda q^t}{2\chi_f}\left(2(\sqrt{2}-1)^{3/2}-\chi_f c(\epsilon)\right)\mathrm{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star)+\frac{\lambda^2q^{2t}}{(1-\epsilon)^6},\end{aligned}$$

where the first line follows from (D.10) and $\chi_f = L/\mu$, and the second line follows from $\eta_t \geq \frac{\lambda q^t}{2L}$ due to (D.8). The induction hypothesis at the t-iteration

$$\mathrm{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star)\leq(1-0.016/\chi_f^2)^{t/2}10^{-3}\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)/\chi_f,$$

combined with the setting of parameters, i.e.

$$\lambda q^t=\frac{(\sqrt{2}-1)^{3/2}}{2}(1-0.016/\chi_f^2)^{t/2}10^{-3}\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)/\chi_f^2,$$

implies that

$$\mathrm{dist}^2(\boldsymbol{F}_{t+1},\boldsymbol{F}_\star)\leq\left(1-\frac{(\sqrt{2}-1)^{3/2}}{4\chi_f^2}\left((\sqrt{2}-1)^{3/2}\left(2-\frac{1}{(1-\epsilon)^6}\right)-\chi_f c(\epsilon)\right)\right)$$

$$(1-0.016/\chi_f^2)^t(10^{-3}\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)/\chi_f)^2,$$

where the contraction rate matches exactly as that using Polyak's stepsize. Therefore, under the condition $\epsilon = 10^{-3}/\chi_f$, we conclude that

$$\mathrm{dist}(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) \le (1-0.016/\chi_f^2)^{\frac{t+1}{2}} 10^{-3}\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)/\chi_f,$$

which is the desired claim.

Proof of (D.5). To prove (D.5), we use the decomposition

$$\begin{aligned}(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\cdot\boldsymbol{\mathcal{C}} - \boldsymbol{\mathcal{X}}_\star = {} & (\boldsymbol{\Delta}_U,\boldsymbol{V}_\star,\boldsymbol{W}_\star)\cdot\boldsymbol{\mathcal{C}}_\star + (\boldsymbol{U}_\star,\boldsymbol{\Delta}_V,\boldsymbol{W}_\star)\cdot\boldsymbol{\mathcal{C}}_\star + (\boldsymbol{U}_\star,\boldsymbol{V}_\star,\boldsymbol{\Delta}_W)\cdot\boldsymbol{\mathcal{C}}_\star + (\boldsymbol{U}_\star,\boldsymbol{V}_\star,\boldsymbol{W}_\star)\cdot\boldsymbol{\Delta}_{\mathcal{C}} \\ & + (\boldsymbol{\Delta}_U,\boldsymbol{\Delta}_V,\boldsymbol{W}_\star)\cdot\boldsymbol{\mathcal{C}}_\star + (\boldsymbol{\Delta}_U,\boldsymbol{V}_\star,\boldsymbol{\Delta}_W)\cdot\boldsymbol{\mathcal{C}}_\star + (\boldsymbol{U},\boldsymbol{\Delta}_V,\boldsymbol{\Delta}_W)\cdot\boldsymbol{\mathcal{C}}_\star \\ & + (\boldsymbol{\Delta}_U,\boldsymbol{V}_\star,\boldsymbol{W}_\star)\cdot\boldsymbol{\Delta}_{\mathcal{C}} + (\boldsymbol{U},\boldsymbol{\Delta}_V,\boldsymbol{W}_\star)\cdot\boldsymbol{\Delta}_{\mathcal{C}} + (\boldsymbol{U},\boldsymbol{V},\boldsymbol{\Delta}_W)\cdot\boldsymbol{\Delta}_{\mathcal{C}}, \qquad \text{(D.11)}\end{aligned}$$

and then invoke Lemma 42 to see

$$\begin{aligned}&\Big\langle \boldsymbol{\mathcal{G}}_t, \boldsymbol{\mathcal{X}}_t - \boldsymbol{\mathcal{X}}_\star - (\boldsymbol{\Delta}_U,\boldsymbol{V}_\star,\boldsymbol{W}_\star)\cdot\boldsymbol{\mathcal{C}}_\star - (\boldsymbol{U}_\star,\boldsymbol{\Delta}_V,\boldsymbol{W}_\star)\cdot\boldsymbol{\mathcal{C}}_\star - (\boldsymbol{U}_\star,\boldsymbol{V}_\star,\boldsymbol{\Delta}_W)\cdot\boldsymbol{\mathcal{C}}_\star - (\boldsymbol{U}_\star,\boldsymbol{V}_\star,\boldsymbol{W}_\star)\cdot\boldsymbol{\Delta}_{\mathcal{C}}\Big\rangle \\ &\quad \le L\Big(\|(\boldsymbol{\Delta}_U,\boldsymbol{\Delta}_V,\boldsymbol{W}_\star)\cdot\boldsymbol{\mathcal{C}}_\star\|_{\mathsf{F}} + \|(\boldsymbol{\Delta}_U,\boldsymbol{V}_\star,\boldsymbol{\Delta}_W)\cdot\boldsymbol{\mathcal{C}}_\star\|_{\mathsf{F}} + \|(\boldsymbol{U},\boldsymbol{\Delta}_V,\boldsymbol{\Delta}_W)\cdot\boldsymbol{\mathcal{C}}_\star\|_{\mathsf{F}} \\ &\qquad\quad \|(\boldsymbol{\Delta}_U,\boldsymbol{V}_\star,\boldsymbol{W}_\star)\cdot\boldsymbol{\Delta}_{\mathcal{C}}\|_{\mathsf{F}} + \|(\boldsymbol{U},\boldsymbol{\Delta}_V,\boldsymbol{W}_\star)\cdot\boldsymbol{\Delta}_{\mathcal{C}}\|_{\mathsf{F}} + \|(\boldsymbol{U},\boldsymbol{V},\boldsymbol{\Delta}_W)\cdot\boldsymbol{\Delta}_{\mathcal{C}}\|_{\mathsf{F}} \Big) \\ &\quad \le L\left(\epsilon\,\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + (2\epsilon+\epsilon^2)\,\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}} + (3\epsilon+3\epsilon^2+\epsilon^3)\|\boldsymbol{\Delta}_{\mathcal{C}}\|_{\mathsf{F}}\right),\end{aligned}$$

where the details in the last inequality are:

$$\begin{aligned}\|(\boldsymbol{\Delta}_U,\boldsymbol{\Delta}_V,\boldsymbol{W}_\star)\cdot\boldsymbol{\mathcal{C}}_\star\|_{\mathsf{F}} &\le \|\boldsymbol{\Delta}_V\mathcal{M}_2(\boldsymbol{\mathcal{C}}_\star)\|_{\mathsf{F}}\|\boldsymbol{W}_\star\|\|\boldsymbol{\Delta}_U\| \le \epsilon\|\boldsymbol{\Delta}_V\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}; \\ \|(\boldsymbol{\Delta}_U,\boldsymbol{V}_\star,\boldsymbol{\Delta}_W)\cdot\boldsymbol{\mathcal{C}}_\star\|_{\mathsf{F}} &\le \|\boldsymbol{\Delta}_W\mathcal{M}_3(\boldsymbol{\mathcal{C}}_\star)\|_{\mathsf{F}}\|\boldsymbol{V}_\star\|\|\boldsymbol{\Delta}_U\| \le \epsilon\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}; \\ \|(\boldsymbol{U},\boldsymbol{\Delta}_V,\boldsymbol{\Delta}_W)\cdot\boldsymbol{\mathcal{C}}_\star\|_{\mathsf{F}} &\le \|\boldsymbol{\Delta}_W\mathcal{M}_3(\boldsymbol{\mathcal{C}}_\star)\|_{\mathsf{F}}\|\boldsymbol{\Delta}_V\|\|\boldsymbol{U}\| \le (1+\epsilon)\epsilon\|\boldsymbol{\Delta}_W\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}; \\ \|(\boldsymbol{\Delta}_U,\boldsymbol{V}_\star,\boldsymbol{W}_\star)\cdot\boldsymbol{\Delta}_{\mathcal{C}}\|_{\mathsf{F}} &\le \|\boldsymbol{\Delta}_U\|\|\boldsymbol{V}_\star\|\|\boldsymbol{W}_\star\|\|\boldsymbol{\Delta}_{\mathcal{C}}\|_{\mathsf{F}} \le \epsilon\|\boldsymbol{\Delta}_{\mathcal{C}}\|_{\mathsf{F}};\end{aligned}$$

$$\|(\boldsymbol{U}, \boldsymbol{\Delta}_V, \boldsymbol{W}_\star) \cdot \boldsymbol{\Delta}_{\mathcal{C}}\|_{\mathsf{F}} \le \|\boldsymbol{U}\| \|\boldsymbol{\Delta}_V\| \|\boldsymbol{W}_\star\| \|\boldsymbol{\Delta}_{\mathcal{C}}\|_{\mathsf{F}} \le (1+\epsilon)\epsilon \|\boldsymbol{\Delta}_{\mathcal{C}}\|_{\mathsf{F}};$$
$$\|(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{\Delta}_W) \cdot \boldsymbol{\Delta}_{\mathcal{C}}\|_{\mathsf{F}} \le \|\boldsymbol{U}\| \|\boldsymbol{V}\| \|\boldsymbol{\Delta}_W\| \|\boldsymbol{\Delta}_{\mathcal{C}}\|_{\mathsf{F}} \le (1+\epsilon)^2 \epsilon \|\boldsymbol{\Delta}_{\mathcal{C}}\|_{\mathsf{F}}.$$

Finally use decomposition other than (D.11), and take an average to balance the coefficients of factors $\boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}, \boldsymbol{\Delta}_V \boldsymbol{\Sigma}_{\star,2}, \boldsymbol{\Delta}_W \boldsymbol{\Sigma}_{\star,3}$ and $\boldsymbol{\Delta}_{\mathcal{C}}$ as

$$\begin{aligned}
&\Big\langle \boldsymbol{\mathcal{G}}_t, \boldsymbol{\mathcal{X}}_t - \boldsymbol{\mathcal{X}}_\star - (\boldsymbol{\Delta}_U, \boldsymbol{V}_\star, \boldsymbol{W}_\star) \cdot \boldsymbol{\mathcal{C}}_\star - (\boldsymbol{U}_\star, \boldsymbol{\Delta}_V, \boldsymbol{W}_\star) \cdot \boldsymbol{\mathcal{C}}_\star - (\boldsymbol{U}_\star, \boldsymbol{V}_\star, \boldsymbol{\Delta}_W) \cdot \boldsymbol{\mathcal{C}}_\star - (\boldsymbol{U}_\star, \boldsymbol{V}_\star, \boldsymbol{W}_\star) \cdot \boldsymbol{\Delta}_{\mathcal{C}} \Big\rangle \\
&\quad \le L\big(\frac{3}{2}\epsilon + \epsilon^2 + \frac{1}{4}\epsilon^3\big) \left(\|\boldsymbol{\Delta}_U \boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_V \boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_W \boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_{\mathcal{C}}\|_{\mathsf{F}}\right).
\end{aligned}$$

Proof of (D.8). The proof is established by repeatedly applying Lemma 42 to each term in (5.9). For example, the first term in (5.9) can be written in the variational form as

$$\left\|\mathcal{M}_1(\boldsymbol{\mathcal{G}}) \breve{\boldsymbol{U}} (\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1/2}\right\|_{\mathsf{F}} = \left\langle \boldsymbol{\mathcal{G}}, (\widetilde{\boldsymbol{U}} (\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1/2}, \boldsymbol{V}, \boldsymbol{W}) \cdot \boldsymbol{\mathcal{C}} \right\rangle,$$

for some $\widetilde{\boldsymbol{U}} \in \mathbb{R}^{n_1 \times r}$ with $\|\widetilde{\boldsymbol{U}}\|_{\mathsf{F}} = 1$. Since $(\widetilde{\boldsymbol{U}} (\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1/2}, \boldsymbol{V}, \boldsymbol{W}) \cdot \boldsymbol{\mathcal{C}}$ has multilinear rank at most $\boldsymbol{r}$, Lemma 42 tells that

$$\begin{aligned}
\left\|\mathcal{M}_1(\boldsymbol{\mathcal{G}}) \breve{\boldsymbol{U}} (\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1/2}\right\|_{\mathsf{F}} &\le L \left\|(\widetilde{\boldsymbol{U}} (\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1/2}, \boldsymbol{V}, \boldsymbol{W}) \cdot \boldsymbol{\mathcal{C}}\right\|_{\mathsf{F}} \\
&\le L \left\|\breve{\boldsymbol{U}} (\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1/2}\right\| = L,
\end{aligned}$$

where the last equality follows from

$$\|\breve{\boldsymbol{U}} (\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1/2}\|^2 = \|\breve{\boldsymbol{U}} (\breve{\boldsymbol{U}}^\top \breve{\boldsymbol{U}})^{-1} \breve{\boldsymbol{U}}^\top\| = 1.$$

D.2 Proof of Proposition 5

We shall prove a more detailed statement: for all tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ with multilinear rank at most $\boldsymbol{r}$, with probability exceeding $1-\rho$, one has:

$$\left(\sqrt{\frac{2}{\pi}} - \delta\right) \|\boldsymbol{\mathcal{X}}\|_{\mathsf{F}} \le \|\mathcal{A}(\boldsymbol{\mathcal{X}})\|_1 \le \left(\sqrt{\frac{2}{\pi}} + \delta\right) \|\boldsymbol{\mathcal{X}}\|_{\mathsf{F}}, \quad \text{for any} \quad 0 < \delta < \sqrt{\frac{2}{\pi}}; \tag{D.12}$$

$$\|\mathcal{A}_{\mathcal{S}^c}(\boldsymbol{\mathcal{X}})\|_1 - \|\mathcal{A}_{\mathcal{S}}(\boldsymbol{\mathcal{X}})\|_1 \ge \left((1-2p_s)\sqrt{\frac{2}{\pi}} - \delta\right) \|\boldsymbol{\mathcal{X}}\|_{\mathsf{F}}, \quad \text{for any} \quad 0 < \delta < (1-2p_s)\sqrt{\frac{2}{\pi}}, \tag{D.13}$$

as long as

$$m \ge \frac{(3nr + r^3)\log(120/\delta) + \log(2/\rho)}{c\delta^2},$$

where $c > 0$ is some constant. A key ingredient is the following result on covering number of the set of unit Frobenius norm low-rank tensors.

Lemma 43 ([RSS17, Lemma 2]). *Denote the set of unit Frobenius norm rank-$\boldsymbol{r}$ tensors as*

$$\mathbb{S}_{\boldsymbol{r}} = \{\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1 \times n_3 \times n_3} : \operatorname{rank}(\boldsymbol{\mathcal{X}}) \le \boldsymbol{r}, \|\boldsymbol{\mathcal{X}}\|_{\mathsf{F}} = 1\}.$$

The ϵ-covering number of $\mathbb{S}_{\boldsymbol{r}}$ with respect to the Frobenius norm is bounded by

$$|\mathbb{S}_{\boldsymbol{r},\epsilon}| \le \left(\frac{12}{\epsilon}\right)^{n_1 r_1 + n_2 r_2 + n_3 r_3 + r_1 r_2 r_3},$$

where $\mathbb{S}_{\boldsymbol{r},\epsilon}$ denotes the ϵ-net of $\mathbb{S}_{\boldsymbol{r}}$.

The proof to (D.12) essentially repeats [LZMCSV20, Proposition 1]. First fix any $\boldsymbol{\mathcal{X}} \in \mathbb{S}_{\boldsymbol{r}}$. Since $\boldsymbol{\mathcal{A}}_i$ has i.i.d. standard Gaussian entries, $\langle \boldsymbol{\mathcal{A}}_i, \boldsymbol{\mathcal{X}} \rangle$ obeys standard Gaussian. $\mathbb{E}|\langle \boldsymbol{\mathcal{A}}_i, \boldsymbol{\mathcal{X}} \rangle| = \sqrt{\frac{2}{\pi}}$, and $|\langle \boldsymbol{\mathcal{A}}_i, \boldsymbol{\mathcal{X}} \rangle| - \sqrt{\frac{2}{\pi}}$ is sub-Gaussian. Hoeffding inequality for sub-Gaussian random variables tells that

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m} |\langle \boldsymbol{\mathcal{A}}_i, \boldsymbol{\mathcal{X}} \rangle| - \sqrt{\frac{2}{\pi}}\right| \le \frac{\delta}{2}\right) \ge 1 - 2\exp(-cm\delta^2),$$

for some constant $c > 0$.

Next apply the covering argument. For any $\boldsymbol{\mathcal{X}} \in \mathbb{S}_{\boldsymbol{r}}$, there exists a $\bar{\boldsymbol{\mathcal{X}}} \in \mathbb{S}_{\boldsymbol{r},\epsilon}$ such that $\|\boldsymbol{\mathcal{X}} - \bar{\boldsymbol{\mathcal{X}}}\|_{\mathsf{F}} \le \epsilon$. Invoke the triangle inequality to see

$$\left| \|\mathcal{A}(\boldsymbol{\mathcal{X}})\|_1 - \sqrt{\frac{2}{\pi}} \right| \le \left| \|\mathcal{A}(\bar{\boldsymbol{\mathcal{X}}})\|_1 - \sqrt{\frac{2}{\pi}} \right| + \|\mathcal{A}(\boldsymbol{\mathcal{X}} - \bar{\boldsymbol{\mathcal{X}}})\|_1.$$

Since $\boldsymbol{\mathcal{X}} - \bar{\boldsymbol{\mathcal{X}}}$ has multilinear rank at most $2\boldsymbol{r}$, we can split it into 8 rank-$\boldsymbol{r}$ tensors: $\boldsymbol{\mathcal{X}} - \bar{\boldsymbol{\mathcal{X}}} = \sum_{i=1}^{8} \boldsymbol{\mathcal{D}}_i$ with each $\boldsymbol{\mathcal{D}}_i$ orthogonal to each other.[1] We can further require that $\|\boldsymbol{\mathcal{D}}_i\|_{\mathsf{F}} \le \epsilon/\sqrt{8}$. It holds that

$$\left| \|\mathcal{A}(\boldsymbol{\mathcal{X}})\|_1 - \sqrt{\frac{2}{\pi}} \right| \le \left| \|\mathcal{A}(\bar{\boldsymbol{\mathcal{X}}})\|_1 - \sqrt{\frac{2}{\pi}} \right| + \sum_{i=1}^{8} \|\mathcal{A}(\boldsymbol{\mathcal{D}}_i)\|_1 \le \left| \|\mathcal{A}(\bar{\boldsymbol{\mathcal{X}}})\|_1 - \sqrt{\frac{2}{\pi}} \right| + \sqrt{8}\epsilon \sup_{\boldsymbol{\mathcal{X}} \in \mathbb{S}_{\boldsymbol{r}}} \|\mathcal{A}(\boldsymbol{\mathcal{X}})\|_1.$$

Take supreme on both sides to see

$$\begin{aligned} &\sup_{\boldsymbol{\mathcal{X}} \in \mathbb{S}_{\boldsymbol{r}}} \left| \|\mathcal{A}(\boldsymbol{\mathcal{X}})\|_1 - \sqrt{\frac{2}{\pi}} \right| \le \sup_{\bar{\boldsymbol{\mathcal{X}}} \in \mathbb{S}_{\boldsymbol{r},\epsilon}} \left| \|\mathcal{A}(\bar{\boldsymbol{\mathcal{X}}})\|_1 - \sqrt{\frac{2}{\pi}} \right| + \sqrt{8}\epsilon \sup_{\boldsymbol{\mathcal{X}} \in \mathbb{S}_{\boldsymbol{r}}} \|\mathcal{A}(\boldsymbol{\mathcal{X}})\|_1 \\ \implies &\sup_{\boldsymbol{\mathcal{X}} \in \mathbb{S}_{\boldsymbol{r}}} \left| \|\mathcal{A}(\boldsymbol{\mathcal{X}})\|_1 - \sqrt{\frac{2}{\pi}} \right| \le \frac{\sup_{\bar{\boldsymbol{\mathcal{X}}} \in \mathbb{S}_{\boldsymbol{r},\epsilon}} \left| \|\mathcal{A}(\bar{\boldsymbol{\mathcal{X}}})\|_1 - \sqrt{\frac{2}{\pi}} \right| + 4\epsilon/\sqrt{\pi}}{1 - \sqrt{8}\epsilon}. \end{aligned}$$

Take the union bound over $\mathbb{S}_{\boldsymbol{r},\epsilon}$ to conclude

$$\mathbb{P}\left(\sup_{\boldsymbol{\mathcal{X}} \in \mathbb{S}_{\boldsymbol{r}}} \left| \|\mathcal{A}(\boldsymbol{\mathcal{X}})\|_1 - \sqrt{\frac{2}{\pi}} \right| \le \frac{\delta/2 + 4\epsilon/\sqrt{\pi}}{1 - \sqrt{8}\epsilon} \right) \ge 1 - \left(\frac{12}{\epsilon}\right)^{3nr+r^3} 2\exp(-cm\delta^2). \tag{D.14}$$

Set $\epsilon = 0.1\delta$ to achieve

$$\frac{\delta/2 + 4\epsilon/\sqrt{\pi}}{1 - \sqrt{8}\epsilon} \le \delta, \quad \text{and} \quad \left(\frac{12}{\epsilon}\right)^{3nr+r^3} 2\exp(-c_1 m\delta^2) \le \rho, \quad \text{if} \quad m \ge \frac{(3nr + r^3)\log(120/\delta) + \log(2/\rho)}{c\delta^2}.$$

The proof of (D.12) is then finished.

[1] Write the rank-$2\boldsymbol{r}$ HOSVD as $\boldsymbol{\mathcal{X}} - \bar{\boldsymbol{\mathcal{X}}} = (\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \cdot \boldsymbol{\mathcal{C}}$, and split $\boldsymbol{U} = [\boldsymbol{U}_1, \boldsymbol{U}_2]$, $\boldsymbol{V} = [\boldsymbol{V}_1, \boldsymbol{V}_2], \boldsymbol{W} = [\boldsymbol{W}_1, \boldsymbol{W}_2]$, then $\boldsymbol{D}_1 = (\boldsymbol{U}_1\boldsymbol{U}_1^\top, \boldsymbol{V}_1\boldsymbol{V}_1^\top, \boldsymbol{W}_1\boldsymbol{W}_1^\top) \cdot (\boldsymbol{\mathcal{X}} - \bar{\boldsymbol{\mathcal{X}}}), \ldots, \boldsymbol{D}_8 = (\boldsymbol{U}_2\boldsymbol{U}_2^\top, \boldsymbol{V}_2\boldsymbol{V}_2^\top, \boldsymbol{W}_2\boldsymbol{W}_2^\top) \cdot (\boldsymbol{\mathcal{X}} - \bar{\boldsymbol{\mathcal{X}}})$.

To prove (D.13), introduce independent sub-Gaussian random variables

$$X_i = \begin{cases} -|\langle \boldsymbol{\mathcal{A}}_i, \boldsymbol{\mathcal{X}} \rangle| + \sqrt{\frac{2}{\pi}}, & i \in \mathcal{S} \\ |\langle \boldsymbol{\mathcal{A}}_i, \boldsymbol{\mathcal{X}} \rangle| - \sqrt{\frac{2}{\pi}}, & i \notin \mathcal{S} \end{cases}.$$

Hoeffding inequality for sub-Gaussian random variables tells that

$$\mathbb{P}\left(\frac{1}{m} \sum_{i=1}^{m} X_i = \|\mathcal{A}_{\mathcal{S}^c}(\boldsymbol{\mathcal{X}})\|_1 - \|\mathcal{A}_{\mathcal{S}}(\boldsymbol{\mathcal{X}})\|_1 - (1-2p_s)\sqrt{\frac{2}{\pi}} \geq -\frac{\delta}{2} \right) \geq 1 - \exp(-cm\delta^2).$$

Next apply the covering argument. For any $\boldsymbol{\mathcal{X}} \in \mathbb{S}_{\boldsymbol{r}}$, there exists a $\bar{\boldsymbol{\mathcal{X}}} \in \mathbb{S}_{\boldsymbol{r},\epsilon}$ such that $\|\boldsymbol{\mathcal{X}} - \bar{\boldsymbol{\mathcal{X}}}\|_{\mathsf{F}} \leq \epsilon$. Invoke the triangle inequality to see

$$\begin{aligned} \|\mathcal{A}_{\mathcal{S}^c}(\boldsymbol{\mathcal{X}})\|_1 \geq \|\mathcal{A}_{\mathcal{S}^c}(\bar{\boldsymbol{\mathcal{X}}})\|_1 - \|\mathcal{A}_{\mathcal{S}^c}(\boldsymbol{\mathcal{X}} - \bar{\boldsymbol{\mathcal{X}}})\|_1, \quad \|\mathcal{A}_{\mathcal{S}}(\boldsymbol{\mathcal{X}})\|_1 \leq \|\mathcal{A}_{\mathcal{S}}(\bar{\boldsymbol{\mathcal{X}}})\|_1 + \|\mathcal{A}_{\mathcal{S}}(\boldsymbol{\mathcal{X}} - \bar{\boldsymbol{\mathcal{X}}})\|_1 \\ \implies \|\mathcal{A}_{\mathcal{S}^c}(\boldsymbol{\mathcal{X}})\|_1 - \|\mathcal{A}_{\mathcal{S}}(\boldsymbol{\mathcal{X}})\|_1 \geq \|\mathcal{A}_{\mathcal{S}^c}(\bar{\boldsymbol{\mathcal{X}}})\|_1 - \|\mathcal{A}_{\mathcal{S}}(\bar{\boldsymbol{\mathcal{X}}})\|_1 - \|\mathcal{A}(\boldsymbol{\mathcal{X}} - \bar{\boldsymbol{\mathcal{X}}})\|_1. \end{aligned}$$

Follow the same argument above to split $\boldsymbol{\mathcal{X}} - \bar{\boldsymbol{\mathcal{X}}}$ into 8 rank-$\boldsymbol{r}$ tensors and obtain

$$\inf_{\boldsymbol{\mathcal{X}} \in \mathbb{S}_{\boldsymbol{r}}} (\|\mathcal{A}_{\mathcal{S}^c}(\boldsymbol{\mathcal{X}})\|_1 - \|\mathcal{A}_{\mathcal{S}}(\boldsymbol{\mathcal{X}})\|_1) \geq \inf_{\bar{\boldsymbol{\mathcal{X}}} \in \mathbb{S}_{\boldsymbol{r},\epsilon}} (\|\mathcal{A}_{\mathcal{S}^c}(\bar{\boldsymbol{\mathcal{X}}})\|_1 - \|\mathcal{A}_{\mathcal{S}}(\bar{\boldsymbol{\mathcal{X}}})\|_1) - \sqrt{8}\epsilon \sup_{\boldsymbol{\mathcal{X}} \in \mathbb{S}_{\boldsymbol{r}}} \|\mathcal{A}(\boldsymbol{\mathcal{X}})\|_1.$$

Take the union bound over $\mathbb{S}_{\boldsymbol{r},\epsilon}$ together with (D.14) to conclude

$$\begin{aligned} \mathbb{P}\left(\inf_{\boldsymbol{\mathcal{X}} \in \mathbb{S}_{\boldsymbol{r}}} (\|\mathcal{A}_{\mathcal{S}^c}(\boldsymbol{\mathcal{X}})\|_1 - \|\mathcal{A}_{\mathcal{S}}(\boldsymbol{\mathcal{X}})\|_1) \geq (1-2p_s)\sqrt{\frac{2}{\pi}} - \frac{\delta}{2} - \sqrt{8}\epsilon \left(\sqrt{\frac{2}{\pi}} + \frac{\delta/2 + 4\epsilon/\sqrt{\pi}}{1 - \sqrt{8}\epsilon} \right) \right) \\ \geq 1 - \left(\frac{12}{\epsilon} \right)^{3nr+r^3} 2\exp(-cm\delta^2). \end{aligned}$$

Set $\epsilon = 0.1\delta$ again to achieve

$$\sqrt{8}\epsilon \left(\sqrt{\frac{2}{\pi}} + \frac{\delta/2 + 4\epsilon/\sqrt{\pi}}{1 - \sqrt{8}\epsilon} \right) \leq \frac{\delta}{2}, \quad \text{and} \quad \left(\frac{12}{\epsilon} \right)^{3nr+r^3} 2\exp(-cm\delta^2) \leq \rho.$$

The proof of (D.13) is then finished.

Bibliography

[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. 3

[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. 4, 69

[ARR14] A. Ahmed, B. Recht, and J. Romberg. Blind deconvolution using convex programming. *IEEE Transactions on Information Theory*, 60(3):1711–1732, 2014. 1

[Bec17] A. Beck. *First-Order Methods in Optimization*, volume 25. SIAM, 2017. 1

[BGW20] S. Buchanan, D. Gilboa, and J. Wright. Deep networks and the multiple manifold problem. *arXiv preprint arXiv:2008.11245*, 2020. 70

[BH89] P. Baldi and K. Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2(1):53–58, 1989. 11

[BJS18] Y. Bai, Q. Jiang, and J. Sun. Subgradient descent learns orthogonal dictionaries. *arXiv preprint arXiv:1810.10702*, 2018. 70

[BKS16] S. Bhojanapalli, A. Kyrillidis, and S. Sanghavi. Dropping convexity for faster semi-definite optimization. In *Conference on Learning Theory*, pages 530–582. PMLR, 2016. 24

[BL20] S. Bahmani and K. Lee. Low-rank matrix estimation from rank-one projections by unlifted convex optimization. *arXiv preprint arXiv:2004.02718*, 2020. 41

[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. 63, 67, 70, 78, 95

[BNS16] S. Bhojanapalli, B. Neyshabur, and N. Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881, 2016. 11

[CC14] Y. Chen and Y. Chi. Robust spectral compressed sensing via structured matrix completion. *IEEE Transactions on Information Theory*, 60(10):6576 – 6601, 2014. 31

[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. 70

[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. 11, 41, 71

[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. 3, 11, 40, 41, 42, 47, 51, 52, 53, 54, 70, 95, 100

[CCF+20] Y. Chen, Y. Chi, J. Fan, C. Ma, and Y. Yan. Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM Journal on Optimization*, 30(4):3098–3121, 2020. 11

[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. 58, 70

[CCFM21] Y. Chen, Y. Chi, J. Fan, and C. Ma. Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806, 2021. 70, 95, 203

[CCG15] Y. Chen, Y. Chi, and A. Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Transactions on Information Theory*, 61(7):4034–4059, July 2015. 51

[CDDD19] V. Charisopoulos, D. Davis, M. Díaz, and D. Drusvyatskiy. Composite optimization for robust blind deconvolution. *arXiv preprint arXiv:1901.01624*, 2019. 95

[CFMY21] Y. Chen, J. Fan, C. Ma, and Y. Yan. Bridging convex and nonconvex optimization in robust PCA: Noise, outliers, and missing data. *The Annals of Statistics*, 49(5):2948–2971, 2021. 11

[Che15] Y. Chen. Incoherence-optimal matrix completion. *IEEE Transactions on Information Theory*, 61(5):2909–2923, 2015. 19, 155

[CHZ20] T. T. Cai, R. Han, and A. R. Zhang. On the non-asymptotic concentration of heteroskedastic Wishart-type matrix. *arXiv preprint arXiv:2008.12434*, 2020. 232

[CL16] Y. Chi and Y. M. Lu. Kaczmarz method for solving quadratic equations. *IEEE Signal Processing Letters*, 23(9):1183–1187, 2016. 41

[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. 11, 147, 207

[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. 1, 11, 25, 39, 41, 64, 71

[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. 70, 76, 95, 203

[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. 11, 38, 70, 147, 155, 207

[CLMW11] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):11:1–11:37, 2011. 1, 17, 95

[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. 4, 70, 92, 95, 203

[CLS15] E. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015. 3, 11, 70

[CP11] E. J. Candès and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011. 15, 16, 121

[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. 70

[CR09] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009. 1, 2, 20

[CRY19] H. Chen, G. Raskutti, and M. Yuan. Non-convex projected gradient descent for generalized low-rank tensor regression. *Journal of Machine Learning Research*, 20(1):172–208, 2019. 67, 69, 95

[CSPW11] V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011. 1, 9, 17

[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. 63

[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. 11, 20, 70, 75, 145

[CWW18] J.-F. Cai, T. Wang, and K. Wei. Spectral compressed sensing via projected gradient descent. *SIAM Journal on Optimization*, 28(3):2625–2653, 2018. 31

[DDKL20] D. Davis, D. Drusvyatskiy, S. Kakade, and J. D. Lee. Stochastic subgradient method converges on tame functions. *Foundations of Computational Mathematics*, 20(1):119–154, 2020. 58

[DDP20] D. Davis, D. Drusvyatskiy, and C. Paquette. The nonsmooth landscape of phase retrieval. *IMA Journal of Numerical Analysis*, 40(4):2652–2695, 2020. 11, 41

[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. 3

[DHL18] S. S. Du, W. Hu, and J. D. Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Advances in Neural Information Processing Systems*, pages 384–395, 2018. 11

[Día19] M. Díaz. The nonsmooth landscape of blind deconvolution. *arXiv preprint arXiv:1911.08526*, 2019. 41

[DLDMV00a] L. De Lathauwer, B. De Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000. 73

[DLDMV00b] L. De Lathauwer, B. De Moor, and J. Vandewalle. On the best rank-1 and rank-

$(R_1, R_2, \ldots, R_N)$ approximation of higher-order tensors. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1324–1342, 2000. 73

[DR16] M. A. Davenport and J. Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):608–622, 2016. 1, 2, 41

[DR19] J. C. Duchi and F. Ruan. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *Information and Inference: A Journal of the IMA*, 8(3):471–529, 2019. 41, 95, 100

[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. 70

[FG20] A. Frandsen and R. Ge. Optimization landscape of Tucker decomposition. *Mathematical Programming*, pages 1–26, 2020. 66, 69

[FL18] S. Friedland and L.-H. Lim. Nuclear norm of higher-order tensors. *Mathematics of Computation*, 87(311):1255–1281, 2018. 63

[GHJY15] R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points-online stochastic gradient for tensor decomposition. In *Conference on Learning Theory (COLT)*, pages 797–842, 2015. 11

[GJZ17] R. Ge, C. Jin, and Y. Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International Conference on Machine Learning*, pages 1233–1242, 2017. 11, 108, 109

[GLM16] R. Ge, J. D. Lee, and T. Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016. 11

[GM20] R. Ge and T. Ma. On the optimization landscape of tensor decompositions. *Mathematical Programming*, pages 1–47, 2020. 70

[Gof77] J.-L. Goffin. On convergence rates of subgradient optimization methods. *Mathematical Programming*, 13(1):329–347, 1977. 45, 98

[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. 70

[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. 70, 95

[GRG14] S. Gunasekar, P. Ravikumar, and J. Ghosh. Exponential family matrix completion under structural constraints. In *International Conference on Machine Learning*, pages 1917–1925, 2014. 23

[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. 63, 69

[Hac12] W. Hackbusch. *Tensor spaces and numerical tensor calculus*, volume 42. Springer, 2012. 63, 73

[Han17] P. Hand. Phaselift is robust to a constant fraction of arbitrary errors. *Applied and Computational Harmonic Analysis*, 42(3):550–562, 2017. 41

[HL13] C. J. Hillar and L.-H. Lim. Most tensor problems are NP-hard. *Journal of the ACM*, 60(6):1–39, 2013. 73

[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. 66, 69, 95

[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. 70

[HW14] M. Hardt and M. Wootters. Fast matrix completion without the condition number. In *Proceedings of The 27th Conference on Learning Theory*, pages 638–678, 2014. 11, 43

[HWZ20] R. Han, R. Willett, and A. Zhang. An optimal statistical and computational framework for generalized tensor estimation. *arXiv preprint arXiv:2002.11255*, 2020. 4, 63, 64, 66, 67, 68, 80, 87, 88, 95, 222, 229

[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. 70

[JGN+17] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan. How to escape saddle points efficiently. In *International Conference on Machine Learning*, pages 1724–1732, 2017. 11

[JK17] P. Jain and P. Kar. Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–336, 2017. 11

[JMD10] P. Jain, R. Meka, and I. S. Dhillon. Guaranteed rank minimization via singular value projection. In *Advances in Neural Information Processing Systems*, pages 937–945, 2010. 12, 43

[JNS13] P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the 45th Annual ACM Symposium on Theory of Computing*, pages 665–674, 2013. 2, 11, 17, 34, 43

[JO14] P. Jain and S. Oh. Provable tensor factorization with missing data. *Advances in Neural Information Processing Systems*, 2:1431–1439, 2014. 70

[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. 72

[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 4th ACM Conference on Recommender Systems*, pages 79–86, 2010. 62

[Kaw16] K. Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pages 586–594, 2016. 11

[KB09] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009. 3, 62, 71, 94

[KC12] A. Kyrillidis and V. Cevher. Matrix ALPS: Accelerated low rank and sparse matrix reconstruction. In *2012 IEEE Statistical Signal Processing Workshop (SSP)*, pages 185–188. IEEE, 2012. 38

[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. 66, 70

[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. 69

[KSV14] D. Kressner, M. Steinlechner, and B. Vandereycken. Low-rank tensor completion by Riemannian optimization. *BIT Numerical Mathematics*, 54(2):447–468, 2014. 70

[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. 70

[Laf15] J. Lafond. Low rank matrix completion with exponential family noise. In *Conference on Learning Theory*, pages 1224–1243, 2015. 23

[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. 42, 54, 56, 95, 99

[LFC+16] C. Lu, J. Feng, Y. Chen, W. Liu, Z. Lin, and S. Yan. Tensor robust principal component analysis: Exact recovery of corrupted low-rank tensors via convex optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5249–5257, 2016. 95, 104

[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. 70

[LHLZ20] Y. Luo, W. Huang, X. Li, and A. R. Zhang. Recursive importance sketching for rank constrained least squares: Algorithms and high-order convergence. *arXiv preprint arXiv:2011.08360*, 2020. 17

[Li13] X. Li. Compressed sensing and matrix completion with constant proportion of corruptions. *Constructive Approximation*, 37(1):73–99, 2013. 42

[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. 11, 71

[LM00] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000. 231

[LM20] A. Liu and A. Moitra. Tensor completion made practical. *Advances in Neural Information Processing Systems*, 33, 2020. 4, 70, 95

[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. 70

[LMCC21] Y. Li, C. Ma, Y. Chen, and Y. Chi. Nonconvex matrix factorization from rank-one measurements. *IEEE Transactions on Information Theory*, 67(3):1928–1950, 2021. 3, 23, 41, 42, 56

[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. 3

[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. 72

[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. 70

[LSC17] Y. Li, Y. Sun, and Y. Chi. Low-rank positive semidefinite matrix recovery from corrupted rank-one measurements. *IEEE Transactions on Signal Processing*, 65(2):397–408, 2017. 41, 95

[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. 66, 67, 69, 87, 229

[LZMCSV20] X. Li, Z. Zhu, A. Man-Cho So, and R. Vidal. Nonconvex robust low-rank matrix recovery. *SIAM Journal on Optimization*, 30(1):660–686, 2020. 41, 42, 53, 54, 58, 95, 243

[MAS12] B. Mishra, K. A. Apuroop, and R. Sepulchre. A Riemannian geometry for low-rank matrix completion. *arXiv preprint arXiv:1211.1550*, 2012. 9, 12, 42

[Maz16] M. Mazeika. The singular value decomposition and low rank approximation. Technical report, University of Chicago, 2016. 113, 115

[MBM18] S. Mei, Y. Bai, and A. Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018. 11

[MF21] J. Ma and S. Fattahi. Implicit regularization of sub-gradient method in robust matrix recovery: Don't be afraid of outliers. *arXiv preprint arXiv:2102.02969*, 2021. 95

[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. 63, 67, 69

[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. 11, 15, 25, 49, 66, 70

[MS16] B. Mishra and R. Sepulchre. Riemannian preconditioning. *SIAM Journal on Optimization*, 26(1):635–660, 2016. 9, 12

[MS18] A. Montanari and N. Sun. Spectral algorithms for tensor completion. *Communications on Pure and Applied Mathematics*, 71(11):2381–2425, 2018. 70, 95

[MWCC19] 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*, pages 1–182, 2019. 3, 11, 30, 38, 41, 42, 70, 71, 92

[NNS+14] P. Netrapalli, U. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain. Non-convex robust PCA. In *Advances in Neural Information Processing Systems*, pages 1107–1115, 2014. 11, 12, 17, 19

[NP06] Y. Nesterov and B. T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006. 11

[Paa00] P. Paatero. Construction and analysis of degenerate PARAFAC models. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 14(3):285–299, 2000. 62

[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. 3

[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. 11, 70

[PKCS18] D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi. Finding low-rank solutions via nonconvex matrix factorization, efficiently and provably. *SIAM Journal on Imaging Sciences*, 11(4):2165–2204, 2018. 24

[PS17] A. Potechin and D. Steurer. Exact tensor completion with sum-of-squares. In *Conference on Learning Theory*, pages 1619–1673. PMLR, 2017. 63, 70

[QZEW19] Q. Qu, Y. Zhang, Y. C. Eldar, and J. Wright. Convolutional phase retrieval via gradient descent. *IEEE Transactions on Information Theory*, 66(3):1785–1821, 2019. 41

[RFP10] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010. 16, 51

[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. 67, 69, 86, 87, 95, 229, 243

[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. 69, 95

[SC21] L. Shi and Y. Chi. Manifold gradient descent solves multi-channel sparse blind deconvolution provably and efficiently. *IEEE Transactions on Information Theory*, 67(7):4784–4811, 2021. 71

[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. 3, 71, 94

[SEC+15] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev. Phase retrieval with application to optical imaging: a contemporary overview. *IEEE Signal Processing Magazine*, 32(3):87–109, 2015. 1

[SL16] R. Sun and Z.-Q. Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016. 11, 70

[SQW15] J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery using nonconvex optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2351–2360, 2015. 11

[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. 70

[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. 70

[SQW18] J. Sun, Q. Qu, and J. Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, 2018. 11

[SWW17] S. Sanghavi, R. Ward, and C. D. White. The local convexity of solving systems of quadratic equations. *Results in Mathematics*, 71(3-4):569–608, 2017. 3, 23

[TBS+16] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via Procrustes flow. In *International Conference Machine Learning*, pages 964–973, 2016. xiii, 2, 9, 10, 11, 15, 17, 30, 41, 108, 109

[TMC21a] T. Tong, C. Ma, and Y. Chi. Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *Journal of Machine Learning Research*, 22(150):1–63, 2021. 3, 5, 42, 95, 207

[TMC21b] T. Tong, C. Ma, and Y. Chi. Low-rank matrix recovery with scaled subgradient methods: Fast and robust convergence without the condition number. *IEEE Transactions on Signal Processing*, 69:2396–2409, 2021. 6, 95, 100

[TMPB+21] T. Tong, C. Ma, A. Prater-Bennette, E. Tripp, and Y. Chi. Scaling and scalability: Provable nonconvex low-rank tensor estimation from incomplete measurements. *arXiv preprint arXiv:2104.14526*, 2021. 6, 95

[Tuc66] L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966. 61

[TW16] J. Tanner and K. Wei. Low rank matrix completion by alternating steepest descent methods. *Applied and Computational Harmonic Analysis*, 40(2):417–429, 2016. 9, 12, 42

[WCCL16] K. Wei, J.-F. Cai, T. F. Chan, and S. Leung. Guarantees of Riemannian optimization for low rank matrix recovery. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1198–1222, 2016. 11, 12

[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. 70

[WGMM13] J. Wright, A. Ganesh, K. Min, and Y. Ma. Compressive principal component pursuit. *Information and Inference*, 2(1):32–68, 2013. 42

[WS13] L. Wang and A. Singer. Exact and stable recovery of rotations for robust synchronization. *Information and Inference: A Journal of the IMA*, 2(2):145–193, 2013. 42

[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. 3

[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. 70

[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. 61, 66, 69, 78

[XYZ21] D. Xia, M. Yuan, and C.-H. Zhang. Statistically optimal and computationally efficient low rank tensor completion from noisy entries. *The Annals of Statistics*, 49(1):76–99, 2021. 69, 76

[XZZ20] D. Xia, A. R. Zhang, and Y. Zhou. Inference for low-rank tensors–no need to debias. *arXiv preprint arXiv:2012.14844*, 2020. 69

[YPCC16] X. Yi, D. Park, Y. Chen, and C. Caramanis. Fast algorithms for robust PCA via gradient descent. In *Advances in Neural Information Processing Systems*, pages 4152–4160, 2016. xiii, 9, 10, 11, 17, 18, 19, 20, 27, 30, 95, 139, 140

[YZ16] M. Yuan and C.-H. Zhang. On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics*, 16(4):1031–1068, 2016. 66, 69, 78, 204

[ZA16] Z. Zhang and S. Aeron. Exact tensor completion using t-SVD. *IEEE Transactions on Signal Processing*, 65(6):1511–1526, 2016. 70

[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. 42, 54, 56, 70, 95, 99

[ZCL18] H. Zhang, Y. Chi, and Y. Liang. Median-truncated nonconvex approach for phase retrieval with outliers. *IEEE Transactions on Information Theory*, 64(11):7287–7310, 2018. 95, 99

[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. 62, 70

[Zha19] A. Zhang. Cross: Efficient low-rank tensor completion. *The Annals of Statistics*, 47(2):936–964, 2019. 69

[ZL15] Q. Zheng and J. Lafferty. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. In *Advances in Neural Information Processing Systems*, pages 109–117, 2015. 11

[ZL16] Q. Zheng and J. Lafferty. Convergence analysis for rectangular matrix completion using Burer-Monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*, 2016. xiii, 10, 11, 22, 146, 147

[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. 87, 95, 229, 231

[ZLTW18] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin. Global optimality in low-rank matrix optimization. *IEEE Transactions on Signal Processing*, 66(13):3614–3628, 2018. 11

[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. 62

[ZQW20] Y. Zhang, Q. Qu, and J. Wright. From symmetry to geometry: Tractable nonconvex problems. *arXiv preprint arXiv:2007.06753*, 2020. 1, 71

[ZX18] A. Zhang and D. Xia. Tensor SVD: Statistical and computational limits. *IEEE Transactions on Information Theory*, 2018. 69

[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. 41, 70