Implicit Regularization of Gradient Descent in Realistic Settings

by

Jianhao Ma

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Industrial and Operations Engineering)
in the University of Michigan
2025

Doctoral Committee:

Assistant Professor Salar Fattahi, Chair
Assistant Professor Albert S. Berahas
Assistant Professor Wei Hu
Professor Jon Lee

Jianhao Ma
jianhao@umich.edu
ORCID iD: 0000-0003-1440-6645

© Jianhao Ma 2025

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Professor Salar Fattahi. Professor Fattahi has taught me how to approach research from the ground up—from structuring proofs and writing papers to effectively presenting our work at conferences. I am particularly grateful for his open-mindedness and the freedom he has given me to explore topics aligned with my interests, coupled with his exceptional guidance on both technical details and impactful research directions.

I sincerely thank my dissertation committee members, Professor Jon Lee, Professor Albert S. Berahas, and Professor Wei Hu, for their invaluable suggestions and continuous support throughout my Ph.D. journey. I am especially indebted to Professor Berahas for his generous assistance with my internship and job search processes, consistently going above and beyond to provide guidance and support.

Special thanks are also due to Dr. Lin Xiao and Professor Anthony Man-Cho So. Their insightful advice has significantly influenced my research scope and taste, and their tremendous support was instrumental during my job search.

I am grateful for the companionship and encouragement from all of my friends: Ruiqi, Jiaye, Qiyuan, Naichen, Geyu, Sheng, Huiwen, Xian, Luze, Zihao, Zhengjia, Zixin, Jiahao, Jikai, Zhongzhu, Jiacheng, Zijin, Jingwen, Kavin, and Hessa, whose presence has brought joy and balance to my Ph.D. life.

Finally, my profound gratitude goes to my partner, Shuo, and my parents, Genhua and Weiping, whose unwavering love, support, and encouragement have made this journey possible.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ii

LIST OF FIGURES viii

LIST OF TABLES xi

ABSTRACT xii

CHAPTER

1 Introduction **1**

1.1 Introduction to Empirical Risk Minimization 1
1.2 Implicit Regularization of Gradient Descent 4
1.3 Our Contribution 5
1.4 Main Proof Idea: Signal-Residual Decomposition 7

2 Robust Matrix Sensing **9**

2.1 Introduction 9
2.1.1 Power of Small Initialization 11
2.1.2 Summary of Results 13
2.2 Related Work 16
2.3 Our Overarching Framework 19
2.3.1 Failure of Existing Techniques 19
2.3.2 Sign-RIP: A New Robust Restricted Isometry Property 20
2.3.3 Choice of Step-size 23
2.3.4 Effect of Over-parameterization 25
2.4 Expected Loss 26
2.5 Empirical Loss with Noiseless Measurements 29
2.6 Empirical Loss with Noisy Measurements 34
2.7 Ground Truths are Strict Saddles in Noisy Setting 38
2.8 Concluding Remarks 40
2.A Proofs of the Main Theorems 41
2.A.1 Proof of Theorem 6 41
2.A.2 Proof of Theorem 7 44
2.A.3 Proof of Theorem 8 47
2.B Proofs of Sign-RIP 52

2.B.1 Preliminary . . . 52
2.B.2 Proof of Theorem 4 . . . 54
2.B.3 Proof of Theorem 5 . . . 59
2.B.4 Proof of Lemma 1 . . . 60
2.B.5 Proof of Lemma 2 . . . 60
2.C Proofs of the Expected Loss . . . 61
2.C.1 Proof of Proposition 3 . . . 61
2.C.2 Proof of Proposition 4 . . . 63
2.C.3 Proof of Lemma 5 . . . 69
2.D Proofs of Empirical Loss with Noiseless Measurements . . . 69
2.D.1 Preliminaries . . . 69
2.D.2 Proof of Proposition 5 . . . 70
2.D.3 Proof of Proposition 6 . . . 73
2.D.4 Proof of Proposition 7 . . . 77
2.D.5 Proof of Lemma 6 . . . 82
2.D.6 Proof of Lemma 7 . . . 83
2.E Proofs of Empirical Loss with Noisy Measurements . . . 86
2.E.1 Preliminaries . . . 86
2.E.2 Proof of Proposition 8 . . . 88
2.E.3 Proof of Proposition 9 . . . 88
2.E.4 Proof of Proposition 10 . . . 89
2.E.5 Proof of Lemma 9 . . . 89
2.F Proofs for Strict Saddles . . . 90
2.F.1 Preliminaries . . . 90
2.F.2 Symmetric Case: Sub-optimality via Parametric Second-order Perturbations . . . 100
2.F.3 Proof of Theorem 9 . . . 102
2.F.4 Matching Lower Bounds . . . 105
2.F.5 Strict Saddle Nature . . . 108
2.G Omitted Proofs . . . 109
2.G.1 Proof of Lemma 19 . . . 109
2.G.2 Proof of Lemma 23 . . . 113
2.G.3 Proof of Lemma 24 . . . 115
2.G.4 Proof of Lemma 27 . . . 116
2.G.5 Proof of Lemma 10 . . . 121
2.G.6 Proof of Lemma 11 . . . 121
2.G.7 Proof of Lemma 12 . . . 123

3 Robust Deep Linear Neural Network . . . 124

3.1 Introduction . . . 124
3.1.1 Our Contributions . . . 125
3.2 Problem Formulation . . . 127
3.3 Main Results . . . 128
3.3.1 Absence of Benign Landscape . . . 128
3.3.2 Convergence of Sub-gradient Method . . . 130

3.3.3 Local Landscape Around Balanced Solution . . . 133
3.4 Proof Techniques . . . 133
3.4.1 Proof Sketch of Theorem 15 . . . 134
3.5 Numerical Experiments: Beyond Linear Regression . . . 137
3.6 Conclusion . . . 139
3.A Related Works . . . 139
3.B Additional Experiments . . . 140
3.B.1 Deeper Linear Models . . . 140
3.B.2 Geometric Step-size . . . 141
3.B.3 Experiments on CIFAR Dataset . . . 141
3.C Proofs of Landscape Analysis . . . 143
3.C.1 Proof of Theorem 14 . . . 143
3.C.2 Proof of Theorem 17 . . . 147
3.D Proofs of Convergence Analysis . . . 150
3.D.1 Proof of Theorem 15 . . . 150
3.D.2 Proof of Theorem 16 . . . 156
3.E Proof of Proposition 11 . . . 162
3.F Auxiliary Lemmas . . . 169
3.G Deferred Proofs . . . 169
3.G.1 Proof of Lemma 58 . . . 169
3.H Preliminaries on the Uniform Concentration Bounds . . . 170

4 Unregularized Matrix Completion . . . 173

4.1 Introduction . . . 173
4.1.1 Summary of Contributions . . . 175
4.2 Problem Setup and Main Results . . . 177
4.3 Proof Outline . . . 179
4.3.1 Dynamic Signal-residual Decomposition . . . 180
4.3.2 Refined Leave-one-out Analysis with Weak Coupling . . . 182
4.3.3 Improved Results for Exact Parameterization . . . 187
4.4 Conclusion and Future Directions . . . 188
4.A Preliminaries . . . 188
4.A.1 Otuline of the Appendix . . . 188
4.A.2 Additional Notations . . . 189
4.A.3 Important Intermediate Lemmas . . . 189
4.B Proofs for Dynamic Signal-residual Decomposition . . . 195
4.B.1 Proof of Signal Dynamic . . . 195
4.B.2 Proof of Residual Dynamic . . . 200
4.B.3 Proof of Error Dynamic . . . 200
4.C Proofs for Main Theorems . . . 204
4.C.1 Proof of Theorem 20 . . . 204
4.C.2 Proof of Theorem 21 . . . 205
4.D Proofs for Incoherence Dynamic . . . 206
4.D.1 Proof of ?? 64 . . . 206
4.D.2 Proof of Proposition 12 . . . 208

4.D.3 Proof of Proposition 13 . . . 211
4.D.4 Proof of Proposition 18 . . . 217
4.E Proofs for Different Initialization Schemes . . . 227
4.F Concentration Inequalities for Matrix Completion . . . 229
4.G Auxiliary Lemmas . . . 232
4.G.1 Concentration Inequalities . . . 232
4.G.2 Matrix Norm Inequalities . . . 232
4.G.3 Other Useful Inequalities . . . 235
5 Function Basis Decomposition . . . 238
5.1 Introduction . . . 238
5.1.1 Main Contributions . . . 239
5.2 General Framework: Function Basis Decomposition . . . 241
5.3 Applications . . . 246
5.3.1 Kernel Regression . . . 246
5.3.2 Symmetric Matrix Factorization . . . 247
5.3.3 Orthogonal Symmetric Tensor Decomposition . . . 248
5.3.4 Empirical Verification on Neural Networks . . . 250
5.A Discussion and Future Direction . . . 252
5.B Additional Experiments . . . 253
5.B.1 Numerical Verification of our Theoretical Results . . . 253
5.B.2 Derivation of Basis Functions for DNNs . . . 254
5.B.3 Further Details on the Experiments . . . 255
5.B.4 Experiments for CIFAR-100 . . . 257
5.B.5 Experiments for Different Losses . . . 257
5.B.6 Experiments for Different Batch Size . . . 258
5.B.7 Experiments for ResNet-18 with SGD on CIFAR-10 . . . 260
5.C Related Work . . . 260
5.D Proofs for General Framework . . . 262
5.D.1 Proof of Proposition 19 . . . 262
5.D.2 Proof of Theorem 23 . . . 263
5.E Proofs for Kernel Regression . . . 267
5.E.1 Proof of Proposition 20 . . . 267
5.E.2 Proof of Theorem 24 . . . 267
5.F Proofs for Symmetric Matrix Factorization . . . 268
5.F.1 Initialization . . . 268
5.F.2 One-step Dynamics . . . 269
5.F.3 Proofs of Proposition 21 and Theorem 25 . . . 271
5.F.4 Analysis of the Logistic Map . . . 278
5.G Proof for Tensor Decomposition . . . 280
5.G.1 Proof of Lemma 99 . . . 283
5.G.2 Proof of Proposition 25 . . . 284
5.G.3 Proof of Proposition 26 . . . 289
5.H Auxiliary Lemmas . . . 292
6 Conclusion . . . 293

BIBLIOGRAPHY 295

LIST OF FIGURES

FIGURE

2.1 (a) The performance of SubGM for different search ranks r'. (b) The performance of SubGM for different corruption probabilities. (c) The performance of SubGM for different values of the initialization scale α. In all of the simulations, the initial point $\boldsymbol{U}_0$ is chosen as $\alpha\boldsymbol{B}$, where $\boldsymbol{B}$ is obtained from Algorithm 2. 11

2.2 (a) The objective value of the solutions obtained via SubGM with and without small initialization. (b) The error of the solutions obtained via SubGM with and without small initialization. In both instances, the initial point is chosen as $\boldsymbol{U}_0 = \alpha\boldsymbol{B}$, where $\boldsymbol{B}$ is obtained from Algorithm 2. The initialization scale α is chosen as $\alpha = 10^{-15}$ and $\alpha = 1$, for SubGM with and without small initialization, respectively. 12

2.3 (a) Performance of SubGM and GD under the Gaussian noise model with ℓ_1- and ℓ_2-loss functions, respectively. (b) The effect of small initialization on the performance of SubGM under the Gaussian noise model. 15

2.4 The number of samples to satisfy ℓ_2-RIP is independent of the noise variance. However, the performance of ℓ_2 highly depends on the noise variance. 19

2.5 The iterations of SubGM for the expected loss equation 2.15 undergo three phases: eigenvalue learning phase, where the eigenvalues of $\boldsymbol{S}_t\boldsymbol{S}_t^\top$ converge to those of $\boldsymbol{X}^*$; global convergence phase, where the generalization error decays linearly; and local convergence phase, where the generalization error decays sub-linearly. 27

2.6 (a) The eigenvalues of $\boldsymbol{X}^\star$ are learned at different rates. (b) During the eigenvalue learning phase, the generalization error is dominated by the signal term. In the global convergence phase, both signal and cross terms decay linearly. Finally, the residual term becomes the dominant term in the local convergence phase, and it governs the generalization error. 28

2.7 (a) The dynamics of the signal, cross, and residual terms for the empirical loss. Unlike the expected loss, the residual term for the empirical loss has a non-monotonic behavior. (b) The dynamics of $\left\|\boldsymbol{F}_t\boldsymbol{F}_t^\top\right\|$ and $\left\|\boldsymbol{G}_t\boldsymbol{G}_t^\top\right\|$. The non-monotonic behavior of the residual term can be attributed to the dynamic of $\left\|\boldsymbol{F}_t\boldsymbol{F}_t^\top\right\|$. 30

3.1 **First row.** Local landscape around the balanced true solution $\boldsymbol{w}^\star = (\sqrt[N]{\boldsymbol{\theta}^\star}, \ldots, \sqrt[N]{\boldsymbol{\theta}^\star})$ for 1-, 2-, and 3-layer models. x and y axis correspond to the points of the form $\boldsymbol{w}^\star + \alpha\boldsymbol{d} + \beta\boldsymbol{h}$, for different values of α and β, where $\boldsymbol{d}$ and $\boldsymbol{h}$ are respectively the most descent sub-gradient direction and the most negatively curved direction of the Hessian after smoothing. **Second row.** Generalization error and the empirical loss of the solutions found by SubGM for 1-, 2-, and 3-layer models. 125

3.2 **Deep matrix recovery (first row).** The ground truth $\boldsymbol{X}^\star \in \mathbb{R}^{20\times 20}$ with $\text{rank}(\boldsymbol{X}^\star) = 3$ is chosen randomly. The elements of the measurement matrices are selected from $\mathcal{N}(0,1)$, and the sample size is set to $m = 180$. The corruption probability is set to $p = 0.05$ with distribution $\mathcal{N}(0, 100)$. We use SubGM with step-size $\eta = 0.001$ and Gaussian initialization with an initialization scale $\alpha = 1 \times 10^{-3}$. **ReLU models (second row).** The samples are chosen from $y_i = \sin(\theta^{\star\top}\boldsymbol{x}_i) + \epsilon_i$ where $\boldsymbol{\theta}^\star \in \mathbb{R}^{50}$ is randomly generated with $\|\boldsymbol{\theta}^\star\|_0 = 2$, $\boldsymbol{x}_i \sim \mathcal{N}(0, I_{50})$, and $\epsilon_i \sim \mathcal{N}(0, 25)$ with corruption probability $p = 0.05$. The sample size is set to be $m = 1500$. We use SubGM with step-size $\eta = 0.001$ and Gaussian initialization $\mathcal{N}(0, \alpha^{2/N}/d)$. 137

3.3 The optimization trajectories of deep models ($N = 4, 5, 6$). 140

3.4 The optimization trajectories of 2 and 3-layer models with different initialization size $\alpha = 0.1, 0.01, 0.001$ using exponentially decayed step-size. 141

3.5 We apply ResNet-18, 34, 50 on noisy CIFAR-10 with both ℓ_1-loss and cross-entropy loss (CE). For the training dataset, we randomly choose 10% samples and replace their labels with uniform random labels. We use SGD with initial learning rate $\eta = 0.1$, momentum 0.9, batch size $B = 32$. For every 80 epochs, we decay the learning rate by a factor 0.33. We use standard data augmentation. The initialization is set by default in PyTorch. 142

3.6 We apply ResNet-18, 34, 50 on noisy CIFAR-100 with both ℓ_1-loss and cross-entropy loss (CE). The data generator and optimizer are the same as those in CIFAR-10 experiment. For each trial, we set the initial learning rate to be $\eta = 0.05$ and run 260 epochs and decay the learning rate with factor 0.33 for each 50 epochs. 143

4.1 The performance of vanilla GD with small initialization and constant step-size on MC without any explicit regularization. In all experiments, we set the dimension $d = 100$ and the true rank $r = 3$. (a): Higher sampling rates improve the final accuracy of GD for the over-parameterized MC, but have no impact on the exactly-parameterized MC. (b): Increasing the initialization scale hampers the final accuracy of GD for the over-parameterized MC, but has no impact on the exactly-parameterized MC. 175

5.1 The solution trajectories of LARS on AlexNet and ResNet-18 and AdamW on ViT with ℓ_2-loss after projecting onto two different orthonormal bases. The first row shows the trajectories of the top-5 coefficients after projecting onto a randomly generated orthonormal basis. The second row shows the trajectories of the top-5 coefficients after projecting onto the eigenvectors of the conjugate kernel evaluated at the last epoch. More detail on our implementation can be found in Appendix 5.B. 240

5.2 The conditions of Theorem 23 are approximately satisfied for LARS on a 2-layer CNN with MNIST dataset. Here, the coefficients are obtained by projecting the solution trajectory onto the eigenvectors of the conjugate kernel after training. (a) The top-20 basis coefficients at a small random initial point along with residual energy on the remaining coefficients. (b) The maximum value of $|\cos(\nabla\beta_i(\boldsymbol{\theta}), \nabla\beta_j(\boldsymbol{\theta}))|$ for $1 \le i < j \le 10$ along the solution trajectory. (c) The scaling of $\|\nabla\beta_i(\boldsymbol{\theta}_t)\|$ with respect to $|\beta_i(\boldsymbol{\theta}_t)|$ for the top-4 coefficients. 245

5.3 (First row) the projected trajectory of LARS on CNNs with MNIST dataset. The test accuracies for 2-Block, 3-Block, and 4-block CNN are $96.10\%, 96.48\%, 96.58\%$. (Second row) the projected trajectories of different optimizers on AlexNet with the CIFAR-10 dataset. The test accuracies for LARS, SGD, and AdamW are 90.54%, 91.03%, and 90.26%, respectively. We use the following settings for each optimizer: (d) LARS: learning rate of 2, Nesterov momentum of 0.9, and weight decay of 1×10^{-4}. (e) SGD: learning rate of 2 with "linear warm-up", Nesterov momentum of 0.9, weight decay of 1×10^{-4}. (f) AdamW: learning rate of 0.01. 251
5.4 Experimental verification to support Theorems 24, 25, and 26 in Section 5.3. The first row shows the projected trajectories of GD onto the specific basis functions we defined for each problem. The second row shows the estimation error. 254
5.5 Solution trajectories for different architectures trained on CIFAR-10. The test accuracies for Alexnet, VGG-11 and ViT are 90.54%, 91.11%, 77.99%, respectively. The test accuracies for ResNet-18, ResNet-34 and ResNet-50 are 88.31%, 93.93% and 94.06%, respectively. 256
5.6 Solution trajectories for different architectures trained on CIFAR-100. The test accuracies for AlexNet, VGG-11, and ViT are 61.20%, 50.64%, and 50.41%, respectively. The test accuracies for ResNet-18, ResNet-34 and ResNet-50 are 78.11%, 74.26% and 80.05%, respectively. . . . 258
5.7 Solution trajectories of LARS on the CIFAR-10 dataset with ℓ_2-loss and CE loss. The test accuracies are 90.54% for ℓ_2-loss and 91.09% for CE loss. 259
5.8 Solution trajectories for AlexNet on the CIFAR-10 dataset with different batch sizes. The test accuracies for batch size $32, 64$ and 256 are 91.50%, 91.79%, and 90.12%, respectively. 259
5.9 Solution trajectories for ResNet-18 on the CIFAR-10 dataset with SGD solver. The best test accuracy is 94.45%. 260

LIST OF TABLES

TABLE

2.1 A comparison between the **iteration complexity** of different techniques. We show that SubGM with small initialization and exponentially decaying step-size converges near-linearly to: (i) an arbitrary accuracy in the outlier noise model, and (ii) a nearly-minimax optimal error in the Gaussian noise model. Our derived iteration complexities are obtained from Theorems 2 and 3 after choosing an appropriate value for the step-size; see Corollaries 2 and 3 for the precise statements. 17

2.2 A comparison between the **sample complexity** of different techniques. Our results provide the best sample complexity bounds in the over-parameterized setting where $r' \gg r$, under both outlier and Gaussian noise models. For simplicity, we hide the dependency on the logarithmic factors. * Under an outlier noise model, the results of [Li, Zhu, Man-Cho So, and Vidal, 2020b, Ding, Jiang, Chen, Qu, and Zhu, 2021] holds under the assumption $p \lesssim 1/\sqrt{r'}$. In contrast, our result relaxes this assumption to $p < 1$. 18

3.1 Test accuracy for ResNets on CIFAR-10 and CIFAR-100 datasets with 10% label noise. 138

4.1 Comparisons between different algorithms for matrix completion without explicit regularization. ✱ The result only holds for $r' = r = 1$. 176

5.1 The summary of our experiments. 253

ABSTRACT

Gradient descent (GD) is the backbone of modern machine learning optimization, and its success is often attributed not just to efficiency but also to an intriguing phenomenon known as implicit regularization—the tendency of GD to favor solutions that generalize well, even without explicit constraints. While existing theoretical studies have shed light on this behavior, they frequently rely on idealized assumptions, such as isotropic data distributions or benign optimization landscapes, which rarely hold in practice. This gap highlights the need to understand implicit regularization in more realistic and challenging settings that better reflect practical machine learning problems.

This dissertation explores implicit regularization in gradient-based learning under three key challenges: (1) robustness to heavy-tailed outlier noise, (2) learning with non-isotropic input distributions, and (3) developing a general theoretical framework for characterizing optimization trajectories. First, we study robust learning with ℓ_1-loss, including robust matrix sensing and robust deep linear networks, and show that GD implicitly preserves low-rank structures throughout training, achieving near-linear convergence under near-optimal sample complexity. Interestingly, we prove that the ground truth corresponds to a strict saddle point, countering the conventional wisdom that saddle points inherently impede optimization. Second, we analyze learning with non-isotropic input distributions—a broad category that includes unregularized matrix completion as a special case. We introduce a novel statistical decoupling technique that establishes near-optimal sample complexity guarantees for GD without relying on standard isotropic assumptions. Lastly, we develop a unified framework based on structured basis function decomposition, revealing that GD trajectories, when projected onto a suitable function basis, exhibit an almost monotonic progression despite their apparent complexity.

Altogether, this work advances the theoretical understanding of implicit regularization under realistic conditions, offering rigorous insights into how gradient descent selects generalizable solutions. These results have broad implications for designing robust and efficient learning algorithms across diverse machine learning tasks.

CHAPTER 1

Introduction

1.1 Introduction to Empirical Risk Minimization

Machine learning (ML), especially deep learning (DL), has achieved remarkable success across various domains, including computer vision [Krizhevsky, Sutskever, and Hinton, 2012, He, Zhang, Ren, and Sun, 2016], natural language processing [Devlin, Chang, Lee, and Toutanova, 2019, Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin, 2017], and robotics [Levine, Finn, Darrell, and Abbeel, 2016, Levine, Pastor, Krizhevsky, and Quillen, 2018].

Most machine learning problems can be categorized within the empirical risk minimization (ERM) paradigm [Devroye, Györfi, and Lugosi, 1996], which provides a systematic approach to learning from data. Specifically, we consider a training dataset $S_n = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ consisting of n data points, where $\boldsymbol{x}_i$ represents the input and y_i represents the output. We assume that each data point $(\boldsymbol{x}_i, y_i)$ is independently and identically distributed (i.i.d.) from an underlying distribution $\mathcal{D}$. However, real-world data is often noisy. In particular, we consider the label noise setting, where the output is generated as $y_i = f^*(\boldsymbol{x}_i) + \epsilon_i$, where f^* is the ground truth function, and ϵ_i follows some noise distribution.

The objective is to approximate f^* using a model $f_{\boldsymbol{\theta}}$ from a predefined hypothesis class $\mathcal{F}_\Theta$, parameterized by $\boldsymbol{\theta}$. Ideally, f^* belongs to or is well-approximated by $\mathcal{F}_\Theta$, ensuring that our model can capture the underlying data distribution effectively. To measure how well $f_{\boldsymbol{\theta}}$ approximates f^*, we introduce a loss function $\ell(\cdot, \cdot)$, which quantifies the discrepancy between the predicted output $f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$ and the observed label y_i. Common choices for $\ell(\cdot, \cdot)$ include the squared loss $\ell(y, \hat{y}) = (y - \hat{y})^2$ and the absolute loss $\ell(y, \hat{y}) = |y - \hat{y}|$.

Given this setup, empirical risk minimization seeks to minimize the empirical loss

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \mathcal{L}_n(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^n \ell(y_i, f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)), \tag{ERM}$$

where $\mathcal{L}_n(\cdot)$ is the empirical loss. Since this optimization problem is typically high-dimensional and nonconvex, it is solved using first-order optimization methods. A widely used approach is gradient descent (GD), which updates parameters iteratively

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \nabla \mathcal{L}_n(\boldsymbol{\theta}_t). \tag{GD}$$

Here, $\eta_t > 0$ is the step size, and $\nabla \mathcal{L}_n(\boldsymbol{\theta}_t)$ is the gradient of $\mathcal{L}_n$ at $\boldsymbol{\theta}_t$.[1] Starting from an initial parameter $\boldsymbol{\theta}_0$, GD generates an optimization trajectory $\{\boldsymbol{\theta}_t\}_{t=0}^T$ and finally output a solution $\boldsymbol{\theta}_{\texttt{GD}}$, which is typically either the last iterate, $\boldsymbol{\theta}_{\texttt{GD}} = \boldsymbol{\theta}_T$, or a weighted average of all the iterates, $\boldsymbol{\theta}_{\texttt{GD}} = \sum_{t=0}^T w_t \boldsymbol{\theta}_t$, where the weights $\{w_t\}_{t=0}^T$ satisfy $w_t \geq 0$ and $\sum_{t=0}^T w_t = 1$.

While ERM focuses on minimizing the empirical loss, the ultimate goal is to ensure that $\boldsymbol{\theta}_{\texttt{GD}}$ generalizes well to unseen data. Specifically, we seek to minimize the noiseless population loss

$$\mathcal{L}(\boldsymbol{\theta}) := \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_{\text{clean}}} \left[\ell(y, f_{\boldsymbol{\theta}}(\boldsymbol{x}))\right].$$

Here, we define the clean distribution as follows: first, we sample $\boldsymbol{x} \sim \mathcal{D}_{\boldsymbol{x}}$, the marginal distribution of $\mathcal{D}$, and then generate the noiseless output as $y = f^*(\boldsymbol{x})$.

Example 1 (Low-Rank Matrix Recovery)**.** *Consider a dataset $\{(\boldsymbol{A}_i, y_i)\}_{i=1}^n$, where each input matrix $\boldsymbol{A}_i \in \mathbb{R}^{d\times d}$ and the corresponding output is generated as $y_i = \langle \boldsymbol{A}_i, \boldsymbol{X}^* \rangle + \epsilon_i$, where $\boldsymbol{X}^*$ is an unknown positive semidefinite (PSD) low-rank ground truth matrix with rank $r \ll d$, and ϵ_i represents noise.*

The objective is to estimate $\boldsymbol{X}^*$ from the given observations. A common approach is the Burer-Monteiro (BM) factorization [Burer and Monteiro, 2003], which represents $\boldsymbol{X}$ as $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{U}^\top$, where $\boldsymbol{U} \in \mathbb{R}^{d\times r'}$ and the search rank $r' \geq r$. Since the true rank r is unknown, we choose a search rank $r' \geq r$ to ensure that the model includes a valid ground truth solution, $\boldsymbol{\Theta}^* = \{\boldsymbol{U}^* \in \mathbb{R}^{d\times r'} : \boldsymbol{U}^*\boldsymbol{U}^{*\top} = \boldsymbol{X}^*\}$.

To estimate $\boldsymbol{X}^*$, we minimize the empirical risk using an ℓ_p-loss function, leading to the following nonconvex optimization problem

$$\min_{\boldsymbol{U}\in\mathbb{R}^{d\times r'}} \frac{1}{n}\sum_{i=1}^n \left|y_i - \left\langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{U}^\top \right\rangle\right|^p. \tag{1.1}$$

When the noises ϵ_i follow a dense light-tailed distribution such as a sub-Gaussian distribution, the ℓ_2-loss is typically used. In more challenging settings, such as those involving outlier noise—where a fraction of the samples are grossly corrupted by extreme values—the ℓ_1-loss

[1] When $\mathcal{L}_n$ is nonsmooth, we can use the subgradient instead of the gradient, leading to the subgradient method.

is preferred to enhance robustness. Notably, when ℓ_1-loss is applied, the optimization problem becomes nonsmooth due to the nondifferentiability of the ℓ_1-loss function at the origin.

There are two major challenges for ERM to be effective.

- **Optimization challenge.** Usually, $f_{\boldsymbol{\theta}}$ is a highly complex function, which can be nonconvex and nonsmooth. As a result, the empirical loss $\mathcal{L}_n(\boldsymbol{\theta})$ in equation `ERM` is generally a nonconvex optimization problem, making it difficult to find the global minimum. First-order methods like gradient descent can only guarantee convergence to a stationary point, which may not be globally optimal. Additionally, in high-dimensional spaces, the loss landscape can have multiple local minima, saddle points, and flat regions, further complicating optimization.

- **Statistical challenge.** Our goal is to find a solution that optimizes the population loss $\mathcal{L}(\boldsymbol{\theta})$, but we only have access to a finite, noisy sample set drawn from the true distribution. This discrepancy between the empirical and population objectives introduces statistical challenges. Moreover, in modern deep learning, we often work in the *overparameterized regime*, where the number of model parameters exceeds what is statistically necessary. Since we do not know the true underlying function f^*, we typically use a highly expressive model class $\mathcal{F}_\Theta$ to ensure it contains a good approximation of f^*. However, this flexibility introduces additional challenges: there are usually multiple solutions that minimize the empirical loss, yet not all of them generalize well to unseen data. In other words, overparameterization allows models to interpolate training data perfectly while still failing to achieve low population loss. This phenomenon makes it crucial to understand the implicit biases of optimization algorithms and develop methods to ensure generalization.

Despite these challenges, ERM combined with first-order optimization has demonstrated remarkable empirical success [Krizhevsky et al., 2012, He et al., 2016, Devlin et al., 2019, Vaswani et al., 2017, Levine et al., 2016, 2018]. A key explanation is implicit regularization [Neyshabur, Bhojanapalli, and Chakrabarti, 2017, Soudry, Hoffer, Nacson, Gunasekar, and Srebro, 2018, Zhang, Bengio, Hardt, Recht, and Vinyals, 2017], where optimization algorithms exhibit an inherent preference for solutions with favorable generalization properties. Instead of selecting an arbitrary empirical minimizer, certain optimization dynamics bias solutions toward structured forms that generalize well. This phenomenon plays a crucial role in deep learning, where large overparameterized models can still perform well on unseen data despite perfectly fitting the training set. Understanding the nature of implicit regularization is essential for designing more efficient and robust learning algorithms.

1.2 Implicit Regularization of Gradient Descent

Formally, the implicit regularization theory suggests that, despite minimizing an unregularized empirical loss function, GD effectively optimizes a **regularized** loss function of the form

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathcal{L}_n(\boldsymbol{\theta}) + \lambda R(\boldsymbol{\theta}), \tag{1.2}$$

where $\lambda > 0$ is the regularization strength, and $R(\cdot)$ is a regularizer. Common choices of $R(\cdot)$ include the ℓ_p-regularizer $\|\cdot\|_p^p$ and the nuclear norm regularizer $\|\cdot\|_\star$. Let us revisit the low-rank matrix recovery example to better understand this theory.

The implicit regularization property of GD provides several important benefits:

- **Optimization benefit.** Implicit regularization can improve the convergence properties of GD by biasing the iterates toward those regions with benign landscape properties. This is particularly beneficial in high-dimensional problems, where the global loss landscape can be very bad. For instance, in overparameterized models such as low-rank matrix recovery problems, GD naturally biases solutions toward low-rank structures so that .

- **Statistical benefit.** Implicit regularization helps **control model complexity**, thereby reducing overfitting and enhancing generalization. By favoring solutions with desirable structures—such as sparsity in high-dimensional regression or low-rankness in matrix estimation—GD naturally mitigates the risk of overfitting even in regimes where the number of parameters exceeds the number of samples. This effect has been rigorously studied in settings like linear regression, where GD with small initialization implicitly prefers solutions with minimal ℓ_2-norm, leading to lower variance and improved predictive performance.

- **Adaptivity of the implicit regularizer.** Unlike explicit regularization, which requires manually choosing the regularization strength λ, implicit regularization is **self-adaptive**—the algorithm automatically adjusts its trajectory to align with the problem's intrinsic structure. A poorly chosen explicit regularization strength can introduce significant bias, harming performance. In contrast, GD dynamically discovers the correct form of regularization. For example, it **naturally induces sparsity** when solving high-dimensional optimization problems via ℓ_1-like implicit biases or **encourages low-rank structures** in matrix recovery problems through nuclear norm regularization.

Gradient descent and its variants have been shown to exhibit implicit regularization in a wide range of applications, including sparse recovery [Vaskevicius, Kanade, and Rebeschini,

2019, Zhao, Yang, and He, 2019], matrix recovery [Ma, Wang, Chi, and Chen, 2017, Arora, Cohen, Hu, and Luo, 2019a], and deep neural networks [Gunasekar, Woodworth, Bhojanapalli, Neyshabur, and Srebro, 2018]. This phenomenon provides a compelling explanation for why these methods achieve both efficient optimization and strong generalization, even in overparameterized regimes.

Despite substantial theoretical progress, key gaps remain between existing theory and practical observations. First, much of the analysis is confined to idealized settings, where training data follows well-behaved distributions—such as isotropic Gaussian inputs and outputs perturbed by light-tailed noise (e.g., sub-Gaussian noise) [Ma et al., 2017, Arora et al., 2019a, Vaskevicius et al., 2019, Razin and Cohen, 2020, Razin, Maman, and Cohen, 2021]. However, real-world data is often heavy-tailed, highly structured, or subject to complex noise patterns, making these assumptions unrealistic in many applications. Second, most theoretical frameworks are tailored to specific problem settings and require case-by-case adjustments, limiting their applicability to broader machine learning and optimization problems. Developing more general and robust theoretical insights remains an important challenge.

1.3 Our Contribution

As discussed in the previous section, a significant gap remains between existing theoretical understanding and practical observations regarding the implicit regularization of gradient descent. To bridge this gap, this dissertation investigates the implicit regularization properties of gradient descent in more **realistic** settings, moving beyond the conventional assumptions of well-behaved training data. Specifically, we examine scenarios where outputs are heavily corrupted by outliers and inputs deviate from isotropy, such as in matrix completion. Additionally, we develop a unified framework for studying implicit regularization, encompassing a wide range of important cases, including practical neural networks.

Below, we summarize the main contributions:

- **Chapter 2: Robust matrix sensing [Ma and Fattahi, 2023a,b].** We study the sub-gradient method (SubGM) for robust low-rank matrix recovery under an ℓ_1-loss formulation, where a subset of measurements may be grossly corrupted by noise. Our focus is on an over-parameterized setting, where the rank of the true solution is unknown and over-estimated, potentially leading to overfitting or degraded algorithmic performance. We show that a simple SubGM with small initialization is inherently robust to both over-parameterization and noise. Specifically, we prove that small initialization nullifies the effect of over-parameterization, leading to an exponential

improvement in the convergence rate. Moreover, we introduce a unified framework that analyzes SubGM under both outlier and Gaussian noise models, demonstrating that it converges to the true solution even in the presence of arbitrarily large and dense noise. Notably, this holds even when globally optimal solutions do not correspond to the ground truth. A key technical contribution is our introduction of a robust variant of the restricted isometry property, termed Sign-RIP, which characterizes the deviation of the ℓ_1-loss sub-differential from the ideal expected loss. As a byproduct, we show that in the case of Gaussian measurements, the sample complexity required for global convergence of SubGM is independent of the over-parameterized rank.

- **Chapter 3: Robust deep linear networks [Ma and Fattahi, 2022a].** We investigate the impact of depth on the optimization landscape of over-parameterized linear regression using deep diagonal linear networks in a robust setting, where a subset of measurements is heavily corrupted by noise. While deep models are generally non-convex, we show that their optimization landscapes exhibit desirable properties. On the negative side, we establish that for any depth $N \geq 1$, there exist solutions corresponding to the ground truth that are neither local nor global minima, indicating a non-benign landscape. However, on the positive side, we prove that for any $N \geq 2$, a simple sub-gradient method is oblivious to such problematic solutions and instead converges to a balanced solution that is both close to the ground truth and enjoys a flat local landscape, thereby eliminating the need for early stopping. Additionally, we empirically validate that these optimization properties extend beyond linear regression to other robust learning tasks, such as deep matrix recovery and deep ReLU networks trained with an ℓ_1-loss.

- **Chapter 4: Unregularized matrix completion [Ma and Fattahi, 2024].** We analyze the problem of symmetric matrix completion, where the goal is to recover a positive semidefinite matrix $\boldsymbol{X}^\star \in \mathbb{R}^{d\times d}$ of rank r, parameterized as $\boldsymbol{U}\boldsymbol{U}^\top$, from a subset of its observed entries. Our key result shows that vanilla gradient descent (GD) with small initialization converges to the ground truth $\boldsymbol{X}^\star$ without requiring explicit regularization, accurate rank estimation, or strong initialization conditions. In the over-parameterized setting where the search rank r' is overestimated ($r' \gg r$), we prove that GD converges nearly linearly to an ϵ-neighborhood of $\boldsymbol{X}^\star$ with $\widetilde{\Omega}(dr^9)$ observations, independent of the over-parameterized rank r'. In the exactly-parameterized setting ($r' = r$), we further refine this result, showing that GD achieves arbitrarily small accuracy $\epsilon > 0$ at a faster convergence rate, provided the initialization satisfies $\|\boldsymbol{U}_0\| = \mathcal{O}(1/d)$. Our analysis is enabled by a novel *weakly-coupled leave-one-*

out technique, which extends beyond the classical leave-one-out method to establish global convergence guarantees for GD.

- **Chapter 5: Function basis decomposition [Ma, Guo, and Fattahi, 2023].** We introduce a novel *basis function decomposition* framework for analyzing the solution trajectory of gradient-based algorithms. Our key insight is that, despite variations across different learning tasks, solution trajectories exhibit an approximately monotonic behavior when projected onto an appropriate orthonormal function basis. This decomposition enables a systematic analysis of the learning dynamics. Using this framework, we provide improved convergence guarantees for gradient descent (GD) in symmetric matrix factorization and establish a new convergence result for orthogonal symmetric tensor decomposition. Additionally, we empirically demonstrate that this decomposition extends to deep neural networks (DNNs), revealing that gradient-based algorithms learn the coefficients of a particular orthonormal function basis defined by the eigenvectors of the conjugate kernel after training. Our findings provide a fresh perspective on the dynamics of gradient-based optimization across different architectures, solvers, and datasets.

1.4 Main Proof Idea: Signal-Residual Decomposition

In this section, we outline the main proof idea underlying all of our results.

In many applications, the set of ground truth solutions, defined as

$$\boldsymbol{\Theta}^* = \left\{ \boldsymbol{\theta}^* \in \mathbb{R}^d : \boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \mathcal{L}(\boldsymbol{\theta}) \right\},$$

is typically contained within a low-dimensional subset $\mathcal{M}$, where $\dim(\mathcal{M}) \ll d$. For instance, in the low-rank matrix recovery example introduced in Example 1, the ground truth solutions lie within the set of low-rank matrices: $\mathcal{M} = \left\{ \boldsymbol{U} \in \mathbb{R}^{d \times r'} : \operatorname{rank}(\boldsymbol{U}) \leq r \right\}$.

To analyze the optimization trajectory $\{\boldsymbol{\theta}_t\}_{t=0}^T$, we decompose each iterate $\boldsymbol{\theta}_t$ into two components relative to the low-dimensional structure $\mathcal{M}$:

$$\boldsymbol{\theta}_t = \underbrace{\mathcal{P}_{\mathcal{M}}(\boldsymbol{\theta}_t)}_{\text{signal}} + \underbrace{\mathcal{P}_{\mathcal{M}}^{\perp}(\boldsymbol{\theta}_t)}_{\text{noise}}. \tag{1.3}$$

Here, we define $\mathcal{P}_{\mathcal{M}}(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}' \in \mathcal{M}} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$ as the projection onto $\mathcal{M}$, and $\mathcal{P}_{\mathcal{M}}^{\perp}(\boldsymbol{\theta}) = \boldsymbol{\theta} - \mathcal{P}_{\mathcal{M}}(\boldsymbol{\theta})$ as the residual component orthogonal to $\mathcal{M}$. Then, we can characterize the performance metric $\operatorname{dist}(\boldsymbol{\theta}_{\mathtt{GD}}, \boldsymbol{\Theta}^*)$ in the following signal-residual decomposition.

Proposition 1 (Signal-residual decomposition)**.** *For any* $\boldsymbol{\theta} \in \mathbb{R}^d$*, we have the following decomposition*

$$\operatorname{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta}^*) \leq \operatorname{dist}\left(\mathcal{P}_{\mathcal{M}}(\boldsymbol{\theta}), \boldsymbol{\Theta}^*\right) + \operatorname{dist}\left(\mathcal{P}_{\mathcal{M}}^{\perp}(\boldsymbol{\theta}), \boldsymbol{\Theta}^*\right). \tag{1.4}$$

This decomposition allows us to separately analyze how the optimization dynamics evolve within the low-dimensional structure and how much deviation occurs outside of it. Then, the success of implicit regularization in GD typically depends on the following three key conditions:

- **Good initialization.** The initialization point should be close to the low-dimensional structure $\mathcal{M}$, which includes the ground truth solutions. Although the exact structure of $\mathcal{M}$ is generally unknown, a common heuristic is to use a small initialization close to the origin[2]. This choice is effective in many cases since real-world models often lie low-dimensional sets such as sparse vectors, low-rank matrices, or low-rank tensors which all include origin. By starting in this region, GD is more likely to converge to a well-generalizing solution rather than an arbitrary minimizer of the empirical loss.

- **Controlled residual dynamic.** Throughout the training process, we want the residual term $\boldsymbol{E}_t$—capturing deviations from the desired low-dimensional set $\mathcal{M}$—to remain small. The specific form of the residual dynamic depends on the problem setting, but in many cases, its growth rate is inversely proportional to the sample size. This means that as the number of training samples increases, the residual grows more slowly. In fact, when the sample size exceeds a certain threshold (which often aligns with the information-theoretic limit), we can show that the residual remains small throughout training, ensuring stability and good generalization.

- **Fast convergence of the signal dynamics.** The signal component, which evolves within the low-dimensional set $\mathcal{M}$, typically converges rapidly to the ground truth solution. This is because, when restricted to $\mathcal{M}$, the empirical loss often exhibits a well-behaved landscape properties such as local strong convexity and sharpness, enabling efficient optimization via first-order methods. More importantly, the global minimizer of the empirical loss within $\mathcal{M}$ is usually close to the minimizer of the population loss. In other words, GD implicitly guides the optimization toward a solution that generalizes well, rather than just overfitting to the training data.

[2]The reason that we do not use the exact origin is that it is usually a saddle point of the empirical loss. Hence, we want to slightly perturb it to escape that saddle point efficiently.

CHAPTER 2

Robust Matrix Sensing

This chapter is based on the following two papers:

- Jianhao Ma and Salar Fattahi. Global convergence of sub-gradient method for robust matrix recovery: Small initialization, noisy measurements, and over-parameterization. *Journal of Machine Learning Research*, 24(96):1–84, 2023a.
- Jianhao Ma and Salar Fattahi. Can Learning Be Explained By Local Optimality In Low-rank Matrix Recovery? *arXiv preprint arXiv:2302.10963*, 2023b.

2.1 Introduction

This chapter investigates the problem of *robust matrix sensing*, where the goal is to recover a low-rank positive semidefinite matrix $\boldsymbol{X}^\star \in \mathbb{R}^{d\times d}$ from a limited number of linear measurements of the form $\boldsymbol{y} = \mathcal{A}(\boldsymbol{X}^\star) + \boldsymbol{s}$, where $\boldsymbol{y} = [y_1, y_2, \ldots, y_m]^\top$ is a vector of measurements, $\mathcal{A}$ is a linear operator defined as $\mathcal{A}(\cdot) = [\langle \boldsymbol{A}_1, \cdot\rangle, \langle \boldsymbol{A}_2, \cdot\rangle, \ldots, \langle \boldsymbol{A}_m, \cdot\rangle]^\top$ with measurement matrices $\{\boldsymbol{A}_i\}_{i=1}^m$, and $\boldsymbol{s} = [s_1, s_2, \ldots, s_m]^\top$ is a noise vector. More formally, the robust matrix sensing is defined as

$$\text{find} \quad \boldsymbol{X}^\star \qquad \text{subject to:} \quad \boldsymbol{y} = \mathcal{A}(\boldsymbol{X}^\star) + \boldsymbol{s}, \quad \text{rank}(\boldsymbol{X}^\star) = r, \tag{2.1}$$

where $r \leq d$ is the rank of $\boldsymbol{X}^\star$. Robust matrix sensing plays a central role in many contemporary machine learning problems, including motion detection in video frames [Bouwmans and Zahzah, 2014], face recognition [Luan, Fang, Liu, Yang, and Qian, 2014], and collaborative filtering in recommender systems [Luo, Zhou, Xia, and Zhu, 2014]. Despite its widespread applications, it is well-known that solving equation 2.1 is a daunting task since it amounts to an NP-hard problem in its worst case [Natarajan, 1995, Recht, Fazel, and Parrilo, 2010]. What makes this problem particularly difficult is the nonconvexity stemming from the rank constraint. The classical methods for solving low-rank matrix sensing problem are based on

convexification techniques, which suffer from notoriously high computational cost. To alleviate this issue, a far more practical approach is to resort to the following natural *nonconvex* and *nonsmooth* formulation

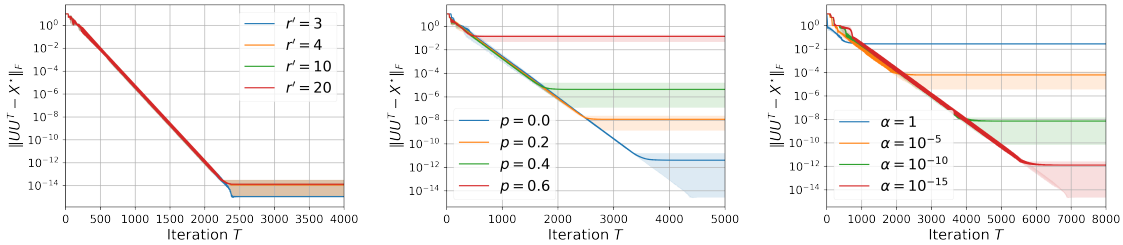
$$\min_{\boldsymbol{U}\in\mathbb{R}^{d\times r'}} \quad f_{\ell_1}(\boldsymbol{U}) := \frac{1}{m}\left\|\boldsymbol{y}-\mathcal{A}\left(\boldsymbol{U}\boldsymbol{U}^\top\right)\right\|_1, \tag{2.2}$$

where $r' \geq r$ is the search rank. The ℓ_1-loss is used to robustify the solution against noisy measurements. The above formulation is inspired by the celebrated Burer-Monteiro approach [Burer and Monteiro, 2003], which circumvents the explicit rank constraint by optimizing directly over the factorized model $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{U}^\top$.

Perhaps the most significant breakthrough result in this line of research was presented by Bhojanapalli, Neyshabur, and Srebro [2016], showing that, when the rank of the true solution is known and the measurements are noiseless, the nonconvex formulation of the problem with a smooth ℓ_2-loss has a *benign landscape*, i.e., it is devoid of undesirable local solutions; as a result, simple local-search algorithms are guaranteed to converge to the globally optimal solution. Such benign landscape seems to be omnipresent in other variants of low-rank matrix sensing, including matrix completion [Ge, Jin, and Zheng, 2017, Ge, Lee, and Ma, 2016], robust PCA [Ge et al., 2017, Fattahi and Sojoudi, 2020], sparse dictionary learning [Sun, Qu, and Wright, 2016, Qu, Zhai, Li, Zhang, and Zhu, 2019], linear neural networks [Kawaguchi, 2016], among others; see recent survey papers [Chi, Lu, and Chen, 2019, Zhang, Qu, and Wright, 2020].

A recurring assumption for the absence of spurious local minima is the exact parameterization of the rank: it is often presumed that the *exact* rank of the true solution is known *a priori*. However, the rank of the true solution is rarely known in many applications. Therefore, it is reasonable to choose the rank of $\boldsymbol{U}\boldsymbol{U}^\top$ conservatively as $r' > r$, leading to an *over-parameterized* model. This challenge is further compounded in the noisy regime, where the injected noise in the measurements can be "absorbed" as a part of the solution, due to the additional degrees of freedom in the model. Evidently, the existing proof techniques face major breakdowns in this setting, *as the problem may no longer enjoy a benign landscape*. Moreover, over-parameterization may lead to a dramatic, exponential slow-down of the local-search algorithms—both theoretically and practically [Zhuo, Kwon, Ho, and Caramanis, 2021, Zhang, Fattahi, and Zhang, 2021, 2022].

In this work, we study the performance of a simple sub-gradient method (SubGM) on $f_{\ell_1}(\boldsymbol{U})$. We prove that small initialization nullifies the effect of over-parameterization on its performance—as if the search rank r' were set to the true (but unknown) rank r. Moreover, we show that SubGM converges to the ground truth at a near-linear rate even if local, or



(a) (b) (c)

Figure 2.1: (a) The performance of SubGM for different search ranks r'. (b) The performance of SubGM for different corruption probabilities. (c) The performance of SubGM for different values of the initialization scale α. In all of the simulations, the initial point $\boldsymbol{U}_0$ is chosen as $\alpha \boldsymbol{B}$, where $\boldsymbol{B}$ is obtained from Algorithm 2.

even global, spurious minima exist. Our proposed overarching framework is based on a novel signal-residual decomposition of the solution trajectory: we decompose the iterations of SubGM into *low-rank (signal)* and *residual* terms, and show that small initialization keeps the residual term small throughout the solution trajectory, while enabling the low-rank term to converge to the ground truth exponentially fast.

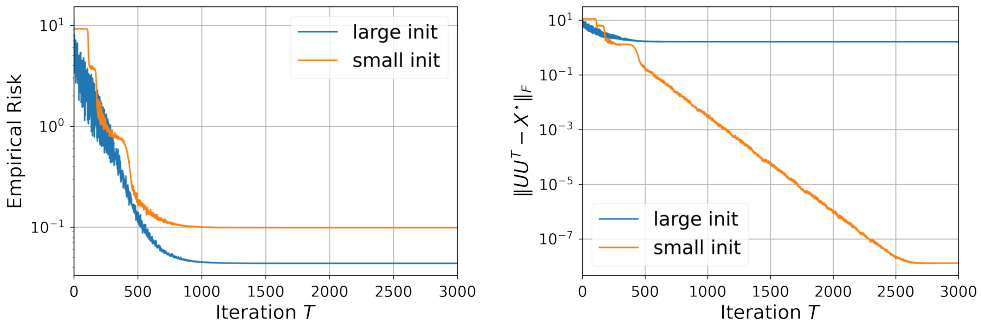
2.1.1 Power of Small Initialization

In this section, we shed light on the power of small initialization on the performance of SubGM for robust matrix sensing. Given an initial point $\boldsymbol{U}_0$ and at every iteration t, SubGM selects an arbitrary direction $\boldsymbol{D}_t$ from the (Clarke) sub-differential [Clarke, 1990] of the ℓ_1-loss function $\partial f_{\ell_1}(\boldsymbol{U}_t)$. Due to local Lipschitzness of the ℓ_1-loss, the Clarke sub-differential exists and can be obtained via chain rule (see [Clarke, 1990]):

$$\partial f_{\ell_1}(\boldsymbol{U}_t) = \frac{1}{m}\sum_{i=1}^{m} \operatorname{sign}\left(\langle \boldsymbol{A}_i, \boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{X}^\star\rangle\right)\left(\boldsymbol{A}_i + \boldsymbol{A}_i^\top\right)\boldsymbol{U}_t. \tag{2.3}$$

At every iteration, SubGM updates the solution by moving towards $-\boldsymbol{D}_t$, where $\boldsymbol{D}_t \in \partial f_{\ell_1}(\boldsymbol{U}_t)$, with a step-size η_t. To showcase the effect of small initialization on the performance of SubGM, we consider an instance of robust matrix sensing, where the true solution $\boldsymbol{X}^*$ is a randomly generated matrix with rank $r = 3$ and dimension $d = 20$. Furthermore, we consider $m = 500$ measurements, where the measurement matrices $\{\boldsymbol{A}_i\}_{i=1}^m$ have i.i.d. standard Gaussian entries.

Property 1: Small initialization makes SubGM agnostic to over-parameterization. Figure 2.1a shows the performance of SubGM with small initialization



(a) (b)

Figure 2.2: (a) The objective value of the solutions obtained via SubGM with and without small initialization. (b) The error of the solutions obtained via SubGM with and without small initialization. In both instances, the initial point is chosen as $\boldsymbol{U}_0 = \alpha \boldsymbol{B}$, where $\boldsymbol{B}$ is obtained from Algorithm 2. The initialization scale α is chosen as $\alpha = 10^{-15}$ and $\alpha = 1$, for SubGM with and without small initialization, respectively.

for both exact ($r' = 3$) and over-parameterized ($r' > 3$) settings, where 10% of the measurements are grossly corrupted with noise. Our simulations uncover an intriguing property of small initialization: neither the convergence rate nor the final error of SubGM is affected by the over-estimation of the rank. Moreover, Figure 2.1b depicts the performance of SubGM for the fully over-parameterized problem (i.e., $r' = d = 20$) with different levels of corruption probability (the fraction of measurements that are corrupted with noise). The added noise values are i.i.d. and drawn from a Gaussian distribution with mean and variance equal to 0 and 100, respectively. It can be seen that, even in the fully over-parameterized setting, SubGM is robust against large corruption probabilities.

Property 2: Small initialization improves convergence. It is known that different variants of the (sub-)gradient method converge linearly to the true solution, provided that the search rank coincides with the true rank ($r' = r$) [Tu, Boczar, Simchowitz, Soltanolkotabi, and Recht, 2016, Zheng and Lafferty, 2015, Tong, Ma, and Chi, 2021, Li et al., 2020b]. However, these methods suffer from a dramatic, exponential slow-down in over-parameterized models with noisy measurements [Zhuo et al., 2021]. Our simulations reveal that small initialization can restore the convergence back to linear, even in the over-parameterized and noisy settings. Figure 2.1c shows that SubGM converges linearly to an error that is proportional to the norm of the initial point: smaller initial points lead to more accurate solutions at the expense of a slightly larger number of iterations.

Property 3: Emergence of "spurious" global minima. Inspired by these simulations, a natural approach to explaining the desirable performance of SubGM is by showing that the robust matrix sensing problem enjoys a benign landscape. We refute this conjecture by

showing that, not only does the robust matrix sensing with over-parameterized rank have sub-optimal solutions, but also its globally optimal solutions may be "spurious", i.e., they do not correspond to the ground truth $\boldsymbol{X}^*$. Figure 2.2 shows the performance of SubGM with and without small initialization. It can be seen that SubGM converges to the ground truth which is a local solution for the ℓ_1-loss with sub-optimal objective value. On the other hand, SubGM without small initialization converges to a high-rank solution with a strictly smaller objective value. In other words, the ground truth is not necessarily a globally optimal solution, and conversely, globally optimal solutions do not necessarily correspond to the ground truth.

From a statistical perspective, our simulations support the common empirical observation that first-order methods "generalize well". In particular, SubGM converges to a low-rank solution that is close to the ground truth—i.e., has a better *generalization error*—rather than recovering a high-rank solution with a smaller objective value (or better *training error*). The smaller objective values for higher rank solutions are precisely due to the overfitting phenomenon: it is possible that the globally optimal solution to equation 2.2 achieves a zero objective value by absorbing the noise into its redundant ranks. To circumvent the issue of overfitting, a common approach is to regularize the high-rank solutions in favor of the low-rank ones via different regularization techniques. Therefore, the desirable performance of SubGM with small initialization can be attributed to its *implicit regularization* property. In particular, we show that small initialization of SubGM is akin to implicitly regularizing the redundant rank of the over-parameterized model, thereby avoiding overfitting; a recent work [Stöger and Soltanolkotabi, 2021] has shown a similar property for the gradient descent algorithm on the noiseless matrix sensing with ℓ_2-loss.

2.1.2 Summary of Results

In this part, we present a summary of our results. Let σ_1 and σ_r be the largest and smallest (nonzero) eigenvalues of $\boldsymbol{X}^\star$, and define the condition number κ as σ_1/σ_r.

Theorem 1 (Convergence of SubGM; Informal)**.** *Suppose that the measurements satisfy a direction-preserving property delineated in Section 2.3.2. Suppose that the initial point is chosen as $\boldsymbol{U}_0 = \alpha\boldsymbol{B}$, for a special choice of $\boldsymbol{B}$ and an initialization scale α. Consider the iterations $\{\boldsymbol{U}_t\}_{t=0}^T$ generated by SubGM applied to the robust matrix sensing with step-size $\eta_t = \eta\rho^t$, for an appropriate choice of $0 < \rho < 1$ and sufficiently small η. Then, for any arbitrary accuracy $\epsilon > 0$ and initialization scale $\alpha = \mathcal{O}((\epsilon/d)^{1/\beta})$, we have*

$$\left\|\boldsymbol{U}_T\boldsymbol{U}_T^\top - \boldsymbol{X}^\star\right\|_{\mathrm{F}} \leq \epsilon \tag{2.4}$$

after $T = \mathcal{O}\left(\frac{\kappa \log^2(d/\epsilon)}{\beta\eta}\right)$ *iterations, where* $0 < \beta \leq 2$ *is a constant depending on the parameters of the problem.*

The above result characterizes the performance of SubGM for the robust matrix sensing with ℓ_1-loss. In particular, it shows that SubGM converges almost *linearly* to the true low-rank solution $\boldsymbol{X}^\star$, with a final error that is proportional to the initialization scale. Surprisingly, the required number of iterations is independent of the search rank r' and depends only logarithmically on d.

At the crux of our analysis lies a new restricted isometry property of the sub-differentials, which we call Sign-RIP. Under Sign-RIP, the sub-differentials of the ℓ_1-loss are δ-away from the sub-differentials of an ideal, expected loss function (see Section 2.3.2 for precise definitions). We will show that the classical notions of ℓ_2-RIP [Recht et al., 2010] and ℓ_1/ℓ_2-RIP [Li et al., 2020b] face major breakdowns in the presence of noise. In contrast, Sign-RIP provides much better robustness against noisy measurements, while being no more restrictive than its classical counterparts. We will show that, with Gaussian measurements, the Sign-RIP holds with an overwhelming probability under two popular noise models, namely *outlier noise model* and *Gaussian noise model*.

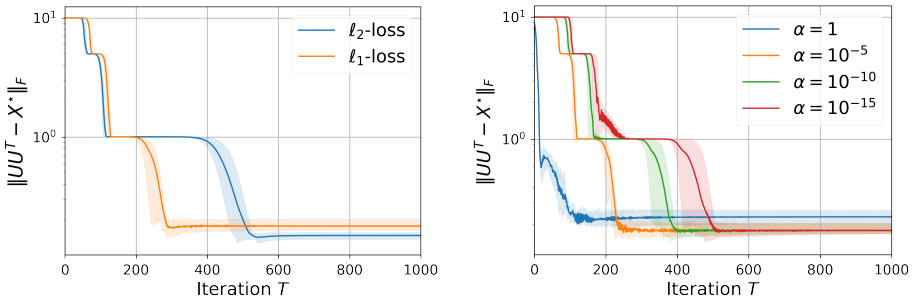
Our next theorem establishes the convergence of SubGM under the outlier noise model. To streamline the presentation, we use $\tilde{\mathcal{O}}(\cdot)$ and $\tilde{\Omega}(\cdot)$ to hide the dependency on logarithmic factors.

Theorem 2 (Convergence of SubGM under Outlier Noise Model; Informal)**.** *Suppose that the measurement matrices* $\{\boldsymbol{A}_i\}_{i=1}^m$ *have i.i.d. standard Gaussian entries, and a fraction* $p < 1$ *of the measurements are corrupted with arbitrarily large noise values. Suppose that the initial point is chosen as* $\boldsymbol{U}_0 = \alpha\boldsymbol{B}$*, for a special choice of* $\boldsymbol{B}$ *and a sufficiently small initialization scale* α*. Consider the iterations* $\{\boldsymbol{U}_t\}_{t=0}^T$ *generated by SubGM applied to the robust matrix sensing with an exponentially decaying step-size* $\eta_t = \eta\rho^t$*, for an appropriate choice of* $0 < \rho < 1$ *and sufficiently small* η*. Finally, suppose that the number of measurements satisfies* $m = \tilde{\Omega}\left(\kappa^4 dr^2/(1-p)^2\right)$*. Then, for any arbitrary accuracy* $\epsilon > 0$ *and initialization scale* $\alpha = \epsilon/d$*, and with an overwhelming probability, we have*

$$\left\|\boldsymbol{U}_T\boldsymbol{U}_T^\top - \boldsymbol{X}^\star\right\|_{\mathrm{F}} \leq \epsilon, \tag{2.5}$$

after $T = \mathcal{O}\left(\frac{\kappa \log^2(d/\epsilon)}{\eta}\right)$ *iterations.*

We compare Theorem 2 with a recent result by Ding et al. [2021] on the convergence of SubGM for robust matrix sensing. Theorem 2 shows that small initialization enables SubGM to converge almost linearly, which is exponentially faster than the sublinear rate



(a) (b)

Figure 2.3: (a) Performance of SubGM and GD under the Gaussian noise model with ℓ_1- and ℓ_2-loss functions, respectively. (b) The effect of small initialization on the performance of SubGM under the Gaussian noise model.

$\tilde{\mathcal{O}}(1/\epsilon)$ introduced by Ding et al. [2021]. Second, Ding et al. [2021] show that SubGM requires $\tilde{\Omega}(\kappa^{12} d r'^3)$ samples to converge, which depends on the search rank r'. In the over-parameterized regime, where the true rank is small (i.e., $r = \mathcal{O}(1)$) and the search rank is large (i.e., $r' = \Omega(d)$), our result leads to *three orders of magnitude* improvement in the required number of samples (modulo the dependency on κ). Moreover, Ding et al. [2021] crucially rely on the equivalence between globally optimal solutions and the ground truth, which only holds when $p \leq 1/\sqrt{r'}$. We relax this assumption and show that SubGM converges to the ground truth, even if p is arbitrarily close to 1.

Next, we turn our attention to the Gaussian noise model, and show that SubGM converges even if the measurements are corrupted with dense Gaussian noises.

Theorem 3 (Convergence of SubGM under Gaussian Noise Model; Informal)**.** *Suppose that the measurement matrices $\{\boldsymbol{A}_i\}_{i=1}^m$ have i.i.d. standard Gaussian entries, and each measurement is corrupted with a zero-mean Gaussian noise with a variance of at most ν^2. Suppose that the initial point is chosen as $\boldsymbol{U}_0 = \alpha \boldsymbol{B}$, for a special choice of $\boldsymbol{B}$ and a sufficiently small initialization scale α. Consider the iterations $\{\boldsymbol{U}_t\}_{t=0}^T$ generated by SubGM applied to the robust matrix sensing with exponentially decaying step-sizes $\eta_t = \eta\rho^t$, for an appropriate choice of $0 < \rho < 1$ and sufficiently small η. Finally, suppose that the number of measurements satisfies $m \gtrsim \tilde{\Omega}(\nu^2\kappa^4 d r^2)$. Then, with an overwhelming probability, we have*

$$\left\|\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{X}^\star\right\|_{\mathrm{F}} = \tilde{\mathcal{O}}\left(\sqrt{\frac{\nu^2 d r^2}{m}}\right), \tag{2.6}$$

after $T = \mathcal{O}\left(\frac{\kappa}{\eta}\log^2\left(\frac{m\kappa}{\nu r}\right)\right)$ iterations.

Traditionally, ℓ_2-loss has been used for recovering the ground truth under the Gaussian

noise model due to its correspondence to the so-called maximum likelihood estimation. Our paper extends the application of ℓ_1-loss to this setting, proving that SubGM is robust against not only the outlier, but also Gaussian noise levels. More precisely, Theorem 3 shows that SubGM outputs a solution with an estimation error of $\tilde{\mathcal{O}}(\sqrt{\nu^2 dr^2/m})$, which is again independent of the search rank r'. To the best of our knowledge, the sharpest known estimation error for gradient descent (GD) [Zhuo et al., 2021] and its variants [Zhang et al., 2021] on ℓ_2-loss is $\mathcal{O}(\sqrt{\nu^2 dr'/m})$, which scales with the search rank r'; in the fully over-parameterized regime, our provided bound improves upon this error by a factor of $\mathcal{O}(\sqrt{d/r})$. Figure 2.3 compares the performance of SubGM and GD on ℓ_1- and ℓ_2-losses, when the measurements are corrupted with Gaussian noise. Candes and Plan [2011] showed that *any* estimate $\widehat{\boldsymbol{U}}$ suffers from a minimax error of $\|\widehat{\boldsymbol{U}}\widehat{\boldsymbol{U}}^\top - \boldsymbol{X}^\star\|_{\mathrm{F}} = \Omega(\sqrt{\nu^2 dr/m})$. Compared to this information-theoretic lower bound, our provided final error is sub-optimal only by a factor of $\sqrt{r}$.

2.2 Related Work

Landscape v.s. Trajectory Analysis: It has been recently shown that different variants of low-rank matrix sensing (e.g., matrix completion [Ge et al., 2016], matrix sensing [Ge et al., 2017], robust PCA [Fattahi and Sojoudi, 2020]) enjoy benign landscape. In particular, it is shown that low-rank matrix sensing with ℓ_2-loss and noiseless measurements has a benign landscape in both exact [Ge et al., 2017, 2016, Zhang, Sojoudi, and Lavaei, 2019] and over-parameterized [Zhang, 2021] settings. On the other hand, it is known that ℓ_1-loss possesses better robustness against outlier noise. However, there are far fewer results characterizing the landscape of low-rank matrix sensing with ℓ_1-loss. Fattahi and Sojoudi [2020] and Josz, Ouyang, Zhang, Lavaei, and Sojoudi [2018] prove that robust matrix sensing with ℓ_1-loss has no spurious local solution, provided that with $r' = r = 1$ and the measurement matrices correspond to element-wise projections. This result was later extended to Gaussian matrices by Ma and Fattahi [2021b]. However, it is unclear whether these results extend to higher ranks or more general measurement matrices.

Despite its theoretical significance, benign landscape is too restrictive to hold in practice: Zhang et al. [2019] and Zhang [2021] show that spurious local minima are ubiquitous in the low-rank matrix sensing, even under fairly mild conditions. On the other hand, our experiments in Subsection 2.1.1 reveal that local-search algorithms may be able to avoid spurious local/global solutions with proper initialization. An alternative approach to explain the desirable performance of local-search algorithms is via *trajectory analysis*. It has been recently shown that the trajectories picked up by gradient-based algorithms benefit from implicit regularization [Gunasekar et al., 2018, Ma and Fattahi, 2022b], and incremental

Algorithm	Outlier noise model		Gaussian noise model	
Rank	$r' = r$	$r' \geq r$	$r' = r$	$r' \geq r$
GD+ℓ_2-loss	N/A	N/A	$\kappa \log\left(\frac{m}{\nu d r}\right)$ [25]	$\frac{\sigma_r}{\nu}\sqrt{\frac{m}{d}}$ [155]
SubGM+ℓ_1-loss	$\frac{r\kappa}{\sigma_r}\log\left(\frac{\sigma_r}{\epsilon}\right)$ [82]	$\frac{\sigma_1\kappa}{\epsilon}$ [39]	N/A	N/A
Our results	$\kappa^2 \log^3\left(\frac{d}{\epsilon}\right)$ (see Corollary 2)		$\kappa^2 \log^3\left(\frac{\kappa m}{\nu r}\right)$ (see Corollary 3)	

Table 2.1: A comparison between the **iteration complexity** of different techniques. We show that SubGM with small initialization and exponentially decaying step-size converges near-linearly to: (i) an arbitrary accuracy in the outlier noise model, and (ii) a nearly-minimax optimal error in the Gaussian noise model. Our derived iteration complexities are obtained from Theorems 2 and 3 after choosing an appropriate value for the step-size; see Corollaries 2 and 3 for the precise statements.

learning [Ma et al., 2023], or behave non-monotonically over short timescales, yet consistently improve over long timescales [Cohen, Kaur, Li, Kolter, and Talwalkar, 2021]. In the context of over-parameterized low-rank matrix sensing with ℓ_2-loss, Li, Ma, and Zhang [2018] and Stöger and Soltanolkotabi [2021] use trajectory analysis to show that GD with small initialization can recover the ground truth, provided that the measurements are noiseless. Zhuo et al. [2021] extend this result to the noisy setting, showing that GD converges to a minimax optimal solution at a sublinear rate, and with a number of samples that scale with the search rank.

Iteration and Sample Complexity: Despite their guaranteed convergence, local-search algorithms may suffer from notoriously slow convergence rates: whereas 10 digits of accuracy can be expected in just a few hundred iterations of GD when $r' = r$, tens of thousands of iterations might produce just 1-2 accurate digits once $r' > r$ [Zhang et al., 2021]. Table 2.1 shows the iteration complexity of the existing algorithms with different loss functions, compared to our proposed method. Evidently, under the outlier noise model, GD does not perform well due to the sensitivity of the ℓ_2-loss to outliers. On the other hand, SubGM converges linearly in the exact setting ($r' = r$), and at a significantly slower (sublinear) rate in the over-parameterized regime ($r' > r$). In contrast, our proposed SubGM algorithm with small initialization converges near-linearly in *both* exact and over-parameterized regimes. In the Gaussian noise model, it is known that GD converges linearly to a minimax optimal solution in the exact setting, but suffers from a drastic, exponential slow-down in the over-parameterized regime. In contrast, our proposed SubGM algorithm with small initialization is not affected by the over-parameterization, and maintains its desirable convergence rate in both settings.

Another important aspect of local-search algorithms is their sample complexity. Table 2.2 provides a comparison between the sample complexity of the existing algorithms, and our proposed method. In the outlier noise model, Li, Ma, Chen, and Chi [2019] show that SubGM

Algorithm	Outlier noise model		Gaussian noise model	
Rank	$r' = r$	$r' \geq r$	$r' = r$	$r' \geq r$
GD+ℓ_2-loss	N/A	N/A	$\nu^2\kappa^2 dr$ [25]	$\nu^2\kappa^2 dr'$ [155]
SubGM+ℓ_1-loss	$\kappa^2 dr^2$ [82]*	$\kappa^{12} dr'^3$ [39]*	N/A	N/A
Our results	$\frac{\kappa^4 dr^2}{(1-p)^2}$ (see Corollary 2)		$\nu^2\kappa^4 dr^2$ (see Corollary 3)	

Table 2.2: A comparison between the **sample complexity** of different techniques. Our results provide the best sample complexity bounds in the over-parameterized setting where $r' \gg r$, under both outlier and Gaussian noise models. For simplicity, we hide the dependency on the logarithmic factors. * Under an outlier noise model, the results of [Li et al., 2020b, Ding et al., 2021] holds under the assumption $p \lesssim 1/\sqrt{r'}$. In contrast, our result relaxes this assumption to $p < 1$.

with spectral initialization on ℓ_1-loss requires $\mathcal{O}(dr^2)$ samples (modulo the condition number), provided that the true rank is known ($r' = r$), and the corruption probability is upper bounded as $p \lesssim 1/\sqrt{r'}$. Ding et al. [2021] extend this result to the over-parameterized regime, showing that SubGM with spectral initialization requires $\mathcal{O}(dr'^3)$ samples to converge, under the same assumption $p \lesssim 1/\sqrt{r'}$. In both works, the upper bound $p \lesssim 1/\sqrt{r'}$ is imposed to guarantee that the global minima of the ℓ_1-loss correspond to the true solution. Our result relaxes this upper bound on the corruption probability by showing that SubGM converges to the ground truth, even if the ground truth is not globally optimal. In the Gaussian noise model, Zhuo et al. [2021] shows that GD recovers the true solution with $\mathcal{O}(dr')$ samples. In the over-parameterized regime, our result reduces the sample complexity to $\mathcal{O}(dr^2)$, showing that the sample complexity of SubGM is independent of the search rank r'.

Notations

For a rank-r matrix $\boldsymbol{M} \in \mathbb{R}^{m\times n}$, its singular values are denoted as $\|\boldsymbol{M}\| = \sigma_1(\boldsymbol{X}) \geq \sigma_2(\boldsymbol{X}) \geq \cdots \geq \sigma_r(\boldsymbol{X}) := \sigma_{\min}(\boldsymbol{X})$. For a square matrix $\boldsymbol{X} \in \mathbb{R}^{n\times n}$, its eigenvalues are defined as $\lambda_1(\boldsymbol{X}) \geq \lambda_2(\boldsymbol{X}) \geq \cdots \geq \lambda_n(\boldsymbol{X}) := \lambda_{\min}(\boldsymbol{X})$. For two matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ of the same size, their inner product is defined as $\langle \boldsymbol{X}, \boldsymbol{Y}\rangle = \mathrm{Tr}(\boldsymbol{X}^\top \boldsymbol{Y})$, where $\mathrm{Tr}(\cdot)$ is the trace operator. For a matrix $\boldsymbol{X}$, its operator and Frobenius norms are denoted as $\|\boldsymbol{X}\|$ and $\|\boldsymbol{X}\|_{\mathrm{F}}$, respectively. The unit rank-r sphere is defined as $\mathbb{S}_r^d = \{X \in \mathbb{R}^{d\times d} : \|\boldsymbol{X}\|_{\mathrm{F}} = 1, \mathrm{rank}(\boldsymbol{X}) \leq r\}$. We also omit the superscript d when there is no ambiguity. We define $\mathcal{P}_{\boldsymbol{V}}$ as the projection operator onto the row space of the matrix $\boldsymbol{V}$. The notation $\mathbb{B}(\boldsymbol{X}, \epsilon)$ refers to a ball of radius ϵ centered at $\boldsymbol{X}$, i.e., $\mathbb{B}(\boldsymbol{X}, \epsilon) = \{\boldsymbol{X}' : \|\boldsymbol{X}' - \boldsymbol{X}\|_{\mathrm{F}} \leq \epsilon\}$. The ℓ_q norm of a vector $\boldsymbol{x}$ is defined as $\|\boldsymbol{x}\|_q = (\sum |x_i|^q)^{1/q}$. Given two sequences $f(n)$ and $g(n)$, the notation $f(n) \lesssim g(n)$ implies that there exists a constant $\boldsymbol{C} < \infty$ satisfying $f(n) \leq Cg(n)$. Moreover, the notation $f(n) \asymp g(n)$ implies that $f(n) \lesssim g(n)$ and $g(n) \lesssim f(n)$. Throughout the paper, the symbols

$\boldsymbol{C}, c_1, c_2, \ldots$ refer to universal constants whose precise value may change according to the context. The sign function $\operatorname{sign}(\cdot)$ is defined as $\operatorname{sign}(x) = x/|x|$ if $x \neq 0$, and $\operatorname{sign}(0) = [-1, 1]$. For two sets $\mathcal{X}$ and $\mathcal{Y}$, the notation $\mathcal{X} + \mathcal{Y}$ refers to their Minkowski sum. Given two scalars a and b, the symbols $a \wedge b$ and $a \vee b$ are used to denote their minimum and maximum, respectively.

2.3 Our Overarching Framework

In this section, we present our overarching framework for the analysis of SubGM. To this goal, we first explain why the existing techniques for studying the smooth variants of the low-rank matrix sensing cannot be extended to their robust counterparts.

2.3.1 Failure of Existing Techniques

The majority of existing methods study the behavior of the gradient descent on ℓ_2-loss $f_{\ell_2}(\boldsymbol{U}) = \frac{1}{m}\left\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{U}\boldsymbol{U}^\top)\right\|^2$ by analyzing its deviation from an "ideal", noiseless loss function $\bar{f}_{\ell_2}(\boldsymbol{U}) = \|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{X}^\star\|_{\mathrm{F}}^2$. It is known that $\bar{f}_{\ell_2}(\boldsymbol{U}) = \|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{X}^\star\|_{\mathrm{F}}^2$ is devoid of spurious local minima, and its saddle points are strict, and hence, escapable (see [Zhang et al., 2020, Appendix A] for a simple proof). Therefore, by controlling the deviation of $f_{\ell_2}(\boldsymbol{U})$ and its gradients from $\bar{f}_{\ell_2}(\boldsymbol{U})$, one can show that $f_{\ell_2}(\boldsymbol{U})$ inherits the desirable properties of $\bar{f}_{\ell_2}(\boldsymbol{U})$. More concretely, the gradient of $f_{\ell_2}(\boldsymbol{U})$ can be written as $\nabla f_{\ell_2}(\boldsymbol{U}) = Q(\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{X}^\star)\boldsymbol{U}$, where $\boldsymbol{Q}(\boldsymbol{X}) = (2/m)\sum_{i=1}^m \left(\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle - s_i\right)\left(\boldsymbol{A}_i + \boldsymbol{A}_i^\top\right)$. One sufficient condition for $\nabla f_{\ell_2}(\boldsymbol{U}) \approx \nabla \bar{f}_{\ell_2}(\boldsymbol{U})$ is to ensure that $\boldsymbol{Q}(\boldsymbol{M})$ remains uniformly close to $\boldsymbol{M}$ for every rank-$(r + r')$ matrix $\boldsymbol{X}$. In the noiseless setting, this condition can be guaranteed via ℓ_2-RIP:

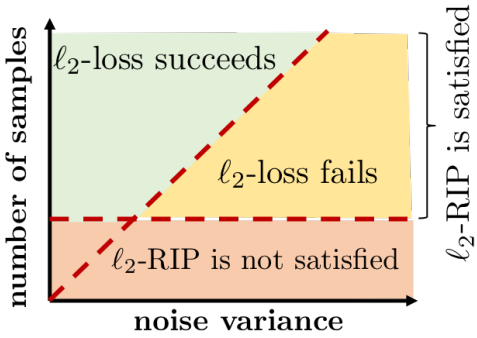


Figure 2.4: The number of samples to satisfy ℓ_2-RIP is independent of the noise variance. However, the performance of ℓ_2 highly depends on the noise variance.

Definition 1 (ℓ_2-RIP, Recht et al. [2010])**.** *The linear operator $\mathcal{A}(\cdot)$ satisfies ℓ_2-RIP with parameters (k, δ) if, for every rank-k matrix $\boldsymbol{M}$, we have $(1-\delta)\|\boldsymbol{M}\|_{\mathrm{F}}^2 \leq \frac{1}{m}\|\mathcal{A}(\boldsymbol{M})\|^2 \leq (1+\delta)\|\boldsymbol{M}\|_{\mathrm{F}}^2$.*

Roughly speaking, ℓ_2-RIP entails that the linear operator $\mathcal{A}(\cdot)$ is nearly "norm-preserving" for every rank-k matrix. In the noiseless setting, this implies that $\boldsymbol{Q}(\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{X}^\star) \approx 4(\boldsymbol{U}\boldsymbol{U}^\top -$

$\boldsymbol{X}^\star$), which in turn leads to $\nabla f_{\ell_2}(\boldsymbol{U}) \approx \nabla \bar{f}_{\ell_2}(\boldsymbol{U})$. On the other hand, it is known that ℓ_2-RIP is satisfied under mild conditions. For instance, (k,δ)-ℓ_2-RIP holds with $m = \Omega(dk/\delta^2)$ Gaussian measurements [Recht et al., 2010]. However, the next proposition shows that ℓ_2-RIP is not enough to guarantee $\boldsymbol{Q}(\boldsymbol{M}) \approx 4\boldsymbol{M}$ when the measurements are subject to noise.

Proposition 2 (Ma and Fattahi [2021a])**.** *Suppose that $r' = d$ and the measurement matrices $\{\boldsymbol{A}_i\}_{i=1}^m$ have i.i.d. standard Gaussian entries. Moreover, suppose that the noise vector $\boldsymbol{s}$ satisfies $s_i \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$ with probability p, and $s_i = 0$ with probability $1-p$, for every $i = 1,\ldots,m$. Then we have*

$$\mathbb{P}\left(\sup_{M\in\mathbb{S}} \|\boldsymbol{Q}(\boldsymbol{M}) - 4\boldsymbol{M}\|_{\mathrm{F}} \gtrsim \sqrt{\frac{(1+p\sigma^2)d^2}{m}}\right) \geq \frac{1}{2}.$$

Proposition 2 sheds light on a fundamental shortcoming of ℓ_2-RIP: in the presence of noise, it is possible for the measurements to satisfy ℓ_2-RIP, yet $\nabla f_{\ell_2}(\boldsymbol{U})$ may be far from $\nabla \bar{f}_{\ell_2}(\boldsymbol{U})$. In particular, we show that, in order to have $\nabla f_{\ell_2}(\boldsymbol{U}) \approx \nabla \bar{f}_{\ell_2}(\boldsymbol{U})$, the number of measurements must grow with the noise variance. On the other hand, for any fixed δ, ℓ_2-RIP is guaranteed to be satisfied with a number of measurements that is *independent* of the noise variance. Figure 2.4 shows that ℓ_2-RIP cannot capture the behavior of ℓ_2-loss in the high noise regime. Other notions of RIP, such as ℓ_1/ℓ_2-RIP [Li et al., 2020b], are also oblivious to the nature of the noise.

Algorithm 1 Subgradient Method

Input: measurement matrices $\{\boldsymbol{A}_i\}_{i=1}^m$, measurement vector $\boldsymbol{y} = [y_1,\cdots,y_m]^\top$, number of iterations $\boldsymbol{T}$, the initial point $\boldsymbol{U}_0$;
Output: Solution $\hat{\boldsymbol{X}}_T = \boldsymbol{U}_T\boldsymbol{U}_T^\top$ to equation 2.2;
for $t \leq T$ **do**
  Compute a sub-gradient $\boldsymbol{D}_t \in \partial f_{\ell_1}(\boldsymbol{U}_t)$;
  Select the step-size η_t (see equation 2.10);
  Set $\boldsymbol{U}_{t+1} \leftarrow \boldsymbol{U}_t - \eta_t \boldsymbol{D}_t$;
end for

2.3.2 Sign-RIP: A New Robust Restricted Isometry Property

To address the aforementioned challenges, we argue that, while the measurements may not be norm-preserving in the presence of noise, they may still enjoy a "direction-preserving" property. At the heart of our analysis lies the following decomposition of the sub-differential

of the ℓ_1-loss:

$$\partial f_{\ell_1}(\boldsymbol{U}) = \underbrace{\gamma \cdot \partial \bar{f}_{\ell_1}(\boldsymbol{U})}_{\text{expected sub-differential}} + \underbrace{\big(\partial f_{\ell_1}(\boldsymbol{U}) - \gamma \cdot \partial \bar{f}_{\ell_1}(\boldsymbol{U})\big)}_{\text{sub-differential deviation}},$$

where γ is a strictly positive number. In the above decomposition, the function $\bar{f}_{\ell_1}(\boldsymbol{U})$ is called the *expected loss*, and it is defined as $\left\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{X}^*\right\|_{\mathrm{F}}$. As will be shown later, $\bar{f}_{\ell_1}(\boldsymbol{U})$ captures the expectation of the *empirical loss* $f_{\ell_1}(\boldsymbol{U})$, when the measurement matrices have i.i.d. Gaussian entries. To analyze the behavior of SubGM on $f_{\ell_1}(\boldsymbol{U})$ (Algorithm 3), we first study the ideal scenario, where the loss deviation is zero, and hence, $f_{\ell_1}(\boldsymbol{U})$ coincides with its expectation. Under such an ideal scenario, we establish the global convergence of SubGM with small initialization. We then extend our result to the general case by carefully controlling the effect of sub-differential deviation. More specifically, we show that the desirable performance of SubGM extends to the empirical loss $f_{\ell_1}(\boldsymbol{U})$, provided that the sub-differentials are "direction-preserving", that is, $\boldsymbol{D} \approx \gamma \bar{\boldsymbol{D}}$ for every $\boldsymbol{D} \in \partial f_{\ell_1}(\boldsymbol{U})$ and $\bar{\boldsymbol{D}} \in \partial \bar{f}_{\ell_1}(\boldsymbol{U})$, where

$$\partial f_{\ell_1}(\boldsymbol{U}) = \left\{ \left(\boldsymbol{Q} + \boldsymbol{Q}^\top\right) \boldsymbol{U} : Q \in \mathcal{Q}(\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{X}^\star) \right\}, \text{ with } \mathcal{Q}(\boldsymbol{X}) = \frac{1}{m} \sum_{i=1}^{m} \operatorname{sign}(\langle \boldsymbol{A}_i, \boldsymbol{X} \rangle - s_i) \boldsymbol{A}_i. \tag{2.7}$$

Definition 2 (ϵ-approximate rank-k matrix)**.** *We say matrix $\boldsymbol{X}$ is ϵ-approximate rank-k if there exists a matrix $\boldsymbol{X}'$ with* $\operatorname{rank}(\boldsymbol{X}') \le k$*, such that* $\|\boldsymbol{X} - \boldsymbol{X}'\|_{\mathrm{F}} \le \epsilon$.

Definition 3 (Sign-RIP)**.** *The measurements are said to satisfy Sign-RIP with parameters $(k, \delta, \epsilon, \mathcal{S})$ and a uniformly positive and bounded scaling function $\varphi : \mathcal{S} \to \mathbb{R}$ over the set $\mathcal{S}$ if for every nonzero ϵ-approximate rank-k $\boldsymbol{X}, \boldsymbol{Y} \in \mathcal{S}$, and every $\boldsymbol{Q} \in \mathcal{Q}(\boldsymbol{X})$, we have*

$$\left\langle \boldsymbol{Q} - \varphi(\boldsymbol{X}) \frac{\boldsymbol{X}}{\|\boldsymbol{X}\|_{\mathrm{F}}}, \frac{\boldsymbol{Y}}{\|\boldsymbol{Y}\|_{\mathrm{F}}} \right\rangle \le \varphi(\boldsymbol{X}) \delta. \tag{2.8}$$

According to our definition, the scaling function satisfies $\underline{\varphi} \le \varphi(\boldsymbol{X}) \le \bar{\varphi}, \forall X \in \mathcal{S}$, for some constants $0 < \underline{\varphi} \le \bar{\varphi} < \infty$. Without loss of generality, we assume that $\underline{\varphi} \le 1 \le \bar{\varphi}$. Later, we will show that this assumption is satisfied for Gaussian measurements and different noise models. Whenever there is no ambiguity, we say the measurements satisfy $(k, \delta, \epsilon, \mathcal{S})$-Sign-RIP if they satisfy Sign-RIP with parameters $(k, \delta, \epsilon, \mathcal{S})$ and a (possibly unknown) uniformly positive and bounded scaling function $\varphi : \mathcal{S} \to \mathbb{R}$.

Next, we provide the intuition behind Sign-RIP. For any $\boldsymbol{U} \in \mathbb{R}^{d \times r'}$, the rank of $\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{X}^*$ is at most $r + r'$. Now, suppose that the measurements satisfy $(r' + r, \delta, \epsilon, \mathcal{S})$-Sign-RIP with

small δ and suitable choices of $\epsilon, \mathcal{S}$. Then, upon defining $\gamma = \varphi(\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{X}^\star) \leq \bar{\varphi}$, we have $\left\|\boldsymbol{D} - \gamma\bar{\boldsymbol{D}}\right\| \leq 2\bar{\varphi}\left\|\boldsymbol{U}\right\|\delta$ for every $\boldsymbol{D} \in \partial f_{\ell_1}(\boldsymbol{U})$ and $\bar{\boldsymbol{D}} \in \partial \bar{f}_{\ell_1}(\boldsymbol{U})$. In other words, for sufficiently small δ, $\partial f_{\ell_1}(\boldsymbol{U})$ and $\partial \bar{f}_{\ell_1}(\boldsymbol{U})$ are almost aligned under $(r'+r, \delta, \epsilon, \mathcal{S})$-Sign-RIP. A caveat of this analysis is that the required parameters of Sign-RIP depend on the search rank r'. One of the important contributions of this work is to relax this dependency by showing that every matrix in the sequence $\{\boldsymbol{U}_t\boldsymbol{U}_t{}^\top - \boldsymbol{X}^*\}_{t=0}^T$ generated by SubGM is ϵ-approximate rank-r, for some small $\epsilon > 0$.

At the first glance, one may speculate that Sign-RIP is extremely restrictive: it requires the uniform concentration of the set-valued function $\mathcal{Q}(\boldsymbol{X})$ over ϵ-approximate rank-k matrices. However, we show that, statistically, Sign-RIP is no more restrictive than ℓ_2- [Recht et al., 2010] and ℓ_1/ℓ_2-RIP [Li et al., 2020b], and—unlike its classical counterparts—holds under different noise models.

Definition 4 (Outlier Noise Model)**.** *With probability p, each entry of the noise vector $\mathbf{s}$ is independently drawn from a zero mean distribution $\mathbb{P}$; otherwise, it is set to zero.*

Notice that our proposed noise model does not impose any assumption on the magnitude of the nonzero elements of $\boldsymbol{s}$, or the specific form of their distribution, which makes it particularly suitable for modeling outliers with arbitrary magnitudes.

Definition 5 (Gaussian Noise Model)**.** *Each element of the noise vector $\mathbf{s}$ is independently drawn from a Gaussian distribution with zero mean and variance $\nu_g^2 < \infty$.*

Our next two theorems characterize the sample complexity of Sign-RIP under the outlier and Gaussian noise models.

Theorem 4 (Sign-RIP under Outlier Noise Model)**.** *Assume that the measurement matrices $\{\boldsymbol{A}_i\}_{i=1}^m$ defining the linear operator $\mathcal{A}(\cdot)$ have i.i.d. standard Gaussian entries, and that the noise vector $\boldsymbol{s}$ follows the outlier noise model with $0 \leq p < 1$ (Definition 4). Then, with probability of at least $1 - C_1 e^{-C_2 m(1-p)^2\delta^2}$, $(k, \delta, \epsilon, \mathcal{S})$-Sign-RIP holds with parameters $k \leq d$, $\delta \leq 1$, $\mathcal{S} = \{\boldsymbol{X} : \zeta \leq \|\boldsymbol{X}\|_{\mathrm{F}} \leq R\}$ for some arbitrary $R \geq \zeta > 0$, $\epsilon \lesssim \zeta\sqrt{k/m}$, and a scaling function $\varphi(\boldsymbol{X}) = \sqrt{\frac{2}{\pi}}\left(1 - p + p\mathbb{E}\left[e^{-s^2/(2\|\boldsymbol{X}\|_{\mathrm{F}}^2)}\right]\right)$, provided that the number of samples satisfies $m \gtrsim \frac{dk\log^2(m)\log(R/\zeta)}{(1-p)^2\delta^2}$.*

The proof of the above theorem is provided in Appendix 2.B.2. Theorem 4 shows that, for any fixed R, ζ, p, and δ, Sign-RIP is satisfied with $\tilde{\mathcal{O}}(dk)$ number of Gaussian measurements, which has the same order as ℓ_2- [Recht et al., 2010] and ℓ_1/ℓ_2-RIP [Li et al., 2020b] (modulo logarithmic factors). However, unlike ℓ_2- and ℓ_1/ℓ_2-RIP, Sign-RIP is *not* oblivious to noise. In particular, our theorem shows that Sign-RIP holds with a number of samples that scales

with $(1-p)^{-2}$, ultimately alleviating the issue raised in Subsection 2.3.1. Moreover, our result does not impose any restriction on p, which improves upon the assumption $p < 1/\sqrt{r'}$ made by Li et al. [2020b] and Ding et al. [2021].

Theorem 5 (Sign-RIP for Gaussian noise model)**.** *Assume that the measurement matrices $\{\boldsymbol{A}_i\}_{i=1}^m$ defining the linear operator $\mathcal{A}(\cdot)$ have i.i.d. standard Gaussian entries, and that the noise vector $\boldsymbol{s}$ follows the Gaussian noise model (Definition 5). Then, with probability of at least $1 - C_1 e^{-C_2 m\zeta^2\delta^2/\nu_g^2}$ $(k, \delta, \epsilon, \mathcal{S})$-Sign-RIP holds with parameters $k \leq m$, $\delta \leq 1$, $\mathcal{S} = \{\boldsymbol{X} : \zeta \leq \|\boldsymbol{X}\|_{\mathrm{F}} \leq R\}$ for some arbitrary $R \geq \zeta > 0$, $\epsilon \lesssim \zeta\sqrt{k/m}$, and a scaling function $\varphi(\boldsymbol{X}) = \sqrt{\frac{2}{\pi}} \frac{\|\boldsymbol{X}\|_{\mathrm{F}}}{\sqrt{\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \nu_g^2}}$, provided that the number of samples satisfies $m \gtrsim \frac{\nu_g^2 dk \log^2(m) \log(R/\zeta)}{\zeta^2\delta^2}$.*

The proof of the above theorem is provided in Appendix 2.B.3. Theorem 5 extends Sign-RIP beyond the outlier noise model, showing that it holds even when all measurements are corrupted with Gaussian noise. However, unlike the outlier noise model, the sample complexity of Sign-RIP scales with the noise variance.

2.3.3 Choice of Step-size

Next, we discuss our choice of the step-size, and its effect on the performance of SubGM. For simplicity, let $\boldsymbol{\Delta}_t = \boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{X}^\star$ and $\varphi_t = \varphi(\boldsymbol{U}_t\boldsymbol{U}_t - \boldsymbol{X}^\star)$. Under Sign-RIP, we have $\boldsymbol{D}_t \approx (\varphi_t/\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}) \cdot \boldsymbol{\Delta}_t\boldsymbol{U}_t$ for every $\boldsymbol{D}_t \in \partial f_{\ell_1}(\boldsymbol{U}_t)$. Therefore, the iterations of SubGM can be approximated as $\boldsymbol{U}_{t+1} \approx \boldsymbol{U}_t - (\eta_t\varphi_t/\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}) \cdot \boldsymbol{\Delta}_t\boldsymbol{U}_t$. Consequently, with the choice of $\eta_t = \eta\varphi_t^{-1}\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}$, the iterations of SubGM reduce to

$$U_{t+1} = \boldsymbol{U}_t - \eta \cdot \boldsymbol{\Delta}_t\boldsymbol{U}_t + \mathsf{deviation}.$$

Ignoring the deviation term, the above update coincides with the iterations of GD with a constant step-size η, applied to the expected loss function $\bar{f}_{\ell_2}(\boldsymbol{U}) = \|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{X}^\star\|_{\mathrm{F}}^2$. By controlling the effect of the deviation term, we show that SubGM on $\bar{f}_{\ell_2}(\boldsymbol{U})$ behaves similarly to GD with a constant step-size. A caveat of this analysis is that the proposed step-size $\eta_t = \eta\varphi_t^{-1}\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}$ depends on $\boldsymbol{X}^\star$, which is not known *a priori*. In the noiseless scenario, Sign-RIP can be invoked to show that $\varphi_t^{-1}\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}$ can be accurately estimated by $f_{\ell_1}(\boldsymbol{U}_t)$, as shown in the following lemma.

Lemma 1. *Suppose that the measurements are noiseless, and satisfy $(k, \delta, \epsilon, \mathcal{S})$-Sign-RIP for some $\delta \leq 1$, $k \leq d$, $\epsilon \geq 0$, $\mathcal{S} \neq \emptyset$, and uniformly positive and bounded scaling function $\varphi(\cdot)$. Moreover, suppose that $\boldsymbol{\Delta}_t$ is ϵ-approximate rank-k and $\boldsymbol{\Delta}_t \in \mathcal{S}$. Then, we have*

$$(1-\delta)\varphi_t \|\boldsymbol{\Delta}_t\|_{\mathrm{F}} \leq f_{\ell_1}(\boldsymbol{U}_t) \leq (1+\delta)\varphi_t \|\boldsymbol{\Delta}_t\|_{\mathrm{F}}. \tag{2.9}$$

The above lemma is the byproduct of a more general result presented in Appendix 2.B.4. It implies that, for small δ, the step-size $\eta_t = \eta f_{\ell_1}(\boldsymbol{U}_t)$ satisfies $\eta_t \approx \eta\varphi_t \left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}$, and hence, $\boldsymbol{U}_{t+1} \approx \boldsymbol{U}_t - \eta\varphi_t^2 \boldsymbol{\Delta}_t \boldsymbol{U}_t$, which again resembles the iterations of GD on $\bar{f}_{\ell_2}(\boldsymbol{U})$ with the "effective" step-size $\eta\varphi_t^2$.

In the noisy setting, the value of $\varphi_t^{-1}\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}$ *cannot* be estimated merely based on $f_{\ell_1}(\boldsymbol{U}_t)$, since $f_{\ell_1}(\boldsymbol{U}_t)$ is highly sensitive to the magnitude of the noise. To alleviate this issue, we propose an exponentially decaying step-size that circumvents the need for an accurate estimate of $\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}$. In particular, consider the following choice of step-size

$$\eta_t = \frac{\eta}{\|\boldsymbol{Q}_t\|} \cdot \rho^t, \quad \text{where} \quad \boldsymbol{Q}_t \in \mathcal{Q}(\boldsymbol{\Delta}_t), \tag{2.10}$$

for appropriate values of η and $0 < \rho < 1$. We note that the set $\mathcal{Q}(\boldsymbol{\Delta}_t)$ can be explicitly characterized without any prior knowledge on $\boldsymbol{\Delta}_t$:

$$\mathcal{Q}(\boldsymbol{\Delta}_t) = \frac{1}{m}\sum_{i=1}^{m} \operatorname{sign}\left(\langle \boldsymbol{A}_i, \boldsymbol{\Delta}_t\rangle - s_i\right) \boldsymbol{A}_i = \frac{1}{m}\sum_{i=1}^{m} \operatorname{sign}\left(\langle \boldsymbol{A}_i, \boldsymbol{U}_t\boldsymbol{U}_t^\top\rangle - y_i\right) \boldsymbol{A}_i.$$

Our next lemma shows that the above choice of step-size is well-defined (i.e., $\boldsymbol{Q}_t \neq 0$), so long as $\boldsymbol{\Delta}_t$ is not too small and the measurements satisfy $(k, \delta, \epsilon, \mathcal{S})$-Sign-RIP.

Lemma 2. *Suppose that the measurements satisfy $(k, \delta, \epsilon, \mathcal{S})$-Sign-RIP with $\delta < 2/(1+5\sqrt{k})$, $k \leq d$, $\epsilon > 0$, $\mathcal{S} \neq \emptyset$, and a uniformly positive and bounded scaling function $\varphi(\cdot)$. Moreover, suppose that $\boldsymbol{\Delta}_t$ is ϵ-approximate rank-k, $\|\boldsymbol{\Delta}_t\| \geq 4\epsilon$, and $\boldsymbol{\Delta}_t \in \mathcal{S}$. Then, we have*

$$\left(1 - \left(\frac{1+5\sqrt{k}}{2}\right)\delta\right) \frac{\eta\rho^t}{\varphi_t} \frac{\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}}{\|\boldsymbol{\Delta}_t\|} \leq \eta_t \leq \left(1 + \left(\frac{1+5\sqrt{k}}{2}\right)\delta\right) \frac{\eta\rho^t}{\varphi_t} \frac{\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}}{\|\boldsymbol{\Delta}_t\|}. \tag{2.11}$$

The proof of the above lemma can be found in Appendix 2.B.5. Lemma 2 implies that the chosen step-size remains close to $(\eta\rho^t/\varphi(\boldsymbol{\Delta}_t))(\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}/\|\boldsymbol{\Delta}_t\|)$, as long as the error is not close to zero. Due to Lemma 2, the iterations of SubGM with exponentially-decaying step-size can be approximated as

$$U_{t+1} = \boldsymbol{U}_t - \left(\frac{\eta\rho^t}{\|\boldsymbol{\Delta}_t\|}\right) \boldsymbol{\Delta}_t \boldsymbol{U}_t + \mathsf{deviation}. \tag{2.12}$$

In other words, SubGM selects an approximately correct direction of descent, while the exponentially decaying step-size ensures convergence to the ground truth.

2.3.4 Effect of Over-parameterization

At every iteration of SubGM, the rank of the error matrix $\boldsymbol{\Delta}_t = \boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{X}^\star$ may be as large as $r + r'$. Therefore, in order to guarantee the direction-preserving property of the sub-differentials, a sufficient condition is to satisfy Sign-RIP for every rank-$(r + r')$ matrix. Such crude analysis implies that the performance of SubGM may depend on the search rank r'. In particular, with Gaussian measurements, this would increase the required number of samples to $\tilde{\mathcal{O}}\left(\frac{dr'}{(1-p)^2\delta^2}\right)$, which scales linearly with the over-parameterized rank. To address this issue, we provide a finer analysis of the iterations. Consider the eigen-decomposition of $\boldsymbol{X}^\star$, given as

$$\boldsymbol{X}^\star = \begin{bmatrix}\boldsymbol{V} & \boldsymbol{V}_\perp\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{V}^\top\\ \boldsymbol{V}_\perp^\top\end{bmatrix}^\top = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^\top,$$

where $\boldsymbol{V} \in \mathbb{R}^{d\times r}$ and $\boldsymbol{V}_\perp \in \mathbb{R}^{d\times(d-r)}$ are (column) orthonormal matrices satisfying $\boldsymbol{V}^\top\boldsymbol{V}_\perp = \boldsymbol{0}$, and $\boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}$ is a diagonal matrix collecting the nonzero eigenvalues of $\boldsymbol{X}^*$. We assume that the diagonal entries of $\boldsymbol{\Sigma}$ are in decreasing order, i.e., $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. Moreover, without loss of generality, we assume that $\sigma_1 \geq 1 \geq \sigma_r$. Based on this eigen-decomposition, we introduce the *signal-residual decomposition* of $\boldsymbol{U}_t$ as follows:

$$\boldsymbol{U}_t = \boldsymbol{V}\boldsymbol{S}_t + \boldsymbol{V}_\perp\underbrace{(\boldsymbol{F}_t + \boldsymbol{G}_t)}_{\boldsymbol{E}_t}, \quad \text{where} \quad \boldsymbol{S}_t = \boldsymbol{V}^\top\boldsymbol{U}_t,\ \boldsymbol{E}_t = \boldsymbol{V}_\perp^\top\boldsymbol{U}_t,\ \boldsymbol{F}_t = \boldsymbol{E}_t\mathcal{P}_{\boldsymbol{S}_t},\ \boldsymbol{G}_t = \boldsymbol{E}_t\mathcal{P}_{\boldsymbol{S}_t}^\perp. \tag{2.13}$$

In the above expression, $\mathcal{P}_{\boldsymbol{S}_t}$ is the orthogonal projection onto the row space of $\boldsymbol{S}_t$, and $\mathcal{P}_{\boldsymbol{S}_t}^\perp$ is its orthogonal complement. It is easy to see that $\boldsymbol{U}_t\boldsymbol{U}_t^\top = \boldsymbol{X}^\star$ if and only if $\boldsymbol{S}_t\boldsymbol{S}_t^\top = \boldsymbol{\Sigma}$ and $\boldsymbol{E}_t\boldsymbol{E}_t^\top = 0$. Therefore, our goal is to show that $\boldsymbol{S}_t\boldsymbol{S}_t^\top$ and $\boldsymbol{E}_t\boldsymbol{E}_t^\top$ converge to $\boldsymbol{\Sigma}$ and 0, respectively. Based on the above signal-residual decomposition, one can write

$$\begin{aligned}\boldsymbol{\Delta}_t &= \boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{X}^\star\\ &= \underbrace{\boldsymbol{V}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right)\boldsymbol{V}^\top + \boldsymbol{V}\boldsymbol{S}_t\boldsymbol{E}_t^\top\boldsymbol{V}_\perp^\top + \boldsymbol{V}_\perp\boldsymbol{E}_t\boldsymbol{S}_t^\top\boldsymbol{V}^\top + \boldsymbol{V}_\perp\boldsymbol{F}_t\boldsymbol{F}_t^\top\boldsymbol{V}_\perp^\top}_{\text{rank-}4r} + \underbrace{\boldsymbol{V}_\perp\boldsymbol{G}_t\boldsymbol{G}_t^\top\boldsymbol{V}_\perp^\top}_{\text{small norm}}.\end{aligned}$$

An important implication of the above equation is that $\boldsymbol{\Delta}_t$ can be treated as an ϵ-approximate rank-$4r$ matrix, where $\epsilon = \left\|\boldsymbol{V}_\perp\boldsymbol{G}_t\boldsymbol{G}_t^\top\boldsymbol{V}_\perp^\top\right\|_{\mathrm{F}}$. We show that $\left\|\boldsymbol{V}_\perp\boldsymbol{G}_t\boldsymbol{G}_t^\top\boldsymbol{V}_\perp^\top\right\|_{\mathrm{F}} = \mathcal{O}(\sqrt{d}\alpha)$ throughout the solution trajectory, and hence, $\boldsymbol{\Delta}_t$ is approximately rank-$4r$, provided that the initialization scale α is small enough. To this goal, we first characterize the generalization error $\|\boldsymbol{\Delta}_t\|$ in terms of the *signal term* $\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{X}^\star\right\|$, *cross term* $\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\|$, and the *residual term* $\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|$.

Lemma 3. *We have*

$$\|\boldsymbol{\Delta}_t\| \leq \left\|\boldsymbol{\Sigma} - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| + 2\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\| + \left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|. \tag{2.14}$$

The proof of the above lemma follows directly from the signal-residual decomposition equation 2.13, and is omitted for brevity. Motivated by the above lemma, we will study the dynamics of the signal, cross, and residual terms under different settings.

2.4 Expected Loss

In this section, we consider a special scenario, where the measurements are noiseless, the measurement matrices $\{\boldsymbol{A}_i\}_{i=1}^m$ have i.i.d. standard Gaussian entries, and the number of measurements m approaches infinity. Evidently, these assumptions do not hold in practice. Nonetheless, our analysis of this ideal scenario will be the building block for our subsequent analysis. Since the number of measurements approaches infinity, the uniform law of large numbers implies that $f_{\ell_1}(\boldsymbol{U})$ converges to its expectation $\mathbb{E}[f_{\ell_1}(\boldsymbol{U})]$ almost surely, over any compact set of $\boldsymbol{U}$ [Wainwright, 2019]. The next lemma provides the explicit form of $\mathbb{E}[f_{\ell_1}(\boldsymbol{U})]$.

Lemma 4. *Suppose that the measurements are noiseless and the measurement matrices* $\{\boldsymbol{A}_i\}_{i=1}^m$ *have i.i.d. standard Gaussian entries. Then, we have*

$$\mathbb{E}[f_{\ell_1}(\boldsymbol{U})] = \sqrt{\frac{2}{\pi}}\left\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{X}^*\right\|_{\mathrm{F}}. \tag{2.15}$$

Proof. Due to the i.i.d. nature of $\{\boldsymbol{A}_i\}_{i=1}^m$ and the absence of noise, one can write $\mathbb{E}[f_{\ell_1}(\boldsymbol{U})] = \mathbb{E}\left[\left|\langle A, \boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{X}^*\rangle\right|\right]$, where A is random matrix with i.i.d. standard Gaussian entries. It is easy to see that $\langle A, \boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{X}^*\rangle$ is a Gaussian random variable with variance $\left\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{X}^*\right\|_{\mathrm{F}}^2$. The proof is completed by noting that, for a zero-mean Gaussian random variable $\boldsymbol{X}$ with variance σ^2, we have $\mathbb{E}[|X|] = \sqrt{\frac{2}{\pi}}\sigma$. □

Next, we study the performance of SubGM with small initialization for $\bar{f}_{\ell_1}(\boldsymbol{U}) = \sqrt{\pi/2}\mathbb{E}[f_{\ell_1}(\boldsymbol{U})]$. First, it is easy to see that $\partial\bar{f}_{\ell_1}(\boldsymbol{U}) = \frac{2\left(\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{X}^\star\right)\boldsymbol{U}}{\left\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{X}^\star\right\|_{\mathrm{F}}}$ for $\boldsymbol{U}\boldsymbol{U}^\top \neq X^*$. Moreover, $\partial\bar{f}_{\ell_1}(\boldsymbol{U})$ is nonempty and bounded for every $\boldsymbol{U}$ that satisfies $\boldsymbol{U}\boldsymbol{U}^\top = X^*$. Therefore, upon choosing the step-size as $\eta_t = (\eta/2)\left\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{X}^\star\right\|_{\mathrm{F}}$, the update rule for SubGM reduces to $\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta_t\boldsymbol{D}_t = \eta\left(\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{X}^\star\right)\boldsymbol{U}_t$, for any $\boldsymbol{D}_t \in \partial\bar{f}_{\ell_1}(\boldsymbol{U}_t)$. In other words, the iterations of SubGM with step-size $\eta_t = (\eta/2)\left\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{X}^\star\right\|_{\mathrm{F}}$ on $\bar{f}_{\ell_1}(\boldsymbol{U})$ are

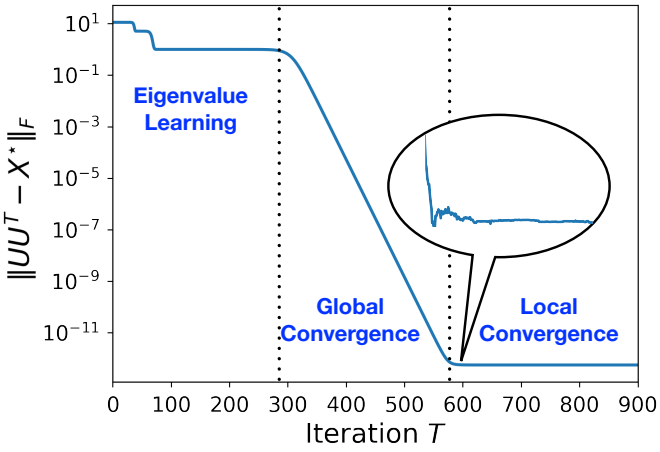


Figure 2.5: The iterations of SubGM for the expected loss equation 2.15 undergo three phases: eigenvalue learning phase, where the eigenvalues of $\boldsymbol{S}_t\boldsymbol{S}_t^\top$ converge to those of $\boldsymbol{X}^*$; global convergence phase, where the generalization error decays linearly; and local convergence phase, where the generalization error decays sub-linearly.

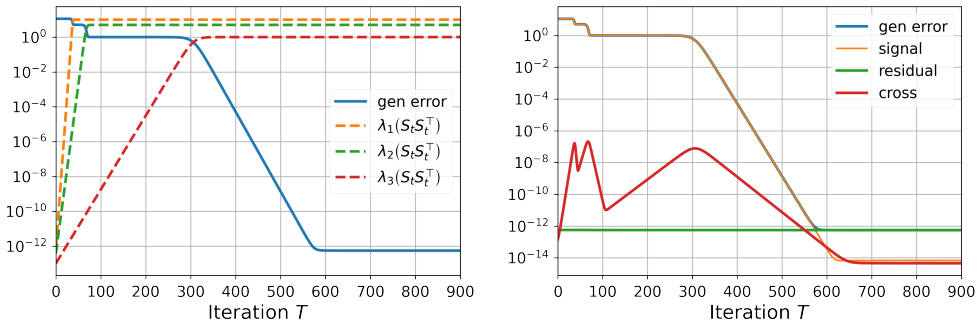
equivalent to the iterations of GD with constant step-size η on the expected ℓ_2-loss function $\bar{f}_{\ell_2}(\boldsymbol{U}) = (1/4)\left\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{X}^\star\right\|_{\mathrm{F}}^2$.

Due to this equivalence, we instead study the behavior of GD on $\bar{f}_{\ell_2}(\boldsymbol{U})$. Based on the decomposition of the generalization error in Lemma 3, we show that the iterations of SubGM on the expected loss undergo three phases:

- *Eigenvalue learning:* Due to small initialization, the signal, residual, and cross terms are small at the initial point. Therefore, the generalization error is dominated by the signal term $\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right\| \approx \|\boldsymbol{\Sigma}\|$. We show that, in the first phase, SubGM improves the generalization error by *learning the eigenvalues of* $\boldsymbol{X}^\star$, i.e., by reducing $\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right\|$. During this phase, the residual term $\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|$ will decrease at a sublinear rate.

- *Global convergence:* Once the eigenvalues are learned to a certain accuracy, both signal and cross terms $\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right\|$ and $\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\|$ start to decay at a linear rate, while the residual term maintains its sublinear decay rate.

- *Local convergence:* The discrepancy between the decay rates of the signal and cross terms, and that of the residual term implies that, at some point, the residual term becomes the dominant term, and hence, the generalization error starts to decay at a sublinear rate.



(a) (b)

Figure 2.6: (a) The eigenvalues of $\boldsymbol{X}^\star$ are learned at different rates. (b) During the eigenvalue learning phase, the generalization error is dominated by the signal term. In the global convergence phase, both signal and cross terms decay linearly. Finally, the residual term becomes the dominant term in the local convergence phase, and it governs the generalization error.

Figure 2.5 illustrates the three phases of SubGM on the expected loss $\bar{f}_{\ell_1}(\boldsymbol{U})$ with a rank-3 ground truth $\boldsymbol{X}^*$. Here, we assume that the problem is fully over-parameterized, i.e., $r' = d = 20$. A closer look at the first phase of the algorithm reveals that SubGM learns the eigenvalues of $\boldsymbol{X}^*$ at different rates: the larger eigenvalues are learned faster than the small ones (Figure 2.6a). A similar observation has been made for gradient flow applied to low-rank matrix factorization [Li, Luo, and Lyu, 2020c], and is referred to as *incremental learning* [Gidel, Bach, and Lacoste-Julien, 2019]. Finally, Figure 2.6b illustrates the dynamics of the signal, cross, and residual terms.

Proposition 3 (Minimum eigenvalue dynamic)**.** *Consider the iterations of SubGM for the expected loss $\bar{f}_{\ell_1}(\boldsymbol{U})$ with the step-size $\eta_t = (\eta/2)\bar{f}_{\ell_1}(\boldsymbol{U}_t)$. Suppose that $\eta \lesssim 1/\sigma_1$, $\boldsymbol{S}_t\boldsymbol{S}_t^\top \succ \boldsymbol{0}$, $\|\boldsymbol{E}_t\boldsymbol{E}_t\| \leq \sigma_1$, and $\|\boldsymbol{S}_t\boldsymbol{S}_t\| \leq 2\sigma_1$. Then, we have*

$$\lambda_{\min}\left(\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right) \geq \left((1+\eta\sigma_r)^2 - 2\eta\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|\right)\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right) - 2\eta\left(1+\eta\sigma_r\right)\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^2. \tag{2.16}$$

The proof of Proposition 3 can be found in Appendix 2.C.1. The above proposition shows that the minimum eigenvalue of $\boldsymbol{S}_t\boldsymbol{S}_t^\top$ grows exponentially fast at a rate of $1 + \Theta(\eta\sigma_r)$, provided that η and $\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|$ are small. This implies that the minimum eigenvalue satisfies $\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right) \gtrsim \sigma_r$ after $\mathcal{O}\left(\log\left(1/\lambda_{\min}(\boldsymbol{S}_0\boldsymbol{S}_0^\top)\right)/(\eta\sigma_r)\right)$ iterations.

Proposition 4 (Signal, cross, and residual dynamics)**.** *Consider the iterations of SubGM for the expected loss $\bar{f}_{\ell_1}(\boldsymbol{U})$, and with the step-size $\eta_t = (\eta/2)\bar{f}_{\ell_1}(\boldsymbol{U}_t)$. Suppose that $\eta \lesssim 1/\sigma_1$,*

$\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| \leq 1.01\sigma_1$ *and* $\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| \leq \sigma_1$. *Then, we have*

$$\left\|\boldsymbol{\Sigma}-\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right\| \leq \left(1-\eta\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right)\left\|\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| + 5\eta\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\|^2, \tag{2.17}$$

$$\left\|\boldsymbol{S}_{t+1}\boldsymbol{E}_{t+1}^\top\right\| \leq \left(1-\eta\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)+2\eta\left\|\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|+2\eta\left\|\boldsymbol{E}_t\boldsymbol{E}_t\right\|\right)\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\|, \tag{2.18}$$

$$\left\|\boldsymbol{E}_{t+1}\boldsymbol{E}_{t+1}^\top\right\| \leq \left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| - \eta\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|^2, \tag{2.19}$$

$$\left\|\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right\| \leq 1.01\sigma_1. \tag{2.20}$$

The proof of Proposition 4 can be found in Appendix 2.C.2. The above proposition shows that, once the minimum eigenvalue of $\boldsymbol{S}_t\boldsymbol{S}_t^\top$ approaches σ_r, the iterations enter the second phase, in which the signal and cross terms start to decay exponentially fast at the rate of $1-\Theta(\eta\sigma_r)$. Moreover, it shows that the residual term is independent of $\lambda_{\min}(\boldsymbol{S}_t\boldsymbol{S}_t^\top)$, and decreases sublinearly throughout the entire solution path. Given these dynamics, we present our main result.

Theorem 6 (Global convergence of SubGM for expected loss)**.** *Consider the iterations of SubGM for the expected loss* $\bar{f}_{\ell_1}(\boldsymbol{U})$ *with the step-size* $\eta_t = (\eta/2)\bar{f}_{\ell_1}(\boldsymbol{U}_t)$. *Suppose that* $\eta \lesssim 1/\sigma_1$ *and the initial point is selected such that* $\left\|\boldsymbol{U}_0\boldsymbol{U}_0^\top - 2\alpha^2 X^\star\right\| \leq \alpha^2\sigma_r$ *for some* $\alpha \lesssim \sqrt{\sigma_r}$. *Then, the following statements hold:*

- ***Linear convergence:*** *After* $\bar{T} \lesssim \frac{\log(\sigma_1/\alpha)}{\eta\sigma_r}$ *iterations, we have*

$$\left\|\boldsymbol{U}_{\bar{T}}\boldsymbol{U}_{\bar{T}}^\top - \boldsymbol{X}^*\right\| \lesssim \alpha^2.$$

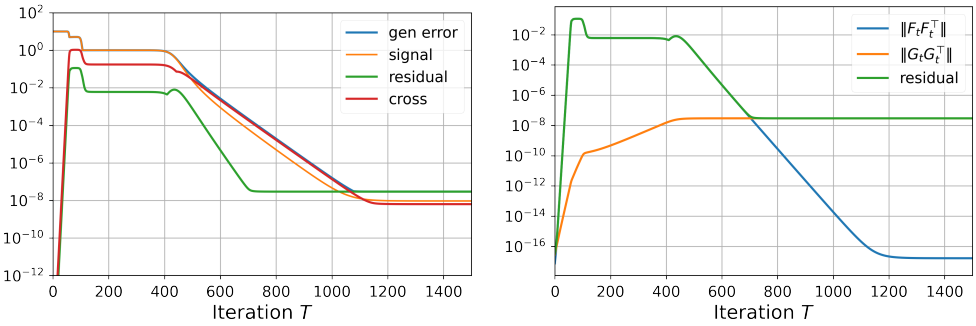
- ***Sub-linear convergence:*** *For every* $t \geq \bar{T}$, *we have*

$$\left\|\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{X}^\star\right\| \lesssim \frac{\alpha^2}{\eta\alpha^2 t+1}.$$

The detailed proof of Theorem 6 is presented in Section 2.A.1. According to the above theorem, for any accuracy $\epsilon > 0$, one can guarantee $\left\|\boldsymbol{U}_{\bar{T}}\boldsymbol{U}_{\bar{T}}^\top - \boldsymbol{X}^*\right\| \leq \epsilon$ after $\mathcal{O}\left(\log(\sigma_1/\epsilon)/(\eta\sigma_r)\right)$ iterations, provided that $\alpha \lesssim \sqrt{\epsilon} \wedge \sqrt{\sigma_r}$. In Section 2.5, we will show how to obtain an initial point that satisfies the conditions of Theorem 6.

2.5 Empirical Loss with Noiseless Measurements

A key difference between the behavior of SubGM for the empirical loss $f_{\ell_1}(\boldsymbol{U})$ and its expected counterpart $\bar{f}_{\ell_1}(\boldsymbol{U})$ is the fact that the residual term $\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|$ no longer enjoys a



(a) (b)

Figure 2.7: (a) The dynamics of the signal, cross, and residual terms for the empirical loss. Unlike the expected loss, the residual term for the empirical loss has a non-monotonic behavior. (b) The dynamics of $\left\|\boldsymbol{F}_t\boldsymbol{F}_t^\top\right\|$ and $\left\|\boldsymbol{G}_t\boldsymbol{G}_t^\top\right\|$. The non-monotonic behavior of the residual term can be attributed to the dynamic of $\left\|\boldsymbol{F}_t\boldsymbol{F}_t^\top\right\|$.

monotonically decreasing behavior. In particular, Figure 2.7a shows that, even with an infinitesimal initialization scale α, the residual term grows to a non-negligible value, before decaying linearly to a small level. In order to analyze this behavior, we further decompose $\boldsymbol{E}_t$ as

$$\boldsymbol{E}_t = \boldsymbol{F}_t + \boldsymbol{G}_t, \qquad \text{where} \qquad \boldsymbol{F}_t = \boldsymbol{E}_t\mathcal{P}_{\boldsymbol{S}_t}, \quad \text{and} \quad \boldsymbol{G}_t = \boldsymbol{E}_t\mathcal{P}_{\boldsymbol{S}_t}^{\perp}.$$

Based on the above decomposition and Lemma 3, the generalization error can be written as:

$$\left\|\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{X}^\star\right\| \leq \left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right\| + 2\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\| + \left\|\boldsymbol{F}_t\boldsymbol{F}_t^\top\right\| + \left\|\boldsymbol{G}_t\boldsymbol{G}_t^\top\right\|. \tag{2.21}$$

This decomposition plays a key role in characterizing the behavior of the residual term: we show that the increasing nature of $\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|$ in the initial stage of the algorithm can be attributed to the dynamic of $\left\|\boldsymbol{F}_t\boldsymbol{F}_t^\top\right\|$. During this phase, the term $\left\|\boldsymbol{G}_t\boldsymbol{G}_t^\top\right\|$ also increases, but at a much slower rate. In particular, we show that $\left\|\boldsymbol{G}_t\boldsymbol{G}_t^\top\right\|$ remains in the order of α^γ for some $0 < \gamma \leq 2$ throughout the entire solution path. In the second phase, $\left\|\boldsymbol{G}_t\boldsymbol{G}_t^\top\right\|$ remains roughly in the same order, while $\left\|\boldsymbol{F}_t\boldsymbol{F}_t^\top\right\|$ decays linearly until it is dominated by $\left\|\boldsymbol{G}_t\boldsymbol{G}_t^\top\right\|$. At the end of this phase, the overall error will be in the order of $\mathcal{O}(\alpha^\gamma)$. Figure 2.7b illustrates the behavior of $\left\|\boldsymbol{F}_t\boldsymbol{F}_t^\top\right\|$ and $\left\|\boldsymbol{G}_t\boldsymbol{G}_t^\top\right\|$, together with $\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|$.

Similar to our analysis for the expected loss, our first step towards analyzing the behavior of SubGM is to characterize the dynamic of the minimum eigenvalue of $\boldsymbol{S}_t\boldsymbol{S}_t^\top$. For simplicity of notation, we define $\bar{\eta}_t = \eta\varphi(\boldsymbol{\Delta}_t)^2$ in the sequel. Recall that, due to our assumption on $\varphi(\boldsymbol{\Delta}_t)^2$, we have $\underline{\varphi}^2\eta \leq \bar{\eta}_t \leq \bar{\varphi}^2\eta$, provided that $\boldsymbol{\Delta}_t \in \mathcal{S}$.

Proposition 5 (Minimum eigenvalue dynamic)**.** *Consider the iterations of SubGM for the*

empirical loss $f_{\ell_1}(\boldsymbol{U})$ *with the step-size* $\eta_t = (\eta/2) f_{\ell_1}(\boldsymbol{U}_t)$. *Suppose that the measurements are noiseless and satisfy* $(4r, \delta, \epsilon, \mathcal{S})$*-Sign-RIP with* $\delta \lesssim 1/\sqrt{r}$, $\epsilon = \sqrt{d}\,\|\boldsymbol{G}_t\|^2$, *and* $\mathcal{S} = \{\boldsymbol{X} : \zeta \le \|\boldsymbol{X}\|_{\mathrm{F}} \le R\}$ *for* $\zeta = \epsilon\left(1/\delta \vee \sqrt{d}\right)$ *and* $R = 5\sqrt{r}\sigma_1$. *Moreover, suppose that* $\eta \lesssim 1/(\bar{\varphi}^2\sigma_1)$, $\boldsymbol{S}_t\boldsymbol{S}_t^\top \succ \boldsymbol{0}$, $\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| \le \sigma_1$, $\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| \le 2\sigma_1$, $\boldsymbol{\Delta}_t \in \mathcal{S}$ *is* ϵ*-approximate rank-*$4r$, *and* $\left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top(\boldsymbol{S}_t\boldsymbol{S}_t^\top)^{-1}\right\| \le 1/3$. *Then, we have*

$$\begin{aligned}\lambda_{\min}\left(\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right) \ge \left((1+\bar{\eta}_t\sigma_r)^2 - 2\bar{\eta}_t\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| - 72\bar{\eta}_t\delta\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}\right)\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right) \\ - 2\bar{\eta}_t\left(1+\bar{\eta}_t\sigma_r\right)\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^2.\end{aligned}$$

The proof of Proposition 5 can be found in Appendix 2.D.2. Later, we will show that the conditions of Proposition 5 are satisfied with a sufficiently small initial point. The above proposition shows that, in the first phase of the algorithm, $\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)$ grows exponentially with a rate of at least $1 + \Omega(\eta\underline{\varphi}^2\sigma_r)$. Comparing this result with Proposition 3 reveals that $\lambda_{\min}\left(\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right)$ for the empirical loss behaves almost the same as its expected counterpart. This will play an important role in establishing the linear convergence of SubGM for the empirical loss. Finally, note that Sign-RIP must be satisfied for every ϵ-approximate rank-$4r$ matrix, where $\epsilon = \sqrt{d}\,\|\boldsymbol{G}_t\|^2$. Later, we will show that, with small initialization, the value of $\sqrt{d}\,\|\boldsymbol{G}_t\|^2$ scales with α, and hence, can be kept small throughout the iterations. Our next proposition characterizes the behavior of the signal and cross terms for the empirical loss.

Proposition 6 (Signal and cross dynamics)**.** *Consider the iterations of SubGM for the empirical loss* $f_{\ell_1}(\boldsymbol{U})$ *with the step-size* $\eta_t = (\eta/2) f_{\ell_1}(\boldsymbol{U}_t)$. *Suppose that the measurements are noiseless and satisfy* $(4r, \delta, \epsilon, \mathcal{S})$*-Sign-RIP with* $\delta \lesssim 1/\sqrt{r}$, $\epsilon = \sqrt{d}\,\|\boldsymbol{G}_t\|^2$, *and* $\mathcal{S} = \{\boldsymbol{X} : \zeta \le \|\boldsymbol{X}\|_{\mathrm{F}} \le R\}$ *for* $\zeta = \epsilon\left(1/\delta \vee \sqrt{d}\right)$ *and* $R = 5\sqrt{r}\sigma_1$. *Moreover, suppose that* $\eta \lesssim 1/(\bar{\varphi}^2\sigma_1)$, $\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| \le 1.01\sigma_1$, $\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| \le \sigma_1$, $\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|_{\mathrm{F}} \le \sqrt{r}\sigma_1$, *and* $\boldsymbol{\Delta}_t \in \mathcal{S}$ *is* ϵ*-approximate rank-*$4r$. *Then, we have*

$$\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right\| \le \left(1 - \bar{\eta}_t\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right)\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| + 5\bar{\eta}_t\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\|^2 + 37\bar{\eta}_t\delta\sigma_1\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}, \tag{2.22}$$

$$\begin{aligned}\left\|\boldsymbol{S}_{t+1}\boldsymbol{E}_{t+1}^\top\right\| \le &\left(1 - \bar{\eta}_t\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right) + 2\bar{\eta}_t\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| + 2\bar{\eta}_t\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|\right)\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\| \\ &+ 22\bar{\eta}_t\delta\sigma_1\|\boldsymbol{\Delta}_t\|_{\mathrm{F}},\end{aligned} \tag{2.23}$$

$$\left\|\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right\| \le 1.01\sigma_1. \tag{2.24}$$

The proof of this proposition is presented in Appendix 2.D.3. Proposition 6 shows that, under Sign-RIP, the one-step dynamics of the signal and cross terms behave almost the same as their expected counterparts, provided that δ is sufficiently small.

Finally, we provide the one-step dynamic of the residual term. To this goal, we will

Algorithm 2 Initialization Scheme

Input: initialization scale α, measurement matrices $\{\boldsymbol{A}_i\}_{i=1}^m$, measurement vector $\boldsymbol{y} = [y_1, \cdots, y_m]^\top$, and the search rank r';
Output: An initialization matrix $\boldsymbol{U}_0 \in \mathbb{R}^{d\times r'}$;
Obtain $\boldsymbol{C} \in \frac{1}{m}\sum_{i=1}^m \operatorname{sign}(y_i)\frac{\boldsymbol{A}_i+\boldsymbol{A}_i^\top}{2}$;
Compute the eigenvalue decomposition $\boldsymbol{C} = \boldsymbol{\Lambda}\boldsymbol{D}\boldsymbol{\Lambda}^\top$;
Define $\boldsymbol{D}_+^{r'}$ as the top $r'\times r'$ sub-matrix of $\boldsymbol{D}$ corresponding to the r' largest eigenvalues of $\boldsymbol{C}$, whose negative values are replaced by 0;
Set $\boldsymbol{U}_0 = \alpha\boldsymbol{\Lambda}\left(\boldsymbol{D}_+^{r'}\right)^{1/2}$.

separately analyze $\boldsymbol{F}_t$ and $\boldsymbol{G}_t$, i.e., the projection of $\boldsymbol{E}_t$ onto the row space of $\boldsymbol{S}_t$ and its orthogonal complement. This together with $\boldsymbol{E}_t = \boldsymbol{F}_t + \boldsymbol{G}_t$ characterizes the dynamic of the residual term.

Proposition 7 (Residual dynamic)**.** *Consider the iterations of SubGM for the empirical loss $f_{\ell_1}(\boldsymbol{U})$ with the step-size $\eta_t = (\eta/2)f_{\ell_1}(\boldsymbol{U}_t)$. Suppose that the measurements are noiseless and satisfy $(4r,\delta,\epsilon,\mathcal{S})$-Sign-RIP with $\delta \lesssim 1/\sqrt{r}$, $\epsilon = \sqrt{d}\left\|\boldsymbol{G}_t\right\|^2$, and $\mathcal{S} = \{\boldsymbol{X} : \zeta \le \|\boldsymbol{X}\|_{\mathrm{F}} \le R\}$ for $\zeta = \epsilon\left(1/\delta \vee \sqrt{d}\right)$ and $R = 5\sqrt{r}\sigma_1$. Moreover, suppose that $\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| \le 1.01\sigma_1$, $\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| \le \sigma_1$, $\boldsymbol{\Delta}_t \in \mathcal{S}$ is ϵ-approximate rank-$4r$, and $\|\boldsymbol{E}_t\boldsymbol{S}_t(\boldsymbol{S}_t\boldsymbol{S}_t)^{-1}\| \le 1/3$. Then, the following statements hold:*

- *If $\bar{\eta}_t \lesssim 1/\left\|\boldsymbol{\Delta}_t\right\|$, we have*

$$\left\|\boldsymbol{G}_{t+1}\right\| \le \left(1+\bar{\eta}_t^2\left(2\left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top\right\|^2+\left\|\boldsymbol{E}_t\right\|^4+2\left\|\boldsymbol{\Delta}_t\right\|\left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top\right\|\right)+7\bar{\eta}_t\delta\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}\right)\left\|\boldsymbol{G}_t\right\|.$$

- *If $\eta \lesssim 1/(\bar{\varphi}^2\sigma_1)$, we have*

$$\left\|\boldsymbol{F}_{t+1}\right\| \le \left(1-\bar{\eta}_t\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)+3\bar{\eta}_t\delta\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}\right)\left\|\boldsymbol{F}_t\right\|+3\bar{\eta}_t\delta\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}\left\|\boldsymbol{S}_t\right\|+6\bar{\eta}_t\left\|\boldsymbol{\Delta}_t\right\|\left\|\boldsymbol{G}_t\right\|.$$

The proof of the above proposition can be found in Appendix 2.D.4. Note that the condition $\eta \lesssim 1/(\bar{\varphi}^2\sigma_1)$ for the dynamic of $\|\boldsymbol{F}_t\|$ readily implies $\bar{\eta}_t \lesssim 1/\|\boldsymbol{\Delta}_t\|$. Therefore, the one-step dynamic of $\|\boldsymbol{G}_t\|$ holds under a milder condition on the step-size. Moreover, unlike $\|\boldsymbol{F}_t\|$, the dynamic of $\|\boldsymbol{G}_t\|$ is independent of $\lambda_{\min}(\boldsymbol{S}_t\boldsymbol{S}_t^\top)$. At the early stages of the algorithm, the term $\bar{\eta}_t^2\left(2\left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top\right\|^2+\left\|\boldsymbol{E}_t\right\|^4+2\left\|\boldsymbol{\Delta}_t\right\|\left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top\right\|\right) = \mathcal{O}(\alpha^2)$ is dominated by $\bar{\eta}_t\delta\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}} \approx \bar{\eta}_t\delta\left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}}$. Therefore, $\|\boldsymbol{G}_t\|$ grows at a slow rate of $1+\mathcal{O}(1)\eta\delta\bar{\varphi}^2\left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}}$. As the algorithm makes progress, $\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}$ decreases, leading to an even slower growth rate for $\|\boldsymbol{G}_t\|$. This is in line with the behavior of $\|\boldsymbol{G}_t\|$ in Figure 2.7b: the growth rate of $\|\boldsymbol{G}_t\|$ decreases as SubGM makes progress towards the ground truth, and it eventually "flattens

out" at a level proportional to the initialization scale. However, unlike $\|\boldsymbol{G}_t\|$, the term $\|\boldsymbol{F}_t\|$ does not have a monotonic behavior. In particular, according to Proposition 7, $\|\boldsymbol{F}_t\|$ may increase at the early stages of the algorithm, where $\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)$ is negligible compared to $\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}$. However, $\|\boldsymbol{F}_t\|$ will start decaying as soon as $\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right) \gtrsim \delta\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}$, which, according to Proposition 5, is guaranteed to happen after certain number of iterations. The non-monotonic behavior of $\|\boldsymbol{F}_t\|$ is also observed in practice (see Figure 2.7b).

Before presenting the main result, we provide our proposed initialization scheme in Algorithm 2. The presented initialization method is analogous to the classical spectral initialization in the noiseless matrix sensing problems, with a key difference that we scale down the norm of the initial point by a factor of α^2. As will be shown later, the scaling of the initial point is crucial for establishing the linear convergence of SubGM; without such scaling, both GD and SubGM suffer from sublinear convergence rates, as evidenced by the recent works [Zhuo et al., 2021, Ding et al., 2021].

Theorem 7 (Global Convergence of SubGM with Noiseless Measurements)**.** *Consider the iterations of SubGM for the empirical loss $f_{\ell_1}(\boldsymbol{U})$ with the step-size $\eta_t = (\eta/2)f_{\ell_1}(\boldsymbol{U}_t)$. Suppose that the initial point $\boldsymbol{U}_0$ is obtained from Algorithm 2 with an initialization scale that satisfies $\alpha \lesssim 1/(\bar{\varphi}\sqrt{d}) \wedge 1/\kappa$. Suppose that the measurements are noiseless and satisfy $(4r, \delta, \epsilon, \mathcal{S})$-Sign-RIP with $\delta \lesssim 1/\left(\sqrt{r}\kappa^2\bar{\varphi}^4\log^2(1/\alpha)\right)$, $\epsilon \asymp \sqrt{d}\alpha^{2-\mathcal{O}\left(\sqrt{r}\kappa^2\delta\right)}\delta$, and $\mathcal{S} = \{\boldsymbol{X} : \zeta \le \|\boldsymbol{X}\|_{\mathrm{F}} \le R\}$ for $\zeta = \epsilon\left(1/\delta \vee \sqrt{d}\right)$ and $R = 5\sqrt{r}\sigma_1$. Finally, suppose that $\eta \lesssim 1/(\bar{\varphi}^2\sigma_1)$. Then, after $t = T_{end} \lesssim \log(1/\alpha)/(\eta\sigma_r\underline{\varphi}^2)$ iterations, we have*

$$\left\|\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{X}^\star\right\|_{\mathrm{F}} \lesssim d\alpha^{2-\mathcal{O}\left(\sqrt{r}\kappa^2\delta\right)}.$$

The proof of Theorem 7 is presented in Subsection 2.A.2, and follows the same structure as the proof of Theorem 6. However, unlike the expected loss, the final error will be in the order of $\alpha^{\beta(\delta)}$, for some $0 < \beta(\delta) \le 2$ that is a decreasing function of δ. Indeed, smaller δ will improve the dependency of the final generalization error on α. Moreover, for an arbitrarily small $\epsilon > 0$, one can guarantee $\left\|\boldsymbol{U}_T\boldsymbol{U}_T^\top - \boldsymbol{X}^\star\right\|_{\mathrm{F}} \le \epsilon$ within $T \lesssim \log(d/\epsilon)/(\beta(\delta)\eta\sigma_r\underline{\varphi}^2)$ iterations, provided that the initialization scale satisfies $\alpha \lesssim (\epsilon/d)^{1/\beta(\delta)}$.

Finally, we characterize the sample complexity of SubGM with noiseless, Gaussian measurements.

Corollary 1 (Gaussian Measurements)**.** *Suppose that the measurement matrices $\{\boldsymbol{A}_i\}_{i=1}^m$ have i.i.d. standard Gaussian entries. Consider the iterations of SubGM for the empirical loss $f_{\ell_1}(\boldsymbol{U})$, with the step-size $\eta_t = (\eta/2)f_{\ell_1}(\boldsymbol{U}_t)$ and $\eta \lesssim 1/\sigma_1$. Suppose that the initial point $\boldsymbol{U}_0$ is obtained from Algorithm 2 with an initialization scale that satisfies $\alpha \lesssim 1/\sqrt{d} \wedge 1/\kappa$. Finally, suppose that the number of measurements satisfies $m \gtrsim \kappa^4 dr^2\log^5(1/\alpha^2)\log^2(m)$.*

Then, after $t = T_{end} \lesssim \log(1/\alpha)/(\eta\sigma_r)$ *iterations, and with an overwhelming probability, we have*

$$\left\|\boldsymbol{U}_T\boldsymbol{U}_T^\top - \boldsymbol{X}^\star\right\|_{\mathrm{F}} \lesssim d\alpha^{2-\mathcal{O}\left(\sqrt{\frac{\kappa^4 dr^2\log(1/\alpha)\log^2(m)}{m}}\right)}.$$

The above corollary is a direct consequence of Theorem 4 after setting the corruption probability p to zero. To the best of our knowledge, Corollary 1 is the first result showing that, with Gaussian measurements, the sample complexity of SubGM is *independent* of the search rank, provided that the initial point is sufficiently close to the origin.

2.6 Empirical Loss with Noisy Measurements

In this section, we establish the convergence of SubGM with small initialization and noisy measurements. A key difference compared to our previous analysis is in the choice of the step-size: in the presence of noise, the value of $\eta_t = (\eta/2)f_{\ell_1}(\boldsymbol{U}_t)$ can be arbitrarily far from the error $\eta\varphi_t^{-1}\|\boldsymbol{\Delta}_t\|$. To circumvent this issue, we instead propose to use the following geometric step-size:

$$\eta_t = \eta\cdot\frac{\rho^t}{\|\boldsymbol{Q}_t\|}, \quad \text{where} \quad \boldsymbol{Q}_t \in \mathcal{Q}(\boldsymbol{\Delta}_t). \tag{2.25}$$

Our first goal is to show that, under a similar Sign-RIP condition, our previous guarantees on SubGM extend to geometric step-size. Then, we show how our general result can be readily tailored to specific noise models. Our next result characterizes the dynamic of $\lambda_{\min}(\boldsymbol{S}_t\boldsymbol{S}_t^\top)$ with the above choice of step-size.

Proposition 8 (Minimum eigenvalue dynamic)**.** *Consider the iterations of SubGM on* $f_{\ell_1}(\boldsymbol{U})$ *with the step-size defined as equation 2.25. Suppose that the measurements satisfy* $(4r,\epsilon,\delta,\mathcal{S})$*-Sign-RIP with* $\delta \lesssim 1/\sqrt{r}$*,* $\epsilon = \sqrt{d}\,\|\boldsymbol{G}_t\|^2$*, and* $\mathcal{S} = \{\boldsymbol{X} : \zeta \le \|\boldsymbol{X}\|_{\mathrm{F}} \le R\}$ *for* $\zeta = \epsilon\left(1/\delta \vee \sqrt{d}\right)$ *and* $R = 5\sqrt{r}\sigma_1$*. Moreover, suppose that* $\boldsymbol{S}_t\boldsymbol{S}_t^\top \succ \boldsymbol{0}$*,* $\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| \le \sigma_1$*,* $\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| \le 2\sigma_1$*,* $\left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\right\| \le 1/3$*,* $\frac{\eta\rho^t}{\|\boldsymbol{\Delta}_t\|} \lesssim \frac{1}{\sigma_1}$*, and* $\boldsymbol{\Delta}_t \in \mathcal{S}$ *is* ϵ*-approximate rank-*$4r$*. Then, we have*

$$\begin{aligned}\lambda_{\min}\left(\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right) \ge \left(\left(1+\frac{\eta\rho^t}{\|\boldsymbol{\Delta}_t\|}\sigma_r\right)^2 - \frac{2\eta\rho^t}{\|\boldsymbol{\Delta}_t\|}\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| - 384\sqrt{r}\eta\rho^t\delta\right)\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\\ - 2\frac{\eta\rho^t}{\|\boldsymbol{\Delta}_t\|}\left(1+\frac{\eta\rho^t}{\|\boldsymbol{\Delta}_t\|}\sigma_r\right)\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^2.\end{aligned}$$

The proof of the above proposition is presented in Appendix 2.E.2. Recalling our discussion in Section 2.3.3, SubGM with geometric step-size moves towards a direction close to

$\boldsymbol{\Delta}_t / \|\boldsymbol{\Delta}_t\|$ with an "effective" step-size of $\eta\rho^t/\|\boldsymbol{\Delta}_t\|$. In light of this, the above proposition is analogous to Proposition 5, with an additional assumption that the effective step-size is upper bounded by $1/\sigma_1$. Proposition 8 can be used to show the exponential growth of $\lambda_{\min}(\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top)$ in the first phase of the algorithm. To see this, note that, due to small initialization, we have $\|\boldsymbol{\Delta}_t\| \approx \|\boldsymbol{X}^\star\| = \sigma_1$, $\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^2 \ll \lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)$, and $\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| \approx 0$ at the early stages of the algorithm. This implies that the minimum eigenvalue dynamic can be accurately approximated as $\lambda_{\min}\left(\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right) \geq (1+\Omega(\eta/\kappa))\,\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)$, which grows exponentially fast. We next characterize the dynamics of the signal and cross terms.

Proposition 9 (Signal and cross dynamics)**.** *Consider the iterations of SubGM on $f_{\ell_1}(\boldsymbol{U})$ with the step-size defined as equation 2.25. Suppose that the measurements satisfy $(4r,\epsilon,\delta,\mathcal{S})$-Sign-RIP with $\delta \lesssim 1/\sqrt{r}$, $\epsilon = \sqrt{d}\,\|\boldsymbol{G}_t\|^2$, and $\mathcal{S} = \{\boldsymbol{X} : \zeta \leq \|\boldsymbol{X}\|_{\mathrm{F}} \leq R\}$ for $\zeta = \epsilon\left(1/\delta \vee \sqrt{d}\right)$ and $R = 5\sqrt{r}\sigma_1$. Moreover, suppose that $\boldsymbol{S}_t\boldsymbol{S}_t^\top \succ \boldsymbol{0}$, $\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| \leq \sigma_1$, $\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| \leq 1.01\sigma_1$, $\left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\right\| \leq 1/3$, $\frac{\eta\rho^t}{\|\boldsymbol{\Delta}_t\|} \lesssim \frac{1}{\sigma_1}$, and $\boldsymbol{\Delta}_t \in \mathcal{S}$ is ϵ-approximate rank-$4r$. Then, we have*

$$\left\|\boldsymbol{\Sigma}-\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right\| \leq \left(1-\frac{\eta\rho^t}{\|\boldsymbol{\Delta}_t\|}\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right)\left\|\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| + 5\frac{\eta\rho^t}{\|\boldsymbol{\Delta}_t\|}\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\|^2 + 193\sqrt{r}\eta\rho^t\delta\sigma_1, \tag{2.26}$$

$$\left\|\boldsymbol{S}_{t+1}\boldsymbol{E}_{t+1}^\top\right\| \leq \left(1-\frac{\eta\rho^t}{\|\boldsymbol{\Delta}_t\|}\left(\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)-2\left\|\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|-2\|\boldsymbol{E}_t\boldsymbol{E}_t\|\right)\right)\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\| + 113\sqrt{r}\eta\rho^t\delta\sigma_1, \tag{2.27}$$

$$\left\|\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right\| \leq 1.01\sigma_1. \tag{2.28}$$

The proof of Proposition 9 is analogous to Proposition 6, and can be found in Appendix 2.E.3. Assuming $\|\boldsymbol{\Delta}_t\| \asymp \rho^t$, the above proposition shows that both signal and cross terms behave similarly to their expected counterparts in Proposition 4, and their deviation diminishes exponentially fast.

Proposition 10. *Consider the iterations of SubGM on $f_{\ell_1}(\boldsymbol{U})$ with the step-size defined as equation 2.25. Suppose that the measurements satisfy $(4r,\epsilon,\delta,\mathcal{S})$-Sign-RIP with $\delta \lesssim 1/\sqrt{r}$, $\epsilon = \sqrt{d}\,\|\boldsymbol{G}_t\|^2$, and $\mathcal{S} = \{\boldsymbol{X} : \zeta \leq \|\boldsymbol{X}\|_{\mathrm{F}} \leq R\}$ for $\zeta = \epsilon\left(1/\delta \vee \sqrt{d}\right)$ and $R = 5\sqrt{r}\sigma_1$. Moreover, suppose that $\boldsymbol{S}_t\boldsymbol{S}_t^\top \succ \boldsymbol{0}$, $\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| \leq \sigma_1$, $\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| \leq 1.1\sigma_1$, $\left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\right\| \leq 1/3$, and $\boldsymbol{\Delta}_t \in \mathcal{S}$ is ϵ-approximate rank-$4r$. Then, the following statements hold:*

- *If $\eta \lesssim \frac{1}{\|\boldsymbol{\Delta}_t\|}$, we have*

$$\|\boldsymbol{G}_{t+1}\| \leq \left(1+\frac{\eta^2\rho^{2t}}{\|\boldsymbol{\Delta}_t\|^2}\left(2\left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top\right\|^2+\|\boldsymbol{E}_t\|^4+2\|\boldsymbol{\Delta}_t\|\left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top\right\|\right)\right)\|\boldsymbol{G}_t\| + 49\sqrt{r}\eta_0\rho^t\delta\|\boldsymbol{G}_t\|,$$

which can be further simplified as

$$\|\boldsymbol{G}_{t+1}\| \leq \left(1 + 5\eta^2\rho^{2t} + 49\sqrt{r}\eta_0\rho^t\delta\right)\|\boldsymbol{G}_t\|\,. \tag{2.29}$$

- *If* $\frac{\eta\rho^t}{\|\boldsymbol{\Delta}_t\|} \lesssim \frac{1}{\sigma_1}$, *we have*

$$\|\boldsymbol{F}_{t+1}\| \leq \left(1 - \frac{\eta\rho^t}{\|\boldsymbol{\Delta}_t\|}\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right) + 16\sqrt{r}\eta\rho^t\delta\right)\|\boldsymbol{F}_t\| + 16\sqrt{r}\eta\rho^t\delta\,\|\boldsymbol{S}_t\| + 6\eta\rho^t\,\|\boldsymbol{G}_t\|\,.$$

The proof of Proposition 10 follows that of Proposition 7, and can be found in Appendix 2.E.4. Inequality equation 2.29 implies that the growth rate of $\|\boldsymbol{G}_t\|$ diminishes with t. We will use this property to show that $\|\boldsymbol{G}_t\|$ remains proportional to the initialization scale α throughout the solution trajectory, which will be used to control the final generalization error. Moreover, unlike the dynamic of $\|\boldsymbol{F}_t\|$, equation 2.29 holds even when $\|\boldsymbol{\Delta}_t\|$ decays faster than $\eta\sigma_1\rho^t$; this will play a key role in the proof of our next theorem.

Theorem 8 (Global Convergence of SubGM with Noisy Measurements)**.** *Consider the iterations of SubGM on* $f_{\ell_1}(\boldsymbol{U})$ *with the step-size defined as equation 2.25, and parameters* $\eta \lesssim 1/\left(\kappa\log(1/\alpha)\right)$ *and* $\rho = 1 - \Theta\left(\eta/(\kappa\log(1/\alpha))\right)$. *Suppose that the initial point* $\boldsymbol{U}_0$ *is obtained from Algorithm 2 with an initialization scale that satisfies* $\alpha \lesssim 1/(\sqrt{d}) \wedge 1/\kappa \wedge \underline{\varphi}$. *Suppose that the measurements satisfy* $(4r, \delta, \epsilon, \mathcal{S})$*-Sign-RIP with* $\delta \lesssim 1/\left(\sqrt{r}\kappa^2\bar{\varphi}^4\log^2(1/\alpha)\right)$, $\epsilon \asymp \sqrt{d}\alpha^{2-\mathcal{O}\left(\sqrt{r}\kappa\delta\right)}\delta$, *and* $\mathcal{S} = \{\boldsymbol{X} : \zeta \leq \|\boldsymbol{X}\|_{\mathrm{F}} \leq R\}$ *for* $\zeta \geq \epsilon\left(1/\delta \vee \sqrt{d}\right)$ *and* $R = 5\sqrt{r}\sigma_1$. *Then, for any* $t \geq T_{end} = \mathcal{O}((\kappa/\eta)\log^2(1/\alpha))$, *we have*

$$\left\|\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{X}^\star\right\|_{\mathrm{F}} \lesssim d\alpha^{2-\mathcal{O}\left(\sqrt{r}\kappa\delta\right)} \vee \zeta.$$

The proof of the above theorem can be found in Section 2.A.3. Upon defining $\beta(\delta) = 2 - \mathcal{O}(\sqrt{r}\kappa\delta)$, the above result implies that, for any arbitrary accuracy $\epsilon \geq \zeta$, SubGM converges to a solution that satisfies $\left\|\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{X}^\star\right\|_{\mathrm{F}} \leq \epsilon$ within $\mathcal{O}(\kappa\log^2(d/\epsilon)/(\beta(\delta)\eta))$ iterations, provided that $\alpha \lesssim (\epsilon/d)^{1/\beta(\delta)}$. Compared to the noiseless setting, the final error in Theorem equation 8 has an additional term ζ. This is due to the fact that we only require a lower bound on the choice of ζ; as will be explained later, this additional freedom will be used to show the convergence of SubGM under the Gaussian noise model. Moreover, compared to the noiseless setting, the iteration complexity of SubGM in the noisy regime is higher by a factor of $\log(d/\epsilon)$, and its step-size must be chosen more conservatively. The higher iteration complexity is due to the lack of a prior estimate of $\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}$; to alleviate this issue, we proposed a geometric step-size, which inevitably leads to a slightly higher iteration complexity.

In the proof of the above theorem, we show that SubGM already achieves a small error after T_{end} iterations, and any additional iterations may in fact worsen the final error. In fact, this is the main reason behind the early stopping of gradient descent with constant step-size, as shown by Li et al. [2018], Stöger and Soltanolkotabi [2021]. However, we no longer require an early stopping for SubGM. This is due to the fact that, because of their exponentially-decaying nature, the step-sizes become so small that SubGM stops making any meaningful progress after T_{end} iterations. More precisely, we control the dynamics of SubGM in the last $t - T_{end}$ iterations, showing that it does not deviate too much from the solution at iteration T_{end}.

Finally, according to Theorem 8, ρ must satisfy $1-\rho = \Theta\left(\eta/(\kappa \log(1/\alpha))\right)$, which requires a prior knowledge of the condition number κ. However, our analysis in fact guarantees the convergence of SubGM for any $1 - \Theta\left(\eta/(\kappa \log(1/\alpha))\right) \leq \rho < 1$. Therefore, if κ is not known (or only a conservative upper bound of κ is known), ρ can be conservatively close to one at the expense of a slower convergence rate.

Equipped with the above theorem and Theorems 4 and 5, we next characterize the behavior of SubGM under both outlier and Gaussian noise regimes.

Corollary 2 (Outlier Noise Model)**.** *Suppose that the measurement matrices $\{\boldsymbol{A}_i\}_{i=1}^m$ have i.i.d. standard Gaussian entries, and the noise vector $\boldsymbol{s}$ follows an outlier noise model with a corruption probability $0 \leq p < 1$ (Definition 4). Consider the iterations of SubGM on $f_{\ell_1}(\boldsymbol{U})$ with the step-size defined as equation 2.10, and parameters $\eta \lesssim 1/\left(\kappa \log(1/\alpha)\right)$ and $\rho = 1 - \Theta\left(\eta/(\kappa \log(1/\alpha))\right)$. Suppose that the initial point $\boldsymbol{U}_0$ is obtained from Algorithm 2 with an initialization scale that satisfies $\alpha \lesssim 1/\sqrt{d} \wedge 1/\kappa \wedge (1-p)$. Suppose that the number of measurements satisfies $m \gtrsim \kappa^4 dr^2 \log^5(1/\alpha^2) \log^2(m)/(1-p)^2$. Then, for any $t \geq T_{end} = \mathcal{O}((\kappa/\eta)\log^2(1/\alpha))$ and with an overwhelming probability, we have*

$$\left\|\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{X}^\star\right\|_{\mathrm{F}} \lesssim d\alpha^{2-\mathcal{O}\left(\sqrt{\frac{\kappa^2 dr^2 \log(1/\alpha)\log^2(m)}{(1-p)^2 m}}\right)}.$$

Corollary 3 (Gaussian Noise Model)**.** *Suppose that the measurement matrices $\{\boldsymbol{A}_i\}_{i=1}^m$ have i.i.d. standard Gaussian entries, and the noise vector $\boldsymbol{s}$ follows a Gaussian noise model with a variance $\nu_g < \infty$ (Definition 5). Consider the iterations of SubGM on $f_{\ell_1}(\boldsymbol{U})$ with the step-size defined as equation 2.10, and parameters $\eta \lesssim 1/\left(\kappa \log(1/\alpha)\right)$ and $\rho = 1 - \Theta\left(\eta/(\kappa \log(1/\alpha))\right)$. Suppose that the initial point $\boldsymbol{U}_0$ is obtained from Algorithm 2 with an initialization scale that satisfies $\alpha \lesssim 1/\sqrt{d} \wedge \sqrt{dr/m} \wedge 1/\kappa$. Then, for any $t \geq T_{end} =$*

$\mathcal{O}((\kappa/\eta)\log^2(1/\alpha))$ *and with an overwhelming probability, we have*

$$\left\|\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{X}^\star\right\|_{\mathrm{F}} = \mathcal{O}\left(\sqrt{\frac{\nu_g^2\kappa^4 dr^2\log^5(1/\alpha)\log^2(m)}{m}}\right).$$

The proof of Corollary 3 follows directly from Theorems 5 and 8 after choosing

$$\zeta = C\sqrt{\nu_g^2\kappa^4 dr^2\log^5(1/\alpha)\log^2(m)/m},$$

for sufficiently large constant C. The details are omitted for brevity.

Remark 1. *Our result can be readily extended to settings where the measurements are corrupted with both outlier and Gaussian noise values. Consider measurements of the form* $y_i = \langle \boldsymbol{A}_i, \boldsymbol{X}^\star\rangle + s_i^{(1)} + s_i^{(2)}$*, where* $s_i^{(1)}$ *and* $s_i^{(2)}$ *follow the outlier and Gaussian noise models delineated in Definitions 4 and 5. In this setting, Corollaries 2 and 3 can be combined to show that, with* $m = \tilde{\Omega}\left(\nu_g^2\kappa^4 dr^2/(1-p)^2\right)$ *samples, SubGM with small initialization and geometric step-size achieves the error* $\left\|\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{X}^\star\right\|_{\mathrm{F}}^2 = \tilde{\mathcal{O}}\left(\nu_g^2\kappa^4 dr^2/((1-p)^2 m)\right)$ *(modulo logarithmic factors).*

Remark 2. *We highlight that our required measurement number for the convergence of SubGM must scale poly-logarithmically with the initialization scale. According to Theorems 7 and 8, the final error of SubGM depends on the initialization scale, which in turn leads to the dependency of the final error on the measurement number. This dependency is inevitable in the Gaussian noise regime (Corollary 3) due to the information-theoretical lower bounds. Moreover, for the noiseless (Corollary 1) and outlier noise settings (Corollary 2), the dependency of the final error on the measurement number is of the form* $\exp(-\Omega(\mathrm{poly}(m)))$*. As such, to obtain an* ϵ*-accurate solution, the measurement number must scale as* $\mathrm{poly}\log(1/\epsilon)$*. While we note that this dependency is mild, as a future research direction, we will investigate the possibility of removing such a dependency.*

2.7 Ground Truths are Strict Saddles in Noisy Setting

In this section, we show that under some mild conditions, the true solutions $\mathcal{U}^* := \{\boldsymbol{U}^* \in \mathbb{R}^{d\times r'} : \boldsymbol{U}^*\boldsymbol{U}^{*\top} = \boldsymbol{X}^*\}$ are strict saddle points of $f_{\ell_1}(\boldsymbol{U})$. To this goal, we first consider the following noise model.

Assumption 1 (Noise Model)**.** *Each measurement is independently corrupted with noise with a corruption probability* $0 \le p \le 1$*. Let the set of noisy measurements be denoted as*

$\mathcal{S}$. *For each entry* $i \in \mathcal{S}$, *the value of* $\boldsymbol{\epsilon}_i$ *is drawn from a distribution* P_o. *Moreover, a random variable* ζ *under the distribution* P_o *satisfies* $\mathbb{P}(|\zeta| \geq t_0) \geq p_0$ *for some constants* $0 < t_0, p_0 \leq 1$.

Our considered noise model is generic and includes all the "typical" noise models, including Gaussian and outlier noise models. Intuitively, it only requires a nonzero mass at a nonzero value. For example, suppose that for every $i \in \mathcal{S}$, $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \sigma^2)$. A basic anti-concentration inequality implies that $\mathbb{P}(|\boldsymbol{\epsilon}_i| \geq t_0) = 1 - \mathbb{P}(|\boldsymbol{\epsilon}_i|/\sigma < t_0/\sigma) \geq 1 - 2t_0/\sigma$, which satisfies Assumption 2 with $(t_0, p_0) = ((4\sigma)^{-1}, 1/2)$. Similarly, a randomized sparse outlier noise is expected to satisfy Assumption 2.

Our next two theorems establish the necessary and sufficient conditions for determining the local optimality, or lack thereof, of true solutions in symmetric matrix sensing.

Theorem 9 (Sub-optimality of true solutions for symmetric matrix sensing)**.** *Suppose that the noise satisfies Assumption 2. Suppose that* $r' > r$ *and* $m \lesssim pp_0 d(r' - r)$. *With probability at least* $1 - \exp(-\Omega(d(r' - r))) - \exp(-\Omega(pp_0 m))$, *none of the true solutions in* $\mathcal{U}^*$ *are local optima of* $f_{\ell_1}(\boldsymbol{U})$. *More precisely, for every* $\boldsymbol{U}^\star \in \mathcal{U}^*$, *there exists a descent direction* $\Delta\boldsymbol{U}$ *satisfying* $\|\Delta\boldsymbol{U}\|_{\mathrm{F}} = 1$ *such that, for every* $0 < \gamma^2 \lesssim \frac{t_0}{\sqrt{d}}$, *we have*

$$f_{\ell_1}(\boldsymbol{U}^\star + \gamma\Delta\boldsymbol{U}) - f_{\ell_1}(\boldsymbol{U}^\star) \lesssim -\gamma^2. \tag{2.30}$$

According to Theorem 9, none of the true solutions emerge as local optima of f_{ℓ_1}, unless the number of measurements scale with the search rank r'. In particular, suppose that $2r \leq k$ and the parameters of the noise model are fixed. Then, Theorem 9 shows that, with high probability, none of the true solutions are local optima, so long as $m \lesssim dk$.

Theorem 10 (Optimality of true solutions for symmetric matrix sensing)**.** *Suppose that the noise satisfies Assumption 2 with a corruption probability* $0 \leq p < 1/2$. *Moreover, suppose that* $r' \geq r$. *The following statements hold:*

- *Suppose that* $m \gtrsim \frac{dr}{(1-2p)^4}$. *With probability at least* $1 - \exp(-\Omega(dr))$, *for every* $\boldsymbol{U}^\star \in \mathcal{U}^*$ *and* $\gamma \geq 0$, *we have*

$$\min_{\|\Delta\boldsymbol{U}\|_{\mathrm{F}}=1} \{f_{\ell_1}(\boldsymbol{U}^\star + \gamma\Delta\boldsymbol{U}) - f_{\ell_1}(\boldsymbol{U}^\star)\} \gtrsim -\left(\sqrt{\frac{2}{\pi}} + \sqrt{\frac{dr'}{m}}\right)\gamma^2.$$

- *Suppose that* $m \gtrsim \frac{dr'}{(1-2p)^4}$. *With probability at least* $1 - \exp(-\Omega(dr'))$, *all true solutions*

coincide with the global optima. More precisely, we have

$$\boldsymbol{U}^\star \in \mathcal{U}^* \quad \iff \quad \boldsymbol{U}^\star \in \arg\min_{\boldsymbol{U}} f_{\ell_1}(\boldsymbol{U}).$$

We note that the second statement of the above theorem has been previously proven by Ding et al. [2021]. However, we include it here for completeness.

Combining Theorem 9 with the first statement of Theorem 10 leads to the following key corollary.

Corollary 4. *Suppose that the noise satisfies Assumption 2 with* $p_0 = \Theta(1)$ *and* $p = \Theta(1) < 1/2$. *Additionally, suppose that* $k \gtrsim r$ *and* $dr \lesssim m \lesssim dr'$. *Then, with probability at least* $1 - \exp(-\Omega(dr')) - \exp(-\Omega(m))$, *all true solutions are critical and strict saddle points.*

The above results show that the landscape of symmetric matrix sensing with ℓ_1-loss and outlier noise (with $p < 1/2$) undergoes two critical transitions: When $m \lesssim dr$, exact recovery is impossible due to the existing information theoretical barriers [Candes and Plan, 2011]. When $dr \lesssim m \lesssim dr'$, all true solutions emerge as strict saddle points. We note that this is precisely the regime where the sub-gradient method empirically converges to one of the true solutions [Ma and Fattahi, 2023a]. As soon as $m \gtrsim dr'$, all true solutions coincide with the global optima.

2.8 Concluding Remarks

In this work, we study the performance of sub-gradient method (SubGM) on a nonconvex and nonsmooth formulation of the robust matrix sensing with noisy measurements, where the rank of the true solution r is unknown and over-estimated instead with $r' \geq r$. We prove that the over-estimation of the rank has no effect on the performance of SubGM, provided that the initial point is sufficiently close to the origin. Moreover, we prove that SubGM is robust against outlier and Gaussian noise values. In particular, we show that SubGM provably converges to the ground truth, even if the globally optimal solutions of the problem are "spurious", i.e., they do not correspond to the ground truth. At the heart of our method lies a new notion of restricted isometry property, called Sign-RIP, which guarantees a direction-preserving property for the sub-differentials of the ℓ_1-loss. We show that, while the classical notions of restricted isometry property face major breakdowns in the face of noise, Sign-RIP can handle a wide range of noisy measurements, and hence, is better-suited for analyzing the robust variants of low-rank matrix sensing. A few remarks are in order next:

Spectral vs. random initialization: In our work, we assume that the initial point is obtained via a special form of the spectral method, followed by a norm reduction. A natural question thus arises as to whether the spectral method can be replaced by small random initialization. Here we briefly mention the potential technical difficulty in extending our theoretical result to the random initialization. According to the proofs of Theorems 7 and 8, we require the minimum eigenvalue of the signal term to dominate the maximum eigenvalues of the cross and residual terms at the initial point. If the initial point is chosen randomly, it is likely that such an assumption is violated. As a consequence, it is unclear how to ensure the initial exponential growth of the signal term under our current error perturbation analysis. Despite this theoretical hurdle, we observed in our simulations that SubGM with small random initialization behaves almost the same as SubGM with spectral initialization. Therefore, we conjecture that a small random initialization followed by a few iterations of SubGM is in fact equivalent to spectral initialization; a similar result has been recently proven by Stöger and Soltanolkotabi [2021] for gradient descent on ℓ_2-loss. We consider a rigorous verification of this conjecture as an enticing challenge for future research.

Beyond Sign-RIP: Another natural question pertains to the performance of SubGM on problems that *do not* satisfy Sign-RIP. An important and relevant example is *over-parameterized matrix completion*, where the linear measurement operator is an element-wise projector that reveals partial and potentially noisy observations of a low-rank matrix. Indeed, the performance SubGM on problems of this type requires a more refined analysis, which is left as future work.

2.A Proofs of the Main Theorems

2.A.1 Proof of Theorem 6

Before delving into the details, we first present the general overview of our proof technique for Theorem 6. First, we prove that the conditions of Propositions 3 and 4 hold for every $0 \leq t < \infty$. Then, we use the minimum eigenvalue dynamic in Proposition 4 to show that $\lambda_{\min}(\boldsymbol{S}_t\boldsymbol{S}_t^\top) \geq 0.98\sigma_r$ after $\mathcal{O}\left(\log(1/\alpha^2)/(\eta\sigma_r)\right)$ iterations. In the second phase, we leverage the lower bound $\lambda_{\min}(\boldsymbol{S}_t\boldsymbol{S}_t^\top) \geq 0.98\sigma_r$ to further simplify the one-step dynamics in Proposition 3, and show that both signal and cross terms decay linearly, while the residual term remains in the order of α^2. This phase lasts for $\mathcal{O}\left(\log(\sigma_1/\alpha^2)/(\eta\sigma_r)\right)$ iterations, and the generalization error can be upper bounded by α^2 at the end of this phase. Finally, in the third phase, we show that the residual term will dominate the signal and cross terms, and the generalization error will decay at a sublinear rate.

Lemma 5. *The conditions of Propositions 3 and 4 are satisfied for every $0 \leq t < \infty$. In particular, for any $0 \leq t < \infty$, we have*

$$\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| \leq \frac{\alpha^2}{\eta\alpha^2 t+1}, \tag{2.31}$$

$$\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| \leq 1.01\sigma_1, \tag{2.32}$$

$$\boldsymbol{S}_t\boldsymbol{S}_t^\top \succ \boldsymbol{0}. \tag{2.33}$$

The proof of the above lemma can be found in Appendix 2.C.3. Given Lemma 5, we proceed to prove Theorem 6.

Phase 1: Eigenvalue Learning. Due to Proposition 3 and Lemma 5, we have

$$\begin{aligned}\lambda_{\min}\left(\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right) &\geq \left(1+2\eta\sigma_r - 2\eta\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|\right)\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right) - 2.01\eta\lambda_{\min}^2\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\\ &\geq \left(1+1.99\eta\sigma_r\right)\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right) - 2.01\eta\lambda_{\min}^2\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right),\end{aligned} \tag{2.34}$$

where we used the assumption $2\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| \leq 2\alpha^2 \leq 0.01\sigma_r$ due to our choice of α. Now, we consider two cases:

- Suppose that T_1 is the largest iteration such that $\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right) \leq \sigma_r/2.01$ for every $t \leq T_1$. According to equation 2.34, we have

  $$\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right) \geq (1+0.99\eta\sigma_r)^t\lambda_{\min}\left(\boldsymbol{S}_0\boldsymbol{S}_0^\top\right) \geq (1+0.99\eta\sigma_r)^t\alpha^2\sigma_r.$$

  This implies that, after $\mathcal{O}\left(\log(1/\alpha^2)/(\eta\sigma_r)\right)$ iterations, we have $\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right) > \sigma_r/2.01$, and hence, $T_1 = \mathcal{O}\left(\log(1/\alpha^2)/(\eta\sigma_r)\right)$.

- For $t > T_1$, let $x_t = \sigma_r - \lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)$. Then, according to equation 2.34, we have

  $$\begin{aligned}x_{t+1} &\leq \left(1-2.03\eta\sigma_r\right)x_t + 2.01\eta x_t^2 + 0.02\eta\sigma_r^2\\ &\leq \left(1-1.02\eta\sigma_r\right)x_t + 0.02\eta\sigma_r^2,\end{aligned} \tag{2.35}$$

  where in the second inequality, we used the fact that $x_t = \sigma_r - \lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right) \leq 1.01\sigma_r/2.01$. The above inequality implies

  $$\begin{aligned}&x_{t+1} - 0.0196\sigma_r \leq \left(1-1.02\eta\sigma_r\right)\left(x_t - 0.0196\sigma_r\right)\\ \implies\ &x_{t+1} - 0.0196\sigma_r \leq \left(1-1.02\eta\sigma_r\right)^{t-T_1+1}\left(x_{T_1} - 0.0196\sigma_r\right).\end{aligned}$$

  Hence, we have $x_t \leq 0.02\sigma_r$ after $T_3 = T_1 + T_2$ iterations, where $T_2 = \mathcal{O}\left(1/\eta\sigma_r\right)$, which in turn shows that $\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right) \geq 0.98\sigma_r$.

The above analysis shows that $\lambda_{\min}(\boldsymbol{S}_t\boldsymbol{S}_t^\top) \geq 0.98\sigma_r$ for every $t \geq T_3 = T_1 + T_2 = \mathcal{O}\left(\log(1/\alpha^2)/(\eta\sigma_r)\right)$.

Phase 2: Global Convergence. We have $0.98\sigma_r \leq \lambda_{\min}(\boldsymbol{S}_t\boldsymbol{S}_t^\top) \leq 1.01\sigma_1$ for every $t \geq T_3$. This, combined with the one-step signal dynamic equation 2.17, implies that

$$\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right\| \leq (1 - 0.98\eta\sigma_r)\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| + 5\eta\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\|^2.$$

On the other hand, due to Lemma 5, we have

$$\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\|^2 \leq \left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| \leq (1.01\sigma_1)\alpha^2.$$

This implies that

$$\begin{aligned}
&\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right\| \leq (1 - 0.98\eta\sigma_r)\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| + 6\eta\sigma_1\alpha^2 \qquad (2.36)\\
\implies &\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right\| - \frac{6\sigma_1\alpha^2}{0.98\sigma_r} \leq (1 - 0.98\eta\sigma_r)\left(\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| - \frac{6\sigma_1\alpha^2}{0.98\sigma_r}\right)\\
\implies &\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right\| - \frac{6\sigma_1\alpha^2}{0.98\sigma_r} \leq (1 - 0.98\eta\sigma_r)^{t-T_3+1}\left(\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_{T_3}\boldsymbol{S}_{T_3}^\top\right\| - \frac{6\sigma_1\alpha^2}{0.98\sigma_r}\right).
\end{aligned}$$

Therefore,

$$\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| \leq 7\kappa\alpha^2 \quad \text{for} \quad t \geq T_5 = T_3 + T_4, \quad \text{where} \quad T_4 = \mathcal{O}\left(\frac{\log\left(\frac{\sigma_1}{\kappa\alpha^2}\right)}{\eta\sigma_r}\right).$$

Here, we use the inequality $\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_{T_3}\boldsymbol{S}_{T_3}^\top\right\| \leq \|\boldsymbol{\Sigma}\| + \left\|\boldsymbol{S}_{T_3}\boldsymbol{S}_{T_3}^\top\right\| \leq 2.01\sigma_1$. On the other hand, the one-step dynamics for the cross term equation 2.18 implies that

$$\begin{aligned}
\left\|\boldsymbol{S}_{t+1}\boldsymbol{E}_{t+1}^\top\right\| &\leq \left(1 - \eta\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right) + 2\eta\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| + 2\eta\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|\right)\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\|\\
&\leq (1 - 0.5\eta\sigma_r)\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\|\\
\implies \left\|\boldsymbol{S}_{t+1}\boldsymbol{E}_{t+1}^\top\right\| &\leq (1 - 0.5\eta\sigma_r)^{t-T_5+1}\left\|\boldsymbol{S}_{T_5}\boldsymbol{E}_{T_5}^\top\right\|\\
&\leq (1 - 0.5\eta\sigma_r)^{t-T_5+1}\left(1.01\alpha\sqrt{\sigma_1}\right), \qquad (2.37)
\end{aligned}$$

where the second inequality follows from the proven upper bound $\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| \leq 7\kappa\alpha^2$ and Lemma 5. Moreover, the last inequality is due to the fact that

$$\left\|\boldsymbol{S}_{T_5}\boldsymbol{E}_{T_5}^\top\right\| \leq \|\boldsymbol{S}_{T_5}\|\,\|\boldsymbol{E}_{T_5}\| \leq 1.01\alpha\sqrt{\sigma_1}.$$

The inequality equation 2.37 results in

$$\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\| \leq \alpha^2 \quad \text{for} \quad t \geq T_7 = T_6 + T_5 \quad \text{where} \quad T_6 = \mathcal{O}\left(\frac{\log\left(\frac{\sqrt{\sigma_1}}{\alpha}\right)}{\eta\sigma_r}\right).$$

This upper bound can in turn be used in equation 2.36 further to strengthen the upper bound on the signal term as follows

$$\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| \leq \alpha^2 \qquad \text{for} \qquad t \geq T_8 = T_7 + T_6 = \mathcal{O}\left(\frac{\log\left(\frac{\sigma_1}{\alpha^2}\right)}{\eta\sigma_r}\right).$$

Finally, invoking the signal-residual decomposition in Lemma 3, we have

$$\left\|\boldsymbol{U}_t{\boldsymbol{U}_t}^\top - \boldsymbol{X}^*\right\| \leq \left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right\| + 2\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\| + \left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| \lesssim \alpha^2$$

for $t \geq T_8 = \mathcal{O}\left(\frac{\log\left(\frac{\sigma_1}{\alpha^2}\right)}{\eta\sigma_r}\right)$.

Phase 3: Sublinear convergence. Once both signal and cross terms are in the order of α^2, the residual term becomes the dominant term, while both signal and cross terms maintain their linear decay rates. Therefore, we have

$$\left\|\boldsymbol{U}_t{\boldsymbol{U}_t}^\top - \boldsymbol{X}^*\right\| \lesssim \left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| \lesssim \frac{\alpha^2}{\eta\alpha^2 t + 1}.$$

This completes the proof. □

2.A.2 Proof of Theorem 7

The proof of Theorem 7 follows the same structure as the proof of Theorem 6: first, we use Proposition 5 to show that $\lambda_{\min}(\boldsymbol{S}_t\boldsymbol{S}_t^\top)$ reaches $0.98\sigma_r$ after $T_1 \lesssim \log(1/\alpha)/(\eta\sigma_r\underline{\varphi}^2)$ iterations. Given this inequality and equipped with the one-step dynamics of the signal, cross, and residual terms (Propositions 6 and 7), we then establish the linear convergence of SubGM to the ground truth. As a first step, we show an important property of the proposed initialization scheme.

Lemma 6. *Suppose that the measurements are noiseless and satisfy $(4r, \delta, \epsilon, \mathcal{S})$-Sign-RIP where $\epsilon = 0$, and $\boldsymbol{X}^\star \in \mathcal{S}$. Then, the initial point $\boldsymbol{U}_0 = \boldsymbol{V}\boldsymbol{S}_0 + \boldsymbol{V}_\perp\boldsymbol{E}_0$ generated from Algorithm 2 satisfies*

$$\left\|\boldsymbol{U}_0\boldsymbol{U}_0^\top - \alpha^2\varphi(\boldsymbol{X}^\star)\frac{\boldsymbol{X}^\star}{\|\boldsymbol{X}^\star\|_{\mathrm{F}}}\right\| \leq 2\alpha^2\varphi(\boldsymbol{X}^\star)\delta,$$

The proof can be found in Appendix 2.D.5. An immediate consequence of the above lemma is the following inequality:

$$\left\|\boldsymbol{S}_0\boldsymbol{S}_0^\top - \alpha^2\varphi(\boldsymbol{X}^\star)\frac{\boldsymbol{\Sigma}}{\|\boldsymbol{X}^\star\|_{\mathrm{F}}}\right\| \vee \left\|\boldsymbol{S}_0\boldsymbol{E}_0^\top\right\| \vee \left\|\boldsymbol{E}_0\boldsymbol{E}_0^\top\right\| \leq 2\alpha^2\varphi(\boldsymbol{X}^\star)\delta.$$

Given this property of the proposed initialization scheme, we next show that the conditions of Propositions 5, 6, and 7 are satisfied throughout solution path.

Lemma 7. *We either have* $\|\boldsymbol{\Delta}_t\|_{\mathrm{F}} \lesssim d\alpha^{2-\mathcal{O}(\sqrt{r}\kappa^2\delta)}$ *for some* $0 \leq t \leq T_{end}$*, or the conditions of Propositions 5, 6, and 7 are satisfied for every* $0 \leq t \leq T_{end}$*. In particular, for any* $0 \leq t \leq T_{end}$*, we have*

$$\|\boldsymbol{F}_t\| \leq \eta\bar{\varphi}^2\left(100\sqrt{r}\sigma_1^{1.5}\delta + 30\sigma_1\sqrt{\alpha\bar{\varphi}\delta}\right)(t+1), \tag{2.38}$$

$$\|\boldsymbol{G}_t\| \leq \alpha^{1-\mathcal{O}\left(\sqrt{r}\kappa^2\delta\right)}\sqrt{\bar{\varphi}\delta}, \tag{2.39}$$

$$\|\boldsymbol{\Delta}_t\| \leq 5\sigma_1, \tag{2.40}$$

$$\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| \leq 1.01\sigma_1, \tag{2.41}$$

$$\boldsymbol{S}_t\boldsymbol{S}_t^\top \succ \boldsymbol{0}, \tag{2.42}$$

$$\left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\right\| \leq 1/3. \tag{2.43}$$

The proof of Lemma 7 is provided in Appendix 2.D.6. Note that the inequality $\|\boldsymbol{\Delta}_t\|_{\mathrm{F}} \lesssim d\alpha^{2-\mathcal{O}(\sqrt{r}\kappa^2\delta)}$ for some $0 \leq t \leq T_{end}$ readily implies the final result. On the the other hand, if $\|\boldsymbol{\Delta}_t\|_{\mathrm{F}} \gtrsim d\alpha^{2-\mathcal{O}(\sqrt{r}\kappa^2\delta)}$, Lemma 7 implies that Propositions 5, 6, and 7 hold for every $0 \leq t \leq T_{end}$.

Phase 1: Eigenvalue Learning. Based on Lemma 7, the conditions of Proposition 5 are satisfied for $T_1 \lesssim \log(1/\alpha)/(\eta\sigma_r\underline{\varphi}^2)$, and we have

$$\begin{aligned}\lambda_{\min}\left(\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right) \geq \left((1+2\bar{\eta}_t\sigma_r) - 2\bar{\eta}_t\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| - 72\bar{\eta}_t\delta\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}\right)\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right) \\ - 2\bar{\eta}_t\left(1+\bar{\eta}_t\sigma_r\right)\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^2.\end{aligned} \tag{2.44}$$

Due to equation 2.38 and equation 2.39, we have $\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| \leq \|\boldsymbol{F}_t\|^2 + \|\boldsymbol{G}_t\|^2 \leq 0.005\sigma_r$ for every $t \lesssim \log(1/\alpha)/(\eta\sigma_r\underline{\varphi}^2)$, where we used Lemma 7 and the assumed upper bounds on δ and α. Moreover, $72\delta\|\boldsymbol{\Delta}_t\|_{\mathrm{F}} \leq 150\sigma_1\delta \leq 0.005\sigma_r$, where again we used the assumed upper bound on δ and Lemma 7. These two inequalities together with equation 2.44 lead to

$$\lambda_{\min}\left(\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right) \geq \left(1+1.99\bar{\eta}_t\sigma_r\right)\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right) - 2.01\bar{\eta}_t\lambda_{\min}^2\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right).$$

The above inequality is identical to equation 2.34, after noticing that $\eta\underline{\varphi}^2 \le \bar{\eta}_t \le \eta\bar{\varphi}^2$. On the other hand, Lemma 6 shows that $\lambda_{\min}\left(\boldsymbol{S}_0\boldsymbol{S}_0^\top\right) \ge \alpha^2\varphi(\boldsymbol{X}^\star)\left(\frac{\sigma_r}{\|\boldsymbol{X}^\star\|_{\mathrm{F}}} - 2\delta\right) \ge \alpha^2\underline{\varphi}\frac{1}{2\sqrt{r}\kappa}$. Therefore, using an argument analogous to the proof of Theorem 6, we have $\lambda_{\min}(\boldsymbol{S}_t\boldsymbol{S}_t^\top) \ge 0.98\sigma_r$ for $t \ge T_1 = \mathcal{O}(\log(1/\alpha)/(\eta\sigma_r\underline{\varphi}^2))$. The details are omitted for brevity.

Phase 2: Global convergence. Recall the signal-residual decomposition

$$\left\|\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{X}^\star\right\| \le \left\|\boldsymbol{\Sigma} - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| + 2\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\| + \left\|\boldsymbol{F}_t\boldsymbol{F}_t^\top\right\| + \left\|\boldsymbol{G}_t\boldsymbol{G}_t^\top\right\|.$$

In what follows, we show that once $\lambda_{\min}(\boldsymbol{S}_t\boldsymbol{S}_t^\top) \ge 0.98\sigma_r$, all terms in the above inequality decay at a linear rate, except for $\left\|\boldsymbol{G}_t\boldsymbol{G}_t^\top\right\|$. To this goal, first note that Propositions 6 and 7 together with $0.98\sigma_r \le \lambda_{\min}(\boldsymbol{S}_t\boldsymbol{S}_t^\top) \le 1.01\sigma_1$ lead to the following one-step dynamics:

$$\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right\| \le (1 - 0.98\bar{\eta}_t\sigma_r)\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| + 5\bar{\eta}_t\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\|^2 + 37\bar{\eta}_t\delta\bar{\varphi}^2\sigma_1\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}, \quad (2.45)$$

$$\left\|\boldsymbol{S}_{t+1}\boldsymbol{E}_{t+1}^\top\right\| \le \left(1 - 0.98\bar{\eta}_t\sigma_r + 2\bar{\eta}_t\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| + 2\bar{\eta}_t\left\|\boldsymbol{E}_t\boldsymbol{E}_t\right\|\right)\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\| + 22\bar{\eta}_t\delta\sigma_1\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}, \quad (2.46)$$

$$\left\|\boldsymbol{F}_{t+1}\right\| \le (1 - 0.98\bar{\eta}_t\sigma_r)\left\|\boldsymbol{F}_t\right\| + 5\sqrt{\sigma_1}\bar{\eta}_t\delta\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}} + 6\bar{\eta}_t\left\|\boldsymbol{\Delta}_t\right\|\left\|\boldsymbol{G}_t\right\|. \quad (2.47)$$

Note that, unlike $\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|$ and $\|\boldsymbol{F}_t\|$, the cross term $\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\|$ enjoys linear decay only under the condition $\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| < 0.98\sigma_r$, which is not necessarily satisfied in the eigenvalue learning phase. Our next lemma shows that this condition is satisfied shortly after the eigenvalue learning phase.

Lemma 8. *We have $\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| < 0.03\sigma_r$, for every $T_{end} \ge t \ge T_1 + T_2$, where $T_2 = \mathcal{O}\left(\log(\kappa)/(\eta\sigma_r\underline{\varphi}^2\right)$.*

Proof. It is easy to see that

$$5\bar{\eta}_t\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\|^2 \le 5.05\sigma_1\bar{\eta}_t\left\|\boldsymbol{E}_t\right\|^2 \le 5.05\bar{\eta}_t\sigma_1\left(\|\boldsymbol{F}_t\| + \|\boldsymbol{G}_t\|\right)^2 \le 0.01\bar{\eta}_t\sigma_r^2, \quad (2.48)$$

where the last inequality is due to our choice of δ and Lemma 7. Similarly, we can show that $37\bar{\eta}_t\delta\sigma_1\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}} \le 0.01\bar{\eta}_t\sigma_r^2$. These two inequalities combined with equation 2.45 lead to

$$\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right\| \le (1 - 0.98\bar{\eta}_t\sigma_r)\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| + 0.02\bar{\eta}_t\sigma_r^2.$$

This implies that $\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| \le 0.03\sigma_r$ after $\mathcal{O}\left(\log(\kappa)/(\eta\sigma_r\underline{\varphi}^2)\right)$ iterations. □

The above lemma shows that the one-step dynamic of the cross term can be simplified as

$$\left\|\boldsymbol{S}_{t+1}\boldsymbol{E}_{t+1}^\top\right\| \le (1 - 0.49\bar{\eta}_t\sigma_r)\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\| + 22\bar{\eta}_t\delta^2\sigma_1\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}. \quad (2.49)$$

Moreover, recall that

$$\begin{aligned}\|\boldsymbol{G}_{t+1}\| &\leq \left(1+\bar{\eta}_t^2\left(2\left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top\right\|^2+\|\boldsymbol{E}_t\|^4+2\|\boldsymbol{\Delta}_t\|\left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top\right\|\right)+7\bar{\eta}_t\delta\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}\right)\|\boldsymbol{G}_t\| \\ &\leq \left(1+5\bar{\eta}_t^2\|\boldsymbol{\Delta}_t\|^2+7\bar{\eta}_t\delta\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}\right)\|\boldsymbol{G}_t\|. \end{aligned} \tag{2.50}$$

Here we use the fact that $\left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top\right\| \vee \|\boldsymbol{E}_t\|^2 \leq \|\boldsymbol{\Delta}_t\|$. Now, let us define $\gamma_t := \left\|\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| + 2\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\| + \|\boldsymbol{F}_t\|^2 + \|\boldsymbol{G}_t\|^2$ as an upper bound for the generalization error $\left\|\boldsymbol{U}_t{\boldsymbol{U}_t}^\top - \boldsymbol{X}^\star\right\|$. Combining equation 2.45, equation 2.47, equation 2.49, and equation 2.50, we have

$$\begin{aligned}\gamma_{t+1} &= \left\|\boldsymbol{\Sigma}-\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right\| + 2\left\|\boldsymbol{S}_{t+1}\boldsymbol{E}_{t+1}^\top\right\| + \|\boldsymbol{F}_{t+1}\|^2 + \|\boldsymbol{G}_{t+1}\|^2 \\ &\leq (1-C_1\bar{\eta}_t\sigma_r)\,\gamma_t + C_2\bar{\eta}_t\delta\sigma_1\|\boldsymbol{\Delta}_t\|_{\mathrm{F}} + C_3\bar{\eta}_t\sigma_r\|\boldsymbol{G}_t\|^2 \\ &\leq \left(1-C_1\bar{\eta}_t\sigma_r + C_2\bar{\eta}_t\sqrt{r}\sigma_1\delta\right)\gamma_t + C_4\bar{\eta}_t\sigma_r\sqrt{d}\,\|\boldsymbol{G}_t\|^2 \\ &\leq (1-C_5\bar{\eta}_t\sigma_r)\,\gamma_t + C_4\bar{\eta}_t\sigma_r\sqrt{d}\,\|\boldsymbol{G}_t\|^2,\end{aligned}$$

for some universal constants $C_1, C_2, C_3, C_4, C_5 > 0$. This implies that

$$\begin{aligned}&\gamma_{t+1} - C_6\sqrt{d}\,\|\boldsymbol{G}_t\|^2 \leq (1-C_5\bar{\eta}_t\sigma_r)\left(\gamma_t - C_6\sqrt{d}\,\|\boldsymbol{G}_t\|^2\right) \\ \Longrightarrow\ &\gamma_{t+1} - C_6\sqrt{d}\,\|\boldsymbol{G}_t\|^2 \leq (1-C_5\bar{\eta}_t\sigma_r)^{t-(T_1+T_2)}\left(\gamma_{T_1+T_2} - C_6\sqrt{d}\,\|\boldsymbol{G}_t\|^2\right).\end{aligned}$$

Note that $\gamma_{T_1+T_2} \leq 0.1\sigma_r$, according to Lemma 8, inequality equation 2.48, and the upper bounds on $\|\boldsymbol{F}_t\|^2$ and $\|\boldsymbol{G}_t\|^2$. On the other hand, $\|\boldsymbol{G}_t\| \leq \alpha^{1-\mathcal{O}\left(\sqrt{r}\kappa^2\delta\right)}\sqrt{\bar{\varphi}\delta}$, according to Lemma 7. Therefore, after $T_1+T_2+T_3$ iterations with $T_3 = \mathcal{O}\left(\log(1/\alpha)/(\eta\sigma_r\underline{\varphi}^2)\right)$, we have $\|\boldsymbol{\Delta}_t\|_{\mathrm{F}} \leq \sqrt{d}\gamma_t \lesssim d\alpha^{2-\mathcal{O}\left(\sqrt{r}\kappa^2\delta\right)}\bar{\varphi}\delta$. This completes the proof. □

2.A.3 Proof of Theorem 8

Without loss of generality, we assume that $\|\boldsymbol{\Delta}_t\|_{\mathrm{F}} \geq \zeta$ for every $0 \leq t \leq T_{end}$; otherwise, the final bound for the generalization error holds, and the proof is complete. Before delving into the details, we first provide a general overview of our approach. The proof of Theorem 8 is similar to that of Theorem 7 with a key difference that we divide our analysis into two parts depending on the value of $\|\boldsymbol{\Delta}_t\|$: if $\eta\rho^t/\|\boldsymbol{\Delta}_t\| \lesssim 1/\sigma_1$, then $\|\boldsymbol{\Delta}_t\|$ decays slower than $\sigma_1\eta\rho^t$. Under this assumption, we will use the one-step dynamics of the signal, cross, and residual terms in Propositions 8, 9, and 10 to prove that $\|\boldsymbol{\Delta}_t\|$ decays exponentially fast. Alternatively, $\eta\rho^t/\|\boldsymbol{\Delta}_t\| \gtrsim 1/\sigma_1$ implies that $\|\boldsymbol{\Delta}_t\| \lesssim \sigma_1\eta\rho^t$, which readily establishes the exponential decay of the generalization error. Indeed, the above two cases may occur alternatively, which requires a more delicate analysis.

Similar to the proof of Theorem 7, we show that SubGM undergoes two phases: (1) eigenvalue learning phase, and (2) global convergence phase. In the first phase, we show that $\lambda_{\min}(\boldsymbol{S}_t\boldsymbol{S}_t^\top)$ and $\|\boldsymbol{\Delta}_t\|$ converge to $0.98\sigma_r$ and $0.02\sigma_r$, respectively. The main difference between the proofs of Theorems 7 and 8 is the fact that the one-step dynamics of the signal, cross, and residual terms may not hold during the entire solution trajectory. However, our next lemma shows that they indeed hold in the first phase of our analysis.

Lemma 9. *Suppose that $\|\boldsymbol{\Delta}_t\| \geq 0.02\sigma_r$ for every $0 \leq t \leq \bar{T} \leq T_{end}$. Then, the assumptions of Propositions 8, 9, and 10 are satisfied for every $0 \leq t \leq \bar{T}$. In particular, for any $0 \leq t \leq \bar{T}$, we have*

$$\|\boldsymbol{F}_t\| \leq \left(15\sqrt{r}\eta\rho^t\sqrt{\sigma_1}\delta + 12\sqrt{2}\eta\sqrt{\bar{\varphi}\alpha\delta}\right)(t+1), \tag{2.51}$$

$$\|\boldsymbol{G}_t\| \leq 2\sqrt{2}\alpha^{1-\mathcal{O}\left(\sqrt{r}\kappa\delta\right)}\sqrt{\bar{\varphi}\delta}, \tag{2.52}$$

$$\|\boldsymbol{\Delta}_t\| \leq 5\sigma_1, \tag{2.53}$$

$$\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| \leq 1.1\sigma_1, \tag{2.54}$$

$$\boldsymbol{S}_t\boldsymbol{S}_t^\top \succ \boldsymbol{0}, \tag{2.55}$$

$$\left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\right\| \leq 1/3. \tag{2.56}$$

Moreover, for any $t \geq \bar{T}$, we also have

$$\|\boldsymbol{G}_t\| \leq 2\sqrt{2}\alpha^{1-\mathcal{O}\left(\sqrt{r}\kappa\delta\right)}\sqrt{\bar{\varphi}\delta}. \tag{2.57}$$

The proof of Lemma 9 is analogous to that of Lemma 7, and can be found in Appendix 2.E.5. Equipped with this lemma, we provide proof for the eigenvalue learning phase.

Phase 1: Eigenvalue Learning. We will show that $\lambda_{\min}(\boldsymbol{S}_t\boldsymbol{S}_t^\top) \geq 0.98\sigma_r$ within $t \leq T_1 = \mathcal{O}((\kappa/\eta)\log(1/\alpha))$ iterations. After that, we prove that we only need $T_2 = \mathcal{O}((\kappa/\eta)\log(\kappa))$ additional iterations to ensure that $\|\boldsymbol{\Delta}_t\| \leq 0.02\sigma_r$; this marks the end of Phase 1. Without loss of generality, we may assume that $\|\boldsymbol{\Delta}_t\| \geq 0.02\sigma_r$ for every $t \leq T_1$. To see this, suppose that $\|\boldsymbol{\Delta}_t\| \leq 0.02\sigma_r$ for some $t \leq T_1$. This implies that $\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| \leq \|\boldsymbol{\Delta}_t\| \leq 0.02\sigma_r$, which in turn leads to $\lambda_{\min}(\boldsymbol{S}_t\boldsymbol{S}_t^\top) \geq 0.98\sigma_r$. On the other hand, $\|\boldsymbol{\Delta}_t\| \geq 0.02\sigma_r$ together

with Lemma 9 implies that the one-step dynamic in Proposition 8 holds and we have

$$\begin{aligned}
&\lambda_{\min}\left(\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right) \\
&\geq \left(\left(1+\frac{\eta\rho^t}{\|\boldsymbol{\Delta}_t\|}\sigma_r\right)^2 - \frac{2\eta\rho^t}{\|\boldsymbol{\Delta}_t\|}\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| - 384\sqrt{r}\eta\rho^t\delta\right)\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right) \\
&\quad - 2\frac{\eta\rho^t}{\|\boldsymbol{\Delta}_t\|}\left(1+\frac{\eta\rho^t}{\|\boldsymbol{\Delta}_t\|}\sigma_r\right)\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^2 \\
&\geq \left(1+\frac{2\eta\rho^t}{\|\boldsymbol{\Delta}_t\|}\left(\sigma_r - \mathcal{O}(\sqrt{r}\sigma_1\delta)\right)\right)\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right) - \frac{2.01\eta\rho^t}{\|\boldsymbol{\Delta}_t\|}\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^2 \\
&\geq \left(1+\frac{1.99\eta\rho^t\sigma_r}{\|\boldsymbol{\Delta}_t\|}\right)\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right) - \frac{2.01\eta\rho^t}{\|\boldsymbol{\Delta}_t\|}\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^2 \qquad (2.58)
\end{aligned}$$

where the second inequality follows from $\|\boldsymbol{\Delta}_t\| \geq 0.02\sigma_r$ and $\eta \lesssim 1$, and the last inequality follows from $\delta \lesssim 1/(\sqrt{r}\sigma_1)$. The rest of the proof for the eigenvalue learning phase is similar to the arguments made after equation 2.34 in the proof of Theorem 6. Suppose that T_1' is the largest iteration such that $\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right) \leq \sigma_r/2.01$ for every $t \leq T_1'$. We show that $T_1' = \mathcal{O}((\kappa/\eta)\log(1/\alpha))$. To see this, note that for every $t \leq T_1'$, the above inequality can be simplified as

$$\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right) \geq \left(1+\frac{0.99\eta\rho^t\sigma_r}{\|\boldsymbol{\Delta}_t\|}\right)\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right) \geq \lambda_{\min}\left(\boldsymbol{S}_0\boldsymbol{S}_0^\top\right)\prod_{s=0}^{t}\left(1+\frac{\eta\rho^s}{6\kappa}\right), \qquad (2.59)$$

where we used the assumption $\delta \lesssim 1/(\sqrt{r}\kappa)$ and $\|\boldsymbol{\Delta}_t\| \leq 5\sigma_1$. To proceed with the proof, we need the following technical lemma.

Lemma 10. *For any $\alpha > 0, 0 < \rho < 1$ and $T \in \mathbb{N}_+$, we have*

$$\exp\left(\frac{\alpha T\rho^T}{1+\alpha}\right) \leq \prod_{t=0}^{T}\left(1+\alpha\rho^t\right) \leq \exp\left(\frac{\alpha}{1-\rho}\right).$$

The proof of Lemma 10 can be found in Appendix 2.G.5. Lemma 10 together with Lemma 6 and equation 2.59 leads to

$$\begin{aligned}
\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right) &\geq \exp\left(\frac{\eta t\rho^t}{6\kappa+\eta}\right)\alpha^2\varphi(\boldsymbol{X}^\star)\left(\frac{\sigma_r}{\|\boldsymbol{X}^\star\|_{\mathrm{F}}} - \delta\right) \\
&\geq \exp\left(\frac{\eta t\rho^t}{6\kappa+\eta}\right)\alpha^2\varphi(\boldsymbol{X}^\star)\left(\frac{\sigma_r}{\sqrt{r}\sigma_1} - \delta\right) \\
&\geq \exp\left(\frac{\eta t\rho^t}{6\kappa+\eta}\right)\frac{\alpha^2\varphi(\boldsymbol{X}^\star)}{2\sqrt{r}\kappa}.
\end{aligned}$$

Due to our assumption $\rho = 1 - \Theta(\eta/(\kappa \log(1/\alpha)))$, we have $\rho^t \geq 1 - \mathcal{O}(\eta t/(\kappa \log(1/\alpha)))$. This together with the assumption $t \leq \mathcal{O}((\kappa/\eta)\log(1/\alpha))$ leads to $\rho^t \geq 1 - \mathcal{O}(1) \geq \Omega(1)$. Therefore, one can write

$$\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right) \geq \exp\left(\Omega(1)\frac{\eta t}{\kappa}\right)\frac{\alpha^2\varphi(\boldsymbol{X}^\star)}{2\sqrt{r}\kappa}.$$

Given the above equation, it is easy to verify that after $\mathcal{O}((\kappa/\eta)\log(1/\alpha))$ iterations, we have $\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right) \geq \sigma_r/2.01$. This implies that $T_1' = \mathcal{O}((\kappa/\eta)\log(1/\alpha))$.

For $t > T_1'$, define $x_t = \sigma_r - \lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)$. Then, arguments analogous to the proof of Theorem 6 can be used to write

$$\begin{aligned}
x_{t+1} - 0.0196\sigma_r &\leq \left(1 - 1.02\frac{\eta\rho^t\sigma_r}{\|\boldsymbol{\Delta}_t\|}\right)(x_t - 0.0196\sigma_r) \\
&\leq \left(1 - \frac{\eta\rho^t}{5\kappa}\right)(x_t - 0.0196\sigma_r) \\
&\leq (x_{T_1'} - 0.0196\sigma_r)\prod_{s=T_1'}^{t}\left(1 - \frac{\eta\rho^t}{5\kappa}\right).
\end{aligned}$$

Recall that $\rho^t = \Omega(1)$ for every $t \leq \mathcal{O}((\kappa/\eta)\log(1/\alpha))$. Therefore, for every $T_1' < t \leq \mathcal{O}((\kappa/\eta)\log(1/\alpha))$, we have

$$x_{t+1} - 0.0196\sigma_r \leq (x_{T_1'} - 0.0196\sigma_r)\left(1 - \Omega(1)\frac{\eta\sigma_r}{\kappa}\right)^{t-T_1'+1}.$$

This in turn implies that after additional $T_1'' = \mathcal{O}(\kappa/\eta)$ iterations, we have $x_{t+1} \leq 0.02\sigma_r$. Therefore, we have $\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right) \geq 0.98\sigma_r$ after $T_1 = T_1' + T_1'' = \mathcal{O}((\kappa/\eta)\log(1/\alpha))$ iterations. Now, it suffices to show that, after additional $T_2 = \mathcal{O}((\kappa/\eta)\log(\kappa))$ iterations, we have $\|\boldsymbol{\Delta}_t\| \leq 0.02\sigma_r$. To this goal, suppose that $\|\boldsymbol{\Delta}_t\| \geq 0.02\sigma_r$ for every $T_1 \leq t \leq T_1 + T_2$. Recalling the signal-residual decomposition equation 2.21, one can write

$$\begin{aligned}
\|\boldsymbol{\Delta}_t\| &\leq \left\|\boldsymbol{\Sigma} - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| + 2\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\| + \|\boldsymbol{F}_t\|^2 + \|\boldsymbol{G}_t\|^2 \\
&\leq \left\|\boldsymbol{\Sigma} - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| + 0.01\sigma_r,
\end{aligned}$$

where the second inequality follows from Lemma 9. Given the above inequality, $\|\boldsymbol{\Delta}_t\| \geq 0.02\sigma_r$ implies that $\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| \geq 0.01\sigma_r$ for every $T_1 \leq t \leq T_1 + T_2$. Combined with the one-step dynamic for the signal term, we have

$$
\begin{aligned}
\left\|\boldsymbol{\Sigma}-\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right\| &\le \left(1-\frac{\eta\rho^t}{\|\boldsymbol{\Delta}_t\|}\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right)\left\|\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| + 5\frac{\eta\rho^t}{\|\boldsymbol{\Delta}_t\|}\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\|^2 \\
&\quad + 193\sqrt{r}\eta\rho^t\delta\sigma_1 \\
&\overset{(a)}{\le} \left(1-\frac{0.98\eta\rho^t}{\|\boldsymbol{\Delta}_t\|}\sigma_r\right)\left\|\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| + \mathcal{O}\left(\sqrt{r}\eta\rho^t\delta\sigma_1\right) \\
&\le \left\|\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| - 0.98\eta\rho^t\sigma_r\frac{\left\|\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|}{\left\|\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|+0.01\sigma_r} \\
&\quad + \mathcal{O}\left(\sqrt{r}\eta\rho^t\delta\sigma_1\right) \\
&\le \left\|\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| - 0.49\eta\rho^t\sigma_r + \mathcal{O}\left(\sqrt{r}\eta\rho^t\delta\sigma_1\right) \\
&\le \left\|\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| - \Omega\left(\eta\rho^t\sigma_r\right),
\end{aligned}
\tag{2.60}
$$

where in (a) we used the fact that $\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\| \lesssim \sqrt{\sigma_1}\|\boldsymbol{F}_t\| \lesssim \sigma_1\sqrt{r}\eta_0\rho^t\delta t$, $\delta \lesssim 1/\left(\sqrt{r}\kappa^2\bar{\varphi}^4\log^2(1/\alpha)\right)$, and $t \lesssim (\kappa/\eta)\log(1/\alpha)$. The above inequality leads to

$$
\left\|\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| \le \left\|\boldsymbol{\Sigma}-\boldsymbol{S}_{T_1}\boldsymbol{S}_{T_1}^\top\right\| - \Omega\left(\sum_{s=T_1}^{t-1}\eta\rho^t\sigma_r\right).
$$

Therefore, after $T_3 = T_1 + \mathcal{O}((\kappa/\eta)\log\kappa)$ iterations, we have $\left\|\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| \le 0.01\sigma_r$. This completes the proof of the first phase.

Phase 2: Global Convergence. In the second phase, we show that, once $\|\boldsymbol{\Delta}_t\| \le 0.02\sigma_r$, $\|\boldsymbol{\Delta}_t\|$ starts to decay linearly until it is dominated by $\sqrt{d}\|\boldsymbol{G}_t\|^2$. Similar to before, we define $\gamma_t = \left\|\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| + 2\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\| + \|\boldsymbol{F}_t\|^2 + \|\boldsymbol{G}_t\|^2$. Our next lemma plays a central role in our subsequent arguments:

Lemma 11. *Suppose that $T_3 \le t \le T_{end}$ is chosen such that $\gamma_s \le 0.1\sigma_r$ and $\|\boldsymbol{\Delta}_s\| \ge \sqrt{d}\|\boldsymbol{G}_s\|^2$, for every $T_3 \le s \le t$. Then, we have*

- $\gamma_{t+1} \le 0.1\sigma_r$.

Moreover, at least one of the following statements is satisfied:

- $\|\boldsymbol{\Delta}_{t+1}\| \ge \sqrt{d}\|\boldsymbol{G}_{t+1}\|^2$,
- $\|\boldsymbol{\Delta}_{t+1}\| \lesssim \sqrt{d}\alpha^{2-\mathcal{O}(\sqrt{r}\kappa\delta)}$.

To streamline the presentation, we defer the proof of the above lemma to Appendix 2.G.6. According to the above lemma, we may assume that $\gamma_t \le 0.1\sigma_r$ and $\|\boldsymbol{\Delta}_t\| \ge \sqrt{d}\|\boldsymbol{G}_t\|^2$ for all

iterations $T_3 \le t \le T_{end}$; otherwise, we have $\|\boldsymbol{\Delta}_t\| \lesssim \sqrt{d}\alpha^{2-\mathcal{O}(\sqrt{r}\kappa\delta)}$ for some $T_3 \le t \le T_{end}$, which readily completes the proof. On the other hand, the assumptions $\gamma_t \le 0.1\sigma_r$ and $\|\boldsymbol{\Delta}_t\| \ge \sqrt{d}\|\boldsymbol{G}_t\|^2$ lead to

$$\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right) \ge 0.9\sigma_r, \quad \left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| \le 1.1\sigma_1,$$

and

$$\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| \le 0.1\sigma_r, \quad \left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\right\| \le 0.2. \tag{2.61}$$

Together with our analysis in Phase 1, this implies that the one-step dynamic of $\boldsymbol{G}_t$ holds for every $0 \le t \le T_{end}$, and we have $\|\boldsymbol{G}_t\| \le 2\sqrt{2}\alpha^{1-\mathcal{O}(\sqrt{r}\kappa\delta)}\sqrt{\bar{\varphi}\delta}$, for every $0 \le t \le T_{end}$.

Under the assumption $\|\boldsymbol{\Delta}_t\| \ge \sqrt{d}\|\boldsymbol{G}_t\|^2$, our next lemma shows that, if $\|\boldsymbol{\Delta}_{t+T_3}\| \ge 0.02\sigma_r\rho^t$ for some $t \ge 0$, then there exists t' satisfying $t' - t = \mathcal{O}(1/\eta)$ such that $\|\boldsymbol{\Delta}_{t'+T_3}\| \le 0.02\sigma_r\rho^{t'}$. This in turn ensures that the generalization error decays by a constant factor every $\mathcal{O}(1/\eta)$ iterations until it reaches the same order as $\sqrt{d}\|\boldsymbol{G}_t\|^2$. In particular, we have the following lemma:

Lemma 12. *Suppose that $\|\boldsymbol{\Delta}_t\| \ge \sqrt{d}\|\boldsymbol{G}_t\|^2$ for every $T_3 \le t \le T_{end}$. Suppose that $t_0 \ge 1$ satisfies $\|\boldsymbol{\Delta}_{t_0+T_3-1}\| \le 0.02\sigma_r\rho^{t_0-1}$ and $\|\boldsymbol{\Delta}_{t_0+T_3}\| > 0.02\sigma_r\rho^{t_0}$. Then, after at most $\Delta t = \mathcal{O}(1/\eta)$ iterations, we have $\|\boldsymbol{\Delta}_{t_0+\Delta t+T_3}\| \le 0.02\sigma_r\rho^{t_0+\Delta t}$.*

The proof of the above lemma is presented in Appendix 2.G.7. We show how Lemma 12 can be used to finish the proof of Theorem 8. Recall that $\rho = 1 - \Theta(\eta/(\kappa\log(1/\alpha)))$. Let us pick $T_4 = \mathcal{O}((\kappa/\eta)\log^2(1/\alpha))$. Simple calculation reveals that $0.02\sigma_r\rho^{T_4} \lesssim \alpha^2$. According to Lemma 12, if $\|\boldsymbol{\Delta}_{T_4+T_3}\| > 0.02\sigma_r\rho^{T_4}$, then there exists $\Delta t = \mathcal{O}(1/\eta)$ such that $\|\boldsymbol{\Delta}_{T_4+T_3+\Delta t}\| \le 0.02\sigma_r\rho^{T_4+\Delta t} \lesssim \alpha^2$. Combined with Lemma 11, we have $\|\boldsymbol{\Delta}_t\|_{\mathrm{F}} \le \sqrt{d}\|\boldsymbol{\Delta}_t\| \lesssim d\|\boldsymbol{G}_t\|^2 \vee \zeta \lesssim d\alpha^{2-\mathcal{O}(\sqrt{r}\kappa\delta)} \vee \zeta$ after at most $T_{end} = T_3 + T_4 + \Delta t = \mathcal{O}((\kappa/\eta)\log^2(1/\alpha))$ iterations. Last, note that $\|\boldsymbol{G}_t\| \le 2\sqrt{2}\alpha^{1-\mathcal{O}(\sqrt{r}\kappa\delta)}\sqrt{\bar{\varphi}\delta}, \forall t \ge 0$ as shown in Lemma 9. We immediately derive $\|\boldsymbol{\Delta}_t\|_{\mathrm{F}} \le \sqrt{d}\|\boldsymbol{\Delta}_t\| \lesssim d\|\boldsymbol{G}_t\|^2 \vee \zeta \lesssim d\alpha^{2-\mathcal{O}(\sqrt{r}\kappa\delta)} \vee \zeta$. This completes the proof of Theorem 8. □

2.B Proofs of Sign-RIP

2.B.1 Preliminary

We first provide the preliminary probability tools for proving Theorems 4 and 5.

Definition 6 (Sub-Gaussian random variable)**.** *We say a random variable $X \in \mathbb{R}$ with expectation $\mathbb{E}[X] = \mu$ is σ^2-sub-Gaussian if for all $\lambda \in \mathbb{R}$, we have $\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \le e^{\frac{\lambda^2\sigma^2}{2}}$.*

Moreover, the sub-Gaussian norm of X is defined as $\|X\|_{\psi_2} := \sup_{p\in\mathbb{N}_+} \left\{p^{-1/2}(\mathbb{E}[|X|^p])^{1/p}\right\}$.

According to [Wainwright, 2019], the following statements are equivalent:

- X is σ^2-sub-Gaussian.
- (Tail bound) For any $t > 0$, we have $\mathbb{P}(|X - \mu| \geq t) \leq 2e^{-\frac{t^2}{2\sigma^2}}$.
- (Moment bound) We have $\|X\|_{\psi_2} \lesssim \sigma$.

Next, we provide the definitions of the sub-Gaussian process, ξ-net, and the covering number.

Definition 7 (Sub-Gaussian process). *A zero mean stochastic process $\{\mathcal{X}_\theta, \theta \in \mathbb{T}\}$ is a σ^2-sub-Gaussian process with respect to a metric d on a set $\mathbb{T}$, if for every $\theta, \theta' \in \mathbb{T}$, the random variable $\mathcal{X}_\theta - \mathcal{X}_{\theta'}$ is $(\sigma d(\theta, \theta'))^2$-sub-Gaussian.*

Definition 8 (ξ-net and covering number). *A set $\mathcal{N}$ is called an ξ-net on $(\mathbb{T}, d)$ if for every $t \in \mathbb{T}$, there exists $\pi(t) \in \mathcal{N}$ such that $d(t, \pi(t)) \leq \xi$. The covering number $N(\mathbb{T}, d, \xi)$ is defined as the smallest cardinality of an ξ-net for $(\mathbb{T}, d)$:*

$$N(\mathbb{T}, d, \xi) := \inf\{|\mathcal{N}| : \mathcal{N} \text{ is an } \xi\text{-net for } (\mathbb{T}, d)\}.$$

Next, we introduce some additional notations which will be used throughout our arguments. Define the rank-k and ϵ-approximate rank-k unit balls as:

$$\begin{aligned}
\mathbb{S}_k &= \{\boldsymbol{X} \in \mathbb{R}^{d\times d} : \operatorname{rank}(\boldsymbol{X}) \leq k, \|\boldsymbol{X}\|_{\mathrm{F}} = 1\}, \\
\mathbb{S}_{k,\epsilon} &= \{\boldsymbol{X} \in \mathbb{R}^{d\times d} : \|\boldsymbol{X}\|_{\mathrm{F}} = 1 \text{ and } \exists \boldsymbol{X}' \text{ s.t. } \operatorname{rank}(\boldsymbol{X}') \leq k, \|\boldsymbol{X} - \boldsymbol{X}'\|_{\mathrm{F}} \leq \epsilon\}.
\end{aligned}$$

For simplicity of notation, we use $N_{k,\xi}$ to denote $N(\mathbb{S}_k, \|\cdot\|_{\mathrm{F}}, \xi)$. Moreover, we define $\mathcal{S}_{k,\epsilon}$ as the restriction of the set $\mathcal{S}$ to the set of ϵ-approximate rank-k matrices, i.e., $\mathcal{S}_{k,\epsilon} = \{\boldsymbol{X} : \boldsymbol{X} \in \mathcal{S}, \boldsymbol{X} \text{ is } \epsilon\text{-approximate rank-}k\}$. The following lemma characterizes the covering number of the set of low-rank matrices with the unit norm.

Lemma 13 (Li et al. [2020b]). *We have $N_{k,\xi} = N(\mathbb{S}_k, \|\cdot\|_{\mathrm{F}}, \xi) \leq \left(\frac{9}{\xi}\right)^{(2d+1)k}$.*

The following well-known result characterizes a concentration bound on the supremum of a sub-Gaussian process.

Theorem 11 (Corollary 5.25 and Theorem 5.29 in [Van Handel, 2014]). *Let $\{X_t\}_{t\in\mathbb{T}}$ be a separable sub-Gaussian process on $(\mathbb{T}, d)$. Then, the following statements hold:*

- *We have*
$$\mathbb{E}\left[\sup_{t\in\mathbb{T}} X_t\right] \leq 12 \int_0^\infty \sqrt{\log N(\mathbb{T}, d, \xi)} d\xi.$$

- *For all $t_0 \in \mathbb{T}$ and $x \geq 0$, we have*

$$\mathbb{P}\left(\sup_{t\in\mathbb{T}}\{X_t - X_{t_0}\} \geq C\int_0^\infty \sqrt{\log N(\mathbb{T}, d, \xi)}d\xi + x\right) \leq Ce^{-x^2/C\,\mathrm{diam}(\mathbb{T})^2},$$

where $C < \infty$ is a universal constant, and $\mathrm{diam}(\mathbb{T}) = \sup_{t,t'\in\mathbb{T}}\{d(t,t')\}$ is the diameter of $\mathbb{T}$.

Equipped with these preliminary results, we proceed with the proof of Theorem 4.

2.B.2 Proof of Theorem 4

To provide the proof of Theorem 4, we first define the following stochastic process:

$$\mathcal{H}_{\boldsymbol{X},\boldsymbol{Y}} = \sup\left\{\frac{1}{m}\sum_{i=1}^m \mathrm{sign}(\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle - s_i)\langle \boldsymbol{A}_i, \boldsymbol{Y}\rangle - \varphi(\boldsymbol{X})\left\langle \frac{\boldsymbol{X}}{\|\boldsymbol{X}\|_{\mathrm{F}}}, \boldsymbol{Y}\right\rangle\right\},$$

where $\varphi(\boldsymbol{X}) := \sqrt{\frac{2}{\pi}}\left(1 - p + p\mathbb{E}\left[e^{-s^2/(2\|\boldsymbol{X}\|_{\mathrm{F}}^2)}\right]\right)$, and the supremum is taken over the set-valued function $\mathrm{sign}(\cdot)$. Moreover, to streamline the presentation and whenever there is no ambiguity, we drop the supremum when it is taken with respect to the set-valued function $\mathrm{sign}(\cdot)$. Our next lemma provides a sufficient condition for Sign-RIP.

Lemma 14. *Sign-RIP holds with parameters delineated in Theorem 4, if*

$$\sup_{\boldsymbol{X}\in\mathcal{S}_{k,\epsilon},\boldsymbol{Y}\in\mathbb{S}_{k,\epsilon/\zeta}} \mathcal{H}_{\boldsymbol{X},\boldsymbol{Y}} \lesssim \inf_{\boldsymbol{X}\in\mathcal{S}_{k,\epsilon}} \varphi(\boldsymbol{X})\delta. \tag{2.62}$$

Proof. According to the definition, Sign-RIP is satisfied if, for every $\boldsymbol{X}, \boldsymbol{Y} \in \mathcal{S}_{k,\epsilon}$ and $\boldsymbol{Q} \in \mathcal{Q}(\boldsymbol{X})$, we have

$$\left\langle \boldsymbol{Q} - \varphi(\boldsymbol{X})\frac{\boldsymbol{X}}{\|\boldsymbol{X}\|_{\mathrm{F}}}, \frac{\boldsymbol{Y}}{\|\boldsymbol{Y}\|_{\mathrm{F}}}\right\rangle \leq \varphi(\boldsymbol{X})\delta. \tag{2.63}$$

Recall that $\mathcal{S}_{k,\epsilon} = \{\boldsymbol{X} : \zeta \leq \|\boldsymbol{X}\|_{\mathrm{F}} \leq R, \boldsymbol{X} \text{ is } \epsilon\text{-approximate rank-}k\}$. This implies that $\boldsymbol{Y} \in \mathcal{S}_{k,\epsilon}$ if $\boldsymbol{Y}/\|\boldsymbol{Y}\|_{\mathrm{F}} \in \mathbb{S}_{k,\epsilon/\zeta}$. Hence, it suffices to restrict $\boldsymbol{Y} \in \mathbb{S}_{k,\epsilon/\zeta}$. Therefore, Sign-RIP is satisfied if

$$\mathcal{H}_{\boldsymbol{X},\boldsymbol{Y}} \leq \varphi(\boldsymbol{X})\delta, \quad \forall \boldsymbol{X} \in \mathcal{S}_{k,\epsilon}, \boldsymbol{Y} \in \mathbb{S}_{k,\epsilon/\zeta}.$$

Hence, to guarantee Sign-RIP, it suffices to have

$$\sup_{\boldsymbol{X}\in\mathcal{S}_{k,\epsilon},\boldsymbol{Y}\in\mathbb{S}_{k,\epsilon/\zeta}} \mathcal{H}_{\boldsymbol{X},\boldsymbol{Y}} \leq \inf_{\boldsymbol{X}\in\mathcal{S}_{k,\epsilon}} \varphi(\boldsymbol{X})\delta.$$

□

Relying on the above lemma, we instead focus on analyzing the stochastic process $\{\mathcal{H}_{\boldsymbol{X},\boldsymbol{Y}}\}_{\boldsymbol{X}\in\mathcal{S}_{k,\epsilon},\boldsymbol{Y}\in\mathbb{S}_{k,\epsilon/\zeta}}$. As a first step towards this goal, we show that the scaling function can be used to characterize $\mathbb{E}[\langle \boldsymbol{Q},\boldsymbol{Y}\rangle]$, for every $\boldsymbol{Q}\in\mathcal{Q}(\boldsymbol{X})$.

Lemma 15. *Suppose that the matrix A has i.i.d. standard Gaussian entries and the noise satisfies Assumption 4. Then, for every $\boldsymbol{Q}\in\mathcal{Q}(\boldsymbol{X})$, we have*

$$\mathbb{E}\left[\langle \boldsymbol{Q},\boldsymbol{Y}\rangle\right] = \sqrt{\frac{2}{\pi}}\left(1-p+p\mathbb{E}\left[e^{-s^2/\left(2\|\boldsymbol{X}\|_F\right)}\right]\right)\left\langle \frac{\boldsymbol{X}}{\|\boldsymbol{X}\|_{\mathrm{F}}},\boldsymbol{Y}\right\rangle.$$

Proof. Without loss of generality, we assume $\|\boldsymbol{X}\|_{\mathrm{F}} = \|\boldsymbol{Y}\|_{\mathrm{F}} = 1$. Let us denote $u := \langle \boldsymbol{A},\boldsymbol{X}\rangle, v := \langle \boldsymbol{A},\boldsymbol{Y}\rangle, \rho := \mathrm{Cov}(u,v) = \langle \boldsymbol{X},\boldsymbol{Y}\rangle$. Then, we have

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{sign}\left(\langle \boldsymbol{A},\boldsymbol{X}\rangle - s\right)\langle \boldsymbol{A},\boldsymbol{Y}\rangle \,|s\neq 0\right]\\
&= \mathbb{E}\left[\mathrm{sign}\left(u-s\right)v|s\neq 0\right]\\
&\overset{(a)}{=} \rho\mathbb{E}\left[\mathrm{sign}(u-s)u|s\neq 0\right]\\
&= \rho\mathbb{E}_{s\sim\mathbb{P}}\left[\int_s^\infty u\frac{1}{\sqrt{2\pi}}e^{-u^2/2}du - \int_{-\infty}^s u\frac{1}{\sqrt{2\pi}}e^{-u^2/2}du\right]\\
&= \rho\mathbb{E}_{s\sim\mathbb{P}}\left[\int_s^\infty u\frac{1}{\sqrt{2\pi}}e^{-u^2/2}du + \int_{-s}^\infty u\frac{1}{\sqrt{2\pi}}e^{-u^2/2}du\right]\\
&= 2\rho\mathbb{E}_{s\sim\mathbb{P}}\left[\int_{|s|}^\infty u\frac{1}{\sqrt{2\pi}}e^{-u^2/2}du\right]\\
&= \sqrt{\frac{2}{\pi}}\mathbb{E}_{s\sim\mathbb{P}}\left[\int_{|s|}^\infty d\left(-e^{-u^2/2}\right)\right]\langle \boldsymbol{X},\boldsymbol{Y}\rangle\\
&= \sqrt{\frac{2}{\pi}}\mathbb{E}_{s\sim\mathbb{P}}\left[e^{-s^2/2}\right]\langle \boldsymbol{X},\boldsymbol{Y}\rangle,
\end{aligned}$$

where, in (a), we used the fact that $v|u,s\sim\mathcal{N}(\rho u,1-\rho^2)$ since s is independent of u,v. Similarly, one can show that $\mathbb{E}\left[\mathrm{sign}\left(\langle \boldsymbol{A},\boldsymbol{X}\rangle\right)\langle \boldsymbol{A},\boldsymbol{Y}\rangle\right] = \sqrt{\frac{2}{\pi}}\langle \boldsymbol{X},\boldsymbol{Y}\rangle$. The proof is completed by noting that

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{sign}\left(s+\langle \boldsymbol{A},\boldsymbol{X}\rangle\right)\langle \boldsymbol{A},\boldsymbol{Y}\rangle\right]\\
&= p\mathbb{E}\left[\mathrm{sign}\left(\langle \boldsymbol{A},\boldsymbol{X}\rangle - s\right)\langle \boldsymbol{A},\boldsymbol{Y}\rangle\,|s\neq 0\right] + (1-p)\mathbb{E}\left[\mathrm{sign}\left(\langle \boldsymbol{A},\boldsymbol{X}\rangle\right)\langle \boldsymbol{A},\boldsymbol{Y}\rangle\right].
\end{aligned}$$

This completes the proof. □

Now, we provide an overview of our proof technique for Theorem 4. Let $\mathcal{G}_{\boldsymbol{Y}} =$

$\sup_{\boldsymbol{X}\in\mathcal{S}_{k,\epsilon}} \mathcal{H}_{\boldsymbol{X},\boldsymbol{Y}}$ and $\bar{\mathcal{G}}_{\boldsymbol{Y}} = \mathcal{G}_{\boldsymbol{Y}} - \mathbb{E}[\mathcal{G}_{\boldsymbol{Y}}]$ be stochastic processes indexed by $\boldsymbol{Y}$. According to Lemma 14, it suffices to control $\sup_{\boldsymbol{Y}\in\mathbb{S}_{k,\epsilon/\zeta}} \mathcal{G}_{\boldsymbol{Y}}$. We consider the following decomposition:

$$\begin{aligned}\sup_{\boldsymbol{Y}\in\mathbb{S}_{k,\epsilon/\zeta}} \mathcal{G}_{\boldsymbol{Y}} &\le \sup_{\boldsymbol{Y}\in\mathbb{S}_{k,\epsilon/\zeta}} \bar{\mathcal{G}}_{\boldsymbol{Y}} + \sup_{\boldsymbol{Y}\in\mathbb{S}_{k,\epsilon/\zeta}} \mathbb{E}[\mathcal{G}_{\boldsymbol{Y}}] \\ &\le \sup_{\boldsymbol{Y}\in\mathbb{S}_{k}} \bar{\mathcal{G}}_{\boldsymbol{Y}} + \frac{\epsilon}{\zeta} \sup_{\boldsymbol{Y}\in\mathbb{S}} \bar{\mathcal{G}}_{\boldsymbol{Y}} + \sup_{\boldsymbol{Y}\in\mathbb{S}_{k}} \mathbb{E}[\mathcal{G}_{\boldsymbol{Y}}] + \frac{\epsilon}{\zeta} \sup_{\boldsymbol{Y}\in\mathbb{S}} \mathbb{E}[\mathcal{G}_{\boldsymbol{Y}}], \end{aligned} \tag{2.64}$$

where the second inequality follows from $\mathbb{S}_{k,\epsilon/\zeta} \subset \mathbb{S}_k + \frac{\epsilon}{\zeta}\mathbb{S}$, and the fact that both $\bar{\mathcal{G}}_{\boldsymbol{Y}}$ and $\mathbb{E}[\mathcal{G}_{\boldsymbol{Y}}]$ are linear functions of $\boldsymbol{Y}$. We control the individual terms in the above decomposition separately. To provide an upper bound for $\sup_{\boldsymbol{Y}\in\mathbb{S}_k} \bar{\mathcal{G}}_{\boldsymbol{Y}}$ and $\sup_{\boldsymbol{Y}\in\mathbb{S}} \bar{\mathcal{G}}_{\boldsymbol{Y}}$, we rely on the following key lemma.

Lemma 16. *The stochastic processes $\{\bar{\mathcal{G}}_{\boldsymbol{Y}}\}_{\boldsymbol{Y}\in\mathbb{S}_k}$ and $\{\bar{\mathcal{G}}_{\boldsymbol{Y}}\}_{\boldsymbol{Y}\in\mathbb{S}}$ are $\mathcal{O}(1/m)$-sub-Gaussian processes.*

Proof. Since $\mathbb{S}_k \subset \mathbb{S}$, it suffices to show that $\{\bar{\mathcal{G}}_{\boldsymbol{Y}}\}_{\boldsymbol{Y}\in\mathbb{S}}$ is $\mathcal{O}(1/m)$-sub-Gaussian. According to Definition 17, the stochastic process $\left\{\bar{\mathcal{G}}_{\boldsymbol{Y}}\right\}_{\boldsymbol{Y}\in\mathbb{S}}$ is sub-Gaussian if for any arbitrary $\boldsymbol{Y}, \boldsymbol{Y}' \in \mathbb{S}$, $\bar{\mathcal{G}}_{\boldsymbol{Y}} - \bar{\mathcal{G}}_{\boldsymbol{Y}'}$ is $\mathcal{O}\left(\|\boldsymbol{Y} - \boldsymbol{Y}'\|_{\mathrm{F}}^2 / m\right)$-sub-Gaussian. Note that $\bar{\mathcal{G}}_{\boldsymbol{Y}} - \bar{\mathcal{G}}_{\boldsymbol{Y}'}$ is sub-Gaussian if and only if $\mathcal{G}_{\boldsymbol{Y}} - \mathcal{G}_{\boldsymbol{Y}'}$ is sub-Gaussian with the same parameter. The latter will be proven by checking the moment bound condition in Definition 16. For arbitrary $\boldsymbol{Y}, \boldsymbol{Y}' \in \mathbb{S}$, denote $\boldsymbol{\Delta Y} = \boldsymbol{Y} - \boldsymbol{Y}'$. Then, for any $p \in \mathbb{N}_+$, we have

$$\begin{aligned}&\mathbb{E}\left[\left|\mathcal{G}_{\boldsymbol{Y}} - \mathcal{G}_{\boldsymbol{Y}'}\right|^{2p}\right] \\ &\le \mathbb{E}\left[\left|\sup_{\boldsymbol{X}\in\mathcal{S}_{k,\epsilon}} \frac{1}{m}\sum_{i=1}^{m} \operatorname{sign}\left(\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle - s_i\right) \langle \boldsymbol{A}_i, \boldsymbol{\Delta Y}\rangle - \varphi(\boldsymbol{X}) \left\langle \frac{\boldsymbol{X}}{\|\boldsymbol{X}\|_{\mathrm{F}}}, \boldsymbol{\Delta Y}\right\rangle\right|^{2p}\right] \\ &\le \mathbb{E}\left[\left(\sup_{\boldsymbol{X}\in\mathcal{S}_{k,\epsilon}} \left|\frac{1}{m}\sum_{i=1}^{m} \operatorname{sign}\left(\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle - s_i\right) \langle \boldsymbol{A}_i, \boldsymbol{\Delta Y}\rangle\right| + \sup_{\boldsymbol{X}\in\mathcal{S}_{k,\epsilon}} \left|\varphi(\boldsymbol{X}) \left\langle \frac{\boldsymbol{X}}{\|\boldsymbol{X}\|_{\mathrm{F}}}, \boldsymbol{\Delta Y}\right\rangle\right|\right)^{2p}\right] \\ &\le \mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^{m} |\langle \boldsymbol{A}_i, \boldsymbol{\Delta Y}\rangle| + \sqrt{\frac{2}{\pi}} \|\boldsymbol{\Delta Y}\|_{\mathrm{F}}\right)^{2p}\right] \\ &\le 2^{2p}\left(\mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^{m} |\langle \boldsymbol{A}_i, \boldsymbol{\Delta Y}\rangle|\right)^{2p}\right] + \left(\frac{2}{\pi}\right)^{p} \|\boldsymbol{\Delta Y}\|_{\mathrm{F}}^{2p}\right),\end{aligned}$$

where in the third inequality, we used $\varphi(\boldsymbol{X}) \le \sqrt{2/\pi}$, which holds for every $X \in \mathbb{R}^{d\times d}$. According to [Li et al., 2020b, Appendix A.2], the random variable $(1/m)\sum_{i=1}^{m} |\langle \boldsymbol{A}_i, \boldsymbol{\Delta Y}\rangle|$

is $\mathcal{O}\left(\|\boldsymbol{\Delta Y}\|_{\mathrm{F}}^2/m\right)$-sub-Gaussian with mean $\sqrt{2/\pi}\,\|\boldsymbol{\Delta Y}\|_{\mathrm{F}}$. Therefore, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^{m}|\langle \boldsymbol{A}_i, \boldsymbol{\Delta Y}\rangle|\right)^{2p}\right] \\
&= \mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^{m}|\langle \boldsymbol{A}_i, \boldsymbol{\Delta Y}\rangle| - \sqrt{\frac{2}{\pi}}\|\boldsymbol{\Delta Y}\|_F + \sqrt{\frac{2}{\pi}}\|\boldsymbol{\Delta Y}\|_F\right)^{2p}\right] \\
&\le 2^{2p}\left(\mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^{m}|\langle \boldsymbol{A}_i, \boldsymbol{\Delta Y}\rangle| - \sqrt{\frac{2}{\pi}}\|\boldsymbol{\Delta Y}\|_F\right)^{2p}\right] + \left(\frac{2}{\pi}\right)^{p}\|\boldsymbol{\Delta Y}\|_F^{2p}\right) \\
&\le 2^{2p}\left((2p-1)!!\frac{1}{m^p}\|\boldsymbol{\Delta Y}\|_F^{2p} + \left(\frac{2}{\pi}\right)^{p}\|\boldsymbol{\Delta Y}\|_F^{2p}\right).
\end{aligned}
$$

Combining the above inequalities leads to

$$
\mathbb{E}\left[|\mathcal{G}_{\boldsymbol{Y}} - \mathcal{G}_{\boldsymbol{Y}'}|^{2p}\right]^{1/2p} \le 4\left((2p-1)!!\frac{1}{m^p}\|\boldsymbol{\Delta Y}\|_F^{2p} + \left(\frac{2}{\pi}\right)^{p}\|\boldsymbol{\Delta Y}\|_F^{2p}\right)^{1/2p} \lesssim \sqrt{\frac{p}{m}}\,\|\boldsymbol{\Delta Y}\|_{\mathrm{F}},
$$

given that $p > m$. Therefore, $\{\mathcal{G}_{\boldsymbol{Y}}\}_{\boldsymbol{Y}\in\mathbb{S}}$ is a $\mathcal{O}\left(1/m\right)$-sub-Gaussian process, which implies that $\{\mathcal{G}_{\boldsymbol{Y}}\}_{\boldsymbol{Y}\in\mathbb{S}_k}$ is also a $\mathcal{O}\left(1/m\right)$-sub-Gaussian process. □

Given that both $\{\bar{\mathcal{G}}_{\boldsymbol{Y}}\}_{\boldsymbol{Y}\in\mathbb{S}_k}$ and $\{\bar{\mathcal{G}}_{\boldsymbol{Y}}\}_{\boldsymbol{Y}\in\mathbb{S}}$ are sub-Gaussian processes, we can readily obtain sharp concentration bounds on their suprema.

Lemma 17. *The following statements hold:*

$$
\mathbb{E}\left[\sup_{\boldsymbol{Y}\in\mathbb{S}_k}\bar{\mathcal{G}}_{\boldsymbol{Y}}\right] \lesssim \sqrt{\frac{dk}{m}}, \qquad \mathbb{P}\left(\sup_{\boldsymbol{Y}\in\mathbb{S}_k}\bar{\mathcal{G}}_{\boldsymbol{Y}} \ge \mathbb{E}\left[\sup_{\boldsymbol{Y}\in\mathbb{S}_k}\bar{\mathcal{G}}_{\boldsymbol{Y}}\right] + \gamma\right) \lesssim e^{-cm\gamma^2}; \tag{2.65}
$$

$$
\mathbb{E}\left[\sup_{\boldsymbol{Y}\in\mathbb{S}}\bar{\mathcal{G}}_{\boldsymbol{Y}}\right] \lesssim \sqrt{\frac{d^2}{m}}, \qquad \mathbb{P}\left(\sup_{\boldsymbol{Y}\in\mathbb{S}}\bar{\mathcal{G}}_{\boldsymbol{Y}} \ge \mathbb{E}\left[\sup_{\boldsymbol{Y}\in\mathbb{S}}\bar{\mathcal{G}}_{\boldsymbol{Y}}\right] + \gamma\right) \lesssim e^{-cm\gamma^2}. \tag{2.66}
$$

Proof. The proof follows directly from Theorem 11. The details are omitted for brevity. □

Equipped with the above lemma, we provide a concentration bound on $\sup_{\boldsymbol{Y}\in\mathbb{S}_{k,\epsilon/\zeta}}\bar{\mathcal{G}}_{\boldsymbol{Y}}$.

Lemma 18. *Assume that $m \gtrsim \sqrt{dk/\gamma^2}$ and $\epsilon/\zeta \lesssim \sqrt{k/d}$. Then, the following inequality holds with probability of at least $1 - Ce^{-cm\gamma^2}$:*

$$
\sup_{\boldsymbol{Y}\in\mathbb{S}_{k,\epsilon/\zeta}}\bar{\mathcal{G}}_{\boldsymbol{Y}} \le \left(3 + \sqrt{\frac{k}{d}}\right)\gamma.
$$

Proof. Based on equation 2.65, we have, with probability of at least $1 - Ce^{-cm\gamma^2}$

$$\sup_{\boldsymbol{Y}\in\mathbb{S}_k} \bar{\mathcal{G}}_{\boldsymbol{Y}} \leq \mathbb{E}\left[\sup_{\boldsymbol{Y}\in\mathbb{S}_k} \bar{\mathcal{G}}_{\boldsymbol{Y}}\right] + \gamma \lesssim \sqrt{\frac{dk}{m}} + \gamma \leq 2\gamma,$$

where the last inequality follows from the assumption $m \gtrsim \sqrt{dk/\gamma^2}$. Similarly, based on equation 2.66, the following inequalities hold with a probability of at least $1 - Ce^{-cm\gamma^2}$:

$$\frac{\epsilon}{\zeta}\sup_{\boldsymbol{Y}\in\mathbb{S}} \bar{\mathcal{G}}_{\boldsymbol{Y}} \leq \frac{\epsilon}{\zeta}\left(\mathbb{E}\left[\sup_{\boldsymbol{Y}\in\mathbb{S}} \bar{\mathcal{G}}_{\boldsymbol{Y}}\right] + \gamma\right) \lesssim \sqrt{\frac{\epsilon^2 d^2}{\zeta^2 m}} + \frac{\epsilon\gamma}{\zeta} \leq \left(1 + \sqrt{\frac{k}{d}}\right)\gamma,$$

where the last inequality follows from $m \gtrsim \sqrt{dk/\gamma^2}$ and $\epsilon/\zeta \lesssim \sqrt{k/d}$. Finally, a simple union bound implies that

$$\sup_{\boldsymbol{Y}\in\mathbb{S}_{k,\epsilon/\zeta}} \bar{\mathcal{G}}_{\boldsymbol{Y}} \leq \sup_{\boldsymbol{Y}\in\mathbb{S}_k} \bar{\mathcal{G}}_{\boldsymbol{Y}} + \frac{\epsilon}{\zeta}\sup_{\boldsymbol{Y}\in\mathbb{S}} \bar{\mathcal{G}}_{\boldsymbol{Y}} \leq \left(3 + \sqrt{\frac{k}{d}}\right)\gamma,$$

with probability of at least $1 - Ce^{-cm\gamma^2}$, thereby completing the proof. □

Recalling equation 2.64, it remains to control the terms $\sup_{\boldsymbol{Y}\in\mathbb{S}_k} \mathbb{E}[\mathcal{G}_{\boldsymbol{Y}}]$ and $\sup_{\boldsymbol{Y}\in\mathbb{S}} \mathbb{E}[\mathcal{G}_{\boldsymbol{Y}}]$.

Lemma 19. *The following inequality holds:*

$$\sup_{\boldsymbol{Y}\in\mathbb{S}_k} \mathbb{E}[\mathcal{G}_{\boldsymbol{Y}}] \leq \sup_{\boldsymbol{Y}\in\mathbb{S}} \mathbb{E}[\mathcal{G}_{\boldsymbol{Y}}] \lesssim \sqrt{\frac{dk}{m}\log^2(m)\log\left(\frac{R}{\zeta}\right)}. \tag{2.67}$$

Due to its length, we defer the proof of Lemma 19 to Appendix 2.G.1. Equipped with Lemmas 18 and 19, we are ready to present the proof of Theorem 4.

Proof of Theorem 4. The inequality equation 2.64 combined with Lemmas 18 and 19 implies that the following inequality holds with probability of at least $1 - Ce^{-cm\gamma^2}$:

$$\sup_{\boldsymbol{Y}\in\mathbb{S}_{k,\epsilon/\zeta}} \mathcal{G}_{\boldsymbol{Y}} \lesssim \sqrt{\frac{dk}{m}\log^2(m)\log\left(\frac{R}{\zeta}\right)} + \gamma.$$

On the other hand, one can write

$$\inf_{\boldsymbol{X}\in\mathcal{S}_{k,\epsilon}} \varphi(\boldsymbol{X}) = \inf_{\boldsymbol{X}\in\mathcal{S}_{k,\epsilon}} \left\{\sqrt{\frac{2}{\pi}}\left(1 - p + p\mathbb{E}\left[e^{-s^2/(2\|\boldsymbol{X}\|_{\mathrm{F}}^2)}\right]\right)\right\} \geq \sqrt{\frac{2}{\pi}}(1-p). \tag{2.68}$$

Therefore, upon choosing

$$m \gtrsim \frac{dk \log^2(m) \log(R/\zeta)}{(1-p)^2\delta^2}, \quad \text{and} \quad \gamma \lesssim (1-p)\delta,$$

we have

$$\sup_{\boldsymbol{Y} \in \mathbb{S}_{k,\epsilon/\zeta}} \mathcal{G}_{\boldsymbol{Y}} = \sup_{\substack{\boldsymbol{X} \in \mathcal{S}_{k,\epsilon} \\ Y \in \mathbb{S}_{k,\epsilon/\zeta}}} \mathcal{H}_{\boldsymbol{X},\boldsymbol{Y}} \leq \inf_{\boldsymbol{X} \in \mathcal{S}_{k,\epsilon}} \varphi(\boldsymbol{X})\delta$$

with probability of at least $1 - Ce^{-cm(1-p)^2\delta^2}$. This completes the proof of Theorem 4. □

2.B.3 Proof of Theorem 5

Recall that, with the outlier noise model, the scaling function takes the form $\varphi(\boldsymbol{X}) = \sqrt{\frac{2}{\pi}}\left(1 - p + p\mathbb{E}\left[e^{-s^2/(2\|\boldsymbol{X}\|_{\mathrm{F}}^2)}\right]\right)$. Setting $p = 1$, and $s_i \sim \mathcal{N}(0, \nu_g^2)$ immediately implies $\varphi(\boldsymbol{X}) = \mathbb{E}_{s\sim\mathcal{N}(0,\nu_g^2)}\left[e^{-s^2/(2\|\boldsymbol{X}\|_{\mathrm{F}}^2)}\right] = \sqrt{\frac{2}{\pi}}\frac{\|\boldsymbol{X}\|_{\mathrm{F}}}{\sqrt{\|\boldsymbol{X}\|_{\mathrm{F}}^2+\nu_g^2}}$. On the other hand, using the same method in the proof of Theorem 4, we can show that the following inequality holds with probability at least $1 - Ce^{-cm\gamma^2}$:

$$\sup_{\boldsymbol{Y} \in \mathbb{S}_{k,\epsilon/\zeta}} \mathcal{G}_{\boldsymbol{Y}} \lesssim \sqrt{\frac{dk}{m} \log^2(m) \log\left(\frac{R}{\zeta}\right)} + \gamma,$$

where $\mathcal{G}_{\boldsymbol{Y}}$ is defined in the proof of Theorem 4. It remains to bound $\inf_{\boldsymbol{X} \in \mathcal{S}_{k,\epsilon}} \varphi(\boldsymbol{X})$. To this goal, note that $\|\boldsymbol{X}\|_{\mathrm{F}} \geq \zeta, \forall X \in \mathcal{S}_{k,\epsilon}$. Hence, we have the following lower bound for the scaling function

$$\inf_{\boldsymbol{X} \in \mathcal{S}_{k,\epsilon}} \varphi(\boldsymbol{X}) = \inf_{\boldsymbol{X} \in \mathcal{S}_{k,\epsilon}} \sqrt{\frac{2}{\pi}} \frac{\|\boldsymbol{X}\|_{\mathrm{F}}}{\sqrt{\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \nu_g^2}} \geq \sqrt{\frac{2}{\pi}} \frac{\zeta}{\sqrt{\zeta^2 + \nu_g^2}} \gtrsim \frac{\zeta}{\nu_g}, \tag{2.69}$$

provided that $\zeta \lesssim \nu_g$. Therefore, upon choosing $m \gtrsim \frac{\nu_g^2 dk \log^2(m) \log(R/\zeta)}{\zeta^2\delta^2}$ and $\gamma \lesssim \frac{\zeta}{\nu_g}\delta$, we have

$$\sup_{\boldsymbol{Y} \in \mathbb{S}_{k,\epsilon/\zeta}} \mathcal{G}_{\boldsymbol{Y}} = \sup_{\substack{\boldsymbol{X} \in \mathcal{S}_{k,\epsilon} \\ Y \in \mathbb{S}_{k,\epsilon/\zeta}}} \mathcal{H}_{\boldsymbol{X},\boldsymbol{Y}} \leq \inf_{\boldsymbol{X} \in \mathcal{S}_{k,\epsilon}} \varphi(\boldsymbol{X})\delta, \tag{2.70}$$

with probability at least $1 - C_1 e^{-C_2 m\zeta^2\delta^2/\nu_g^2}$, which completes the proof. □

2.B.4 Proof of Lemma 1

Lemma 20 is a direct consequence of the following more general result:

Lemma 20. *Suppose that the measurements are noiseless, and satisfy $(k, \delta, \epsilon, \mathcal{S})$-Sign-RIP for any $\epsilon \geq 0$, $\mathcal{S}$, and scaling function $\varphi(\cdot)$. Then, for any ϵ-approximate rank-k matrix $\boldsymbol{X} \in \mathcal{S}$, we have*

$$\left| \frac{1}{m} \sum_{i=1}^{m} |\langle \boldsymbol{A}_i, \boldsymbol{X} \rangle| - \varphi(\boldsymbol{X}) \left\| \boldsymbol{X} \right\|_{\mathrm{F}} \right| \leq \delta \varphi(\boldsymbol{X}) \left\| \boldsymbol{X} \right\|_{\mathrm{F}}.$$

Proof. According to the definition of Sign-RIP, for every ϵ-approximate rank-k matrices $\boldsymbol{X}, \boldsymbol{Y} \in \mathcal{S}$, and every $\boldsymbol{Q} \in \mathcal{Q}(\boldsymbol{X})$, we have $\left| \left\langle \boldsymbol{Q} - \varphi(\boldsymbol{X}) \frac{\boldsymbol{X}}{\|\boldsymbol{X}\|_{\mathrm{F}}}, \frac{\boldsymbol{Y}}{\|Y\|_{\mathrm{F}}} \right\rangle \right| \leq \varphi(\boldsymbol{X}) \delta$. In particular, upon choosing $\boldsymbol{Y} = X$, we have

$$\left| \left\langle \boldsymbol{Q} - \varphi(\boldsymbol{X}) \frac{\boldsymbol{X}}{\|\boldsymbol{X}\|_{\mathrm{F}}}, \boldsymbol{X} \right\rangle \right| = \left| \frac{1}{m} \sum_{i=1}^{m} |\langle \boldsymbol{A}_i, \boldsymbol{X} \rangle| - \varphi(\boldsymbol{X}) \left\| \boldsymbol{X} \right\|_{\mathrm{F}} \right| \leq \varphi(\boldsymbol{X}) \delta \left\| \boldsymbol{X} \right\|_{\mathrm{F}}.$$

This completes the proof. □

Evidently, the result of Lemma 20 can be readily recovered from the above lemma after substituting $\boldsymbol{X}$ with $\boldsymbol{\Delta}_t$. Moreover, it is worth noting that the above lemma recovers the so-called ℓ_1/ℓ_2-RIP introduced in [Li et al., 2020b] with $\epsilon = 0$ and $\varphi(\boldsymbol{X}) = \sqrt{2/\pi}$.

2.B.5 Proof of Lemma 2

Due to Sign-RIP, we have

$$\begin{aligned} & \|\boldsymbol{Q}_t\| \leq \varphi(\boldsymbol{\Delta}_t) \left(\frac{\|\boldsymbol{\Delta}_t\|}{\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}} + \delta \right) \\ \Longrightarrow & \frac{1}{\|\boldsymbol{Q}_t\|} \geq \frac{1}{\varphi(\boldsymbol{\Delta}_t)} \frac{1}{\delta + \frac{\|\boldsymbol{\Delta}_t\|}{\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}}} \\ \Longrightarrow & \frac{1}{\|\boldsymbol{Q}_t\|} \geq \frac{1}{\varphi(\boldsymbol{\Delta}_t)} \frac{\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}}{\|\boldsymbol{\Delta}_t\|} \left(1 - \delta \frac{\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}}{\|\boldsymbol{\Delta}_t\|} \right). \end{aligned} \tag{2.71}$$

By our assumption, $\boldsymbol{\Delta}_t$ is ϵ-approximate rank-k. This implies that, there exists a rank-k matrix $\boldsymbol{X}'$ such that $\|\boldsymbol{\Delta}_t - \boldsymbol{X}'\|_{\mathrm{F}} \leq \epsilon$. This in turn implies that

$$\begin{aligned} \|\boldsymbol{\Delta}_t\|_{\mathrm{F}} &\leq \|\boldsymbol{\Delta}_t - \boldsymbol{X}'\|_{\mathrm{F}} + \|\boldsymbol{X}'\|_{\mathrm{F}} \\ &\leq \epsilon + \|\boldsymbol{X}'\|_{\mathrm{F}} \\ &\leq \epsilon + \sqrt{k} \|\boldsymbol{X}'\| \\ &\leq (1 + \sqrt{k})\epsilon + \sqrt{k} \|\boldsymbol{\Delta}_t\| \end{aligned}$$

and

$$\begin{aligned}\|\boldsymbol{\Delta}_t\| &\geq \|\boldsymbol{X}'\| - \|\boldsymbol{\Delta}_t - \boldsymbol{X}'\| \\ &\geq \|\boldsymbol{X}'\| - \epsilon \\ &\geq \|\boldsymbol{\Delta}_t\| - 2\epsilon \\ &\geq \|\boldsymbol{\Delta}_t\|/2,\end{aligned}$$

where in the last inequality we used the assumption $\|\boldsymbol{\Delta}_t\| \geq 4\epsilon$. Combining the above inequalities implies that

$$\frac{\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}}{\|\boldsymbol{\Delta}_t\|} \leq \frac{2(1+\sqrt{k})\epsilon}{\|\boldsymbol{\Delta}_t\|} + 2\sqrt{k} \leq \frac{1+5\sqrt{k}}{2}.$$

Combining the above inequality with equation 2.71 leads to

$$\eta_t = \frac{\eta\rho^t}{\|\boldsymbol{Q}_t\|} \geq \left(1 - \left(\frac{1+5\sqrt{k}}{2}\right)\delta\right) \frac{\eta\rho^t}{\varphi(\boldsymbol{\Delta}_t)} \frac{\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}}{\|\boldsymbol{\Delta}_t\|}.$$

The upper bound can be derived in a similar fashion. □

2.C Proofs of the Expected Loss

2.C.1 Proof of Proposition 3

According to the update rule $\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta\left(\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{X}^\star\right)\boldsymbol{U}_t$, we have

$$\boldsymbol{S}_{t+1} = \boldsymbol{V}^\top\boldsymbol{U}_{t+1} = \boldsymbol{S}_t - \eta\left(\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right)\boldsymbol{S}_t + \boldsymbol{S}_t\boldsymbol{E}_t^\top\boldsymbol{E}_t\right).$$

Define an auxiliary matrix $\boldsymbol{M}$ as

$$M := (\boldsymbol{I} - \boldsymbol{\Xi})\left(\boldsymbol{S}_t - \eta\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right)\boldsymbol{S}_t\right)\left(\boldsymbol{S}_t^\top - \eta\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right)\right)\left(\boldsymbol{I} - \boldsymbol{\Xi}^\top\right), \tag{2.72}$$

where $\boldsymbol{\Xi} = \eta\boldsymbol{S}_t\boldsymbol{E}_t^\top\boldsymbol{E}_t\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\left(\boldsymbol{I} - \eta\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right)\right)^{-1}$. Note that, due to our assumptions and the choice of η, we have $\boldsymbol{S}_t\boldsymbol{S}_t^\top \succ \boldsymbol{0}$ and $\eta\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right\| < 1$. Therefore, both matrices $\boldsymbol{S}_t\boldsymbol{S}_t^\top$ and $\boldsymbol{I} - \eta\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right)$ are invertible and the matrix $\boldsymbol{\Xi}$ is well-defined. Our goal is to first show that $\lambda_{\min}(\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top) \geq \lambda_{\min}(\boldsymbol{M})$, and then derive a lower bound for $\lambda_{\min}(\boldsymbol{M})$. Based on the definition of $\boldsymbol{\Xi}$, it is easy to verify that $\boldsymbol{\Xi}\left(\boldsymbol{S}_t - \eta\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right)\boldsymbol{S}_t\right) =$

$\boldsymbol{S}_t\boldsymbol{E}_t^\top\boldsymbol{E}_t\mathcal{P}_{\boldsymbol{S}_t}$. Combining this with equation 2.72 reveals that

$$\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top - \boldsymbol{M} = \eta^2\boldsymbol{S}_t\boldsymbol{E}_t^\top\boldsymbol{E}_t\mathcal{P}_{\boldsymbol{S}_t}^\perp\boldsymbol{E}_t^\top\boldsymbol{E}_t\boldsymbol{S}_t^\top \succeq \boldsymbol{0}.$$

Therefore, instead of obtaining a lower bound for $\lambda_{\min}(\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top)$, it suffices to obtain a lower bound for $\lambda_{\min}(\boldsymbol{M})$. One can write

$$\begin{aligned}
\|\boldsymbol{\Xi}\| &\leq \eta\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\boldsymbol{E}_t\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\right\|\left\|\left(\boldsymbol{I}-\eta\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top-\boldsymbol{\Sigma}\right)\right)^{-1}\right\| \\
&\lesssim \eta\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\boldsymbol{E}_t\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\right\| \\
&= \eta\left\|\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1/2}\boldsymbol{S}_t\boldsymbol{E}_t^\top\boldsymbol{E}_t\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1/2}\right\| \\
&\leq \eta\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|\left\|\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1/2}\boldsymbol{S}_t\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1/2}\right\| \\
&= \eta\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| \\
&< 1,
\end{aligned} \tag{2.73}$$

where the last inequality is due to the assumption $\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| \leq \sigma_1$ and the choice of η. Therefore, we have

$$\begin{aligned}
\lambda_{\min}\left(\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right) &\geq \lambda_{\min}\left(\boldsymbol{M}\right) \\
&\geq \left(1-\|\boldsymbol{\Xi}\|\right)^2\lambda_{\min}\left(\left(\boldsymbol{S}_t-\eta\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top-\boldsymbol{\Sigma}\right)\boldsymbol{S}_t\right)\left(\boldsymbol{S}_t^\top-\eta\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top-\boldsymbol{\Sigma}\right)\right)\right).
\end{aligned} \tag{2.74}$$

Now it suffices to bound $\lambda_{\min}\left(\left(\boldsymbol{S}_t-\eta\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top-\boldsymbol{\Sigma}\right)\boldsymbol{S}_t\right)\left(\boldsymbol{S}_t^\top-\eta\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top-\boldsymbol{\Sigma}\right)\right)\right)$. First, note that

$$\boldsymbol{S}_t-\eta\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top-\boldsymbol{\Sigma}\right)\boldsymbol{S}_t = \left(\boldsymbol{I}+\eta\boldsymbol{\Sigma}\left(\boldsymbol{I}-\eta\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\right)\left(\boldsymbol{S}_t-\eta\boldsymbol{S}_t\boldsymbol{S}_t^\top\boldsymbol{S}_t\right).$$

Based on the above equality, one can write

$$
\begin{aligned}
&\lambda_{\min}\left(\left(\boldsymbol{S}_t-\eta\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top-\boldsymbol{\Sigma}\right)\boldsymbol{S}_t\right)\left(\boldsymbol{S}_t^\top-\eta\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top-\boldsymbol{\Sigma}\right)\right)\right)\\
&=\lambda_{\min}\left(\left(\boldsymbol{I}+\eta\boldsymbol{\Sigma}\left(\boldsymbol{I}-\eta\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\right)\left(\boldsymbol{S}_t-\eta\boldsymbol{S}_t\boldsymbol{S}_t^\top\boldsymbol{S}_t\right)\right.\\
&\qquad\left.\left(\boldsymbol{S}_t^\top-\eta\boldsymbol{S}_t^\top\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\left(\boldsymbol{I}+\eta\boldsymbol{\Sigma}\left(\boldsymbol{I}-\eta\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\right)\right)\\
&\geq\lambda_{\min}\left(\left(\boldsymbol{I}+\eta\boldsymbol{\Sigma}\left(\boldsymbol{I}-\eta\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\right)\right)^2\lambda_{\min}\left(\left(\boldsymbol{S}_t-\eta\boldsymbol{S}_t\boldsymbol{S}_t^\top\boldsymbol{S}_t\right)\left(\boldsymbol{S}_t^\top-\eta\boldsymbol{S}_t^\top\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right)\\
&\overset{(a)}{\geq}\left(1+\eta\sigma_r\left(1-\eta\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right)^{-1}\right)^2\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top-2\eta\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^2+\eta^2\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^3\right)\qquad(2.75)\\
&\overset{(b)}{=}\left(1+\eta\sigma_r\left(1-\eta\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right)^{-1}\right)^2\\
&\qquad\times\left(\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)-2\eta\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^2+\eta^2\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^3\right)\\
&\geq\left(1+\eta\sigma_r\left(1-\eta\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right)^{-1}\right)^2\left(\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)-2\eta\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^2\right)\\
&=\left(1+\eta\sigma_r\right)^2\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)-2\eta\left(1+\eta\sigma_r\right)\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^2,
\end{aligned}
$$

where in (a), we used the fact that

$$
\begin{aligned}
\lambda_{\min}\left(\boldsymbol{I}+\eta\boldsymbol{\Sigma}\left(\boldsymbol{I}-\eta\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\right)&=1+\lambda_{\min}\left(\eta\boldsymbol{\Sigma}\left(\boldsymbol{I}-\eta\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\right)\\
&=1+\eta\lambda_{\min}\left(\boldsymbol{\Sigma}^{1/2}\left(\boldsymbol{I}-\eta\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\boldsymbol{\Sigma}^{1/2}\right)\\
&\geq1+\eta\sigma_r\lambda_{\min}\left(\left(\boldsymbol{I}-\eta\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\right)\\
&=1+\eta\sigma_r\left(1-\eta\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right)^{-1},
\end{aligned}
$$

and in (b) we used the fact that the matrices $\boldsymbol{S}_t\boldsymbol{S}_t^\top$, $(\boldsymbol{S}_t\boldsymbol{S}_t^\top)^2$, and $(\boldsymbol{S}_t\boldsymbol{S}_t^\top)^3$ share the same eigenvectors. Combining equation 2.75 with equation 2.73 and equation 2.74 completes the proof. □

2.C.2 Proof of Proposition 4

Before delving into the details, we provide the update rule for $\boldsymbol{S}_t$ and $\boldsymbol{E}_t$, which will be used frequently throughout our proof. Applying the signal-residual decomposition to $\boldsymbol{U}_{t+1}=\boldsymbol{U}_t-\eta\left(\boldsymbol{U}_t\boldsymbol{U}_t^\top-\boldsymbol{X}^\star\right)\boldsymbol{U}_t$ leads to

$$
\boldsymbol{S}_{t+1}=\boldsymbol{V}^\top\boldsymbol{U}_{t+1}=\boldsymbol{S}_t-\eta\left(\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top-\boldsymbol{\Sigma}\right)\boldsymbol{S}_t+\boldsymbol{S}_t\boldsymbol{E}_t^\top\boldsymbol{E}_t\right),\qquad(2.76)
$$

$$
\boldsymbol{E}_{t+1}=\boldsymbol{V}_\perp^\top\boldsymbol{U}_{t+1}=\boldsymbol{E}_t-\eta\boldsymbol{E}_t\left(\boldsymbol{S}_t^\top\boldsymbol{S}_t+\boldsymbol{E}_t^\top\boldsymbol{E}_t\right).\qquad(2.77)
$$

Bounding $\left\|\boldsymbol{\Sigma}-\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^{\top}\right\|$: The update rule for $\boldsymbol{S}_{t+1}$ leads to

$$\begin{aligned}
&\boldsymbol{\Sigma}-\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^{\top}\\
&=\underbrace{\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^{\top}+\eta\boldsymbol{S}_t\boldsymbol{S}_t^{\top}\left(\boldsymbol{S}_t\boldsymbol{S}_t^{\top}-\boldsymbol{\Sigma}\right)+\eta\left(\boldsymbol{S}_t\boldsymbol{S}_t^{\top}-\boldsymbol{\Sigma}\right)\boldsymbol{S}_t\boldsymbol{S}_t^{\top}-\eta^2\left(\boldsymbol{S}_t\boldsymbol{S}_t^{\top}-\boldsymbol{\Sigma}\right)\boldsymbol{S}_t\boldsymbol{S}_t^{\top}\left(\boldsymbol{S}_t\boldsymbol{S}_t^{\top}-\boldsymbol{\Sigma}\right)}_{(A)}\\
&\underbrace{+2\eta\boldsymbol{S}_t\boldsymbol{E}_t^{\top}\boldsymbol{E}_t\boldsymbol{S}_t^{\top}-\eta^2\left(\boldsymbol{S}_t\boldsymbol{S}_t^{\top}-\boldsymbol{\Sigma}\right)\boldsymbol{S}_t\boldsymbol{E}_t^{\top}\boldsymbol{E}_t\boldsymbol{S}_t^{\top}-\eta^2\boldsymbol{S}_t\boldsymbol{E}_t^{\top}\boldsymbol{E}_t\boldsymbol{S}_t^{\top}\left(\boldsymbol{S}_t\boldsymbol{S}_t^{\top}-\boldsymbol{\Sigma}\right)}_{(B)}\\
&-\underbrace{\eta^2\boldsymbol{S}_t\boldsymbol{E}_t^{\top}\boldsymbol{E}_t\boldsymbol{E}_t^{\top}\boldsymbol{E}_t\boldsymbol{S}_t^{\top}}_{(C)}.
\end{aligned}\tag{2.78}$$

First, we provide an upper bound for (B). Recall that $\eta=\mathcal{O}\left(1/\sigma_1\right)$. Moreover, we have $\left\|\boldsymbol{S}_t\boldsymbol{S}_t^{\top}-\boldsymbol{\Sigma}\right\|\leq 2.01\sigma_1$ due to the assumption $\left\|\boldsymbol{S}_t\boldsymbol{S}_t^{\top}\right\|\leq 1.01\sigma_1$. Therefore, one can write

$$\|(B)\|\leq 2\eta\left\|\boldsymbol{S}_t\boldsymbol{E}_t^{\top}\boldsymbol{E}_t\boldsymbol{S}_t^{\top}\right\|+2\eta^2\left\|\boldsymbol{S}_t\boldsymbol{S}_t^{\top}-\boldsymbol{\Sigma}\right\|\left\|\boldsymbol{S}_t\boldsymbol{E}_t^{\top}\boldsymbol{E}_t\boldsymbol{S}_t^{\top}\right\|\leq 4\eta\left\|\boldsymbol{S}_t\boldsymbol{E}_t^{\top}\right\|^2.$$

Similarly, due to our assumption on $\left\|\boldsymbol{E}_t\boldsymbol{E}_t^{\top}\right\|$ and η, we have $\left\|\boldsymbol{E}_t\boldsymbol{E}_t^{\top}\right\|\leq 1/\eta$, which in turn implies

$$\|(C)\|\leq\eta^2\left\|\boldsymbol{E}_t\boldsymbol{E}_t^{\top}\right\|\left\|\boldsymbol{S}_t\boldsymbol{E}_t^{\top}\right\|^2\leq\eta\left\|\boldsymbol{S}_t\boldsymbol{E}_t^{\top}\right\|^2.$$

Finally, we provide an upper bound for (A). First, one can verify that

$$(A)=\underbrace{\left(\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^{\top}\right)\left(0.5\boldsymbol{I}-\eta\boldsymbol{S}_t\boldsymbol{S}_t^{\top}\right)}_{(A_1)}+\underbrace{\left(0.5\boldsymbol{I}-\eta\boldsymbol{S}_t\boldsymbol{S}_t^{\top}+\eta^2\left(\boldsymbol{S}_t\boldsymbol{S}_t^{\top}-\boldsymbol{\Sigma}\right)\boldsymbol{S}_t\boldsymbol{S}_t^{\top}\right)\left(\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^{\top}\right)}_{(A_2)}.$$

For the first term, we have

$$\|(A_1)\|\leq\left\|0.5\boldsymbol{I}-\eta\boldsymbol{S}_t\boldsymbol{S}_t^{\top}\right\|\left\|\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^{\top}\right\|\leq\left(0.5-\eta\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^{\top}\right)\right)\left\|\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^{\top}\right\|.$$

To provide a bound for (A_2), observe that

$$0.5\boldsymbol{I}-\eta\boldsymbol{S}_t\boldsymbol{S}_t^{\top}+\eta^2\left(\boldsymbol{S}_t\boldsymbol{S}_t^{\top}-\boldsymbol{\Sigma}\right)\boldsymbol{S}_t\boldsymbol{S}_t^{\top}=0.5\boldsymbol{I}-\eta\left(\boldsymbol{I}+\eta\left(\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^{\top}\right)\right)\boldsymbol{S}_t\boldsymbol{S}_t^{\top}.$$

The next step in our proof is to show that, with the choice of $\eta\lesssim 1/\sigma_1$, the eigenvalues of $\left(\boldsymbol{I}+\eta\left(\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^{\top}\right)\right)\boldsymbol{S}_t\boldsymbol{S}_t^{\top}$ are nonnegative. We prove this by showing that the eigenvalues of $\left(\boldsymbol{I}+\eta\left(\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^{\top}\right)\right)\boldsymbol{S}_t\boldsymbol{S}_t^{\top}$ are close to those of $\boldsymbol{S}_t\boldsymbol{S}_t^{\top}$. First, note that $\boldsymbol{I}+\eta\left(\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^{\top}\right)$ is positive definite due to our choice of step-size. Therefore, the matrix $\boldsymbol{D}=\left(\boldsymbol{I}+\eta\left(\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^{\top}\right)\right)^{1/2}$ is well-defined and invertable. Hence, for any $1\leq i\leq r$,

we have

$$\lambda_i\left(\left(\boldsymbol{I}+\eta\left(\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right)\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)=\lambda_i\left(\boldsymbol{D}^{-1}\left(\boldsymbol{I}+\eta\left(\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right)\boldsymbol{S}_t\boldsymbol{S}_t^\top\boldsymbol{D}\right)=\lambda_i\left(\boldsymbol{D}\boldsymbol{S}_t\boldsymbol{S}_t^\top\boldsymbol{D}\right).$$

To proceed with our argument, we first present a relative perturbation bound for symmetric matrices.

Lemma 21 (Relative Perturbation Bound, Eisenstat and Ipsen [1995])**.** *Let $\boldsymbol{X}\in\mathbb{R}^{d\times d}$ be a symmetric matrix with eigenvalues $\lambda_1\geq\cdots\geq\lambda_d$. Moreover, suppose $\boldsymbol{R}$ is a non-singular matrix. Let $\hat{\lambda}_1\geq\cdots\geq\hat{\lambda}_d$ be the eigenvalues of $\boldsymbol{Y}=\boldsymbol{R}^\top\boldsymbol{X}\boldsymbol{R}$. Then, we have*

$$\left|\lambda_i-\hat{\lambda}_i\right|\leq|\lambda_i|\left\|\boldsymbol{I}-\boldsymbol{R}^\top\boldsymbol{R}\right\|,\quad \textit{for all } i.$$

Invoking Lemma 21 with $\boldsymbol{X}=\boldsymbol{S}_t\boldsymbol{S}_t$ and $\boldsymbol{Y}=D\boldsymbol{S}_t\boldsymbol{S}_tD$ results in

$$\begin{aligned}
&\left|\lambda_i\left(\left(\boldsymbol{I}+\eta\left(\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right)\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)-\lambda_i\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right|\\
&\leq\left|\lambda_i(\boldsymbol{S}_t\boldsymbol{S}_t^\top)\right|\left\|\boldsymbol{I}-\boldsymbol{D}^2\right\|\\
&=\left|\lambda_i(\boldsymbol{S}_t\boldsymbol{S}_t^\top)\right|\left\|\boldsymbol{I}-\left(\boldsymbol{I}+\eta\left(\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right)\right\|\\
&=\eta\left|\lambda_i(\boldsymbol{S}_t\boldsymbol{S}_t^\top)\right|\left\|\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|\\
&\leq 0.5\left|\lambda_i(\boldsymbol{S}_t\boldsymbol{S}_t^\top)\right|.
\end{aligned}$$

for every $i=1,2,\ldots,r$. The above inequality implies that $\lambda_{\min}\left(\left(\boldsymbol{I}+\eta\left(\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right)\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\asymp\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)$, and therefore, $\lambda_{\min}\left(\left(\boldsymbol{I}+\eta\left(\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right)\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\geq 0$. This leads to

$$\begin{aligned}
&\left\|0.5\boldsymbol{I}-\eta\boldsymbol{S}_t\boldsymbol{S}_t^\top+\eta^2\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top-\boldsymbol{\Sigma}\right)\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|\\
&=\left\|0.5\boldsymbol{I}-\eta\boldsymbol{S}_t\boldsymbol{S}_t^\top+\eta^2\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{1/2}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top-\boldsymbol{\Sigma}\right)\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{1/2}\right\|\\
&\leq 0.5-\lambda_{\min}\left(\eta\boldsymbol{S}_t\boldsymbol{S}_t^\top-\eta^2\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{1/2}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top-\boldsymbol{\Sigma}\right)\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{1/2}\right)\\
&=0.5-\lambda_{\min}\left(\eta\boldsymbol{S}_t\boldsymbol{S}_t^\top-\eta^2\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top-\boldsymbol{\Sigma}\right)\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right)\\
&=0.5-\eta\lambda_{\min}\left(\left(\boldsymbol{I}+\eta\left(\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right)\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\leq 0.5,
\end{aligned}$$

where in the last inequality, we used the fact that $\lambda_{\min}\left(\left(\boldsymbol{I}+\eta\left(\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right)\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\geq 0$. Combined with the definition of (A_2), we have

$$\|(A_2)\|\leq\left\|0.5\boldsymbol{I}-\eta\boldsymbol{S}_t\boldsymbol{S}_t^\top+\eta^2\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top-\boldsymbol{\Sigma}\right)\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|\left\|\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|\leq 0.5\left\|\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|,$$

which in turn implies

$$\|(A)\| \leq \left(1 - \eta\lambda_{\min}(\boldsymbol{S}_t\boldsymbol{S}_t^\top)\right) \left\|\boldsymbol{\Sigma} - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|.$$

Combining the derived upper bounds for (A), (B), and (C) completes the proof for the signal dynamics. □

Bounding $\left\|\boldsymbol{S}_{t+1}\boldsymbol{E}_{t+1}^\top\right\|$: Recalling the update rules equation 2.76 and equation 2.77, we have

$$\begin{aligned}\boldsymbol{S}_{t+1}\boldsymbol{E}_{t+1}^\top = &\underbrace{\boldsymbol{S}_t\boldsymbol{E}_t^\top + \eta(\boldsymbol{\Sigma} - \boldsymbol{S}_t\boldsymbol{S}_t^\top)\boldsymbol{S}_t\boldsymbol{E}_t^\top - \eta\boldsymbol{S}_t(\boldsymbol{S}_t^\top\boldsymbol{S}_t + \boldsymbol{E}_t^\top\boldsymbol{E}_t)\boldsymbol{E}_t^\top}_{(A)} \\ &+ \underbrace{\eta^2\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right)\boldsymbol{S}_t(\boldsymbol{S}_t^\top\boldsymbol{S}_t + \boldsymbol{E}_t^\top\boldsymbol{E}_t)\boldsymbol{E}_t^\top}_{(B)} - \underbrace{\eta\boldsymbol{S}_t\boldsymbol{E}_t^\top\boldsymbol{E}_t\boldsymbol{E}_t^\top}_{(C)} \\ &+ \underbrace{\eta^2\boldsymbol{S}_t\boldsymbol{E}_t^\top\boldsymbol{E}_t\left(\boldsymbol{S}_t^\top\boldsymbol{S}_t + \boldsymbol{E}_t^\top\boldsymbol{E}_t\right)\boldsymbol{E}_t^\top}_{(D)}.\end{aligned}$$

We provide separate bounds for the individual terms in the above equality. First, observe that $\|(C)\| \leq \eta\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\|$. Moreover, one can write

$$\begin{aligned}\|(B)\| &\leq \eta^2\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|\left\|\boldsymbol{S}_t(\boldsymbol{S}_t^\top\boldsymbol{S}_t + \boldsymbol{E}_t^\top\boldsymbol{E}_t)\boldsymbol{E}_t^\top\right\| \\ &\leq \eta^2\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|\left(\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| + \left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|\right)\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\| \\ &\leq \eta\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\|,\end{aligned}$$

where, in the last inequality, we used the assumption that $\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| \lesssim \sigma_1$ and $\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| \lesssim \sigma_1$. Similarly, we have

$$\begin{aligned}\|(D)\| &\overset{(a)}{\leq} \eta^2\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\|\left(\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| + \left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|\right)\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| \\ &\leq \eta\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\|,\end{aligned}$$

where, in (a), we used $\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|\boldsymbol{E}_t\boldsymbol{E}_t^\top \succeq \boldsymbol{E}_t\boldsymbol{S}_t^\top\boldsymbol{S}_t\boldsymbol{E}_t^\top \succeq \boldsymbol{0}$. Finally, we provide an upper bound for (A).

$$\begin{aligned}\|(A)\| &\leq \left(\eta\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| + \left\|\boldsymbol{I} - \eta\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|\right)\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\| \\ &\leq \left(\eta\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| + 1 - \eta\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right)\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\|.\end{aligned}$$

Combining the derived bounds for (A), (B), (C), and (D) concludes the proof. □

Bounding $\left\|\boldsymbol{E}_{t+1}\boldsymbol{E}_{t+1}^\top\right\|$: Due to the update rule for $\boldsymbol{E}_{t+1}$, one can write

$$\begin{aligned}\boldsymbol{E}_{t+1}\boldsymbol{E}_{t+1}^\top &= \boldsymbol{E}_t\boldsymbol{E}_t^\top - 2\eta\boldsymbol{E}_t\left(\boldsymbol{E}_t^\top\boldsymbol{E}_t+\boldsymbol{S}_t^\top\boldsymbol{S}_t\right)\boldsymbol{E}_t^\top + \eta^2\boldsymbol{E}_t\left(\boldsymbol{E}_t^\top\boldsymbol{E}_t+\boldsymbol{S}_t^\top\boldsymbol{S}_t\right)^2\boldsymbol{E}_t^\top \\ &= \boldsymbol{E}_t\left(\boldsymbol{I}-2\eta\left(\boldsymbol{E}_t^\top\boldsymbol{E}_t+\boldsymbol{S}_t^\top\boldsymbol{S}_t\right)+\eta^2\left(\boldsymbol{E}_t^\top\boldsymbol{E}_t+\boldsymbol{S}_t^\top\boldsymbol{S}_t\right)^2\right)\boldsymbol{E}_t^\top.\end{aligned} \tag{2.79}$$

On the other hand, $\eta \lesssim 1/\sigma_1$, $\|\boldsymbol{E}_t\boldsymbol{E}_t\| \lesssim \sigma_1$, and $\|\boldsymbol{S}_t\boldsymbol{S}_t\| \lesssim \sigma_1$ imply that

$$I-\eta\boldsymbol{E}_t^\top\boldsymbol{E}_t \succ I-2\eta\left(\boldsymbol{E}_t^\top\boldsymbol{E}_t+\boldsymbol{S}_t^\top\boldsymbol{S}_t\right)+\eta^2\left(\boldsymbol{E}_t^\top\boldsymbol{E}_t+\boldsymbol{S}_t^\top\boldsymbol{S}_t\right)^2 \succ \boldsymbol{0}. \tag{2.80}$$

Hence, we have

$$\begin{aligned}\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| &= \left\|\boldsymbol{E}_t\left(\boldsymbol{I}-2\eta\left(\boldsymbol{E}_t^\top\boldsymbol{E}_t+\boldsymbol{S}_t^\top\boldsymbol{S}_t\right)+\eta^2\left(\boldsymbol{E}_t^\top\boldsymbol{E}_t+\boldsymbol{S}_t^\top\boldsymbol{S}_t\right)^2\right)\boldsymbol{E}_t^\top\right\| \\ &\le \left\|\boldsymbol{E}_t\left(\boldsymbol{I}-\eta\boldsymbol{E}_t^\top\boldsymbol{E}_t\right)\boldsymbol{E}_t^\top\right\| \\ &= \left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top-\eta\boldsymbol{E}_t\boldsymbol{E}_t^\top\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| \\ &= \left(1-\eta\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|\right)\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|,\end{aligned}$$

where in the last inequality, we used the fact that $\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| \le \sigma_1$.

Bounding $\left\|\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right\|$: First, recall that

$$\begin{aligned}\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top &= \boldsymbol{S}_t\boldsymbol{S}_t^\top - 2\eta\boldsymbol{S}_t\boldsymbol{S}_t^\top\boldsymbol{S}_t\boldsymbol{S}_t^\top + \eta\boldsymbol{\Sigma}\boldsymbol{S}_t\boldsymbol{S}_t^\top + \eta\boldsymbol{S}_t\boldsymbol{S}_t^\top\boldsymbol{\Sigma} - 2\eta\boldsymbol{S}_t\boldsymbol{E}_t\boldsymbol{E}_t^\top\boldsymbol{S}_t^\top \\ &\quad + \eta^2\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top-\boldsymbol{\Sigma}\right)\boldsymbol{S}_t\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top-\boldsymbol{\Sigma}\right) + \eta^2\boldsymbol{S}_t\boldsymbol{E}_t^\top\boldsymbol{E}_t\boldsymbol{E}_t^\top\boldsymbol{E}_t\boldsymbol{S}_t^\top \\ &\quad + \eta^2\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top-\boldsymbol{\Sigma}\right)\boldsymbol{S}_t\boldsymbol{E}_t\boldsymbol{E}_t^\top\boldsymbol{S}_t^\top + \eta^2\boldsymbol{S}_t\boldsymbol{E}_t\boldsymbol{E}_t^\top\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top-\boldsymbol{\Sigma}\right) \\ &\preceq \underbrace{\boldsymbol{S}_t\boldsymbol{S}_t^\top - 2\eta\boldsymbol{S}_t\boldsymbol{S}_t^\top\boldsymbol{S}_t\boldsymbol{S}_t^\top + \eta\boldsymbol{\Sigma}\boldsymbol{S}_t\boldsymbol{S}_t^\top + \eta\boldsymbol{S}_t\boldsymbol{S}_t^\top\boldsymbol{\Sigma} + \eta^2\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top-\boldsymbol{\Sigma}\right)\boldsymbol{S}_t\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top-\boldsymbol{\Sigma}\right)}_{(A)} \\ &\quad + \underbrace{\eta^2\boldsymbol{S}_t\boldsymbol{E}_t^\top\boldsymbol{E}_t\boldsymbol{E}_t^\top\boldsymbol{E}_t\boldsymbol{S}_t^\top + \eta^2\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top-\boldsymbol{\Sigma}\right)\boldsymbol{S}_t\boldsymbol{E}_t\boldsymbol{E}_t^\top\boldsymbol{S}_t^\top + \eta^2\boldsymbol{S}_t\boldsymbol{E}_t\boldsymbol{E}_t^\top\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top-\boldsymbol{\Sigma}\right)}_{(B)},\end{aligned}$$

where the last inequality follows by noting that $-2\eta\boldsymbol{S}_t\boldsymbol{E}_t\boldsymbol{E}_t^\top\boldsymbol{S}_t^\top \preceq 0$. Now, it is easy to see that

$$\begin{aligned}\|(B)\| &\le \eta^2\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|^2\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| + 2\eta^2\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|\left(\sigma_1+\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|\right) \\ &\le 3\eta^2\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|\left(\sigma_1+\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|\right).\end{aligned}$$

It remains to provide an upper bound for (A). One can write

$$(A) = \underbrace{\boldsymbol{S}_t\boldsymbol{S}_t^\top - 2\eta\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^2 + \eta^2\left(\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^3 + \boldsymbol{\Sigma}\boldsymbol{S}_t\boldsymbol{S}_t^\top\boldsymbol{\Sigma}\right)}_{(A_1)} + \underbrace{\eta\boldsymbol{\Sigma}\boldsymbol{S}_t\boldsymbol{S}_t^\top\left(\boldsymbol{I} - \eta\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\eta\boldsymbol{S}_t\boldsymbol{S}_t^\top\left(\boldsymbol{I} - \eta\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\boldsymbol{\Sigma}}_{(A_2)}.$$

For (A_1), we have

$$\begin{aligned}\|(A_1)\| &\le \left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top - 2\eta\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^2 + \eta^2\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^3\right\| + \eta^2\left\|\boldsymbol{\Sigma}\boldsymbol{S}_t\boldsymbol{S}_t^\top\boldsymbol{\Sigma}\right\| \\ &\le \left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top - 2\eta\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^2 + \eta^2\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^3\right\| + \eta^2\sigma_1^2\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| \\ &\overset{(a)}{\le} \left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| - 2\eta\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|^2 + \eta^2\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|^3 + \eta^2\sigma_1^2\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|,\end{aligned}$$

where (a) follows from the fact that $\boldsymbol{S}_t\boldsymbol{S}_t^\top$, $\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^2$, $\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^3$ share the same eigenvectors, and the assumption $\eta \lesssim 1/\sigma_1$. On the other hand, one can easily verify that

$$\|(A_2)\| \le 2\eta\sigma_1\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\left(\boldsymbol{I} - \eta\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right\| \le 2\eta\sigma_1\left(\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| - \eta\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|^2\right).$$

Combining the upper bounds for (A) and (B) leads to

$$\begin{aligned}\left\|\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right\| &\le \left((1+\eta\sigma_1)^2 + 3\eta^2\sigma_1\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|\right)\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| \\ &\quad - 2\eta\left(1 + \eta\sigma_1 - 1.5\eta\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|\right)\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|^2 + \eta^2\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|^3 \\ &\overset{(a)}{\le} \left((1+2.001\sigma_1)\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| - 2\eta\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|^2 + \eta^2\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|^3,\end{aligned}$$

where, in (a), we used the assumption $\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| \lesssim \sigma_1$ and $\eta \lesssim 1/\sigma_1$. Now, let us define the function

$$f(x) := \eta^2x^3 - 2\eta x^2 + (1+2.001\sigma_1)\,x.$$

It is easy to see that $f(x)$ is increasing within the interval $x \le 1/4\eta$. On the other hand, we have $1.01\sigma_1 \le 1/4\eta$ due to our choice of η. Therefore, we have

$$\left\|\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right\| \le f\left(\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|\right) \le f\left(1.01\sigma_1\right).$$

On the other hand, a simple calculation reveals that

$$f(1.01\sigma_1) = \eta^2(1.01\sigma_1)^3 - 2\eta(1.01\sigma_1)^2 + (1+2.001\sigma_1)\,(1.01\sigma_1) \le 1.01\sigma_1.$$

for $\eta \leq c/\sigma_1$ with sufficiently small constant c. This completes the proof. $\square$

2.C.3 Proof of Lemma 5

We prove this lemma by induction on t. First, due to our choice of the initial point in Theorem 6, we have $\left\|\boldsymbol{E}_0\boldsymbol{E}_0^\top\right\| \leq \alpha^2$, $\left\|\boldsymbol{S}_0\boldsymbol{S}_0^\top\right\| \leq 1.01\sigma_1$ and $\boldsymbol{S}_0\boldsymbol{S}_0^\top \succ \boldsymbol{0}$. Now, suppose that equation 2.31, equation 2.32, and equation 2.33 hold for t. Then, it is easy to see that

$$\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| \leq \left\|\boldsymbol{E}_0\boldsymbol{E}_0^\top\right\| \leq \alpha^2,$$

due to the decreasing nature of $\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|$. Therefore, Proposition 4 can be invoked to show that $\left\|\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right\| \leq 1.01\sigma_1$. Moreover, we have

$$\begin{aligned}
\left\|\boldsymbol{E}_{t+1}\boldsymbol{E}_{t+1}^\top\right\| &\leq \left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| - \eta\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|^2 \\
&\leq \frac{1}{\eta t + \frac{1}{\alpha^2}} - \frac{\eta}{\left(\eta t + \frac{1}{\alpha^2}\right)^2} \\
&\leq \frac{1}{\eta(t+1) + \frac{1}{\alpha^2}}.
\end{aligned}$$

On the other hand, Proposition 3 can be used to show that

$$\begin{aligned}
\lambda_{\min}\left(\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right) &\geq \left(1+\eta^2\sigma_r^2\right)\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right) - 2\eta\left(1+\eta\sigma_r\right)\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^2 \\
&\geq \left(1 - 2\eta\sigma_1\left(1+\eta\sigma_r\right) + \eta^2\sigma_r^2\right)\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right) \\
&> 0,
\end{aligned}$$

where in the first inequality, we used $\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| \lesssim \sigma_r$, which follows from $\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| \leq \alpha^2$ and our choice of α. Moreover, the last inequality follows from our choice of η. $\square$

2.D Proofs of Empirical Loss with Noiseless Measurements

2.D.1 Preliminaries

For simplicity, we define $\boldsymbol{\Delta}_t = \boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{X}^\star$. Before presenting the proofs for the empirical loss, we first introduce a key decomposition. Recall the update rule

$$\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - 2\eta_t\boldsymbol{Q}_t\boldsymbol{U}_t := \tilde{\boldsymbol{U}}_{t+1} + \boldsymbol{R}_t\boldsymbol{U}_t, \tag{2.81}$$

where $\eta_t = \frac{\eta}{2}\frac{1}{m}\sum_{i=1}^{m} |\langle \boldsymbol{A}_i, \boldsymbol{\Delta}_t\rangle|$ and $\boldsymbol{Q}_t \in \mathcal{Q}(\boldsymbol{\Delta}_t)$. Moreover, denote $\tilde{\boldsymbol{U}}_{t+1} = \boldsymbol{U}_t - \bar{\eta}_t \boldsymbol{\Delta}_t \boldsymbol{U}_t$, where $\bar{\eta}_t = \eta\varphi(\boldsymbol{\Delta}_t)^2$ is the update rule obtained from the expected loss, and $\boldsymbol{R}_t\boldsymbol{U}_t$ with $\boldsymbol{R}_t\bar{\eta}_t\boldsymbol{\Delta}_t - 2\eta_t\boldsymbol{Q}_t$ is the residual caused by the deviation of the empirical loss from its expectation. Finally, we define $\tilde{\boldsymbol{S}}_t = \boldsymbol{V}^\top\tilde{\boldsymbol{U}}_t$, and $\tilde{\boldsymbol{E}}_t = \boldsymbol{V}_\perp^\top\tilde{\boldsymbol{U}}_t$.

Lemma 22. *Suppose that the measurements satisfy Sign-RIP with parameters* $\left(4r, \delta, \sqrt{d}\,\|\boldsymbol{G}_t\|^2\right)$. *Then, we have* $\|\boldsymbol{R}_t\| \le 3\bar{\eta}_t\delta\,\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}$.

Proof. One can write

$$\boldsymbol{\Delta}_t = \underbrace{\boldsymbol{V}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right)\boldsymbol{V}^\top + \boldsymbol{V}\boldsymbol{S}_t\boldsymbol{E}_t^\top\boldsymbol{V}_\perp^\top + \boldsymbol{V}_\perp\boldsymbol{E}_t\boldsymbol{S}_t^\top\boldsymbol{V}^\top + \boldsymbol{V}_\perp\boldsymbol{G}_t\boldsymbol{G}_t^\top\boldsymbol{V}_\perp^\top}_{\text{rank-}4r} + \underbrace{\boldsymbol{V}_\perp\boldsymbol{G}_t\boldsymbol{G}_t^\top\boldsymbol{V}_\perp^\top}_{\text{small perturbation}}.$$

Note that $\left\|\boldsymbol{V}_\perp\boldsymbol{G}_t\boldsymbol{G}_t^\top\boldsymbol{V}_\perp^\top\right\|_{\mathrm{F}} \le \sqrt{d}\,\|\boldsymbol{G}_t\|^2$. Therefore, $\boldsymbol{\Delta}_t$ is an $\sqrt{d}\,\|\boldsymbol{G}_t\|^2$-approximate rank-$4r$ matrix. One can write

$$\begin{aligned}
\boldsymbol{R}_t =& \eta\varphi(\boldsymbol{\Delta}_t)^2\boldsymbol{\Delta}_t - 2\eta_t\boldsymbol{Q}_t \\
=& \left(\frac{\eta}{2}\varphi(\Delta)\,\|\boldsymbol{\Delta}_t\|_{\mathrm{F}} - \eta_t\right)Q_t + \left(\frac{\eta}{2}\varphi(\Delta)\,\|\boldsymbol{\Delta}_t\|_{\mathrm{F}} - \eta_t\right)Q_t^\top \\
&+ \frac{\eta}{2}\varphi(\boldsymbol{\Delta}_t)\,\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}\left(\varphi(\boldsymbol{\Delta}_t)\frac{\boldsymbol{\Delta}_t}{\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}} - \boldsymbol{Q}_t\right) \\
&+ \frac{\eta}{2}\varphi(\boldsymbol{\Delta}_t)\,\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}\left(\varphi(\boldsymbol{\Delta}_t)\frac{\boldsymbol{\Delta}_t}{\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}} - \boldsymbol{Q}_t^\top\right).
\end{aligned}$$

Due to the above decomposition, one can write

$$\|\boldsymbol{R}_t\| \le 2\left|\frac{\eta}{2}\varphi(\boldsymbol{\Delta}_t)\,\|\boldsymbol{\Delta}_t\|_{\mathrm{F}} - \eta_t\right|\|\boldsymbol{Q}_t\| + \eta\varphi(\boldsymbol{\Delta}_t)\,\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}\left\|\varphi(\boldsymbol{X})\frac{\boldsymbol{\Delta}_t}{\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}} - \boldsymbol{Q}_t\right\|. \tag{2.82}$$

First, note that $|(\eta/2)\varphi(\boldsymbol{\Delta}_t)\,\|\boldsymbol{\Delta}_t\|_{\mathrm{F}} - \eta_t| \le \delta(\eta/2)\varphi(\boldsymbol{\Delta}_t)\,\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}$ due to Lemma 20. Moreover, due to Sign-RIP, we have $\|\boldsymbol{Q}_t\| \le (1+\delta)\varphi(\boldsymbol{\Delta}_t)$ and $\left\|\varphi(\boldsymbol{X})\frac{\boldsymbol{\Delta}_t}{\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}} - \boldsymbol{Q}_t\right\| \le \delta\varphi(\boldsymbol{\Delta}_t)$. Combining these upper bounds with equation 2.82 completes the proof. □

2.D.2 Proof of Proposition 5

Due to the proposed decomposition equation 2.81, one can write

$$\boldsymbol{S}_{t+1} = \tilde{\boldsymbol{S}}_{t+1} + \boldsymbol{V}^\top\boldsymbol{R}_t\boldsymbol{U}_t. \tag{2.83}$$

Our main goal is to show that the minimum eigenvalue of $\boldsymbol{S}_t\boldsymbol{S}_t^\top$ follows that of $\tilde{\boldsymbol{S}}_t\tilde{\boldsymbol{S}}_t^\top$—which has been characterized in Lemma 3—plus an additional deviation caused by the

term $\boldsymbol{V}^\top \boldsymbol{R}_t \boldsymbol{U}_t$. Similar to the proof of Proposition 3, we characterize the growth rate of $\lambda_{\min}(\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top)$ by first resorting to a more tractable lower bound. Adopting the notation introduced in Appendix 2.D, consider the following auxiliary matrix

$$\begin{aligned} M &:= (\boldsymbol{I}+\boldsymbol{\Xi})\,\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\,(\boldsymbol{I}+\boldsymbol{\Xi})^\top \\ &= (\boldsymbol{I}+\boldsymbol{\Xi})\,\boldsymbol{V}^\top(\boldsymbol{I}-\bar{\eta}_t\boldsymbol{\Delta}_t)\,\boldsymbol{U}_t\boldsymbol{U}_t{}^\top\,(\boldsymbol{I}-\bar{\eta}_t\boldsymbol{\Delta}_t)\,(\boldsymbol{I}+\boldsymbol{\Xi})^\top, \end{aligned}$$

where $\boldsymbol{\Xi} := \boldsymbol{V}^\top\boldsymbol{R}_t\boldsymbol{U}_t\tilde{\boldsymbol{S}}_{t+1}^\top\left(\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\right)^{-1}$. Note that, according to Lemma 5, we have $\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top \succ \boldsymbol{0}$ due to our assumption $\boldsymbol{S}_t\boldsymbol{S}_t^\top \succ \boldsymbol{0}$. Therefore, $\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top$ is invertible, and hence, the matrix $\boldsymbol{\Xi}$ is well-defined. On the other hand, it is easy to see that $\boldsymbol{\Xi}$ satisfies

$$\boldsymbol{V}^\top\boldsymbol{R}_t\boldsymbol{U}_t\tilde{\boldsymbol{S}}_{t+1}^\top = \boldsymbol{\Xi}\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top.$$

Based on the above equality, one can write

$$\begin{aligned} \boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top - M &= \left(\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top + \boldsymbol{V}^\top\boldsymbol{R}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{R}_t^\top\boldsymbol{V} + \tilde{\boldsymbol{S}}_{t+1}\boldsymbol{U}_t^\top\boldsymbol{R}_t^\top\boldsymbol{V} + \boldsymbol{V}^\top\boldsymbol{R}_t\boldsymbol{U}_t\tilde{\boldsymbol{S}}_{t+1}^\top\right) \\ &\quad - \left(\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top - \boldsymbol{\Xi}\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\boldsymbol{\Xi}^\top + \boldsymbol{\Xi}\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top + \tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\boldsymbol{\Xi}^\top\right) \\ &= \boldsymbol{V}^\top\boldsymbol{R}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{R}_t^\top\boldsymbol{V} - \boldsymbol{V}^\top\boldsymbol{R}_t\boldsymbol{U}_t\tilde{\boldsymbol{S}}_{t+1}^\top\left(\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\right)^{-1}\tilde{\boldsymbol{S}}_{t+1}\boldsymbol{U}_t^\top\boldsymbol{R}_t^\top\boldsymbol{V} \\ &= \boldsymbol{V}^\top\boldsymbol{R}_t\boldsymbol{U}_t\left(\boldsymbol{I} - \tilde{\boldsymbol{S}}_{t+1}^\top\left(\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\right)^{-1}\tilde{\boldsymbol{S}}_{t+1}\right)\boldsymbol{U}_t^\top\boldsymbol{R}_t^\top\boldsymbol{V} \\ &= \boldsymbol{V}^\top\boldsymbol{R}_t\boldsymbol{U}_t\mathcal{P}_{\tilde{\boldsymbol{S}}_{+1}}^\perp\boldsymbol{U}_t{}^\top\boldsymbol{R}_t^\top\boldsymbol{V} \\ &\succeq \boldsymbol{0}. \end{aligned} \tag{2.84}$$

Therefore, we have $\lambda_{\min}(\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top) \geq \lambda_{\min}(\boldsymbol{M})$. Our next goal is to provide a lower bound for $\lambda_{\min}(\boldsymbol{M})$. To this goal, we will show that the minimum eigenvalue of $\boldsymbol{M}$ is close to that of $\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top$. Combined with the minimum eigenvalue dynamics of $\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top$ presented in Proposition 3, this completes the proof.

To show that $\lambda_{\min}(\boldsymbol{M}) \approx \lambda_{\min}(\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top)$, we will use the relative perturbation bound presented in Lemma 21. To this goal, first we need to provide an upper bound for $\|\boldsymbol{\Xi}\|$.

Bounding $\|\Xi\|$. One can write

$$\begin{aligned}
\|\Xi\| &\leq \|\boldsymbol{R}_t\| \left\| \boldsymbol{U}_t \tilde{\boldsymbol{S}}_{t+1}^\top \left(\tilde{\boldsymbol{S}}_{t+1} \tilde{\boldsymbol{S}}_{t+1}^\top \right)^{-1} \right\| \\
&= \|\boldsymbol{R}_t\| \left\| \left(\boldsymbol{I} - \bar{\eta}_t \left(\boldsymbol{U}_t \boldsymbol{U}_t^\top - \boldsymbol{X}^\star \right) \right)^{-1} \tilde{\boldsymbol{U}}_{t+1} \tilde{\boldsymbol{S}}_{t+1}^\top \left(\tilde{\boldsymbol{S}}_{t+1} \tilde{\boldsymbol{S}}_{t+1}^\top \right)^{-1} \right\| \\
&\leq \|\boldsymbol{R}_t\| \left\| \left(\boldsymbol{I} - \bar{\eta}_t \left(\boldsymbol{U}_t \boldsymbol{U}_t^\top - \boldsymbol{X}^\star \right) \right)^{-1} \right\| \left\| \tilde{\boldsymbol{U}}_{t+1} \tilde{\boldsymbol{S}}_{t+1}^\top \left(\tilde{\boldsymbol{S}}_{t+1} \tilde{\boldsymbol{S}}_{t+1}^\top \right)^{-1} \right\| \qquad (2.85) \\
&\overset{(a)}{\leq} 2 \|\boldsymbol{R}_t\| \left\| \tilde{\boldsymbol{U}}_{t+1} \tilde{\boldsymbol{S}}_{t+1}^\top \left(\tilde{\boldsymbol{S}}_{t+1} \tilde{\boldsymbol{S}}_{t+1}^\top \right)^{-1} \right\| \\
&\overset{(b)}{\leq} 6 \bar{\eta}_t \delta \|\boldsymbol{\Delta}_t\|_{\mathrm{F}} \left\| \tilde{\boldsymbol{U}}_{t+1} \tilde{\boldsymbol{S}}_{t+1}^\top \left(\tilde{\boldsymbol{S}}_{t+1} \tilde{\boldsymbol{S}}_{t+1}^\top \right)^{-1} \right\|,
\end{aligned}$$

where in (a), we used the fact that $\left\| \boldsymbol{U}_t \boldsymbol{U}_t^\top - \boldsymbol{X}^\star \right\| \lesssim \sigma_1$ due to our assumptions on $\left\| \boldsymbol{S}_t \boldsymbol{S}_t^\top \right\|$ and $\left\| \boldsymbol{E}_t \boldsymbol{E}_t^\top \right\|$, and our choice of η. Moreover, (b) follows from Lemma 22. Now, note that

$$\begin{aligned}
\tilde{\boldsymbol{U}}_{t+1} \tilde{\boldsymbol{S}}_{t+1}^\top \left(\tilde{\boldsymbol{S}}_{t+1} \tilde{\boldsymbol{S}}_{t+1}^\top \right)^{-1} &= \boldsymbol{V} \tilde{\boldsymbol{S}}_{t+1} \tilde{\boldsymbol{S}}_{t+1}^\top \left(\tilde{\boldsymbol{S}}_{t+1} \tilde{\boldsymbol{S}}_{t+1}^\top \right)^{-1} + \boldsymbol{V}_\perp \tilde{\boldsymbol{E}}_{t+1} \tilde{\boldsymbol{S}}_{t+1}^\top \left(\tilde{\boldsymbol{S}}_{t+1} \tilde{\boldsymbol{S}}_{t+1}^\top \right)^{-1} \\
&= \boldsymbol{V} + \boldsymbol{V}_\perp \tilde{\boldsymbol{E}}_{t+1} \tilde{\boldsymbol{S}}_{t+1}^\top \left(\tilde{\boldsymbol{S}}_{t+1} \tilde{\boldsymbol{S}}_{t+1}^\top \right)^{-1}.
\end{aligned}$$

Therefore, we have

$$\left\| \tilde{\boldsymbol{U}}_{t+1} \tilde{\boldsymbol{S}}_{t+1}^\top \left(\tilde{\boldsymbol{S}}_{t+1} \tilde{\boldsymbol{S}}_{t+1}^\top \right)^{-1} \right\| \leq 1 + \left\| \tilde{\boldsymbol{E}}_{t+1} \tilde{\boldsymbol{S}}_{t+1}^\top \left(\tilde{\boldsymbol{S}}_{t+1} \tilde{\boldsymbol{S}}_{t+1}^\top \right)^{-1} \right\|.$$

In order to provide an upper bound for $\|\Xi\|$, it suffices to bound $\left\| \tilde{\boldsymbol{E}}_{t+1} \tilde{\boldsymbol{S}}_{t+1}^\top \left(\tilde{\boldsymbol{S}}_{t+1} \tilde{\boldsymbol{S}}_{t+1}^\top \right)^{-1} \right\|$. To this goal, first we present the following technical lemma, the proof of which can be found in Appendix 2.G.2.

Lemma 23. *The following statements hold:*

- $\frac{1}{2} \leq \left\| \boldsymbol{S}_t \boldsymbol{S}_t^\top \left(\tilde{\boldsymbol{S}}_{t+1} \tilde{\boldsymbol{S}}_{t+1}^\top \right)^{-1} \right\| \leq 2$ *and* $\frac{1}{3} \leq \left\| \boldsymbol{S}_t \boldsymbol{S}_t^\top \left(\boldsymbol{S}_{t+1} \boldsymbol{S}_{t+1}^\top \right)^{-1} \right\| \leq 3$*;*
- $\left\| \tilde{\boldsymbol{E}}_{t+1} \tilde{\boldsymbol{S}}_{t+1}^\top \left(\tilde{\boldsymbol{S}}_{t+1} \tilde{\boldsymbol{S}}_{t+1}^\top \right)^{-1} \right\| \leq 3 \left\| \boldsymbol{E}_t \boldsymbol{S}_t^\top \left(\boldsymbol{S}_t \boldsymbol{S}_t^\top \right)^{-1} \right\|$.

The second statement of the above lemma connects $\left\| \tilde{\boldsymbol{E}}_{t+1} \tilde{\boldsymbol{S}}_{t+1}^\top \left(\tilde{\boldsymbol{S}}_{t+1} \tilde{\boldsymbol{S}}_{t+1}^\top \right)^{-1} \right\|$ to $\left\| \boldsymbol{E}_t \boldsymbol{S}_t^\top \left(\boldsymbol{S}_t \boldsymbol{S}_t^\top \right)^{-1} \right\|$ which, according to our assumption, is upper bounded by $1/3$. This

upper bound, together with equation 2.85 implies that

$$\|\boldsymbol{\Xi}\| \leq 12\bar{\eta}_t\delta \left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}.$$

On the other hand, applying the relative perturbation bound to $\boldsymbol{M}$ and $\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top$ implies that

$$\begin{aligned}\left|\lambda_{\min}(\boldsymbol{M}) - \lambda_{\min}(\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top)\right| \leq& \lambda_{\min}(\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top)\left\|\boldsymbol{I} - (\boldsymbol{I}+\boldsymbol{\Xi})(\boldsymbol{I}+\boldsymbol{\Xi})^\top\right\| \\ \leq& 3\left\|\boldsymbol{\Xi}\right\|\lambda_{\min}\left(\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\right) \\ \leq& 36\bar{\eta}_t\delta\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}\lambda_{\min}\left(\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\right).\end{aligned}$$

This in turn implies that

$$\begin{aligned}\lambda_{\min}(\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top) \geq& \lambda_{\min}(\boldsymbol{M}) \\ \geq& (1-36\bar{\eta}_t\delta\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}})\lambda_{\min}\left(\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\right) \\ \geq& \left((1+\bar{\eta}_t\sigma_r)^2 - 2\bar{\eta}_t\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|\right)\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right) \\ &- 72\bar{\eta}_t\delta\left\|\boldsymbol{U}_t\boldsymbol{U}_t{}^\top - \boldsymbol{X}^\star\right\|_{\mathrm{F}}\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right) \\ &- 2\bar{\eta}_t\left(1+\bar{\eta}_t\sigma_r\right)\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^2,\end{aligned}$$

which completes the proof. □

2.D.3 Proof of Proposition 6

Bounding $\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right\|$: Recall that $\boldsymbol{S}_{t+1} = \tilde{\boldsymbol{S}}_{t+1} + \boldsymbol{V}^\top\boldsymbol{R}_t\boldsymbol{U}_t$. Therefore, we have

$$\boldsymbol{\Sigma} - \boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top = \boldsymbol{\Sigma} - \tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top \underbrace{-\boldsymbol{V}^\top\boldsymbol{R}_t\boldsymbol{U}_t\tilde{\boldsymbol{S}}_{t+1}^\top - \tilde{\boldsymbol{S}}_{t+1}\boldsymbol{U}_t^\top\boldsymbol{R}_t^\top\boldsymbol{V} - \boldsymbol{V}^\top\boldsymbol{R}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{R}_t^\top\boldsymbol{V}}_{\text{deviation}}. \tag{2.86}$$

The main idea behind our proof is to first control the norm of the deviation term, and then invoke Proposition 4 for $\boldsymbol{\Sigma} - \tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top$. One can write

$$\left\|\boldsymbol{V}^\top\boldsymbol{R}_t\boldsymbol{U}_t\tilde{\boldsymbol{S}}_{t+1}^\top\right\| \leq \left\|\boldsymbol{R}_t\right\|\left\|\tilde{\boldsymbol{S}}_{t+1}\boldsymbol{U}_t^\top\right\| \leq 3\bar{\eta}_t\delta\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}\left\|\tilde{\boldsymbol{S}}_{t+1}\boldsymbol{U}_t^\top\right\|,$$

where in the last inequality, we used Lemma 22. Moreover, $\tilde{\boldsymbol{S}}_{t+1}\boldsymbol{U}_t^\top$ can be rewritten as

$$\begin{aligned}\tilde{\boldsymbol{S}}_{t+1}\boldsymbol{U}_t^\top &= \left(\left(\boldsymbol{I} - \bar{\eta}_t\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right)\boldsymbol{S}_t + \boldsymbol{S}_t\boldsymbol{E}_t^\top\boldsymbol{E}_t\right)\right)\left(\boldsymbol{S}^\top\boldsymbol{V}^\top + \boldsymbol{E}^\top\boldsymbol{V}_\perp^\top\right)\\ &= \left(\boldsymbol{I} - \bar{\eta}_t\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right)\right)\boldsymbol{S}_t\boldsymbol{S}_t^\top\boldsymbol{V}^\top + \left(\boldsymbol{I} - \bar{\eta}_t\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right)\right)\boldsymbol{S}_t\boldsymbol{E}_t^\top\boldsymbol{V}_\perp^\top\\ &\quad + \boldsymbol{S}_t\boldsymbol{E}_t^\top\boldsymbol{E}_t\boldsymbol{S}_t^\top\boldsymbol{V}^\top + \boldsymbol{S}_t\boldsymbol{E}_t^\top\boldsymbol{E}_t\boldsymbol{E}_t^\top\boldsymbol{V}_\perp^\top.\end{aligned}$$

Note that $\left\|\boldsymbol{I} - \bar{\eta}_t\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right)\right\| \le 2$ due to our choice of η and our assumption on $\boldsymbol{S}_t\boldsymbol{S}_t^\top$. Therefore, we have

$$\left\|\tilde{\boldsymbol{S}}_{t+1}\boldsymbol{U}_t^\top\right\| \le 2\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| + 2\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\| + \left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\|^2 + \left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\|\|\boldsymbol{E}_t\|^2 \le 6\sigma_1, \tag{2.87}$$

where we used our assumptions on $\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|$ and $\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|$. Combining the above two inequalities leads to

$$\left\|\tilde{\boldsymbol{S}}_{t+1}\boldsymbol{U}_t^\top\boldsymbol{R}_t^\top V\right\| = \left\|\boldsymbol{V}^\top\boldsymbol{R}_t\boldsymbol{U}_t\tilde{\boldsymbol{S}}_{t+1}^\top\right\| \le 18\bar{\eta}_t\delta\sigma_1\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}. \tag{2.88}$$

Using a similar technique, we have

$$\left\|\boldsymbol{V}^\top\boldsymbol{R}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{R}_t^\top\boldsymbol{V}\right\| \le 9\bar{\eta}_t^2\delta^2\sigma_1\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}^2 \le \bar{\eta}_t\delta\sigma_1\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}, \tag{2.89}$$

where we used the assumed upper bounds on η and δ, and the fact that $\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| \le \sigma_1$, $\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| \le 1.01\sigma_1$, and $\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|_{\mathrm{F}} \le \sqrt{r}\sigma_1$, which in turn implies $\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}} \lesssim \sqrt{r}\sigma_1$ due to Lemma 3. Moreover, we have already shown in Proposition 4 that

$$\left\|\boldsymbol{\Sigma} - \tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\right\| \le \left(1 - \bar{\eta}_t\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right)\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| + 5\bar{\eta}_t\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\|^2.$$

Hence, combining the above inequalities leads to

$$\begin{aligned}\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right\| &\le \left(1 - \bar{\eta}_t\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right)\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|\\ &\quad + 5\bar{\eta}_t\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\|^2 + 37\bar{\eta}_t\delta\sigma_1\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}.\end{aligned}$$

Bounding $\left\|\boldsymbol{S}_{t+1}\boldsymbol{E}_{t+1}^\top\right\|$: First, note that

$$\boldsymbol{S}_{t+1}\boldsymbol{E}_{t+1}^\top = \tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{E}}_{t+1}^\top + \underbrace{\boldsymbol{V}^\top\boldsymbol{R}_t\boldsymbol{U}_t\tilde{\boldsymbol{E}}_{t+1}^\top + \tilde{\boldsymbol{S}}_{t+1}\boldsymbol{U}_t^\top\boldsymbol{R}_t^\top\boldsymbol{V}_\perp + \boldsymbol{V}^\top\boldsymbol{R}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{R}_t^\top\boldsymbol{V}_\perp}_{\text{deviation}}.$$

Similar to the signal term, the main idea behind our proof is to first control the norm of the deviation term, and then invoke Proposition 4 for $\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{E}}_{t+1}^\top$. We first provide an upper

bound for $\left\|\boldsymbol{V}^\top \boldsymbol{R}_t \boldsymbol{U}_t \tilde{\boldsymbol{E}}_{t+1}^\top\right\|$

$$\left\|\boldsymbol{V}^\top \boldsymbol{R}_t \boldsymbol{U}_t \tilde{\boldsymbol{E}}_{t+1}^\top\right\| \leq \|\boldsymbol{R}_t\| \left\|\boldsymbol{U}_t \tilde{\boldsymbol{E}}_{t+1}^\top\right\| \leq 3\bar{\eta}_t \delta \left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}} \left\|\boldsymbol{U}_t \tilde{\boldsymbol{E}}_{t+1}^\top\right\|.$$

To bound $\left\|\boldsymbol{U}_t \tilde{\boldsymbol{E}}_{t+1}^\top\right\|$, one can write

$$\begin{aligned}\left\|\boldsymbol{U}_t \tilde{\boldsymbol{E}}_{t+1}^\top\right\| &\leq \|\boldsymbol{U}_t\| \left\|\boldsymbol{E}_t - \bar{\eta}_t \boldsymbol{S}_t^\top \boldsymbol{S}_t \boldsymbol{E}_t^\top - \bar{\eta}_t \boldsymbol{E}_t \boldsymbol{E}_t^\top \boldsymbol{E}_t\right\| \\ &\leq 2\left(\|\boldsymbol{S}_t\| + \|\boldsymbol{E}_t\|\right) \|\boldsymbol{E}_t\| \\ &\leq 5\sigma_1,\end{aligned}$$

where we used the assumption $\|\boldsymbol{S}_t\| \leq 1.01\sqrt{\sigma_1}$ and $\|\boldsymbol{E}_t\| \leq \sqrt{\sigma_1}$. Similar to equation 2.89, we have

$$\begin{aligned}\left\|\tilde{\boldsymbol{S}}_{t+1} \boldsymbol{U}_t^\top \boldsymbol{R}_t^\top \boldsymbol{V}_\perp\right\| &\leq \left\|\tilde{\boldsymbol{S}}_{t+1} \boldsymbol{U}_t^\top\right\| \|\boldsymbol{R}_t\| \\ &\leq \left\|\boldsymbol{S}_t - \bar{\eta}_t \left(\left(\boldsymbol{S}_t \boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right) \boldsymbol{S}_t + \boldsymbol{S}_t \boldsymbol{E}_t^\top \boldsymbol{E}_t\right)\right\| \|\boldsymbol{R}_t\| \\ &\leq 6\bar{\eta}_t \delta \sqrt{\sigma_1} \left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}},\end{aligned}$$

and

$$\left\|\boldsymbol{V}^\top \boldsymbol{R}_t \boldsymbol{U}_t \boldsymbol{U}_t^\top \boldsymbol{R}_t^\top \boldsymbol{V}_\perp\right\| \leq \left\|\boldsymbol{U}_t \boldsymbol{U}_t^\top\right\| \|\boldsymbol{R}_t\|^2 \leq 10\sigma_1 \bar{\eta}_t^2 \delta^2 \left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}^2 \leq \bar{\eta}_t \delta \sigma_1 \left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}.$$

Moreover, we have already shown in Proposition 4 that

$$\left\|\tilde{\boldsymbol{S}}_{t+1} \tilde{\boldsymbol{E}}_{t+1}^\top\right\| \leq \left(1 - \bar{\eta}_t \lambda_{\min}\left(\boldsymbol{S}_t \boldsymbol{S}_t^\top\right) + 2\bar{\eta}_t \left\|\boldsymbol{\Sigma} - \boldsymbol{S}_t \boldsymbol{S}_t^\top\right\| + 2\bar{\eta}_t \|\boldsymbol{E}_t \boldsymbol{E}_t\|\right) \left\|\boldsymbol{S}_t \boldsymbol{E}_t^\top\right\|. \tag{2.90}$$

Combining the above inequalities leads to

$$\begin{aligned}\left\|\boldsymbol{S}_{t+1} \boldsymbol{E}_{t+1}^\top\right\| \leq &\left(1 - \bar{\eta}_t \lambda_{\min}\left(\boldsymbol{S}_t \boldsymbol{S}_t^\top\right) + 2\bar{\eta}_t \left\|\boldsymbol{\Sigma} - \boldsymbol{S}_t \boldsymbol{S}_t^\top\right\| + 2\bar{\eta}_t \|\boldsymbol{E}_t \boldsymbol{E}_t\|\right) \left\|\boldsymbol{S}_t \boldsymbol{E}_t^\top\right\| \\ &+ 22\bar{\eta}_t \delta \sigma_1 \left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}.\end{aligned}$$

Bounding $\left\|\boldsymbol{S}_{t+1} \boldsymbol{S}_{t+1}^\top\right\|$: First, note that

$$\boldsymbol{S}_{t+1} \boldsymbol{S}_{t+1}^\top = \tilde{\boldsymbol{S}}_{t+1} \tilde{\boldsymbol{S}}_{t+1}^\top + \underbrace{\boldsymbol{V}^\top \boldsymbol{R}_t \boldsymbol{U}_t \boldsymbol{S}_{t+1}^\top + \boldsymbol{S}_{t+1} \boldsymbol{U}_t^\top \boldsymbol{R}_t^\top \boldsymbol{V} + \boldsymbol{V}^\top \boldsymbol{R}_t \boldsymbol{U}_t \boldsymbol{U}_t^\top \boldsymbol{R}_t^\top \boldsymbol{V}}_{\text{deviation}}.$$

Similar to our previous arguments, we will provide an upper bound on the deviation term, and then resort to Proposition 4 to provide an upper bound for $\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top$. First, note that

$$\begin{aligned}\left\|\boldsymbol{V}^\top \boldsymbol{R}_t \boldsymbol{U}_t \boldsymbol{S}_{t+1}^\top\right\| &\le \left\|\boldsymbol{V}^\top \boldsymbol{R}_t \boldsymbol{V} \boldsymbol{S}_t \boldsymbol{S}_{t+1}^\top\right\| + \left\|\boldsymbol{V}^\top \boldsymbol{R}_t \boldsymbol{V}_\perp \boldsymbol{E}_t \boldsymbol{S}_{t+1}^\top\right\| \\ &\le 3\bar{\eta}_t\delta \left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}} \left(\left\|\boldsymbol{S}_t \boldsymbol{S}_{t+1}^\top\right\| + \left\|\boldsymbol{E}_t \boldsymbol{S}_{t+1}^\top\right\|\right), \end{aligned} \tag{2.91}$$

where we used Lemma 22 in the last inequality. On the other hand

$$\begin{aligned}\left\|\boldsymbol{S}_t \boldsymbol{S}_{t+1}^\top\right\| &= \left\|\boldsymbol{S}_t \left(\boldsymbol{S}_t^\top - \bar{\eta}_t \left(\boldsymbol{S}_t^\top \left(\boldsymbol{S}_t \boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right) + \boldsymbol{E}_t^\top \boldsymbol{E}_t \boldsymbol{S}_t^\top\right) + \boldsymbol{U}_t^\top \boldsymbol{R}_t^\top \boldsymbol{V}\right)\right\| \\ &\le \left\|\boldsymbol{S}_t \boldsymbol{S}_t^\top\right\| \left(1 + \bar{\eta}_t \left\|\boldsymbol{S}_t \boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right\| + \bar{\eta}_t \left\|\boldsymbol{E}_t \boldsymbol{E}_t^\top\right\|\right) + \left\|\boldsymbol{S}_t \boldsymbol{U}_t^\top \boldsymbol{R}_t^\top \boldsymbol{V}\right\| \\ &\le 2\left\|\boldsymbol{S}_t \boldsymbol{S}_t^\top\right\| + \left\|\boldsymbol{S}_t \boldsymbol{S}_t^\top \boldsymbol{V}^\top \boldsymbol{R}_t^\top \boldsymbol{V}\right\| + \left\|\boldsymbol{S}_t \boldsymbol{E}_t^\top \boldsymbol{V}_\perp^\top \boldsymbol{R}_t^\top \boldsymbol{V}\right\| \\ &\le 3\left\|\boldsymbol{S}_t \boldsymbol{S}_t^\top\right\| + 3\bar{\eta}_t\delta \left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}} \left\|\boldsymbol{S}_t \boldsymbol{E}_t^\top\right\|. \end{aligned}$$

Similarly, we have

$$\left\|\boldsymbol{E}_t \boldsymbol{S}_{t+1}^\top\right\| \le 3\left\|\boldsymbol{E}_t \boldsymbol{S}_t^\top\right\| + 3\bar{\eta}_t\delta \left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}} \left\|\boldsymbol{E}_t \boldsymbol{E}_t^\top\right\|.$$

The above two inequalities combined with equation 2.91 results in

$$2\left\|\boldsymbol{V}^\top \boldsymbol{R}_t \boldsymbol{U}_t \boldsymbol{S}_{t+1}^\top\right\| \le 216\bar{\eta}_t\sigma_1^2\sqrt{r}\delta.$$

which follows from our assumption on $\|\boldsymbol{S}_t\|$, $\|\boldsymbol{E}_t\|$ and δ. Similarly, one can show that

$$\left\|\boldsymbol{V}^\top \boldsymbol{R}_t \boldsymbol{U}_t \boldsymbol{U}_t^\top \boldsymbol{R}_t^\top \boldsymbol{V}\right\| \le \bar{\eta}_t\sigma_1^2\sqrt{r}\delta.$$

Therefore, the norm of the deviation term is upper bounded by $217\bar{\eta}_t\sigma_1^2\sqrt{r}\delta$. Moreover, Proposition 4 implies

$$\left\|\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\right\| \le \left(1 + 2\sigma_1\bar{\eta}_t + 4\sigma_1^2\bar{\eta}_t^2\right)\left\|\boldsymbol{S}_t \boldsymbol{S}_t^\top\right\| - 2\bar{\eta}_t\left\|\boldsymbol{S}_t \boldsymbol{S}_t^\top\right\|^2 + \bar{\eta}_t^2\left\|\boldsymbol{S}_t \boldsymbol{S}_t^\top\right\|^3,$$

which in turn leads to

$$\left\|\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right\| \le \left(1 + 2\sigma_1\bar{\eta}_t + 4\sigma_1^2\bar{\eta}_t^2\right)\left\|\boldsymbol{S}_t \boldsymbol{S}_t^\top\right\| - 2\bar{\eta}_t\left\|\boldsymbol{S}_t \boldsymbol{S}_t^\top\right\|^2 + \bar{\eta}_t^2\left\|\boldsymbol{S}_t \boldsymbol{S}_t^\top\right\|^3 + 217\bar{\eta}_t\sigma_1^2\sqrt{r}\delta.$$

The rest of the proof is analogous to that of Proposition 4, and hence, omitted for brevity.□

2.D.4 Proof of Proposition 7

Bounding $\|\boldsymbol{G}_{t+1}\|$: To provide an upper bound for $\|\boldsymbol{G}_{t+1}\| = \left\|\boldsymbol{E}_{t+1}\mathcal{P}^{\perp}_{\boldsymbol{S}_{t+1}}\right\|$ in terms of $\|\boldsymbol{G}_t\| = \left\|\boldsymbol{E}_t\mathcal{P}^{\perp}_{\boldsymbol{S}_t}\right\|$, it is crucial to characterize the relationship between the projection operators $\mathcal{P}^{\perp}_{\boldsymbol{S}_{t+1}}$ and $\mathcal{P}^{\perp}_{\boldsymbol{S}_t}$. To this goal, we decompose $\mathcal{P}^{\perp}_{\boldsymbol{S}_{t+1}}$ as follows

$$\mathcal{P}^{\perp}_{\boldsymbol{S}_{t+1}} = \mathcal{P}_{\boldsymbol{S}_t}\mathcal{P}^{\perp}_{\boldsymbol{S}_{t+1}} + \mathcal{P}^{\perp}_{\boldsymbol{S}_t}\mathcal{P}^{\perp}_{\boldsymbol{S}_{t+1}}.$$

Based on the above decomposition, $\boldsymbol{G}_{t+1}$ can be written as

$$\boldsymbol{G}_{t+1} = \underbrace{\boldsymbol{E}_{t+1}\mathcal{P}_{\boldsymbol{S}_t}\mathcal{P}^{\perp}_{\boldsymbol{S}_{t+1}}}_{(A)} + \underbrace{\boldsymbol{E}_{t+1}\mathcal{P}^{\perp}_{\boldsymbol{S}_t}\mathcal{P}^{\perp}_{\boldsymbol{S}_{t+1}}}_{(B)}.$$

We first study (A). Let $\boldsymbol{M}_t\boldsymbol{D}_t\boldsymbol{N}_t^{\top}$ be the singular value decomposition of $\boldsymbol{S}_t$, where $\boldsymbol{M}_t \in \mathbb{R}^{r\times r}$ and $\boldsymbol{N}_t \in \mathbb{R}^{r'\times r}$ are (row) orthonormal matrices, and $\boldsymbol{D}_t \in \mathbb{R}^{r\times r}$ is a diagonal matrix collecting the singular values of $\boldsymbol{S}_t$. Based on this definition, we have $\mathcal{P}_{\boldsymbol{S}_t} = \boldsymbol{N}_t\boldsymbol{N}_t^{\top}$. On the other hand, we have $\boldsymbol{S}_{t+1}\mathcal{P}_{\boldsymbol{S}_t}\mathcal{P}^{\perp}_{\boldsymbol{S}_{t+1}} = -\boldsymbol{S}_{t+1}\mathcal{P}^{\perp}_{\boldsymbol{S}_t}\mathcal{P}^{\perp}_{\boldsymbol{S}_{t+1}}$, which is equivalent to

$$\boldsymbol{S}_{t+1}\boldsymbol{N}_t\boldsymbol{N}_t^{\top}\mathcal{P}^{\perp}_{\boldsymbol{S}_{t+1}} = -\boldsymbol{S}_{t+1}\mathcal{P}^{\perp}_{\boldsymbol{S}_t}\mathcal{P}^{\perp}_{\boldsymbol{S}_{t+1}}. \tag{2.92}$$

Our next technical lemma shows that $\boldsymbol{S}_{t+1}\boldsymbol{N}_t$ is invertible.

Lemma 24. *The matrix $\boldsymbol{S}_{t+1}\boldsymbol{N}_t$ is invertible.*

The proof of this lemma can be found in Appendix 2.G.3. Lemma equation 24 combined with equation 2.92 leads to

$$\boldsymbol{N}_t^{\top}\mathcal{P}^{\perp}_{\boldsymbol{S}_{t+1}} = -\left(\boldsymbol{S}_{t+1}\boldsymbol{N}_t\right)^{-1}\boldsymbol{S}_{t+1}\mathcal{P}^{\perp}_{\boldsymbol{S}_t}\mathcal{P}^{\perp}_{\boldsymbol{S}_{t+1}}.$$

Therefore, we have

$$\begin{aligned}
(A) &= \boldsymbol{E}_{t+1}\mathcal{P}_{\boldsymbol{S}_t}\mathcal{P}^{\perp}_{\boldsymbol{S}_{t+1}} \\
&= \boldsymbol{E}_{t+1}\boldsymbol{N}_t\boldsymbol{N}_t^{\top}\mathcal{P}^{\perp}_{\boldsymbol{S}_{t+1}} \\
&= -\boldsymbol{E}_{t+1}\boldsymbol{N}_t\left(\boldsymbol{S}_{t+1}\boldsymbol{N}_t\right)^{-1}\boldsymbol{S}_{t+1}\mathcal{P}^{\perp}_{\boldsymbol{S}_t}\mathcal{P}^{\perp}_{\boldsymbol{S}_{t+1}} \\
&= \underbrace{-\boldsymbol{E}_{t+1}\boldsymbol{N}_t\left(\boldsymbol{S}_{t+1}\boldsymbol{N}_t\right)^{-1}\tilde{\boldsymbol{S}}_{t+1}\mathcal{P}^{\perp}_{\boldsymbol{S}_t}\mathcal{P}^{\perp}_{\boldsymbol{S}_{t+1}}}_{(A_1)} \\
&\quad \underbrace{-\boldsymbol{E}_{t+1}\boldsymbol{N}_t\left(\boldsymbol{S}_{t+1}\boldsymbol{N}_t\right)^{-1}\left(\boldsymbol{S}_{t+1}-\tilde{\boldsymbol{S}}_{t+1}\right)\mathcal{P}^{\perp}_{\boldsymbol{S}_t}\mathcal{P}^{\perp}_{\boldsymbol{S}_{t+1}}}_{(A_2)}.
\end{aligned}$$

We first control (A_1). Observe that

$$\tilde{\boldsymbol{S}}_{t+1}\mathcal{P}_{\boldsymbol{S}_t}^{\perp} = \left(\boldsymbol{S}_t - \bar{\eta}_t\left(\boldsymbol{S}_t\boldsymbol{S}_t^{\top} - \boldsymbol{\Sigma}\right)\boldsymbol{S}_t - \bar{\eta}_t\boldsymbol{S}_t\boldsymbol{E}_t^{\top}\boldsymbol{E}_t\right)\mathcal{P}_{\boldsymbol{S}_t}^{\perp} = -\bar{\eta}_t\boldsymbol{S}_t\boldsymbol{E}_t^{\top}\boldsymbol{G}_t.$$

Therefore, we have

$$(A_1) = \bar{\eta}_t\boldsymbol{E}_{t+1}\boldsymbol{N}_t\left(\boldsymbol{S}_{t+1}\boldsymbol{N}_t\right)^{-1}\boldsymbol{S}_t\boldsymbol{E}_t^{\top}\boldsymbol{G}_t\mathcal{P}_{\boldsymbol{S}_{t+1}}^{\perp}. \tag{2.93}$$

On the other hand, note that $\boldsymbol{E}_{t+1} = \boldsymbol{E}_t - \bar{\eta}_t\boldsymbol{E}_t\left(\boldsymbol{S}_t^{\top}\boldsymbol{S}_t + \boldsymbol{E}_t^{\top}\boldsymbol{E}_t\right) + \boldsymbol{V}_{\perp}^{\top}\boldsymbol{R}_t\boldsymbol{U}_t$. Hence, we have

$$(B) = \left(\boldsymbol{I} - \bar{\eta}_t\boldsymbol{E}_t\boldsymbol{E}_t^{\top}\right)\boldsymbol{G}_t\mathcal{P}_{\boldsymbol{S}_{t+1}}^{\perp} + \boldsymbol{V}_{\perp}^{\top}\boldsymbol{R}_t\boldsymbol{V}_{\perp}\boldsymbol{G}_t\mathcal{P}_{\boldsymbol{S}_{t+1}}^{\perp}. \tag{2.94}$$

Combining equations equation 2.93 and equation 2.94, we obtain

$$\boldsymbol{G}_{t+1} = \left(\boldsymbol{I} - \bar{\eta}_t\boldsymbol{E}_t\boldsymbol{E}_t^{\top} + \bar{\eta}_t\boldsymbol{E}_{t+1}\boldsymbol{N}_t\left(\boldsymbol{S}_{t+1}\boldsymbol{N}_t\right)^{-1}\boldsymbol{S}_t\boldsymbol{E}_t^{\top}\right)\boldsymbol{G}_t\mathcal{P}_{\boldsymbol{S}_{t+1}}^{\perp} + \boldsymbol{V}_{\perp}^{\top}\boldsymbol{R}_t\boldsymbol{V}_{\perp}\boldsymbol{G}_t\mathcal{P}_{\boldsymbol{S}_{t+1}}^{\perp} + (A_2),$$

which results in

$$\|\boldsymbol{G}_{t+1}\| \leq \underbrace{\left\|\boldsymbol{I} - \bar{\eta}_t\boldsymbol{E}_t\boldsymbol{E}_t^{\top} + \bar{\eta}_t\boldsymbol{E}_{t+1}\boldsymbol{N}_t\left(\boldsymbol{S}_{t+1}\boldsymbol{N}_t\right)^{-1}\boldsymbol{S}_t\boldsymbol{E}_t^{\top}\right\|}_{(C)}\|\boldsymbol{G}_t\| + \|\boldsymbol{R}_t\|\,\|\boldsymbol{G}_t\| + \|(A_2)\|.$$

Therefore, it remains to control the terms $\|(C)\|$ and $\|(A_2)\|$. First, we provide an upper bound on $\|(C)\|$. Define $\boldsymbol{S}_{t+1}\boldsymbol{N}_t = (\boldsymbol{I} + \boldsymbol{\Xi})\boldsymbol{S}_t\boldsymbol{N}_t$, where $\boldsymbol{\Xi} = \boldsymbol{S}_{t+1}\boldsymbol{N}_t\left(\boldsymbol{S}_t\boldsymbol{N}_t\right)^{-1} - \boldsymbol{I}$ (note that $\boldsymbol{S}_t\boldsymbol{N}_t = M_t\boldsymbol{D}_t$ which implies that $\boldsymbol{S}_t\boldsymbol{N}_t$ is invertible). Hence, we have

$$\begin{aligned}\|(C)\| \leq &\underbrace{\left\|\boldsymbol{I} - \bar{\eta}_t\boldsymbol{E}_t\boldsymbol{E}_t^{\top} + \bar{\eta}_t\boldsymbol{E}_{t+1}\boldsymbol{N}_t\left(\boldsymbol{S}_t\boldsymbol{N}_t\right)^{-1}\boldsymbol{S}_t\boldsymbol{E}_t^{\top}\right\|}_{(C_1)} \\ &+ \underbrace{\bar{\eta}_t\left\|\boldsymbol{E}_{t+1}\boldsymbol{N}_t\left(\boldsymbol{S}_t\boldsymbol{N}_t\right)^{-1}\right\|\left\|(\boldsymbol{I} + \boldsymbol{\Xi})^{-1} - \boldsymbol{I}\right\|\left\|\boldsymbol{S}_t\boldsymbol{E}_t^{\top}\right\|}_{(C_2)}.\end{aligned}$$

It is shown in the proof of Lemma 24 that $\|\boldsymbol{\Xi}\| \leq 3\bar{\eta}_t\|\boldsymbol{\Delta}_t\| \leq 1/2$. Therefore, one can write $\|(\boldsymbol{I} + \boldsymbol{\Xi})^{-1} - \boldsymbol{I}\| \leq \|\boldsymbol{\Xi}\|\left\|(\boldsymbol{I} + \boldsymbol{\Xi})^{-1}\right\| \leq 6\bar{\eta}_t\|\boldsymbol{\Delta}_t\|$. To provide an upper bound for $\left\|\boldsymbol{E}_{t+1}\boldsymbol{N}_t\left(\boldsymbol{S}_t\boldsymbol{N}_t\right)^{-1}\right\|$, one can write

$$\begin{aligned}\boldsymbol{E}_{t+1}\boldsymbol{N}_t\left(\boldsymbol{S}_t\boldsymbol{N}_t\right)^{-1} &= \boldsymbol{E}_{t+1}\boldsymbol{S}_t^{\top}\left(\boldsymbol{S}_t\boldsymbol{S}_t^{\top}\right)^{-1} \\ &= \left(\boldsymbol{E}_t - \bar{\eta}_t\boldsymbol{E}_t\left(\boldsymbol{S}_t^{\top}\boldsymbol{S}_t + \boldsymbol{E}_t^{\top}\boldsymbol{E}_t\right) + \boldsymbol{V}_{\perp}^{\top}\boldsymbol{R}_t\boldsymbol{U}_t\right)\boldsymbol{S}_t^{\top}\left(\boldsymbol{S}_t\boldsymbol{S}_t^{\top}\right)^{-1} \\ &= \left(\boldsymbol{I} - \bar{\eta}_t\boldsymbol{E}_t\boldsymbol{E}_t^{\top}\right)\boldsymbol{H}_t - \bar{\eta}_t\boldsymbol{E}_t\boldsymbol{S}_t^{\top} + \boldsymbol{V}_{\perp}^{\top}\boldsymbol{R}_t\boldsymbol{U}_t\boldsymbol{S}_t^{\top}\left(\boldsymbol{S}_t\boldsymbol{S}_t^{\top}\right)^{-1} \\ &= \left(\boldsymbol{I} - \bar{\eta}_t\boldsymbol{E}_t\boldsymbol{E}_t^{\top}\right)\boldsymbol{H}_t - \bar{\eta}_t\boldsymbol{E}_t\boldsymbol{S}_t^{\top} + \boldsymbol{V}_{\perp}^{\top}\boldsymbol{R}_tV + \boldsymbol{V}_{\perp}^{\top}\boldsymbol{R}_t\boldsymbol{V}_{\perp}\boldsymbol{H}_t,\end{aligned}$$

where $\boldsymbol{H}_t = \boldsymbol{E}_t \boldsymbol{S}_t^\top \left(\boldsymbol{S}_t \boldsymbol{S}_t^\top\right)^{-1}$. Therefore, we have

$$\begin{aligned}\left\|\boldsymbol{E}_{t+1}\boldsymbol{N}_t \left(\boldsymbol{S}_t\boldsymbol{N}_t\right)^{-1}\right\| &\leq \|\boldsymbol{H}_t\| + \bar{\eta}_t \left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top\right\| + \|\boldsymbol{R}_t\| \left(1 + \|\boldsymbol{H}_t\|\right) \\ &\leq \|\boldsymbol{H}_t\| + \bar{\eta}_t \left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top\right\| + 4\bar{\eta}_t\delta \left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}.\end{aligned}$$

Hence, we obtain

$$\|C_2\| \leq 6\bar{\eta}_t^2 \|\boldsymbol{\Delta}_t\| \left(\|\boldsymbol{H}_t\| + \bar{\eta}_t \left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top\right\| + \bar{\eta}_t\delta \left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}\right) \left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top\right\|.$$

To control $\|(C_1)\|$, we use triangle inequality to arrive at the following decomposition

$$\begin{aligned}(C_1) \leq &\underbrace{\left\|\boldsymbol{I} - \bar{\eta}_t\boldsymbol{E}_t\boldsymbol{E}_t^\top + \bar{\eta}_t\boldsymbol{E}_t\boldsymbol{N}_t \left(\boldsymbol{S}_t\boldsymbol{N}_t\right)^{-1} \boldsymbol{S}_t\boldsymbol{E}_t^\top\right\|}_{(C_{11})} + \underbrace{\bar{\eta}_t^2 \left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top \boldsymbol{E}_t\boldsymbol{S}_t^\top \left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1} \boldsymbol{S}_t\boldsymbol{E}_t^\top\right\|}_{(C_{12})} \\ &+ \underbrace{\bar{\eta}_t^2 \left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top \boldsymbol{E}_t\boldsymbol{S}_t^\top \left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1} \boldsymbol{S}_t\boldsymbol{E}_t^\top\right\|}_{(C_{13})} + \underbrace{\bar{\eta}_t \left\|\boldsymbol{R}_t\boldsymbol{U}_t\boldsymbol{N}_t \left(\boldsymbol{S}_t\boldsymbol{N}_t\right)^{-1} \boldsymbol{S}_t\boldsymbol{E}_t^\top\right\|}_{(C_{14})}.\end{aligned}$$

It is easy to see that $\|(C_{11})\| = \left\|\boldsymbol{I} - \bar{\eta}_t\boldsymbol{G}_t\boldsymbol{G}_t^\top\right\| \leq 1$, due to the assumed upper bound on $\bar{\eta}_t$ and $\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|$. Moreover, one can verify that $\|(C_{12})\| \leq \bar{\eta}_t^2 \left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top\right\|^2$, $\|(C_{13})\| \leq \bar{\eta}_t^2 \|\boldsymbol{E}_t\|^4$, and $\|(C_{14})\| \leq 3\bar{\eta}_t^2\delta \|\boldsymbol{\Delta}_t\|_{\mathrm{F}} \left(1 + \|\boldsymbol{H}_t\|\right) \left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\| \leq 6\bar{\eta}_t^2\delta \|\boldsymbol{\Delta}_t\|_{\mathrm{F}} \left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\|$. Combining the derived upper bounds for $\|C_1\|$ and $\|C_2\|$, we have

$$(C) \leq 1 + \bar{\eta}_t^2 \left(2\left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top\right\|^2 + \|\boldsymbol{E}_t\|^4 + 7\delta \|\boldsymbol{\Delta}_t\|_{\mathrm{F}} \left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top\right\| + 6\|\boldsymbol{H}_t\| \|\boldsymbol{\Delta}_t\| \left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top\right\|\right).$$

To complete the proof, it remains to provide an upper bound for $\|(A_2)\|$. First, note that $\left(\boldsymbol{S}_{t+1} - \tilde{\boldsymbol{S}}_{t+1}\right) \mathcal{P}_{\boldsymbol{S}_t}^{\perp} = \boldsymbol{V}^\top \boldsymbol{R}_t \boldsymbol{V}_\perp \boldsymbol{G}_t$. Given this equality, one can write

$$\begin{aligned}\|A_2\| &\leq \left\|\boldsymbol{E}_{t+1}\boldsymbol{N}_t \left(\boldsymbol{S}_{t+1}\boldsymbol{N}_t\right)^{-1} \boldsymbol{V}^\top \boldsymbol{R}_t \boldsymbol{V}_\perp \boldsymbol{G}_t \mathcal{P}_{\boldsymbol{S}_{t+1}}^{\perp}\right\| \\ &\leq \left\|\boldsymbol{E}_{t+1}\boldsymbol{N}_t \left(\boldsymbol{S}_{t+1}\boldsymbol{N}_t\right)^{-1}\right\| \|\boldsymbol{R}_t\| \|\boldsymbol{G}_t\| \\ &\leq 3\bar{\eta}_t\delta \left\|\boldsymbol{E}_{t+1}\boldsymbol{N}_t \left(\boldsymbol{S}_{t+1}\boldsymbol{N}_t\right)^{-1}\right\| \|\boldsymbol{\Delta}_t\|_{\mathrm{F}} \|\boldsymbol{G}_t\|.\end{aligned}$$

Therefore, it suffices to provide an upper bound for $\left\|\boldsymbol{E}_{t+1}\boldsymbol{N}_t \left(\boldsymbol{S}_{t+1}\boldsymbol{N}_t\right)^{-1}\right\|$:

$$\begin{aligned}\left\|\boldsymbol{E}_{t+1}\boldsymbol{N}_t \left(\boldsymbol{S}_{t+1}\boldsymbol{N}_t\right)^{-1}\right\| &\leq \left\|\boldsymbol{E}_{t+1}\boldsymbol{N}_t \left(\boldsymbol{S}_t\boldsymbol{N}_t\right)^{-1}\right\| \left\|\boldsymbol{S}_t\boldsymbol{N}_t \left(\boldsymbol{S}_{t+1}\boldsymbol{N}_t\right)^{-1}\right\| \\ &\leq \left\|\boldsymbol{E}_{t+1}\boldsymbol{N}_t \left(\boldsymbol{S}_t\boldsymbol{N}_t\right)^{-1}\right\| \left\|(\boldsymbol{I} + \boldsymbol{\Xi})^{-1}\right\| \\ &\leq 2\left\|\boldsymbol{E}_{t+1}\boldsymbol{N}_t \left(\boldsymbol{S}_t\boldsymbol{N}_t\right)^{-1}\right\| \\ &\leq 2.\end{aligned}$$

Combining the above inequalities leads to

$$\|(A_2)\| \leq 6\bar{\eta}_t\delta \|\boldsymbol{\Delta}_t\|_{\mathrm{F}} \|\boldsymbol{G}_t\|.$$

Finally, combining the derived upper bounds for (C) and (A_2) gives rise to the following inequalities

$$\begin{aligned}\|\boldsymbol{G}_{t+1}\| &\leq \left(1 + 2\bar{\eta}_t^2 \left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top\right\|^2 + \bar{\eta}_t^2 \|\boldsymbol{E}_t\|^4 + 6\bar{\eta}_t^2 \|\boldsymbol{H}_t\| \|\boldsymbol{\Delta}_t\| \left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top\right\| + 7\bar{\eta}_t\delta \|\boldsymbol{\Delta}_t\|_{\mathrm{F}}\right) \|\boldsymbol{G}_t\| \\ &\leq \left(1 + 2\bar{\eta}_t^2 \left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top\right\|^2 + \bar{\eta}_t^2 \|\boldsymbol{E}_t\|^4 + 2\bar{\eta}_t^2 \|\boldsymbol{\Delta}_t\| \left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top\right\| + 7\bar{\eta}_t\delta \|\boldsymbol{\Delta}_t\|_{\mathrm{F}}\right) \|\boldsymbol{G}_t\|.\end{aligned}$$

Bounding $\|\boldsymbol{F}_{t+1}\|$: Similar to the previous part, we use the decomposition $\mathcal{P}_{\boldsymbol{S}_{t+1}} = \mathcal{P}_{\boldsymbol{S}_t}\mathcal{P}_{\boldsymbol{S}_{t+1}} + \mathcal{P}^{\perp}_{\boldsymbol{S}_t}\mathcal{P}_{\boldsymbol{S}_{t+1}}$ to write

$$\boldsymbol{F}_{t+1} = \boldsymbol{E}_{t+1}\mathcal{P}_{\boldsymbol{S}_{t+1}} = \underbrace{\boldsymbol{E}_{t+1}\mathcal{P}_{\boldsymbol{S}_t}\mathcal{P}_{\boldsymbol{S}_{t+1}}}_{(A)} + \underbrace{\boldsymbol{E}_{t+1}\mathcal{P}^{\perp}_{\boldsymbol{S}_t}\mathcal{P}_{\boldsymbol{S}_{t+1}}}_{(B)}.$$

First, we provide an upper bound for $\|(B)\|$:

$$\begin{aligned}\|(B)\| &= \left\|\boldsymbol{E}_{t+1}\mathcal{P}^{\perp}_{\boldsymbol{S}_t}\left(\mathcal{P}_{\boldsymbol{S}_{t+1}} - \mathcal{P}_{\boldsymbol{S}_t}\right)\right\| \\ &\leq \left\|\boldsymbol{E}_{t+1}\mathcal{P}^{\perp}_{\boldsymbol{S}_t}\right\| \left\|\mathcal{P}_{\boldsymbol{S}_{t+1}} - \mathcal{P}_{\boldsymbol{S}_t}\right\| \\ &= \left\|\left(\boldsymbol{I} - \bar{\eta}_t\boldsymbol{E}_t\boldsymbol{E}_t^\top\right)\boldsymbol{G}_t + \boldsymbol{V}_\perp^\top\boldsymbol{R}_t\boldsymbol{U}_t\mathcal{P}^{\perp}_{\boldsymbol{S}_t}\right\| \left\|\mathcal{P}_{\boldsymbol{S}_{t+1}} - \mathcal{P}_{\boldsymbol{S}_t}\right\| \\ &\leq \left(\|\boldsymbol{G}_t\| + \left\|\boldsymbol{V}_\perp^\top\boldsymbol{R}_t\boldsymbol{V}_\perp\boldsymbol{G}_t\right\|\right) \left\|\mathcal{P}_{\boldsymbol{S}_{t+1}} - \mathcal{P}_{\boldsymbol{S}_t}\right\| \\ &\leq \left(\|\boldsymbol{G}_t\| + 3\bar{\eta}_t\delta \|\boldsymbol{\Delta}_t\|_{\mathrm{F}} \|\boldsymbol{G}_t\|\right) \left\|\mathcal{P}_{\boldsymbol{S}_{t+1}} - \mathcal{P}_{\boldsymbol{S}_t}\right\| \\ &\leq 2\|\boldsymbol{G}_t\| \left\|\mathcal{P}_{\boldsymbol{S}_{t+1}} - \mathcal{P}_{\boldsymbol{S}_t}\right\|. \end{aligned} \tag{2.95}$$

To control $\left\|\mathcal{P}_{\boldsymbol{S}_{t+1}} - \mathcal{P}_{\boldsymbol{S}_t}\right\|$, we use the following technical lemma.

Lemma 25 (Theorem 2.4, Chen, Chen, and Li [2016])**.** *Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, and $\boldsymbol{B} = \boldsymbol{A} + \boldsymbol{E} \in \mathbb{R}^{m\times n}$ have the same rank. Then, we have*

$$\|\mathcal{P}_{\boldsymbol{A}} - \mathcal{P}_{\boldsymbol{B}}\| \leq \left\|\boldsymbol{E}\boldsymbol{A}^\dagger\right\| \vee \left\|\boldsymbol{E}\boldsymbol{B}^\dagger\right\|.$$

Due to Proposition 5 and our assumptions, we have $\lambda_{\min}(\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top) \geq \lambda_{\min}(\boldsymbol{S}_t\boldsymbol{S}_t^\top) > 0$,

and hence, both $\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top$ and $\boldsymbol{S}_t\boldsymbol{S}_t^\top$ are rank-r. Invoking Lemma 25, we have

$$\begin{aligned}
\left\|\mathcal{P}_{\boldsymbol{S}_{t+1}} - \mathcal{P}_{\boldsymbol{S}_t}\right\| &\leq \left\|\left(\boldsymbol{S}_{t+1} - \boldsymbol{S}_t\right)\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\right\| \\
&= \left\|\left(-\bar{\eta}_t\left(\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right)\boldsymbol{S}_t + \boldsymbol{S}_t\boldsymbol{E}_t^\top\boldsymbol{E}_t\right) + \boldsymbol{V}^\top\boldsymbol{R}_t\boldsymbol{U}_t\right)\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\right\| \\
&\leq \bar{\eta}_t\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right\| + \bar{\eta}_t\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\|\|\boldsymbol{H}_t\| + \|\boldsymbol{R}_t\|\left(1 + \|\boldsymbol{H}_t\|\right) \\
&\leq \bar{\eta}_t\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right\| + \bar{\eta}_t\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\| + 6\bar{\eta}_t\delta\|\boldsymbol{\Delta}_t\|_{\mathrm{F}} \\
&\leq 3\bar{\eta}_t\|\boldsymbol{\Delta}_t\|,
\end{aligned} \tag{2.96}$$

where in the last inequality, we used the following auxiliary lemma.

Lemma 26. *If $\|\boldsymbol{\Delta}_t\| \geq \sqrt{d}\,\|\boldsymbol{G}_t\|^2$, then we have $\|\boldsymbol{\Delta}_t\|_{\mathrm{F}} \leq 5\sqrt{r}\,\|\boldsymbol{\Delta}_t\|$.*

Proof. Recall the signal-residual decomposition

$$\boldsymbol{\Delta}_t = \boldsymbol{V}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right)\boldsymbol{V}^\top + \boldsymbol{V}\boldsymbol{S}_t\boldsymbol{E}_t^\top\boldsymbol{V}_\perp^\top + \boldsymbol{V}_\perp\boldsymbol{E}_t\boldsymbol{S}_t^\top\boldsymbol{V}^\top + \boldsymbol{V}_\perp\boldsymbol{F}_t\boldsymbol{F}_t^\top\boldsymbol{V}_\perp^\top + \boldsymbol{V}_\perp\boldsymbol{G}_t\boldsymbol{G}_t^\top\boldsymbol{V}_\perp^\top.$$

One can write

$$\begin{aligned}
\|\boldsymbol{\Delta}_t\|_{\mathrm{F}} &\leq \sqrt{r}\left(\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right\| + 2\left\|\boldsymbol{S}_t\boldsymbol{E}_t^\top\right\| + \left\|\boldsymbol{F}_t\boldsymbol{F}_t^\top\right\|\right) + \sqrt{d}\,\|\boldsymbol{G}_t\|^2 \\
&= 4\sqrt{r}\,\|\boldsymbol{\Delta}_t\| + \|\boldsymbol{\Delta}_t\| \\
&\leq 5\sqrt{r}\,\|\boldsymbol{\Delta}_t\|,
\end{aligned}$$

which completes the proof. □

Combining equation 2.96 with equation 2.95 leads to

$$\|(B)\| \leq 6\bar{\eta}_t\|\boldsymbol{\Delta}_t\|\,\|\boldsymbol{G}_t\|.$$

Next, we will provide an upper bound for (A). One can write

$$\begin{aligned}
\|(\boldsymbol{A})\| &\leq \|\boldsymbol{E}_{t+1}\mathcal{P}_{\boldsymbol{S}_t}\| \\
&\leq \left\|\left(\boldsymbol{E}_t - \bar{\eta}_t\boldsymbol{E}_t\left(\boldsymbol{E}_t^\top\boldsymbol{E}_t + \boldsymbol{S}_t^\top\boldsymbol{S}_t\right) + \boldsymbol{V}_\perp^\top\boldsymbol{R}_t\boldsymbol{U}_t\right)\mathcal{P}_{\boldsymbol{S}_t}\right\| \\
&\leq \left\|\boldsymbol{F}_t - \bar{\eta}_t\boldsymbol{E}_t\boldsymbol{E}_t^\top\boldsymbol{F}_t - \bar{\eta}_t\boldsymbol{F}_t\boldsymbol{S}_t^\top\boldsymbol{S}_t\right\| + \left\|\boldsymbol{V}_\perp^\top\boldsymbol{R}_t\boldsymbol{U}_t\mathcal{P}_{\boldsymbol{S}_t}\right\|.
\end{aligned} \tag{2.97}$$

The first term in the above inequality can be bounded as follows

$$\begin{aligned}\left\|\boldsymbol{F}_t-\bar{\eta}_t\boldsymbol{E}_t\boldsymbol{E}_t^\top\boldsymbol{F}_t-\bar{\eta}_t\boldsymbol{F}_t\boldsymbol{S}_t^\top\boldsymbol{S}_t\right\| &\le \left\|\left(0.5\boldsymbol{I}-\bar{\eta}_t\boldsymbol{E}_t\boldsymbol{E}_t^\top\right)\boldsymbol{F}_t\right\|+\left\|\boldsymbol{F}_t\left(0.5\boldsymbol{I}-\bar{\eta}_t\boldsymbol{S}_t^\top\boldsymbol{S}_t\right)\right\| \\ &\le \left(\left\|0.5\boldsymbol{I}-\bar{\eta}_t\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|+\left\|0.5\boldsymbol{I}-\bar{\eta}_t\boldsymbol{S}_t^\top\boldsymbol{S}_t\right\|\right)\|\boldsymbol{F}_t\| \\ &\le \left(1-\bar{\eta}_t\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right)\|\boldsymbol{F}_t\|.\end{aligned}$$

Moreover, we have

$$\left\|\boldsymbol{V}_\perp^\top\boldsymbol{R}_t\boldsymbol{U}_t\mathcal{P}_{\boldsymbol{S}_t}\right\| \le \|\boldsymbol{R}_t\|\left(\|\boldsymbol{S}_t\|+\|\boldsymbol{F}_t\|\right) \le 3\bar{\eta}_t\delta\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}\left(\|\boldsymbol{S}_t\|+\|\boldsymbol{F}_t\|\right).$$

Therefore, we have

$$\|(\boldsymbol{A})\| \le \left(1-\bar{\eta}_t\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)+3\bar{\eta}_t\delta\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}\right)\|\boldsymbol{F}_t\|+3\bar{\eta}_t\delta\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}\|\boldsymbol{S}_t\|.$$

Finally, combining the derived upper bounds for (A) and (B) leads to

$$\|\boldsymbol{F}_{t+1}\| \le \left(1-\bar{\eta}_t\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)+3\bar{\eta}_t\delta\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}\right)\|\boldsymbol{F}_t\|+3\bar{\eta}_t\delta\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}\|\boldsymbol{S}_t\|+6\bar{\eta}_t\left\|\boldsymbol{\Delta}_t\right\|\|\boldsymbol{G}_t\|,$$

which completes the proof. □

2.D.5 Proof of Lemma 6

Recall that $\boldsymbol{U}_0 = \alpha\boldsymbol{B}$, where $\boldsymbol{B}\boldsymbol{B}^\top$ is the best rank-r' approximation of $\boldsymbol{C} \in \frac{1}{2m}\sum_{i=1}^m \operatorname{sign}(y_i)\left(\boldsymbol{A}_i+\boldsymbol{A}_i^\top\right)$. Since $\operatorname{rank}(\boldsymbol{X}^\star)=r$, Sign-RIP implies

$$\left\|\boldsymbol{C}-\varphi(\boldsymbol{X}^\star)\frac{\boldsymbol{X}^\star}{\|\boldsymbol{X}^\star\|_{\mathrm{F}}}\right\| \le \varphi(\boldsymbol{X}^\star)\delta.$$

Note that $\boldsymbol{B}\boldsymbol{B}^\top$ is the best rank-r' approximation of $\boldsymbol{C}$. We have

$$\begin{aligned}\left\|\boldsymbol{B}\boldsymbol{B}^\top-\varphi(\boldsymbol{X}^\star)\frac{\boldsymbol{X}^\star}{\|\boldsymbol{X}^\star\|_{\mathrm{F}}}\right\| &\le \left\|\boldsymbol{B}\boldsymbol{B}^\top-\boldsymbol{C}\right\|+\left\|\boldsymbol{C}-\varphi(\boldsymbol{X}^\star)\frac{\boldsymbol{X}^\star}{\|\boldsymbol{X}^\star\|_{\mathrm{F}}}\right\| \\ &\le \left|\lambda_{r'+1}(\boldsymbol{C})\right|+\varphi(\boldsymbol{X}^\star)\delta \\ &\le \left\|\boldsymbol{C}-\varphi(\boldsymbol{X}^\star)\frac{\boldsymbol{X}^\star}{\|\boldsymbol{X}^\star\|_{\mathrm{F}}}\right\|+\varphi(\boldsymbol{X}^\star)\delta \\ &\le 2\varphi(\boldsymbol{X}^\star)\delta.\end{aligned}$$

Therefore, based on the definition of $\boldsymbol{U}_0$, we have

$$\left\| \boldsymbol{U}_0 \boldsymbol{U}_0^\top - \alpha^2 \varphi(\boldsymbol{X}^\star) \frac{\boldsymbol{X}^\star}{\|\boldsymbol{X}^\star\|_{\mathrm{F}}} \right\| \leq 2\alpha^2 \varphi(\boldsymbol{X}^\star)\delta.$$

Given this bound, one can write

$$\begin{aligned}
\left\| \boldsymbol{S}_0 \boldsymbol{S}_0^\top - \alpha^2 \varphi(\boldsymbol{X}^\star) \frac{\boldsymbol{\Sigma}}{\|\boldsymbol{X}^\star\|_{\mathrm{F}}} \right\| &= \left\| \boldsymbol{V}^\top \left(\boldsymbol{U}_0 \boldsymbol{U}_0^\top - \alpha^2 \varphi(\boldsymbol{X}^\star) \frac{\boldsymbol{X}^\star}{\|\boldsymbol{X}^\star\|_{\mathrm{F}}} \right) \boldsymbol{V} \right\| \\
&\leq \left\| \boldsymbol{U}_0 \boldsymbol{U}_0^\top - \alpha^2 \varphi(\boldsymbol{X}^\star) \frac{\boldsymbol{X}^\star}{\|\boldsymbol{X}^\star\|_{\mathrm{F}}} \right\| \\
&\leq 2\alpha^2 \varphi(\boldsymbol{X}^\star)\delta.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\left\| \boldsymbol{S}_0 \boldsymbol{E}_0^\top \right\| &= \left\| \boldsymbol{V}^\top \left(\boldsymbol{U}_0 \boldsymbol{U}_0^\top - \alpha^2 \varphi(\boldsymbol{X}^\star) \frac{\boldsymbol{X}^\star}{\|\boldsymbol{X}^\star\|_{\mathrm{F}}} \right) \boldsymbol{V}_\perp \right\| \leq 2\alpha^2 \varphi(\boldsymbol{X}^\star)\delta, \\
\left\| \boldsymbol{E}_0 \boldsymbol{E}_0^\top \right\| &= \left\| \boldsymbol{V}_\perp^\top \left(\boldsymbol{U}_0 \boldsymbol{U}_0^\top - \alpha^2 \varphi(\boldsymbol{X}^\star) \frac{\boldsymbol{X}^\star}{\|\boldsymbol{X}^\star\|_{\mathrm{F}}} \right) \boldsymbol{V}_\perp \right\| \leq 2\alpha^2 \varphi(\boldsymbol{X}^\star)\delta.
\end{aligned}$$

This completes the proof. □

2.D.6 Proof of Lemma 7

We prove this lemma by induction on t. First, due to Lemma 6, it is easy to verify that equation 2.40-equation 2.43 hold for $t = 0$. Now, suppose that equation 2.40-equation 2.42 are satisfied for $t < T_{end}$. Moreover, without loss of generality, we assume that $\|\boldsymbol{\Delta}_t\|_{\mathrm{F}} \gtrsim d\alpha^{2-\mathcal{O}(\sqrt{r}\kappa^2\delta)}$ for every $0 \leq t \leq T_{end}$; otherwise, the statement of the lemma holds. Together with the induction hypothesis on $\|\boldsymbol{G}_t\|$, this implies that $\|\boldsymbol{\Delta}_t\|_{\mathrm{F}} \geq \zeta$, for $\zeta > 0$ defined in Propositions 5, 6, and 7.

Bounding $\|\boldsymbol{F}_{t+1}\|$: In order to apply Proposition 7, first we verify its assumptions. One can write $\left\| \boldsymbol{E}_t \boldsymbol{E}_t^\top \right\| = \|\boldsymbol{F}_t\|^2 + \|\boldsymbol{G}_t\|^2$. Therefore, we have $\left\| \boldsymbol{E}_t \boldsymbol{E}_t^\top \right\| \leq \sigma_1$, due to the induction hypothesis on $\|\boldsymbol{F}_t\|$ and $\|\boldsymbol{G}_t\|$. On the other hand, $\left\| \boldsymbol{S}_t \boldsymbol{S}_t^\top \right\| \leq 1.01\sigma_1$ and $\left\| \boldsymbol{E}_t \boldsymbol{S}_t^\top (\boldsymbol{S}_t \boldsymbol{S}_t^\top)^{-1} \right\| \leq 1/3$ due to our induction hypothesis. It remains to show that $(4r, \delta, \epsilon, \mathcal{S})$-Sign-RIP with $\epsilon \geq \sqrt{d}\bar{\varphi}\alpha$ implies $(4r, \delta, \sqrt{d}\|\boldsymbol{G}_t\|^2, \mathcal{S})$-Sign-RIP. To show this, first note that $(4r, \delta, \epsilon_1, \mathcal{S})$-Sign-RIP implies $(4r, \delta, \epsilon_2, \mathcal{S})$-Sign-RIP, for any $\epsilon_2 \leq \epsilon_1$. Using this fact, it suffices to show that $\sqrt{d}\|\boldsymbol{G}_t\|^2 \leq \sqrt{d}\bar{\varphi}\alpha \lesssim \epsilon$, which is immediate due to our induction hypothesis on $\|\boldsymbol{G}_t\|$,

and our choice of δ. Therefore, Proposition 7 holds and we have

$$\|\boldsymbol{F}_{t+1}\| \leq \left(1 - \bar{\eta}_t \lambda_{\min}\left(\boldsymbol{S}_t \boldsymbol{S}_t^\top\right) + 3\bar{\eta}_t \delta \|\boldsymbol{\Delta}_t\|_{\mathrm{F}}\right) \|\boldsymbol{F}_t\| + 3\bar{\eta}_t \delta \|\boldsymbol{\Delta}_t\|_{\mathrm{F}} \|\boldsymbol{S}_t\| + 6\bar{\eta}_t \|\boldsymbol{\Delta}_t\| \|\boldsymbol{G}_t\|. \tag{2.98}$$

Due to our induction hypothesis, equation 2.98 can be simplified as

$$\begin{aligned}
\|\boldsymbol{F}_{t+1}\| \leq & \left(1 + 75\sqrt{r}\sigma_1 \bar{\eta}_t \delta\right) \|\boldsymbol{F}_t\| + 80\sqrt{r}\sigma_1^{1.5} \bar{\eta}_t \delta + 30\sigma_1 \bar{\eta}_t \|\boldsymbol{G}_t\| \\
\leq & \|\boldsymbol{F}_t\| + 100\sqrt{r}\sigma_1^{1.5} \bar{\eta}_t \delta + 30\sigma_1 \bar{\eta}_t \|\boldsymbol{G}_t\| \\
\leq & \|\boldsymbol{F}_t\| + 100\sqrt{r}\sigma_1^{1.5} \bar{\eta}_t \delta + 30\sigma_1 \bar{\eta}_t \sqrt{\alpha\bar{\varphi}\delta} \\
\leq & \eta\bar{\varphi}^2 \left(100\sqrt{r}\sigma_1^{1.5}\delta + 30\sigma_1\sqrt{\alpha\bar{\varphi}\delta}\right)(t+2),
\end{aligned}$$

where in the first inequality, we use Lemma 26 and the induction hypothesis on $\|\boldsymbol{\Delta}_t\|$. Moreover, in the last inequality, we used the induction hypothesis on $\|\boldsymbol{G}_t\|$.

Bounding $\left\|\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right\|$: Similar to our previous argument, it is easy to verify that the assumptions of Proposition 6 are satisfied at iteration t. Therefore, $\left\|\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right\| \leq 1.01\sigma_1$ readily follows from Proposition 6.

Bounding $\left\|\boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^\top - \boldsymbol{X}^\star\right\|$: Note that

$$\left\|\boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^\top - \boldsymbol{X}^\star\right\| \leq \left\|\boldsymbol{\Sigma} - \boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right\| + 2\left\|\boldsymbol{S}_{t+1}\boldsymbol{E}_{t+1}^\top\right\| + \left\|\boldsymbol{E}_{t+1}\boldsymbol{E}_{t+1}^\top\right\|.$$

On the other hand, we have

$$\left\|\boldsymbol{E}_{t+1}\boldsymbol{E}_{t+1}^\top\right\| \leq \|\boldsymbol{F}_{t+1}\|^2 + \|\boldsymbol{G}_{t+1}\|^2 \leq 0.5\sigma_1.$$

where the last inequality is due to our choice of δ and α. Similarly, we can show that $\left\|\boldsymbol{S}_{t+1}\boldsymbol{E}_{t+1}^\top\right\| \leq \sigma_1$. This together with $\left\|\boldsymbol{\Sigma} - \boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right\| \leq \left\|\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right\| + \sigma_1 \leq 2.01\sigma_1$ leads to

$$\left\|\boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^\top - \boldsymbol{X}^\star\right\| \leq 5\sigma_1.$$

Establishing $\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top \succ \boldsymbol{0}$: The proof follows directly from the application of Proposition 5. The details are omitted due to their similarity to the proof of equation 2.33 in Lemma 5.

Bounding $\left\|\boldsymbol{E}_{t+1}\boldsymbol{S}_{t+1}^\top\left(\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right)^{-1}\right\|$: To streamline the proof, let $\boldsymbol{H}_t = \boldsymbol{E}_t\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}$. Our goal is to prove $\|\boldsymbol{H}_t\| \leq 1/3$ by showing the following recursive relationship.

Lemma 27. *For every $0 \le s \le t$, we have*

$$\|\boldsymbol{H}_{s+1}\| \le (1 - c\bar{\eta}_t\sigma_r)\,\|\boldsymbol{H}_s\| + c'\sqrt{r}\sigma_1\bar{\eta}_t\delta, \tag{2.99}$$

where $c, c' > 0$ are some universal constants.

The proof of Lemma 27 can be found in Appendix 2.G.4. Equipped with this lemma, we are ready to derive the desired result. First, due to Lemma 6, we have

$$\|\boldsymbol{H}_0\| \le \frac{\left\|\boldsymbol{E}_0\boldsymbol{S}_0^\top\right\|}{\lambda_{\min}(\boldsymbol{S}_0\boldsymbol{S}_0^\top)} \le \frac{2\alpha^2\delta\varphi(\boldsymbol{X}^\star)}{\alpha^2\varphi(\boldsymbol{X}^\star)\left(\frac{1}{\sqrt{r}\kappa} - 2\delta\right)} \le 4\sqrt{r}\kappa\delta,$$

provided that $\delta \le \frac{1}{4\sqrt{r}\kappa}$. On the other hand, equation 2.99 implies that

$$\|\boldsymbol{H}_{t+1}\| - \frac{c'}{c}\sqrt{r}\kappa\delta \le (1 - c\bar{\eta}_t\sigma_r)^{t+1}\left(\|\boldsymbol{H}_0\| - \frac{c'}{c}\sqrt{r}\kappa\delta\right) \le \|\boldsymbol{H}_0\| + \frac{c'}{c}\sqrt{r}\kappa\delta.$$

Therefore, due to our choice of δ, we have

$$\|\boldsymbol{H}_{t+1}\| \le \left(4\sqrt{r}\kappa + \frac{2c'}{c}\sqrt{r}\kappa\right)\delta \lesssim \sqrt{r}\kappa\delta \le 1/3.$$

Bounding $\|\boldsymbol{G}_{t+1}\|$: Due to Proposition 7, we have

$$\begin{aligned}\|\boldsymbol{G}_{t+1}\| \le \Big(1 + \bar{\eta}_t^2\Big(2\left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top\right\|^2 + \|\boldsymbol{E}_t\|^4 + 6\left\|\boldsymbol{E}_t\boldsymbol{S}_t(\boldsymbol{S}_t\boldsymbol{S}_t)^{-1}\right\|\|\boldsymbol{\Delta}_t\|\left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top\right\|\Big) \\ +7\bar{\eta}_t\delta\,\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}\Big)\,\|\boldsymbol{G}_t\|\,.\end{aligned}$$

Moreover, one can write

$$\bar{\eta}_t^2\left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top\right\|^2 \lesssim \bar{\eta}_t\,\|\boldsymbol{E}_t\|^2 \lesssim r\sigma_1^3\bar{\eta}_t^2\delta^2t^2 + \sigma_1^2\bar{\eta}_t^2\alpha\bar{\varphi}\delta t^2 \lesssim \sqrt{r}\kappa\delta\log\left(\frac{1}{\alpha}\right).$$

Similarly, it can be shown that

$$\bar{\eta}_t^2\,\|\boldsymbol{E}_t\|^4 \vee \bar{\eta}_t^2\,\|\boldsymbol{H}_t\|\,\|\boldsymbol{\Delta}_t\|\left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top\right\| \vee \bar{\eta}_t\delta\,\|\boldsymbol{\Delta}_t\|_{\mathrm{F}} \lesssim \sqrt{r}\kappa\delta\log\left(\frac{1}{\alpha}\right).$$

Therefore, for some universal constant $\boldsymbol{C} > 0$, we have

$$\|\boldsymbol{G}_{t+1}\| \le \left(1 + C\sqrt{r}\kappa\delta\log\left(\frac{1}{\alpha}\right)\right)\|\boldsymbol{G}_t\|\,.$$

Hence, we have

$$\begin{aligned}
\|\boldsymbol{G}_{t+1}\| &\leq \left(1 + C\sqrt{r}\kappa\delta \log\left(\frac{1}{\alpha}\right)\right)^{t+1} \|\boldsymbol{G}_0\| \\
&\leq \left(1 + C\sqrt{r}\kappa\delta \log\left(\frac{1}{\alpha}\right)\right)^{C'\frac{\log(1/\alpha)}{\bar{\eta}_t\sigma_r}} \|\boldsymbol{G}_0\| \\
&\leq \exp\left(C''\sqrt{r}\kappa^2\delta \log\left(\frac{1}{\alpha}\right)\right) \|\boldsymbol{G}_0\| \\
&\leq \alpha^{-\mathcal{O}\left(\sqrt{r}\kappa^2\delta\right)} \|\boldsymbol{G}_0\| \\
&\leq \alpha^{1-\mathcal{O}\left(\sqrt{r}\kappa^2\delta\right)} \sqrt{\bar{\varphi}\delta},
\end{aligned}$$

where in the last inequality, we used the upper bound $\|\boldsymbol{G}_0\| \leq \|\boldsymbol{E}_0\|$ and Lemma 6.

2.E Proofs of Empirical Loss with Noisy Measurements

2.E.1 Preliminaries

Given the update rule $\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - 2\eta_t \boldsymbol{Q}_t \boldsymbol{U}_t$, we consider the following decomposition

$$U_{t+1} = \tilde{\boldsymbol{U}}_{t+1} + \boldsymbol{R}_t \boldsymbol{U}_t, \text{ where } \tilde{\boldsymbol{U}}_{t+1} = \boldsymbol{U}_t - \frac{2\eta\rho^t}{\|\boldsymbol{\Delta}_t\|}\boldsymbol{\Delta}_t\boldsymbol{U}_t, \ \boldsymbol{R}_t = \frac{2\eta\rho^t}{\|\boldsymbol{\Delta}_t\|}\boldsymbol{\Delta}_t - 2\eta_t\boldsymbol{Q}_t, \quad (2.100)$$

In the above decomposition, $\tilde{\boldsymbol{U}}_{t+1}$ resembles one iteration of GD on $\bar{f}_{\ell_2}(\boldsymbol{U})$ with the "effective" step-size $\frac{\eta\rho^t}{2\|\boldsymbol{\Delta}_t\|}$. Moreover, the term $\boldsymbol{R}_t\boldsymbol{U}_t$ captures the deviation of SubGM and GD. Similar to the noiseless setting, the main idea behind our proof technique is to show that $\boldsymbol{R}_t$ remains small throughout the iterations of SubGM, and consequently, SubGM behaves similarly to GD. To this goal, we first provide an upper bound on $\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}$ in terms of $\|\boldsymbol{\Delta}_t\|$.

Lemma 28. *Suppose that* $\sqrt{d}\,\|\boldsymbol{G}_t\|^2 \leq \|\boldsymbol{\Delta}_t\|$. *Then, we have* $\|\boldsymbol{\Delta}_t\|_{\mathrm{F}} \leq 2(1+\sqrt{r})\,\|\Delta\|$.

Proof. Due to our proposed signal-residual decomposition, one can write

$$
\begin{aligned}
&\|\boldsymbol{\Delta}_t\|_{\mathrm{F}} \qquad (2.101)\\
\leq&\left\|\underbrace{\boldsymbol{V}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top-\boldsymbol{\Sigma}\right)\boldsymbol{V}^\top+\boldsymbol{V}\boldsymbol{S}_t\boldsymbol{E}_t^\top\boldsymbol{V}_\perp^\top+\boldsymbol{V}_\perp\boldsymbol{E}_t\boldsymbol{S}_t^\top\boldsymbol{V}^\top+\boldsymbol{V}_\perp\boldsymbol{F}_t\boldsymbol{F}_t^\top\boldsymbol{V}_\perp^\top}_{\text{rank-}4r}\right\|_{\mathrm{F}}+\left\|\underbrace{\boldsymbol{V}_\perp\boldsymbol{G}_t\boldsymbol{G}_t^\top\boldsymbol{V}_\perp^\top}_{\text{small norm}}\right\|_{\mathrm{F}}\\
\leq&\sqrt{4r}\left\|\boldsymbol{V}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top-\boldsymbol{\Sigma}\right)\boldsymbol{V}^\top+\boldsymbol{V}\boldsymbol{S}_t\boldsymbol{E}_t^\top\boldsymbol{V}_\perp^\top+\boldsymbol{V}_\perp\boldsymbol{E}_t\boldsymbol{S}_t^\top\boldsymbol{V}^\top+\boldsymbol{V}_\perp\boldsymbol{F}_t\boldsymbol{F}_t^\top\boldsymbol{V}_\perp^\top\right\|+\sqrt{d}\left\|\boldsymbol{G}_t\right\|^2\\
\leq&\sqrt{4r}\left\|\boldsymbol{\Delta}_t\right\|+\sqrt{4r}\left\|\boldsymbol{V}_\perp\boldsymbol{G}_t\boldsymbol{G}_t^\top\boldsymbol{V}_\perp^\top\right\|+\sqrt{d}\left\|\boldsymbol{G}_t\right\|^2\\
\leq&2(1+\sqrt{r})\left\|\boldsymbol{\Delta}_t\right\|,
\end{aligned}
$$

where the last inequality follows from the assumption $4r\leq d$ and $\sqrt{d}\left\|\boldsymbol{G}_t\right\|^2\leq\left\|\boldsymbol{\Delta}_t\right\|$. This completes the proof. □

Equipped with this technical lemma, we next provide an upper bound on $\|\boldsymbol{R}_t\|$.

Lemma 29. *Suppose that the measurements satisfy $(4r,\delta,\epsilon,\mathcal{S})$-Sign-RIP with $\delta<\frac{1}{4(1+\sqrt{r})}$, $\epsilon=\sqrt{d}\left\|\boldsymbol{G}_t\right\|^2$, $\mathcal{S}=\{\boldsymbol{X}:\|\boldsymbol{X}\|_{\mathrm{F}}\geq\zeta\}$ for $\zeta=\sqrt{d}\left\|\boldsymbol{G}_t\right\|^2\left(\frac{1}{\delta}\vee\sqrt{d}\right)$, and $\sqrt{d}\left\|\boldsymbol{G}_t\right\|^2\leq\left\|\boldsymbol{\Delta}_t\right\|$. Then, we have $\|\boldsymbol{R}_t\|\leq 8(1+\sqrt{r})\eta\rho^t\delta$.*

Proof. One can write

$$
\begin{aligned}
\boldsymbol{R}_t&=\frac{2\eta\rho^t}{\|\boldsymbol{\Delta}_t\|}\boldsymbol{\Delta}_t-\frac{2\eta\rho^t}{\|\boldsymbol{Q}_t\|}\boldsymbol{Q}_t\\
&=-\frac{2\eta\rho^t}{\|\boldsymbol{Q}_t\|}\left(\boldsymbol{Q}_t-\varphi(\boldsymbol{\Delta}_t)\frac{\boldsymbol{\Delta}_t}{\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}}\right)-\frac{2\eta\rho^t\varphi(\boldsymbol{\Delta}_t)}{\|\boldsymbol{Q}_t\|\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}}\boldsymbol{\Delta}_t+\frac{2\eta\rho^t}{\|\boldsymbol{\Delta}_t\|}\boldsymbol{\Delta}_t\\
&=-\frac{2\eta\rho^t}{\|\boldsymbol{Q}_t\|}\left(\boldsymbol{Q}_t-\varphi(\boldsymbol{\Delta}_t)\frac{\boldsymbol{\Delta}_t}{\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}}\right)+\frac{\|\boldsymbol{Q}_t\|-\varphi(\boldsymbol{\Delta}_t)\frac{\|\boldsymbol{\Delta}_t\|}{\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}}}{\|\boldsymbol{Q}_t\|}\frac{2\eta\rho^t}{\|\boldsymbol{\Delta}_t\|}\boldsymbol{\Delta}_t.
\end{aligned}
$$

The above equality implies that

$$
\begin{aligned}
\|\boldsymbol{R}_t\|&\leq\frac{2\eta\rho^t}{\|\boldsymbol{Q}_t\|}\left\|\boldsymbol{Q}_t-\varphi(\boldsymbol{\Delta}_t)\frac{\boldsymbol{\Delta}_t}{\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}}\right\|+2\eta\rho^t\cdot\frac{\left|\|\boldsymbol{Q}_t\|-\varphi(\boldsymbol{\Delta}_t)\frac{\|\boldsymbol{\Delta}_t\|}{\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}}\right|}{\|\boldsymbol{Q}_t\|} \qquad (2.102)\\
&\leq\left(\frac{1}{\|\boldsymbol{Q}_t\|}\right)2\eta\rho^t\varphi(\boldsymbol{\Delta}_t)\delta,
\end{aligned}
$$

where in the last inequality, we used Sign-RIP. Next, we provide an upper bound for $1/\left\|\boldsymbol{Q}_t\right\|$. Due to Sign-RIP, we have $\|\boldsymbol{Q}_t\|\geq\left(\frac{\|\boldsymbol{\Delta}_t\|}{\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}}-\delta\right)\varphi(\boldsymbol{\Delta}_t)$. On the other hand, due to Lemma 28,

we have $\frac{\|\boldsymbol{\Delta}_t\|}{\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}} \leq \frac{1}{2(1+\sqrt{r})}$. Combining these inequalities with equation 2.102, we have

$$\|\boldsymbol{R}_t\| \leq \frac{2}{\frac{1}{2(1+\sqrt{r})} - \delta} \cdot \eta\rho^t\delta \leq 8(1+\sqrt{r})\eta\rho^t\delta,$$

where the last inequality is due to the assumption $\delta \leq \frac{1}{4(1+\sqrt{r})}$. □

2.E.2 Proof of Proposition 8

The proof is almost a line-by-line reconstruction of the proof of Proposition 5 in Appendix 2.D.2. For brevity, we only provide a sketch of the proof. Similar to equation 2.83, one can write

$$\boldsymbol{S}_{t+1} = \tilde{\boldsymbol{S}}_{t+1} + \boldsymbol{V}^\top \boldsymbol{R}_t \boldsymbol{U}_t. \tag{2.103}$$

Given this decomposition, we characterize the growth rate of $\lambda_{\min}(\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top)$ by first resorting to a more tractable lower bound. In particular, we define $\boldsymbol{M} := (\boldsymbol{I} + \boldsymbol{\Xi})\, \tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\, (\boldsymbol{I} + \boldsymbol{\Xi})^\top$, where $\boldsymbol{\Xi} := \boldsymbol{V}^\top \boldsymbol{R}_t \boldsymbol{U}_t \tilde{\boldsymbol{S}}_{t+1}^\top \left(\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\right)^{-1}$. Based on the definition of $\boldsymbol{M}$, a series of inequalities analogous to equation 2.84 can be used to show that $\lambda_{\min}(\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top) \geq \lambda_{\min}(\boldsymbol{M})$. Therefore, providing a lower bound for $\lambda_{\min}(\boldsymbol{M})$ suffices. Similar to the proof of Proposition 5, we first show that $\lambda_{\min}(\boldsymbol{M}) \approx \lambda_{\min}(\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top)$:

$$\begin{aligned}\left|\lambda_{\min}(\boldsymbol{M}) - \lambda_{\min}\left(\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\right)\right| &\leq 3\,\|\boldsymbol{\Xi}\|\,\lambda_{\min}\left(\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\right) \\ &\leq 192\sqrt{r}\eta\rho^t\lambda_{\min}\left(\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\right).\end{aligned}$$

Combining the above inequality with the one-step dynamic of $\lambda_{\min}\left(\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\right)$ from Proposition 3 completes the proof. □

2.E.3 Proof of Proposition 9

Similar to the minimum eigenvalue dynamics, the proof is identical to the proof of Proposition 6. Hence, we only provide a sketch. Similar to equation 2.86, one can write

$$\boldsymbol{\Sigma} - \boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top = \boldsymbol{\Sigma} - \tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top - \boldsymbol{V}^\top\boldsymbol{R}_t\boldsymbol{U}_t\tilde{\boldsymbol{S}}_{t+1}^\top - \tilde{\boldsymbol{S}}_{t+1}\boldsymbol{U}_t^\top\boldsymbol{R}_t^\top\boldsymbol{V} - \boldsymbol{V}^\top\boldsymbol{R}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{R}_t^\top\boldsymbol{V}.$$

Lemma 29 combined with an argument similar to Appendix 2.D.3 leads to

$$\left\|\boldsymbol{V}^\top\boldsymbol{R}_t\boldsymbol{U}_t\tilde{\boldsymbol{S}}_{t+1}^\top + \tilde{\boldsymbol{S}}_{t+1}\boldsymbol{U}_t^\top\boldsymbol{R}_t^\top\boldsymbol{V} + \boldsymbol{V}^\top\boldsymbol{R}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{R}_t^\top\boldsymbol{V}\right\| \leq 193\sqrt{r}\eta\rho^t\sigma_1\delta.$$

The above bound combined with the one-step dynamics of $\boldsymbol{\Sigma} - \tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top$ in Proposition 4 completes the proof for the one-step dynamics of $\boldsymbol{\Sigma} - \boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top$. The dynamics of the cross term equation 2.27 and the upper bound on $\left\|\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right\|$ equation 2.28 can be deduced in a similar fashion. The details are omitted for brevity. □

2.E.4 Proof of Proposition 10

The proof of Proposition 10 is identical to that of Proposition 7, with a key difference that $\|\boldsymbol{R}_t\| \le 16\sqrt{r}\eta\rho^t$. The details are omitted for brevity. □

2.E.5 Proof of Lemma 9

The proof is based on an inductive argument similar to the proof of Lemma 7 in Appendix 2.D.6. Due to our special initialization, it is easy to verify that the statements of the lemma are satisfied for $t = 0$. Now suppose that equation 2.51-equation 2.56 are satisfied for t. Due to Proposition 10, one can write

$$\begin{aligned}
\|\boldsymbol{G}_{t+1}\| &\le \left(1 + 5\eta^2\rho^{2t} + 49\sqrt{r}\eta_0\rho^t\delta\right)\|\boldsymbol{G}_t\|, \\
&\le \exp\left(5\eta^2\rho^{2t} + 49\sqrt{r}\eta\rho^t\delta\right)\|\boldsymbol{G}_t\|, \\
&\le \|\boldsymbol{G}_0\|\prod_{s=0}^{t}\exp\left(5\eta^2\rho^{2s} + 49\sqrt{r}\eta_0\rho^s\delta\right) \\
&\le \|\boldsymbol{G}_0\|\exp\left(\sum_{s=0}^{t} 5\eta^2\rho^{2s} + 49\sqrt{r}\eta_0\rho^s\delta\right) \\
&\le \|\boldsymbol{G}_0\|\exp\left(\sum_{s=0}^{\infty} 5\eta^2\rho^{2s} + 49\sqrt{r}\eta_0\rho^s\delta\right) \\
&\le \|\boldsymbol{G}_0\|\exp\left(\frac{5\eta^2}{1-\rho^2} + \frac{49\sqrt{r}\eta\delta}{1-\rho}\right).
\end{aligned}$$

Due to $\rho = 1 - \Theta(\eta/(\kappa\log(1/\alpha)))$ and $\eta \lesssim 1/(\kappa\log(1/\alpha))$, we have

$$\frac{5\eta^2}{1-\rho^2} \lesssim \eta\kappa\log(1/\alpha) \le 1/2, \quad \text{and} \quad \frac{49\sqrt{r}\eta\delta}{1-\rho} \lesssim \sqrt{r}\kappa\delta\log(1/\alpha).$$

Combining the above inequalities leads

$$\|\boldsymbol{G}_{t+1}\| \le 2\|\boldsymbol{G}_0\|\alpha^{-\mathcal{O}(\sqrt{r}\kappa\delta)} \le 2\sqrt{2}\alpha^{1-\mathcal{O}(\sqrt{r}\kappa\delta)}\sqrt{\bar{\varphi}\delta},$$

where the last inequality is due to our special initialization technique and Lemma 6. The remaining bounds in Lemma 9 can be established similarly to Lemma 7. The details are omitted for brevity. □

2.F Proofs for Strict Saddles

2.F.1 Preliminaries

In Subsection 2.F.1.1, we review basic definitions and results from variational analysis that will be instrumental in developing our results. Then, in Subsections 2.F.1.2 to 2.F.1.5, we lay the groundwork for certain concentration results for different classes of random variables and random processes that will be used throughout our proofs.

2.F.1.1 Basic Properties from Variational Analysis

For a function $f : \mathbb{R}^n \to \mathbb{R}$, a point $\bar{\boldsymbol{x}}$ is called a *global optimal* if it corresponds to its global minimizer. Moreover, a point $\bar{\boldsymbol{x}}$ is called a *local optimal* if it corresponds to the minimum of $f(\boldsymbol{x})$ within an open ball centered at $\bar{\boldsymbol{x}}$. The directional derivative of f at point $\boldsymbol{x}$ in the feasible direction $\boldsymbol{d}$ is defined as

$$f'(\boldsymbol{x}, \boldsymbol{d}) = \lim_{t \to 0^+} \frac{f(\boldsymbol{x} + t\boldsymbol{d}) - f(\boldsymbol{x})}{t},$$

provided that the limit exists. If $f'(\boldsymbol{x}, \boldsymbol{d}) < 0$, then $\boldsymbol{d}$ is called a *descent direction*. For a locally Lipschitz function f, the *Clarke generalized directional derivative* at the point $\boldsymbol{x}$ in the feasible direction $\boldsymbol{d}$ is defined as

$$f^\circ(\boldsymbol{x}, \boldsymbol{d}) := \limsup_{\substack{\boldsymbol{y} \to \boldsymbol{x} \\ t \to 0^+}} \frac{f(\boldsymbol{y} + t\boldsymbol{d}) - f(\boldsymbol{y})}{t}$$

provided that the limit exists. It is a well-known fact that $f'(\boldsymbol{x}, \boldsymbol{d}) \leq f^\circ(\boldsymbol{x}, \boldsymbol{d})$ for every direction $\boldsymbol{d}$ if both $f'(\boldsymbol{x}, \boldsymbol{d})$ and $f^\circ(\boldsymbol{x}, \boldsymbol{d})$ exist [Clarke, 1975, Proposition 1.4]. A function f is called *subdifferentially regular* if $f'(\boldsymbol{x}, \boldsymbol{d}) = f^\circ(\boldsymbol{x}, \boldsymbol{d})$ for every point $\boldsymbol{x}$ and direction $\boldsymbol{d}$ [Clarke, 1990, Definition 2.3.4]. The *Clarke subdifferential* of f at $\boldsymbol{x}$ is defined as the following set (see [Clarke, 1975, Definition 1.1 and Proposition 1.4]):

$$\partial f(\boldsymbol{x}) := \{\boldsymbol{\psi} | f^\circ(\boldsymbol{x}, \boldsymbol{d}) \geq \langle \boldsymbol{\psi}, \boldsymbol{d} \rangle, \forall d \in \mathbb{R}^n\}. \tag{2.104}$$

A point $\bar{\boldsymbol{x}}$ is called *critical* if $\mathbf{0} \in \partial f(\bar{\boldsymbol{x}})$, or equivalently, $f^\circ(\bar{\boldsymbol{x}}, \boldsymbol{d}) \geq 0$ for every feasible direction $\boldsymbol{d}$. The following properties of the critical points are adapted from Li, So, and Ma [2020a] and will be used in our subsequent arguments.

Lemma 30. *For a real-valued locally Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$, we have:*

1. *Every local minimum of f is critical.*
2. *Given a point $\boldsymbol{x}$, suppose that $f'(\boldsymbol{x}, \boldsymbol{d}) \geq 0$ for every feasible $\boldsymbol{d} \in \mathbb{R}^n$. Then, $\boldsymbol{x}$ is a critical point of f.*
3. *Suppose that $f = \max_{1 \leq i \leq m} g_i$ for some smooth functions $g_1, \ldots, g_m$. Then, $f'(\boldsymbol{x}, \boldsymbol{d})$ exists for every point $\boldsymbol{x}$ and direction $\boldsymbol{d}$. Moreover, given some point $\boldsymbol{x}$, suppose that $f'(\boldsymbol{x}, \boldsymbol{d}) < 0$ for some direction $\boldsymbol{d} \in \mathbb{R}^n$. Then, $\boldsymbol{x}$ is not a critical point of f.*

Proof. The first property is a direct consequence of [Clarke, 1990, Proposition 2.3.2]. The second property follows from the basic inequality $f'(\boldsymbol{x}, \boldsymbol{d}) \leq f^\circ(\boldsymbol{x}, \boldsymbol{d})$ [Clarke, 1975, Proposition 1.4] and the definition of the critical points. Finally, we prove the third property. Note that the existence of $f'(\boldsymbol{x}, \boldsymbol{d})$ for every point $\boldsymbol{x}$ and direction $\boldsymbol{d}$ can be easily verified via the definition of f and directional derivative. Moreover, due to [Rockafellar and Wets, 2009, Example 7.28], f is differentiably regular, and hence, $f^\circ(\boldsymbol{x}, \boldsymbol{d}) = f'(\boldsymbol{x}, \boldsymbol{d})$ for every direction $\boldsymbol{d}$. This implies that x is not critical due to the existence of a direction $\boldsymbol{d}$ for which $f^\circ(\boldsymbol{x}, \boldsymbol{d}) = f'(\boldsymbol{x}, \boldsymbol{d}) < 0$. □

Utilizing the above properties, the following crucial lemma provides a sufficient condition for the non-criticality of a point $\bar{\boldsymbol{U}}$ for the ℓ_1-loss $f_{\ell_1}(\boldsymbol{U})$.

Lemma 31. *The directional derivative $f'_{\ell_1}(\bar{\boldsymbol{U}}, \Delta\boldsymbol{U})$ exists for every point $\boldsymbol{U}$ and direction $\Delta\boldsymbol{U}$. Moreover, given a point $\bar{\boldsymbol{U}}$, suppose that there exists a direction $\Delta\boldsymbol{U}$ such that $f'_{\ell_1}(\bar{\boldsymbol{U}}, \Delta\boldsymbol{U}) < 0$. Then, $\bar{\boldsymbol{U}}$ is not a critical point of $f_{\ell_1}(\boldsymbol{U})$.*

Proof. We first define $\mathcal{M}$ as the class of binary functions $\sigma : \{1, \ldots, m\} \to \{-1, +1\}$. It is easy to see that $f_{\ell_1}(\boldsymbol{U}) = \max_{\sigma \in \mathcal{M}} g_\sigma(\boldsymbol{U})$, where $g_\sigma(\boldsymbol{U}) = \sum_{i \in m} \sigma(i) \left(y_i - \left\langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{U}^\top \right\rangle\right)$. Note that $g_\sigma(\boldsymbol{U})$ is smooth for any choice of $\sigma \in \mathcal{M}$. Therefore, the third property of Lemma 30 can be invoked to complete the proof. □

Finally, following Davis and Drusvyatskiy [2022], we call a critical point $\bar{\boldsymbol{x}}$ of a non-differentiable function f *strict saddle* if there exists a direction v such that f decreases quadratically along v. This is formally defined as follows.

Definition 9 (Strict saddle point)**.** *A critical point $\bar{\boldsymbol{x}}$ of the function f is called strict saddle if there exist $\boldsymbol{v} \in \mathbb{R}^d, c > 0, \bar{\gamma} > 0$ such that, for every $0 \leq \gamma \leq \bar{\gamma}$, we have*

$$f(\bar{\boldsymbol{x}} + \gamma \boldsymbol{v}) - f(\bar{\boldsymbol{x}}) \leq -c\gamma^2.$$

2.F.1.2 Orlicz Random Variables and Processes

First, we define the notion of Orlicz norm (see [Wainwright, 2019, Section 5.6] and [Ledoux and Talagrand, 1991, Section 11.1] for more details).

Definition 10 (Orlicz function and Orlicz norm)**.** *A function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ is called an Orlicz function if ψ is convex, increasing, and satisfies*

$$\psi(0) = 0, \quad \psi(x) \to \infty \text{ as } x \to \infty.$$

For a given Orlicz function ψ, the Orlicz norm of a random variable X is defined as

$$\|X\|_\psi := \inf\{t > 0 : \mathbb{E}[\psi(|X|/t)] \leq 1\}.$$

We will use Orlicz functions and norms to study the concentration of both sub-Gaussian and sub-exponential random variables, which are defined as follows.

Definition 11 (Sub-Gaussian and sub-exponential random variables)**.** *Define the Orlicz function $\psi_p(x) = \exp\{x^p\} - 1$. A random variable X is called sub-Gaussian if $\|X\|_{\psi_2} < \infty$. Moreover, a random variable X is called sub-exponential if $\|X\|_{\psi_1} < \infty$. Accordingly, the quantities $\|X\|_{\psi_2}$ and $\|X\|_{\psi_1}$ refer to the sub-Gaussian norm and sub-exponential norm of X, respectively.*

Lemmas 32-35 establish basic properties of sub-Gaussian and sub-exponential random variables.

Lemma 32 (Tail bounds for sub-Gaussian and sub-exponential random variables [Wainwright, 2019, Section 5.6])**.** *For a random variable X with finite Orlicz norm $\|X\|_{\psi_p} < \infty$ where $\psi_p(x) = \exp\{x^p\} - 1$ and $p \geq 1$, we have*

$$\mathbb{P}\left(|X| \geq \delta\right) \leq 2\exp\left(-\frac{\delta^p}{\|X\|_{\psi_p}^p}\right).$$

Lemma 33 (Sum of independent random variables [Vershynin, 2018, Theorems 2.6.2 and 2.8.1])**.** *Let $X_1, \cdots, X_n$ be zero-mean independent random variables.*

- *If each X_i is sub-Gaussian with sub-Gaussian norm $\|X_i\|_{\psi_2}$, we have*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i\right| \geq \delta\right) \leq 2\exp\left(-\frac{n^2\delta^2}{\sum_{i=1}^n \|X_i\|_{\psi_2}^2}\right).$$

- *If each X_i is sub-exponential with sub-exponential norm $\|X_i\|_{\psi_1}$, we have*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \geq \delta\right) \leq 2\exp\left(-n\min\left\{\frac{\delta^2}{\max_i \|X_i\|_{\psi_1}^2}, \frac{\delta}{\max_i \|X_i\|_{\psi_1}}\right\}\right).$$

Lemma 34 (Lemma 2.6.8 in [Vershynin, 2018]). *Suppose that X is sub-Gaussian. We have $\|X - \mathbb{E}[X]\|_{\psi_2} \leq C\|X\|_{\psi_2}$ for some constant $C > 0$.*

Lemma 35 (Lemma 2.7.7 in [Vershynin, 2018]). *Let X, Y be sub-Gaussian random variables. Then XY is sub-exponential with its sub-exponential norm satisfying $\|XY\|_{\psi_1} \leq \|X\|_{\psi_2}\|Y\|_{\psi_2}$.*

Next, we introduce the notion of Orlicz process.

Definition 12 (Orlicz process, Definition 5.35 in [Wainwright, 2019]). *A set of zero-mean random variables $\{X_t, t \in \mathcal{T}\}$ (also known as a stochastic process) is an Orlicz process with Orlicz norm ψ_p (also known as ψ_p-process) with respect to a metric d if*

$$\|X_t - X_{t'}\|_{\psi_p} \leq d(t, t') \quad \text{for all } t, t' \in \mathcal{T}.$$

According to the above definition, the size of the set $\mathcal{T}$ may be infinite. To control the behavior of a ψ_p-process, we first need the notions of ϵ-covering and covering number.

Definition 13 (Covering and covering number). *A set $\mathcal{N}_\epsilon(\mathcal{T})$ is called an ϵ-covering on $(\mathcal{T}, d)$ if for every $t \in \mathcal{T}$, there exists $\pi(t) \in \mathcal{N}_\epsilon(\mathcal{T})$ such that $d(t, \pi(t)) \leq \epsilon$. The covering number $\mathcal{N}(\mathcal{T}, d, \epsilon)$ is defined as the smallest cardinality of an ϵ-covering for $(\mathcal{T}, d)$:*

$$\mathcal{N}(\mathcal{T}, d, \epsilon) := \inf\{|\mathcal{N}_\epsilon(\mathcal{T})| : \mathcal{N}_\epsilon(\mathcal{T}) \text{ is an } \epsilon\text{-covering}\}.$$

Theorem 12 (Concentration of Orlicz process, Theorem 5.36 in [Wainwright, 2019]). *Let $\{X_t, t \in \mathcal{T}\}$ be a ψ_p-process with respect to the metric d. Then, there is a universal constant $C > 0$ such that*

$$\mathbb{P}\left(\sup_{t\in\mathcal{T}} X_t \geq C\mathcal{J}_p(0, D) + \delta\right) \leq \exp(-\delta^p/D^p).$$

Here $D = \sup_{t,t'\in\mathcal{T}} d(t, t')$ is the diameter of the set $\mathcal{T}$ and $\mathcal{J}_p(\delta, D) = \int_\delta^D \log^{1/p}(1 + \mathcal{N}(\mathcal{T}, d, \epsilon))\, d\epsilon$ is called the generalized Dudley entropy integral.

By providing tractable upper bounds on D and $\mathcal{J}_p(\delta, D)$, one can control the supremum of a ψ_p-process. In our next section, we provide specific sub-classes of ψ_p-process for which these quantities can be controlled efficiently.

2.F.1.3 Concentration Bounds for Gaussian Processes over Grassmannian Manifold

In our subsequent analysis, Gaussian processes defined over the Grassmannian manifold, which are a specific category of Orlicz processes, will play a pivotal role. To set the stage, we start with the definition of a Gaussian process.

Definition 14 (Gaussian process)**.** *A random process $\{X_t\}_{t\in\mathcal{T}}$ is called a Gaussian process if, for any finite subset $\mathcal{T}_0 \subset \mathcal{T}$, the random vector $\{X_t\}_{t\in\mathcal{T}_0}$ has a normal distribution.*

Our next lemma characterizes the expectation of the supremum of a Gaussian process in terms of its covering number.

Theorem 13 (Sudakov's minoration inequality, Theorem 5.30 in [Wainwright, 2019])**.** *Let $\{X_t\}_{t\in\mathcal{T}}$ be a zero-mean Gaussian process. Then, for any $\epsilon \geq 0$, we have*

$$\mathbb{E}\left[\sup_{t\in\mathcal{T}} X_t\right] \geq \frac{\epsilon}{2}\sqrt{\log \mathcal{N}(\mathcal{T}, d, \epsilon)}.$$

Here, the metric d is defined as $d(t_1, t_2) = \left(\mathbb{E}\left[(X_{t_1} - X_{t_2})^2\right]\right)^{1/2}$ for any $t_1, t_2 \in \mathcal{T}$.

Lemma 36 (Concentration of Gaussian Process, Lemma 6.12 in [Van Handel, 2014])**.** *Let $\{X_t\}_{t\in\mathcal{T}}$ be a Gaussian process. Then $\sup_{t\in\mathcal{T}} X_t$ is sub-Gaussian with sub-Gaussian norm $\sup_{t\in\mathcal{T}} \|X_t\|_{\psi_2}$.*

Next, we provide the definition of Grassmanian manifold.

$$\mathcal{G}(n,p) = \{\boldsymbol{P} \in \mathbb{R}^{n\times n} : \boldsymbol{P}^\top = \boldsymbol{P}, \boldsymbol{P}^2 = \boldsymbol{P}, \mathrm{rank}(\boldsymbol{P}) = p\}. \tag{2.105}$$

Our next lemma characterizes the covering number of the Grassmannian manifold $\mathcal{G}(n,p)$.

Lemma 37 (Covering number for Grassmannian manifold, Proposition 8 in [Szarek, 1982])**.** *For a Grassmannian manifold $\mathcal{G}(n,p)$, define the metric $d(\boldsymbol{P}_1, \boldsymbol{P}_2) = \|\boldsymbol{P}_1 - \boldsymbol{P}_2\|_{\mathrm{F}}$ for any $\boldsymbol{P}_1, \boldsymbol{P}_2 \in \mathcal{G}(n,p)$. Then, for any $0 < \epsilon < \sqrt{2\min\{p, n-p\}}$, we have*

$$\left(\frac{c_1\sqrt{p}}{\epsilon}\right)^{p(n-p)} \leq \mathcal{N}(\mathcal{G}(n,p), d, \epsilon) \leq \left(\frac{c_2\sqrt{p}}{\epsilon}\right)^{p(n-p)},$$

for some constants $c_1, c_2 > 0$.

Lemma 38 (Lower bound of Gaussian process over Grassmannian manifold)**.** *Consider the Gaussian process $\sup_{\boldsymbol{X}\in\mathcal{G}(n,p)} \langle \boldsymbol{A}, \boldsymbol{X}\rangle$, where $\boldsymbol{A}$ is a standard Gaussian matrix and $p \leq n/2$.*

Then, with probability at least $1 - \exp(-\Omega(np))$*, we have*

$$\sup_{\boldsymbol{X} \in \mathcal{G}(n,p)} \langle \boldsymbol{A}, \boldsymbol{X} \rangle \geq \frac{c_3}{2} \sqrt{np^2},$$

for $c_3 = \sqrt{\log(c_1)/8}$*, where* c_1 *is the constant appeared in Lemma 37.*

Proof. We first use Sudakov's Minoration Inequality (Theorem 13) to characterize the expectation of the supremum:

$$\begin{aligned}
\mathbb{E}\left[\sup_{\boldsymbol{X} \in \mathcal{G}(n,p)} \langle \boldsymbol{A}, \boldsymbol{X} \rangle\right] &\geq \sup_{0<\epsilon} \frac{\epsilon}{2} \sqrt{\log \mathcal{N}\left(\mathcal{G}(n,p), \|\cdot\|_{\mathrm{F}}, \epsilon\right)} \\
&\overset{(a)}{\geq} \sup_{0<\epsilon \leq \sqrt{2p}} \frac{\epsilon}{2} \sqrt{p(n-p) \log\left(\frac{c_1 \sqrt{p}}{\epsilon}\right)} \\
&\overset{\text{set } \epsilon = \sqrt{p}}{\geq} \frac{1}{2} \sqrt{(n-p)p^2 \log(c_1)} \\
&\geq c_3 \sqrt{np^2}.
\end{aligned}$$

In (a), we invoked the lower bound on the covering number $\mathcal{N}\left(\mathcal{G}(n,p), \|\cdot\|_{\mathrm{F}}, \epsilon\right)$ from Lemma 37. Next, we apply Lemma 36 to control the deviation of the supremum from its expectation. To this goal, we first note that for any $\boldsymbol{X} \in \mathcal{G}(n,p)$, we have $\|\langle \boldsymbol{A}, \boldsymbol{X} \rangle\|_{\psi_2} \leq \|\boldsymbol{X}\|_{\mathrm{F}} = \sqrt{p}$. According to Lemma 36, $\sup_{\boldsymbol{X} \in \mathcal{G}(n,p)} \langle \boldsymbol{A}, \boldsymbol{X} \rangle$ is sub-Gaussian with norm $\sqrt{p}$. Due to Definition 11, we have

$$\begin{aligned}
\mathbb{P}\left(\sup_{\boldsymbol{X} \in \mathcal{G}(n,p)} \langle \boldsymbol{A}, \boldsymbol{X} \rangle - c_3 \sqrt{np^2} \leq -\delta\right) &\leq \mathbb{P}\left(\sup_{\boldsymbol{X} \in \mathcal{G}(n,p)} \langle \boldsymbol{A}, \boldsymbol{X} \rangle - \mathbb{E}\left[\sup_{\boldsymbol{X} \in \mathcal{G}(n,p)} \langle \boldsymbol{A}, \boldsymbol{X} \rangle\right] \leq -\delta\right) \\
&\leq \exp\left\{-\frac{\delta^2}{\left\|\sup_{\boldsymbol{X} \in \mathcal{G}(n,p)} \langle \boldsymbol{A}, \boldsymbol{X} \rangle\right\|_{\psi_2^2}}\right\} \\
&\leq \exp\left\{-\delta^2/p\right\}.
\end{aligned}$$

Upon setting $\delta = \frac{c_3}{2} \sqrt{np^2}$, we know that with probability at least $1 - \exp(-\Omega(np))$

$$\sup_{\boldsymbol{X} \in \mathcal{G}(n,p)} \langle \boldsymbol{A}, \boldsymbol{X} \rangle \geq \frac{c_3}{2} \sqrt{np^2},$$

which completes the proof. □

Our next lemma provides another useful bound on the supremum of another random process over Grassmannian manifold.

Lemma 39. *Suppose that $\{\boldsymbol{A}_i\}_{i=1}^m$ are i.i.d. standard Gaussian matrices and $\{w_i\}_{i=1}^m$ are non-negative weights. Then, we have with probability at least $1-\exp(-\Omega(np))$*

$$\sup_{\boldsymbol{X}\in\mathcal{G}(n,p)} \frac{1}{m}\sum_{i=1}^m w_i \left\langle \boldsymbol{A}_i, \boldsymbol{X}\right\rangle^2 \le \frac{cnp^2}{m}\sum_{i=1}^m w_i.$$

For some constant $c>0$.

Proof. We first verify that $\frac{1}{m}\sum_{i=1}^m w_i \left\langle \boldsymbol{A}_i, \boldsymbol{X}\right\rangle^2$ is indeed a ψ_1-process. To see this, for arbitrary two matrices $\boldsymbol{X}, \boldsymbol{X}' \in \mathcal{G}(n,p)$, we have

$$\begin{aligned}
\left\| \frac{1}{m}\sum_{i=1}^m w_i \left(\left\langle \boldsymbol{A}_i, \boldsymbol{X}\right\rangle^2 - \left\langle \boldsymbol{A}_i, \boldsymbol{X}'\right\rangle^2 \right) \right\|_{\psi_1} &\le \frac{1}{n}\sum_{i=1}^n w_i \left\| \left\langle \boldsymbol{A}_i, \boldsymbol{X}\right\rangle^2 - \left\langle \boldsymbol{A}_i, \boldsymbol{X}'\right\rangle^2 \right\|_{\psi_1} \\
&\le \frac{1}{m}\sum_{i=1}^m w_i \left\| \left\langle \boldsymbol{A}_i, \boldsymbol{X}+\boldsymbol{X}'\right\rangle \left\langle \boldsymbol{A}_i, \boldsymbol{X}-\boldsymbol{X}'\right\rangle \right\|_{\psi_1} \\
&\overset{(a)}{\le} \frac{1}{m}\sum_{i=1}^m w_i \left\| \left\langle \boldsymbol{A}_i, \boldsymbol{X}+\boldsymbol{X}'\right\rangle \right\|_{\psi_2} \left\| \left\langle \boldsymbol{A}_i, \boldsymbol{X}-\boldsymbol{X}'\right\rangle \right\|_{\psi_2} \\
&\le \frac{1}{m}\left(\sum_{i=1}^m w_i \right) \left\| \boldsymbol{X}+\boldsymbol{X}' \right\|_{\mathrm{F}} \left\| \boldsymbol{X}-\boldsymbol{X}' \right\|_{\mathrm{F}} \\
&\le \frac{2\sqrt{p}}{m}\left(\sum_{i=1}^m w_i \right) \left\| \boldsymbol{X}-\boldsymbol{X}' \right\|_{\mathrm{F}},
\end{aligned}$$

where in (a), we used Lemma 35. Therefore, $\frac{1}{m}\sum_{i=1}^m w_i \left\langle \boldsymbol{A}_i, \boldsymbol{X}\right\rangle^2$ is a ψ_1-process. On the other hand, we have $\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^m w_i \left\langle \boldsymbol{A}_i, \boldsymbol{X}\right\rangle^2\right] = \frac{1}{m}\sum_{i=1}^m w_i$. Next, we apply Theorem 12 to control the concentration of $\sup_{\boldsymbol{X}\in\mathcal{G}(n,p)} \frac{1}{m}\sum_{i=1}^m w_i \left\langle \boldsymbol{A}_i, \boldsymbol{X}\right\rangle^2$. For notational simplicity, we denote $\beta = \frac{2\sqrt{p}}{m}\sum_{i=1}^m w_i$. We have

$$\mathbb{P}\left(\sup_{\boldsymbol{X}\in\mathcal{G}(n,p)} \frac{1}{m}\sum_{i=1}^m \left(w_i \left\langle \boldsymbol{A}_i, \boldsymbol{X}\right\rangle^2 - w_i \right) \ge C\beta\mathcal{J}_1(0,D) + \beta\delta \right) \le \exp(-\delta/D), \tag{2.106}$$

where the diameter satisfies $D = \sup_{\boldsymbol{X},\boldsymbol{X}'\in\mathcal{G}(n,p)} \left\| \boldsymbol{X}-\boldsymbol{X}' \right\|_{\mathrm{F}} \le 2\sqrt{p}$ and the generalized

Dudley's integral can be bounded as

$$\begin{aligned}\mathcal{J}_1(0,D) &= \int_0^D \log\left(1+\mathcal{N}(\mathcal{G}(n,p),\|\cdot\|_{\mathrm{F}},\epsilon)\right)d\epsilon \\ &\overset{(a)}{\leq} \int_0^{2\sqrt{p}} 2(n-p)p\log\left(\frac{c_1\sqrt{p}}{\epsilon}\right)d\epsilon \\ &\leq \int_0^2 2(n-p)p^{3/2}\log\left(\frac{c_1}{\epsilon}\right)d\epsilon \\ &\leq c_4 np^{3/2}.\end{aligned}$$

Here $c_4 = 2\int_0^2 \log\left(\frac{c_1}{\epsilon}\right)d\epsilon$ is a universal constant. Then, upon choosing $\delta = c_5 np^{3/2}$ with $c_5 = Cc_4$ we have with probability at least $1 - e^{-\Omega(np)}$

$$\sup_{\boldsymbol{X}\in\mathcal{G}(n,p)} \frac{1}{m}\sum_{i=1}^m w_i \left\langle \boldsymbol{A}_i, \boldsymbol{X}\right\rangle^2 \leq \frac{2c_5 np^2}{m}\sum_{i=1}^m w_i.$$

This completes the proof. □

2.F.1.4 Gaussian Comparison Inequalities

In this subsection, we provide other useful lemmas comparing the extremum of two Gaussian processes.

Lemma 40 (Adapted from [Ledoux and Talagrand, 1991, Corollary 3.13.])**.** *For two Gaussian processes $X_{i,j,k}, Y_{i,j,k}$ with $i \in I, j \in J, k \in K$, suppose that*

- $\mathbb{E}[X_{i,j,k}^2] = \mathbb{E}[Y_{i,j,k}^2]$*, for all* $i \in I, j \in J, k \in K$*;*
- $\mathbb{E}[X_{i,j,k}X_{i,l,k}] \leq \mathbb{E}[Y_{i,j,k}Y_{i,l,k}]$*, for all* $i \in I, j, l \in J, k \in K$*;*
- $\mathbb{E}[X_{i,j,k}X_{l,m,k}] \geq \mathbb{E}[Y_{i,j,k}Y_{l,m,k}]$*, for all* $i \neq l \in I, j, m \in J, k \in K$*.*

Then, we have

$$\mathbb{E}\left[\min_{k\in K}\min_{i\in I}\max_{j\in J} Y_{i,j,k}\right] \leq \mathbb{E}\left[\min_{k\in K}\min_{i\in I}\max_{j\in J} X_{i,j,k}\right].$$

Our next lemma is a standard concentration bound on a Lipschitz function of a Gaussian random variable.

Lemma 41 (Proposition 5.34 in [Vershynin, 2010])**.** *Let f be a real valued Lipschitz function on $\mathbb{R}^n$ with Lipschitz constant K. Let $\boldsymbol{x}$ be the standard normal random vector in $\mathbb{R}^n$. Then for every $t \geq 0$ one has*

$$\mathbb{P}\left(f(\boldsymbol{x}) - \mathbb{E}[f(\boldsymbol{x})] \leq -t\right) \leq \exp(-t^2/2K^2).$$

Equipped with the above two lemmas, we present the following result, which will play an important role in our subsequent analysis. In what follows, a standard Gaussian matrix refers to a matrix whose entries are independently drawn from a standard Gaussian distribution.

Lemma 42. *Suppose that that integers d, d, m, k satisfy $d \geq cd$ and $m \leq \frac{(\sqrt{c}-1)^2}{4c} \cdot dk$ for some $c > 1$. Suppose that $\boldsymbol{A} \in \mathbb{R}^{m \times dd}$ is a standard Gaussian matrix. With probability at least $1 - \exp(-\Omega(dk))$, we have*

$$\min_{\boldsymbol{V} \in \mathcal{O}_{d \times k}} \{\sigma_{\min}(\boldsymbol{A} \cdot (\boldsymbol{V} \otimes \boldsymbol{I}_{d \times d}))\} \geq \frac{\sqrt{c}-1}{4\sqrt{c}} \sqrt{dk}.$$

Proof. Let us denote $\boldsymbol{A}_{\boldsymbol{V}} = \boldsymbol{A} \cdot (\boldsymbol{V} \otimes \boldsymbol{I}_{d \times d})$. By Min-max Theorem for singular values, we have

$$\sigma_{\min}(\boldsymbol{A}_{\boldsymbol{V}}) = \min_{\boldsymbol{u} \in \mathbb{R}^m, \|\boldsymbol{u}\|=1} \max_{W \in \mathbb{R}^{d \times k}, \|W\|_{\mathrm{F}}=1} \langle u, \boldsymbol{A}_{\boldsymbol{V}} \mathrm{vec}(W) \rangle .$$

Define $X_{u,W,V} = \langle u, \boldsymbol{A}_{\boldsymbol{V}} \mathrm{vec}(W) \rangle$ as a Gaussian process. Moreover, consider an auxiliary Gaussian process $Y_{u,W,V} = \frac{\sqrt{2}}{2} (\langle g, u \rangle + \langle GV, W \rangle)$ with $\boldsymbol{u} \in \mathbb{R}^m, V \in \mathbb{R}^{d \times k}, W \in \mathbb{R}^{d \times k}$, where $g \in \mathbb{R}^m$ and $G \in \mathbb{R}^{d \times d}$ are standard Gaussian vector and matrix. It is easy to verify that the Gaussian processes $X_{u,W,V}$ and $Y_{u,W,V}$ satisfy the conditions of Lemma 40. Therefore, we have

$$\begin{aligned}
\mathbb{E}\left[\min_{\boldsymbol{V} \in \mathcal{O}_{d \times k}} \sigma_{\min}(\boldsymbol{A}_{\boldsymbol{V}})\right] &= \mathbb{E}\left[\min_{\boldsymbol{V} \in \mathcal{O}_{d \times k}} \min_{\boldsymbol{u} \in \mathbb{R}^m, \|\boldsymbol{u}\|=1} \max_{W \in \mathbb{R}^{d \times k}, \|W\|_{\mathrm{F}}=1} X_{u,W,V}\right] \\
&\geq \mathbb{E}\left[\min_{\boldsymbol{V} \in \mathcal{O}_{d \times k}} \min_{\boldsymbol{u} \in \mathbb{R}^m, \|\boldsymbol{u}\|=1} \max_{W \in \mathbb{R}^{d \times k}, \|W\|_{\mathrm{F}}=1} Y_{u,W,V}\right] \\
&= \mathbb{E}\left[\min_{\boldsymbol{V} \in \mathcal{O}_{d \times k}} \max_{W \in \mathbb{R}^{d \times k}, \|W\|_{\mathrm{F}}=1} \langle GV, W \rangle\right] + \mathbb{E}\left[\min_{u \in \mathbb{R}^m, \|\boldsymbol{u}\|=1} \langle g, u \rangle\right] \\
&\geq \mathbb{E}\left[\min_{\boldsymbol{V} \in \mathcal{O}_{d \times k}} \|GV\|_{\mathrm{F}}\right] - \mathbb{E}\left[\|g\|\right] \\
&\geq \mathbb{E}\left[\sqrt{k}\sigma_{\min}(G)\right] - \mathbb{E}\left[\|g\|\right] \\
&\geq \sqrt{k}\left(\sqrt{d} - \sqrt{d}\right) - \sqrt{m} \\
&\geq \frac{\sqrt{c}-1}{2\sqrt{c}} \sqrt{dk}.
\end{aligned}$$

Let us define $f(A) = \min_{\boldsymbol{V} \in \mathcal{O}_{d \times k}} \sigma_{\min}(\boldsymbol{A}_{\boldsymbol{V}})$. It is easy to verify that $|f(A) - f(A')| \leq \|A - A'\|_{\mathrm{F}}$, for every $A, A' \in \mathbb{R}^{m \times dd}$. This implies that $f(A)$ is Lipschitz with constant 1. Therefore, Lemma 41 can be applied:

$$\mathbb{P}(f(A) - \mathbb{E}[f(A)] \leq -t) \leq \exp(-t^2/2) \implies \mathbb{P}\left(f(A) \geq \frac{\sqrt{c}-1}{2\sqrt{c}} \sqrt{dk} - t\right) \geq 1 - \exp(-t^2/2).$$

Substituting $t = \frac{\sqrt{c}-1}{4\sqrt{c}}\sqrt{dk}$ in the above inequality completes the proof. □

2.F.1.5 Other Useful Lemmas

Lemma 43. *Suppose that $A_1, \cdots, A_m$ are i.i.d. standard Gaussian matrices. For any sequence of scalars $w_1, \cdots, w_m$, the matrix $B = (\sum_{i=1}^m w_i^2)^{-1/2} \sum_{i=1}^m w_i \boldsymbol{A}_i$ is a standard Gaussian matrix provided that $\sum_{i=1}^m w_i^2 > 0$.*

Proof. It is easy to see that each element of $\sum_{i=1}^m w_i \boldsymbol{A}_i$ is i.i.d. and Gaussian with mean zero and variance $\sum_{i=1}^m w_i^2$. Therefore, the elements of B are all i.i.d. with standard Gaussian distribution. □

Lemma 44. *Suppose that $A_1, \cdots, A_m$ are i.i.d. standard Gaussian matrices and $w_1, \cdots, w_m$ are nonnegative scalars with $\sum_{i=1}^m w_i^2 > 0$. Then, for any $\delta > 0$ and a fixed matrix $X \in \mathbb{R}^{n \times n}$ with $\|\boldsymbol{X}\|_{\mathrm{F}} = 1$, we have*

$$\mathbb{P}\left(\left| \frac{1}{m} \sum_{i=1}^m w_i |\left\langle \boldsymbol{A}_i, \boldsymbol{X} \right\rangle| - \sqrt{\frac{2}{\pi}} \frac{1}{m} \sum_{i=1}^m w_i \right| \geq \delta \right) \leq 2 \exp\left(-\frac{m^2 \delta^2}{C \sum_{i=1}^m w_i^2} \right).$$

Proof. Note that $\left\langle \boldsymbol{A}_i, \boldsymbol{X} \right\rangle \overset{iid}{\sim} \mathcal{N}(0,1)$ and $\mathbb{E}[|\left\langle \boldsymbol{A}_i, \boldsymbol{X} \right\rangle|] = \sqrt{\frac{2}{\pi}}$. Due to Lemma 34, we have $\| |\left\langle \boldsymbol{A}_i, \boldsymbol{X} \right\rangle| - \mathbb{E}[|\left\langle \boldsymbol{A}_i, \boldsymbol{X} \right\rangle|] \|_{\psi_2} \leq \| |\left\langle \boldsymbol{A}_i, \boldsymbol{X} \right\rangle| \|_{\psi_2}$. Then final result follows from Lemma 33. □

Lemma 45. *Suppose that $A_1, \cdots, A_m$ are i.i.d. standard Gaussian matrices of the shape $d \times d$. We have*

$$\mathbb{P}\left(\max_{1 \leq i \leq m} \|\boldsymbol{A}_i\|_{\mathrm{F}} \geq \sqrt{2dd} \right) \leq \exp\left(\log m - \Omega(dd) \right).$$

Proof. For any fixed i and $1 \leq \alpha \leq d, 1 \leq \beta \leq d$, $(\boldsymbol{A}_i)[\alpha, \beta]^2 - 1$ is a sub-exponential random variable with zero mean and parameter 1. Therefore, an application of Lemma 33 implies that $\mathbb{P}(\|\boldsymbol{A}_i\|_{\mathrm{F}} \geq \sqrt{1+\delta}\sqrt{dd}) \leq \exp(-\Omega(dd))$. Setting $\delta = 1$ followed by a union bound leads to the final result. □

Lemma 46 (Corollary 5.35 in [Vershynin, 2010])**.** *Consider a random matrix $X \in \mathbb{R}^{N \times n}$ with i.i.d. entries drawn from $\mathcal{N}(0,1)$ and $N > n$. Then for every $t > 0$, with probability at least $1 - 2\exp\{-t^2/2\}$ one has*

$$\sigma_{\min}(X) \geq \sqrt{N} - \sqrt{n} - t.$$

Lemma 47 (Concentration of binomial random variable, [Chung and Lu, 2006])**.** *Suppose that X has a binomial distribution with parameters (m, p). Then, we have*

$$\mathbb{P}\left(|X - pm| \geq \delta\right) \leq 2\exp(-\delta^2/2pm).$$

In particular, when choosing $\delta = \frac{pm}{2}$, we have that with probability at least $1 - 2\exp(pm/4)$,

$$X \geq \frac{pm}{2}, \quad \textit{and} \quad X \leq \frac{3pm}{2}.$$

Lemma 48. *Suppose that X has a binomial distribution with parameters (m, p) and $0 \leq p < 1/2$. Then, with probability at least $1 - \exp(-\Omega((1-2p)^2 m))$, we have*

$$X \leq \max\left\{\frac{3}{8}, \frac{1-(1-2p)^2}{2}\right\} \cdot m.$$

Proof. We consider two cases. First, suppose that $1/4 \leq p < 1/2$. Define $\eta = 1 - 2p$. According to Lemma 47, we have

$$\begin{aligned}
&\mathbb{P}\left(X \leq pm(1+\eta)\right) \geq 1 - \exp(-\Omega(\eta^2 m)) \\
&\implies \mathbb{P}\left(X \leq \frac{1-(1-2p)^2}{2} m\right) \geq 1 - \exp(-\Omega((1-2p)^2 m)),
\end{aligned}$$

where the last inequality follows from the fact that $pm(1+\eta) = \frac{1-(1-2p)^2}{2} m$. On the other hand, for $0 \leq p \leq 1/4$, Lemma 47 implies that

$$\mathbb{P}\left(X \leq 3m/8\right) \geq 1 - \exp(-\Omega(m)).$$

Combining the above two cases leads to the desired result. □

Armed with the concentration bounds discussed above, we are now prepared to present our main proofs.

2.F.2 Symmetric Case: Sub-optimality via Parametric Second-order Perturbations

At the core of our analysis lies a class of parametric second-order perturbations that can be used to show the sub-optimality of the true solutions for both symmetric matrix sensing and symmetric matrix completion. Let $\boldsymbol{X}^\star = \boldsymbol{V}^\star \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top}$ be the eigen-decomposition of $\boldsymbol{X}^\star$, where $\boldsymbol{V}^\star \in \mathbb{R}^{d\times r}$ is an orthonormal matrix and $\boldsymbol{\Sigma}^\star \in \mathbb{R}^{r\times r}$ is a diagonal matrix collecting the nonzero eigenvalues of $\boldsymbol{X}^\star$. Recall that $\mathcal{O}_{r\times k} := \{\boldsymbol{R} \in \mathbb{R}^{r\times k} : \boldsymbol{R}\boldsymbol{R}^\top = \boldsymbol{I}_{r\times r}\}$. The following

lemma provides another characterization of the set $\mathcal{U} = \{\boldsymbol{U} \in \mathbb{R}^{d\times k} : \boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{X}^\star\}$.

Lemma 49. *We have*

$$\boldsymbol{U} \in \mathcal{U} \qquad \Longleftrightarrow \qquad \boldsymbol{U} = \boldsymbol{V}^*\boldsymbol{\Sigma}^{\star 1/2}\boldsymbol{R} \ \text{ for some } \boldsymbol{R} \in \mathcal{O}_{r\times k}.$$

Proof. If $W = \boldsymbol{V}^*\boldsymbol{\Sigma}^{\star 1/2}\boldsymbol{R}$ for some $\boldsymbol{R} \in \mathcal{O}_{r\times k}$, then $\boldsymbol{U}\boldsymbol{U}^\top = X^\star$ and $\boldsymbol{U} \in \mathcal{U}$. Now, suppose that $\boldsymbol{U} \in \mathcal{U}$. We have

$$\begin{aligned}
\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{V}^\star\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top} &\implies \boldsymbol{\Sigma}^{\star -1/2}\boldsymbol{V}^{\star\top}\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{V}^\star\boldsymbol{\Sigma}^{\star -1/2} = \boldsymbol{I}_{r\times r} \\
&\implies \boldsymbol{\Sigma}^{\star -1/2}\boldsymbol{V}^{\star\top}\boldsymbol{U} = \boldsymbol{R} \text{ for some } \boldsymbol{R} \in \mathcal{O}_{r\times k} \\
&\implies \boldsymbol{V}^\star\boldsymbol{V}^{\star\top}\boldsymbol{U} = \boldsymbol{V}^\star\boldsymbol{\Sigma}^{\star 1/2}\boldsymbol{R} \text{ for some } \boldsymbol{R} \in \mathcal{O}_{r\times k}.
\end{aligned}$$

Moreover, let $\boldsymbol{V}_\perp^\star$ be the orthogonal complement of $\boldsymbol{V}^\star$. We have

$$\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{V}^\star\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top} \implies \boldsymbol{V}_\perp^{\star\top}\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{V}_\perp^\star = \boldsymbol{0}_{(d-r)\times(d-r)} \implies \boldsymbol{V}_\perp^{\star\top}\boldsymbol{U} = \boldsymbol{0} \implies \boldsymbol{V}_\perp^\star\boldsymbol{V}_\perp^{\star\top}\boldsymbol{U} = \boldsymbol{0}.$$

Combining the above two equalities, we have

$$\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{V}^\star\boldsymbol{V}^{\star\top}\boldsymbol{U} + \boldsymbol{V}_\perp^\star\boldsymbol{V}_\perp^{\star\top}\boldsymbol{U} = \boldsymbol{V}^\star\boldsymbol{\Sigma}^{\star 1/2}\boldsymbol{R}$$

for some $\boldsymbol{R} \in \mathcal{O}_{r\times k}$. This completes the proof. □

Based on the above lemma, the set of true solutions can be characterized as $\mathcal{U} = \{\boldsymbol{V}^\star\boldsymbol{\Sigma}^{\star 1/2}\boldsymbol{R} : \boldsymbol{R} \in \mathcal{O}_{r\times k}\}$. This characterization of $\mathcal{U}$ will be useful in our subsequent analysis. Our goal is to show that, for every $\boldsymbol{U}^\star \in \mathcal{U}$, there exists a descent direction $\Delta\boldsymbol{U} \in \mathcal{B}$ such that $f_{\ell_1}(\boldsymbol{U}^\star + \gamma\Delta\boldsymbol{U}) - f_{\ell_1}(\boldsymbol{U}^\star) \le -\Omega(\gamma^2)$ for every $0 < \gamma \le \bar{\gamma}$. To this end, for any $\boldsymbol{U}^\star \in \mathcal{U}$, we consider the following set of *parametric second-order perturbations*:

$$\widehat{\mathcal{U}}_{\boldsymbol{X}^\star} := \left\{U\boldsymbol{R} : U \in \mathbb{R}^{d\times(k-r)}, \|\boldsymbol{U}\|_{\mathrm{F}} \le 1, R \in \mathcal{O}_{(k-r)\times k}, \boldsymbol{U}^\star R^\top = 0\right\}.$$

Note that any perturbation $\Delta\boldsymbol{U} = U\boldsymbol{R} \in \widehat{\mathcal{U}}_{\boldsymbol{X}^\star}$ is indeed second-order since:

$$\left\|(\boldsymbol{U}^\star + \gamma\Delta\boldsymbol{U})(\boldsymbol{U}^\star + \gamma\Delta\boldsymbol{U})^\top - \boldsymbol{U}^\star\boldsymbol{U}^{\star\top}\right\|_{\mathrm{F}} = \gamma^2\left\|\boldsymbol{U}\boldsymbol{U}^\top\right\|_{\mathrm{F}} \le \gamma^2.$$

Evidently, we have $\widehat{\mathcal{U}}_{\boldsymbol{X}^\star} \subset \mathcal{B}$. Moreover, $\widehat{\mathcal{U}}_{\boldsymbol{X}^\star}$ is non-empty since, in light of $\mathcal{U} = \{\boldsymbol{V}^\star\boldsymbol{\Sigma}^{\star 1/2}\boldsymbol{R} : \boldsymbol{R} \in \mathcal{O}_{r\times k}\}$, there always exists $\boldsymbol{R} \in \mathcal{O}_{(k-r)\times k}$ such that $\boldsymbol{U}^\star R^\top = 0$. Therefore, to prove our main result, it suffices to find $\bar{\boldsymbol{U}} \in \mathcal{B}^{k-r} := \{U : U \in \mathbb{R}^{d\times(k-r)}, \|\boldsymbol{U}\|_{\mathrm{F}} \le 1\}$ such that the direction defined as $\Delta\boldsymbol{U} = \bar{\boldsymbol{U}}\boldsymbol{R} \in \widehat{\mathcal{U}}_{\boldsymbol{X}^\star}$ achieves the desired decrease in the objective

function. To this goal, note that, for every $\Delta \boldsymbol{U} = \boldsymbol{U}\boldsymbol{R} \in \widehat{\mathcal{U}}_{\boldsymbol{X}^\star}$, we have

$$f_{\ell_1}(\boldsymbol{U}^\star + \gamma \Delta \boldsymbol{U}) - f_{\ell_1}(\boldsymbol{U}^\star) = \frac{1}{m}\sum_{i=1}^{m} \left|\left\langle \boldsymbol{A}_i, \gamma^2 \boldsymbol{U}\boldsymbol{U}^\top \right\rangle - \epsilon_i\right| - |\epsilon_i|. \tag{2.107}$$

Therefore, we need to prove the existence of $\bar{\boldsymbol{U}} \in \mathcal{B}^{k-r}$ for which the right-hand-side of Equation (2.107) archives the desired bound. This can be easily done for the symmetric matrix completion, as shown below.

2.F.3 Proof of Theorem 9

Recall that $\mathcal{S}$ is the index set of the noisy measurements, and define $\bar{\mathcal{S}} = [m] \backslash \mathcal{S}$ as the set of *clean* measurements. Moreover, define $\mathcal{S}_t = \{i : |\epsilon_i| \geq t_0\}$. The set $\mathcal{S}_t$ denotes the set of measurements for which the magnitude of noise exceeds t_0.

Lemma 50. *Suppose that* $\gamma \leq \sqrt{\frac{t_0}{\max_i\{\|\boldsymbol{A}_i\|_{\mathrm{F}}\}}}$. *For every* $\boldsymbol{U} \in \mathcal{B}^{k-r}$, *we have*

$$\frac{1}{m}\sum_{i=1}^{m} \left|\left\langle \boldsymbol{A}_i, \gamma^2 \boldsymbol{U}\boldsymbol{U}^\top \right\rangle - \epsilon_i\right| - |\epsilon_i| \leq \gamma^2 \cdot \left(A\left(\boldsymbol{U}\boldsymbol{U}^\top\right) + B\left(\boldsymbol{U}\boldsymbol{U}^\top\right)\right),$$

where

$$A\left(\boldsymbol{U}\boldsymbol{U}^\top\right) = \frac{1}{m}\sum_{i\in\mathcal{S}_t} -\operatorname{sign}(\epsilon_i)\left\langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{U}^\top\right\rangle, \qquad B\left(\boldsymbol{U}\boldsymbol{U}^\top\right) = \frac{1}{m}\sum_{i\notin\mathcal{S}_t} \left|\left\langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{U}^\top\right\rangle\right|.$$

Proof. We have

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m} \left|\left\langle \boldsymbol{A}_i, \gamma^2 \boldsymbol{U}\boldsymbol{U}^\top \right\rangle - \epsilon_i\right| - |\epsilon_i| \\
&= \underbrace{\frac{1}{m}\sum_{i\in\mathcal{S}_t} \left|\left\langle \boldsymbol{A}_i, \gamma^2 \boldsymbol{U}\boldsymbol{U}^\top \right\rangle - \epsilon_i\right| - |\epsilon_i|}_{E_1} + \underbrace{\frac{1}{m}\sum_{i\notin\mathcal{S}_t} \left|\left\langle \boldsymbol{A}_i, \gamma^2 \boldsymbol{U}\boldsymbol{U}^\top \right\rangle - \epsilon_i\right| - |\epsilon_i|}_{E_2}.
\end{aligned}$$

To bound E_2, we apply triangle inequality, which leads to

$$\frac{1}{m}\sum_{i\notin\mathcal{S}_t} \left|\left\langle \boldsymbol{A}_i, \gamma^2 \boldsymbol{U}\boldsymbol{U}^\top \right\rangle - \epsilon_i\right| - |\epsilon_i| \leq \frac{1}{m}\sum_{i\notin\mathcal{S}_t} \left|\left\langle \boldsymbol{A}_i, \gamma^2 \boldsymbol{U}\boldsymbol{U}^\top \right\rangle\right| = \gamma^2 \cdot B\left(\boldsymbol{U}\boldsymbol{U}^\top\right).$$

Next, we provide an upper bound for E_1. To this goal, we write, for every $i \in \mathcal{S}_t$,

$$\left|\left\langle \boldsymbol{A}_i, \gamma^2 \boldsymbol{U}\boldsymbol{U}^\top \right\rangle - \epsilon_i\right| - |\epsilon_i| = -\gamma^2 \operatorname{sign}(\epsilon_i)\left\langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{U}^\top\right\rangle,$$

where the equality follows from $\left|\left\langle \boldsymbol{A}_i, \gamma^2 \boldsymbol{U}\boldsymbol{U}^\top \right\rangle\right| \leq \gamma^2 \left\|\boldsymbol{A}_i\right\|_{\mathrm{F}} \left\|\boldsymbol{U}\boldsymbol{U}^\top\right\|_{\mathrm{F}} \leq \max_i\{\|\boldsymbol{A}_i\|_{\mathrm{F}}\}\gamma^2 \leq t_0$. This implies that $E_1 = \gamma^2 \cdot A\left(\boldsymbol{U}\boldsymbol{U}^\top\right)$, thereby completing the proof. □

We next show that $A\left(\boldsymbol{U}\boldsymbol{U}^\top\right) + B\left(\boldsymbol{U}\boldsymbol{U}^\top\right) \leq -\Omega(1)$ for some $\boldsymbol{U} \in \mathcal{B}^{k-r}$. Note that $B\left(\boldsymbol{U}\boldsymbol{U}^\top\right) \geq 0$ for every $\boldsymbol{U} \in \mathcal{B}^{k-r}$, so our hope is to show that $A\left(\boldsymbol{U}\boldsymbol{U}^\top\right)$ can take sufficiently negative value to dominate $B\left(\boldsymbol{U}\boldsymbol{U}^\top\right)$. As will be shown in our next lemma, this can be established with a high probability. In particular, we show that, when the measurement matrices follow the matrix sensing model, there exists, with high probability, $\boldsymbol{U} \in \mathcal{B}^{k-r}$ and coefficients $C_A > C_B \geq 0$ such that

$$A\left(\boldsymbol{U}\boldsymbol{U}^\top\right) \leq -C_A, \quad B\left(\boldsymbol{U}\boldsymbol{U}^\top\right) \leq C_B, \tag{2.108}$$

Lemma 51. *Define $r_0 = \min\{k-r, d/2\}$. Suppose that the measurement matrices are i.i.d. standard Gaussian. Conditioned on the noise $\{\epsilon_i\}_{i=1}^m$ and with probability at least $1 - \exp(-\Omega(dr_0))$ over the randomness of the measurement matrices, there exists $\boldsymbol{U} \in \mathcal{B}^{k-r}$ such that Equation (2.116) holds with*

$$C_A = c\sqrt{\frac{dr_0|\mathcal{S}_t|}{m^2}}, \qquad C_B = 1,$$

for some constant $c > 0$.

The proof makes extensive use of the concentration bounds for Gaussian processes over the Grassmanina manifold, as discussed in Section 2.F.1.3. Before presenting the proof of Lemma 51, we show its application in the proof of Theorem 9.

Proof of Theorem 9. For every $\boldsymbol{U}^\star \in \mathcal{U}$, there exists $\boldsymbol{U} \in \mathcal{B}^{k-r}$ such that

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^m \left|\left\langle \boldsymbol{A}_i, \gamma^2 \boldsymbol{U}\boldsymbol{U}^\top \right\rangle - \epsilon_i\right| - |\epsilon_i| &\overset{(a)}{\leq} \gamma^2 \cdot \left(A\left(\boldsymbol{U}\boldsymbol{U}^\top\right) + B\left(\boldsymbol{U}\boldsymbol{U}^\top\right)\right) \\
&\overset{(b)}{\leq} -(C_A - C_B)\gamma^2 \\
&\overset{(c)}{\leq} -\left(c\sqrt{\frac{dr_0|\mathcal{S}_t|}{m^2}} - 1\right)\gamma^2 \\
&\overset{(d)}{\leq} -\left(c\sqrt{\frac{dr_0 pp_0}{2m}} - 1\right)\gamma^2 \\
&\leq -\frac{c}{2}\sqrt{\frac{dr_0 pp_0}{2m}}\gamma^2
\end{aligned}$$

with probability at least $1 - \exp(-\Omega(dr_0)) - \exp(-\Omega(pp_0m))$, provided that

$$\gamma^2 \leq \frac{t_0}{\max_i\{\|\boldsymbol{A}_i\|_{\mathrm{F}}\}}, \quad \text{and} \quad m \leq \frac{c^2}{8} \cdot pp_0dr_0.$$

In the above inequality, (a) follows from Lemma 50, (b) and (c) follow from Lemma 51, and (d) follows from the concentration of the binomial random variables (Lemma 47), which implies $|\mathcal{S}_t| \geq pp_0m/2$, with probability at least $1 - \exp(-\Omega(pp_0m))$. On the other hand, Lemma 45 implies that $\max_i\{\|\boldsymbol{A}_i\|_{\mathrm{F}}\} \leq \sqrt{2}d$ with probability at least $1-\exp(-\Omega(d^2))$. This implies that

$$\frac{t_0}{\max_i\{\|\boldsymbol{A}_i\|_{\mathrm{F}}\}} \geq \frac{t_0}{\sqrt{2d}} \geq \gamma^2$$

with probability at least $1-\exp(-\Omega(d^2))$, where the last inequality follows from our assumed upper bound on γ^2. For this choice of $\boldsymbol{U}$, the direction $\Delta\boldsymbol{U} = \boldsymbol{U}\boldsymbol{R} \in \widehat{\mathcal{U}}_{\boldsymbol{X}^\star}$ achieves the desired result in light of Equation (2.107), thereby completing the proof. □

Next, we present the proof of Lemma 51.

Proof of Lemma 51. Let $\mathcal{C} = (1/\sqrt{r_0})\mathcal{G}(d, r_0)$, where $\mathcal{G}(d, r_0)$ is the Grassmanian manifold defined as Equation (2.105) and $r_0 = \min\{k - r, d/2\}$. It is easy to see that if $X \in \mathcal{C}$, then $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{U}^\top$ for some $\boldsymbol{U} \in \mathcal{B}^{k-r}$. Consider $\bar{\boldsymbol{X}} = \arg\min_{X\in\mathcal{C}} A(\boldsymbol{X})$ and let $\bar{\boldsymbol{X}} = \bar{\boldsymbol{U}}\bar{\boldsymbol{U}}^\top$. Note that $\bar{\boldsymbol{U}} \in \mathcal{B}^{k-r}$. We prove the desired inequality is attained at $\bar{\boldsymbol{U}}$.

Bounding $B(\bar{\boldsymbol{X}})$: According its definition, $\bar{\boldsymbol{x}}$ is independent of $\{\boldsymbol{A}_i\}_{i\notin\mathcal{S}_t}$. Let $\{\boldsymbol{B}_i\}_{i\in\mathcal{S}_t}$ be independent copies of the measurement matrices. We have

$$B(\bar{\boldsymbol{X}}) = \frac{1}{m}\sum_{i\notin\mathcal{S}_t} \left|\left\langle \boldsymbol{A}_i, \bar{\boldsymbol{X}}\right\rangle\right| \leq \frac{1}{m}\sum_{i\notin\mathcal{S}_t} |\left\langle \boldsymbol{A}_i, \bar{\boldsymbol{X}}\right\rangle| + \frac{1}{m}\sum_{i\in\mathcal{S}_t} |\left\langle \boldsymbol{B}_i, \bar{\boldsymbol{X}}\right\rangle| \leq 1$$

with probablity at least $1-\exp(-\Omega(m))$, where in the last inequality we invoked Lemma 44 with $\delta = 1 - \sqrt{\frac{2}{\pi}}$ and $w_i = 1$.

Bounding $A(\bar{\boldsymbol{X}})$. Recall that

$$A(\bar{\boldsymbol{X}}) = \inf_{\boldsymbol{X}\in\mathcal{C}} \left\{ \frac{1}{m}\sum_{i\in\mathcal{S}_t} -\operatorname{sign}(\epsilon_i)\left\langle \boldsymbol{A}_i, \boldsymbol{X}\right\rangle \right\}.$$

Note that, unlike $B(\bar{\boldsymbol{X}})$, the choice of $\bar{\boldsymbol{X}}$ depends on the measurement matrices involved in $A(\bar{\boldsymbol{X}})$. To control this dependency, consider $\tilde{\boldsymbol{A}} = |\mathcal{S}_t|^{-1/2}\sum_{i\in\mathcal{S}_t}\operatorname{sign}(\epsilon_i)\boldsymbol{A}_i$. Invoking

Lemma 43 with $\omega_i = \mathrm{sign}(\epsilon_i)$, we know that $\tilde{\boldsymbol{A}}$ is a standard Gaussian matrix and hence

$$\begin{aligned} A(\bar{\boldsymbol{X}}) &= \frac{\sqrt{|\mathcal{S}_t|}}{m} \inf_{\boldsymbol{X}\in\mathcal{C}} \left\langle -\tilde{\boldsymbol{A}}, \boldsymbol{X} \right\rangle \\ &= -\frac{\sqrt{|\mathcal{S}_t|}}{m} \sup_{\boldsymbol{X}\in\mathcal{C}} \left\langle \tilde{\boldsymbol{A}}, \boldsymbol{X} \right\rangle \\ &= -\frac{\sqrt{|\mathcal{S}_t|}}{m} \frac{1}{\sqrt{r_0}} \sup_{\boldsymbol{X}\in\mathcal{G}(d,r_0)} \left\langle \tilde{\boldsymbol{A}}, \boldsymbol{X} \right\rangle \\ &\le -c\sqrt{\frac{dr_0|\mathcal{S}_t|}{m^2}}, \end{aligned}$$

with probability at least $1 - \exp -\Omega(dr_0)$, where the last inequality follows from Lemma 38. This completes the proof of this lemma. □

2.F.4 Matching Lower Bounds

In this section, we present the proof of Theorem 10. We start with the proof of the second statement of Theorem 10. In particular, we show that if $m \gtrsim dk/(1-2p)^4$, the true solutions coincide with the global optima. Our proof is nearly the same as the one of [Ding et al., 2021, Theorem 2.3]. We include here for completeness.

It suffices to show that for any $\boldsymbol{U} \notin \mathcal{U}$ and $\boldsymbol{U}^\star \in \mathcal{U}$, we have $f_{\ell_1}(\boldsymbol{U}) - f_{\ell_1}(\boldsymbol{U}^\star) > 0$. To this end, let $\Delta\boldsymbol{X} = \boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^\star\boldsymbol{U}^{\star\top}$. Then, one can write

$$\begin{aligned} f_{\ell_1}(\boldsymbol{U}) - f_{\ell_1}(\boldsymbol{U}^\star) &\ge \frac{1}{m}\sum_{i\in\bar{\mathcal{S}}} |\langle \boldsymbol{A}_i, \Delta\boldsymbol{X}\rangle| + \frac{1}{m}\sum_{i\in\mathcal{S}} |\langle \boldsymbol{A}_i, \Delta\boldsymbol{X}\rangle - \epsilon_i| - |\epsilon_i| \\ &\ge \underbrace{\frac{1}{m}\sum_{i\in\bar{\mathcal{S}}} |\langle \boldsymbol{A}_i, \Delta\boldsymbol{X}\rangle| - \frac{1}{m}\sum_{i\in\mathcal{S}} |\langle \boldsymbol{A}_i, \Delta\boldsymbol{X}\rangle|}_{F} \end{aligned}$$

where the second inequality follows from the triangle inequality. We will show that if the corruption probability p is strictly less than half, then $F > 0$ with high probability. To this goal, we use an important property of Gaussian measurements, called *ℓ_1/ℓ_2-restricted isometry property* (ℓ_1/ℓ_2-RIP), which is provided in the following lemma.

Lemma 52 (ℓ_1/ℓ_2-RIP [Li et al., 2020b, Charisopoulos, Chen, Davis, Díaz, Ding, and Drusvyatskiy, 2021])**.** *Suppose that the measurement matrices are i.i.d. standard Gaussian. For all rank-$2r'$ matrices $\boldsymbol{X} \in \mathbb{R}^{d\times d}$ and any $\delta > 0$[1], we have with probability at*

[1]The original statement only covers the case $0 < \delta < \sqrt{2/\pi}$. However, the proof appeared in [Li et al., 2020b] indeed works for $\delta \ge \sqrt{2/\pi}$.

least $1-\exp(-\Omega(m\delta^2))$*:*

$$\left(\sqrt{\frac{2}{\pi}}-\delta\right)\|\boldsymbol{X}\|_{\mathrm{F}} \le \frac{1}{m}\sum_{i=1}^{m}|\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle| \le \left(\sqrt{\frac{2}{\pi}}+\delta\right)\|\boldsymbol{X}\|_{\mathrm{F}},$$

provided that $m \ge cdr'/\delta^2$ *for some constant* $c > 0$*.*

Equipped with the above lemma, we are ready to present the proof of the second statement of Theorem 10.

Proof of the second statement of Theorem 10. Our goal is to show that, if the measurement matrices follow the matrix sensing model, then with an overwhelming probability, we have $F \gtrsim \|\Delta \boldsymbol{X}\|_{\mathrm{F}}$. This in turn implies that if $\Delta \boldsymbol{X} \neq 0$ (or equivalently $\boldsymbol{U} \notin \mathcal{U}$), then $f_{\ell_1}(\boldsymbol{U}) - f_{\ell_1}(\boldsymbol{U}^\star) > 0$. Notice that $\mathrm{rank}(\Delta \boldsymbol{X}) \le k+r$. Let $\mathcal{S}_{k+r}^{d\times d} = \{\boldsymbol{X} : \boldsymbol{X} \in \mathbb{R}^{d\times d}, \mathrm{rank}(\boldsymbol{X}) \le k+r, \|\boldsymbol{X}\|_{\mathrm{F}} = 1\}$. It is easy to see that a sufficient condition for $F \gtrsim \|\Delta \boldsymbol{X}\|_{\mathrm{F}}$ is to have

$$\min_{\Delta \boldsymbol{X} \in \mathcal{S}_{k+r}^{d\times d}} \left\{ \frac{1}{m}\sum_{i\in\bar{\mathcal{S}}}|\langle \boldsymbol{A}_i, \Delta \boldsymbol{X}\rangle| - \frac{1}{m}\sum_{i\in\mathcal{S}}|\langle \boldsymbol{A}_i, \Delta \boldsymbol{X}\rangle| \right\} \gtrsim 1.$$

To prove the above inequality, we write

$$\min_{\Delta \boldsymbol{X} \in \mathcal{S}_{k+r}^{d\times d}} \left\{ \frac{1}{m}\sum_{i\in\bar{\mathcal{S}}}|\langle \boldsymbol{A}_i, \Delta \boldsymbol{X}\rangle| - \frac{1}{m}\sum_{i\in\mathcal{S}}|\langle \boldsymbol{A}_i, \Delta \boldsymbol{X}\rangle| \right\} \ge \underbrace{\min_{\Delta \boldsymbol{X} \in \mathcal{S}_{k+r}^{d\times d}} \left\{ \frac{1}{m}\sum_{i\in\bar{\mathcal{S}}}|\langle \boldsymbol{A}_i, \Delta \boldsymbol{X}\rangle| \right\}}_{F_1} - \underbrace{\max_{\Delta \boldsymbol{X} \in \mathcal{S}_{k+r}^{d\times d}} \left\{ \frac{1}{m}\sum_{i\in\mathcal{S}}|\langle \boldsymbol{A}_i, \Delta \boldsymbol{X}\rangle| \right\}}_{F_2}.$$

We next show that $F_1 - F_2 > 0$. To this goal, we provide separate lower and upper bounds for F_1 and F_2. Conditioned on $|\bar{\mathcal{S}}|$, ℓ_1/ℓ_2-RIP (Lemma 52) with $\delta = \sqrt{\frac{cd(k+r)}{|\bar{\mathcal{S}}|}}$ can be invoked to show that

$$F_1 = \frac{|\bar{\mathcal{S}}|}{m}\min_{\Delta \boldsymbol{X} \in \mathcal{S}_{k+r}^{d\times d}} \left\{ \frac{1}{|\bar{\mathcal{S}}|}\sum_{i\in\bar{\mathcal{S}}}|\langle \boldsymbol{A}_i, \Delta \boldsymbol{X}\rangle| \right\} \ge \frac{|\bar{\mathcal{S}}|}{m}\left(\sqrt{\frac{2}{\pi}} - \sqrt{\frac{cd(k+r)}{|\bar{\mathcal{S}}|}}\right)$$

with probability at least $1-\exp(-\Omega(d(k+r)))$. With the same probability, we have

$$F_2 = \frac{|\mathcal{S}|}{m}\max_{\Delta \boldsymbol{X} \in \mathcal{S}_{k+r}^{d\times d}} \left\{ \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}|\langle \boldsymbol{A}_i, \Delta \boldsymbol{X}\rangle| \right\} \le \frac{|\mathcal{S}|}{m}\left(\sqrt{\frac{2}{\pi}} + \sqrt{\frac{cd(k+r)}{|\mathcal{S}|}}\right).$$

Combining these two inequalities, we have

$$F_1 - F_2 \geq \sqrt{\frac{2}{\pi}} \left(\frac{|\bar{\mathcal{S}}| - |\mathcal{S}|}{m} \right) - \left(\sqrt{\frac{|\bar{\mathcal{S}}|}{m}} + \sqrt{\frac{|\mathcal{S}|}{m}} \right) \sqrt{\frac{cd(k+r)}{m}}$$
$$\geq \sqrt{\frac{2}{\pi}} \left(\frac{m - 2|\mathcal{S}|}{m} \right) - 2\sqrt{\frac{2cdk}{m}} \tag{2.109}$$

with probability at least $1 - \exp(-\Omega(d(k+r)))$. On the other hand, since $|\mathcal{S}|$ has a binomial distribution with parameters (m, p) and $0 \leq p < 1/2$, Lemma 48 can be invoked to show that $\frac{m-2|\mathcal{S}|}{m} \geq \min\left\{\frac{1}{4}, (1-2p)^2\right\}$ with probability at least $1 - \exp(-\Omega((1-2p)^2 m))$. Combining this inequality with Equation (2.109), we have

$$F_1 - F_2 \geq \sqrt{\frac{2}{\pi}} \min\left\{\frac{1}{4}, (1-2p)^2\right\} - 2\sqrt{\frac{2cdk}{m}} \geq \left(\sqrt{\frac{2}{\pi}} - 1\right) \min\left\{\frac{1}{4}, (1-2p)^2\right\} > 0,$$

with probability at least $1 - \exp(-\Omega(d(k+r))) - \exp(-\Omega((1-2p)^2 m) = 1 - \exp(-\Omega(dk))$ assuming that $m \geq \frac{512cdk}{(1-2p)^4}$. This completes the proof. □

Our next goal is to provide the proof of the first statement of Theorem 10. We first note that in this case, $m \gtrsim \frac{dr}{(1-2p)^4}$. Therefore, the sample size is not sufficiently large for ℓ_1/ℓ_2-RIP to control the deviation of the loss uniformly over the perturbation set $\mathcal{S}^{d\times d}_{k+r}$. To circumvent this issue, we further decompose $\Delta \boldsymbol{X}$ into two parts. Consider $\boldsymbol{U} = \boldsymbol{U}^\star + \gamma \Delta \boldsymbol{U}$ for some $\|\Delta \boldsymbol{U}\|_{\mathrm{F}} \leq 1$. Therefore, we have

$$\Delta \boldsymbol{X} = \boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^\star \boldsymbol{U}^{\star\top} = \gamma \cdot \underbrace{\boldsymbol{U}^\star \Delta \boldsymbol{U}^\top + \Delta \boldsymbol{U} \boldsymbol{U}^{\star\top}}_{\Delta \boldsymbol{X}_1} + \gamma^2 \cdot \underbrace{\Delta \boldsymbol{U} \Delta \boldsymbol{U}^\top}_{\Delta \boldsymbol{X}_2}.$$

We have $\mathrm{rank}(\Delta \boldsymbol{X}_1) \leq 2r$. As a result, the effect of $\Delta \boldsymbol{X}_1$ can be controlled via ℓ_1/ℓ_2-RIP with $m \gtrsim \frac{dr}{(1-2p)^4}$. On the other hand, $\mathrm{rank}(\Delta \boldsymbol{X}_2)$ may be as large as k, but its magnitude is much smaller than $\Delta \boldsymbol{X}_1$ since it is controlled by the second-order term γ. With this insight, we present the proof of the first statement of Theorem 10.

Proof of first statement of Theorem 10. We have

$$\begin{aligned}
&f_{\ell_1}(\boldsymbol{U}^\star+\gamma\Delta\boldsymbol{U}) - f_{\ell_1}(\boldsymbol{U}^\star) \\
&= \frac{1}{m}\sum_{i\in\bar{S}} \left|\left\langle \boldsymbol{A}_i, \gamma\Delta\boldsymbol{X}_1 + \gamma^2\Delta\boldsymbol{X}_2\right\rangle\right| + \frac{1}{m}\sum_{i\in S}\left(\left|\left\langle \boldsymbol{A}_i, \gamma\Delta\boldsymbol{X}_1 + \gamma^2\Delta\boldsymbol{X}_2\right\rangle + \epsilon_i\right| - |\epsilon_i|\right) \\
&\geq \frac{1}{m}\sum_{i\in\bar{S}} \left|\left\langle \boldsymbol{A}_i, \gamma\Delta\boldsymbol{X}_1 + \gamma^2\Delta\boldsymbol{X}_2\right\rangle\right| - \frac{1}{m}\sum_{i\in S}\left|\left\langle \boldsymbol{A}_i, \gamma\Delta\boldsymbol{X}_1 + \gamma^2\Delta\boldsymbol{X}_2\right\rangle\right| \\
&\geq \gamma\left(\frac{1}{m}\sum_{i\in\bar{S}}|\langle \boldsymbol{A}_i, \Delta\boldsymbol{X}_1\rangle| - \frac{1}{m}\sum_{i\in S}\left|\left\langle \boldsymbol{A}_i, \Delta\boldsymbol{X}_1\right\rangle\right|\right) - \frac{\gamma^2}{m}\sum_{i=1}^{m}\left|\left\langle \boldsymbol{A}_i, \Delta\boldsymbol{X}_2\right\rangle\right|,
\end{aligned}$$

where the first and second inequalities follow from triangle inequality. Note that $\mathrm{rank}(\Delta\boldsymbol{X}_1) \leq 2r$. Similarly, we have $\mathrm{rank}(\Delta\boldsymbol{X}_2) \leq k$ and $\|\Delta\boldsymbol{X}_2\|_{\mathrm{F}} \leq 1$. Therefore, we have

$$\begin{aligned}
&\min_{\Delta\boldsymbol{U}\in\mathcal{B}}\left\{f_{\ell_1}(\boldsymbol{U}^\star+\gamma\Delta\boldsymbol{U}) - f_{\ell_1}(\boldsymbol{U}^\star)\right\} \\
&\geq \gamma\cdot \underbrace{\min_{\mathrm{rank}(\Delta\boldsymbol{X}_1)\leq 2r}\left\{\frac{1}{m}\sum_{i\in\bar{S}}|\langle \boldsymbol{A}_i, \Delta\boldsymbol{X}_1\rangle| - \frac{1}{m}\sum_{i\in S}\left|\left\langle \boldsymbol{A}_i, \Delta\boldsymbol{X}_1\right\rangle\right|\right\}}_{:=G_1} \\
&\quad - \gamma^2 \underbrace{\max_{\Delta\boldsymbol{X}_2\in\mathcal{S}_k^{d\times d}}\left\{\frac{1}{m}\sum_{i=1}^{m}|\langle \boldsymbol{A}_i, \Delta\boldsymbol{X}_2\rangle|\right\}}_{:=G_2},
\end{aligned}$$

Similar to the proof of the second argument, one can show that that $G_1 > 0$ with probability at least $1-\exp(-\Omega(dr))$. On the other hand, applying ℓ_1/ℓ_2-RIP to G_2 with $\delta = \sqrt{\frac{cdk}{m}}$, we have

$$G_2 \lesssim \sqrt{\frac{2}{\pi}} + \sqrt{\frac{dk}{m}},$$

with probability at least $1-\exp(-\Omega(dk))$. Combining the above inequalities leads to

$$\min_{\boldsymbol{U}^\star\in\mathcal{U}}\min_{\Delta\boldsymbol{U}\in\mathcal{B}}\left\{f_{\ell_1}(\boldsymbol{U}^\star+\gamma\Delta\boldsymbol{U}) - f_{\ell_1}(\boldsymbol{U}^\star)\right\} \geq \gamma G_1 - \gamma^2 G_2 \gtrsim -\left(\sqrt{\frac{2}{\pi}} + \sqrt{\frac{dk}{m}}\right)\gamma^2,$$

with probability at least $1-\exp(-\Omega(dr))$. This completes the proof. □

2.F.5 Strict Saddle Nature

We finally provide the proof of Corollary 4 by combining Theorems 9 and 10.

Proof of Corollary 4. First, we prove that all true solutions become critical points with high

probability, provided that $m \gtrsim dr$. In light of Lemma 30, it suffices to show that, with high probability, $f'_{\ell_1}(\boldsymbol{U}^\star, \Delta\boldsymbol{U}) \geq 0$ for every $\boldsymbol{U}^\star \in \mathcal{U}$ and every direction $\|\Delta\boldsymbol{U}\|_{\mathrm{F}} = 1$. According to the definition of the directional derivative, we need to show that

$$\lim_{\gamma \to 0^+} \frac{f_{\ell_1}(\boldsymbol{U}^\star + \gamma\Delta\boldsymbol{U}) - f_{\ell_1}(\boldsymbol{U}^\star)}{\gamma} \geq 0, \qquad \forall \boldsymbol{U}^\star \in \mathcal{U}, \forall \Delta\boldsymbol{U} : \|\Delta\boldsymbol{U}\|_{\mathrm{F}} = 1. \tag{2.110}$$

The above inequality is readily implied by the first statement of Theorem 10, provided that $m \gtrsim \frac{dr}{(1-2p)^2} \asymp dr$. On the other hand, when $m \lesssim pp_0 d(k-r) \asymp dk$, Theorem 9 can be invoked to show that, for every $\boldsymbol{U}^\star \in \mathcal{U}$, there exists a direction $\|\Delta\boldsymbol{U}\|_{\mathrm{F}} = 1$, and scalars $c > 0$ and $\bar{\gamma} > 0$ such that, for every $0 \leq \gamma \leq \bar{\gamma}$, we have

$$f_{\ell_1}(\boldsymbol{U}^\star + \gamma\Delta\boldsymbol{U}) - f_{\ell_1}(\boldsymbol{U}^\star) \leq -c\gamma^2. \tag{2.111}$$

By combining the above two statements, we conclude that, when $dr \lesssim m \lesssim dk$, all true solutions $\boldsymbol{U}^\star \in \mathcal{U}$ become critical and satisfy the definition of strict saddle points (Definition 9), with a probability stated in Corollary 4. This completes the proof. □

2.G Omitted Proofs

2.G.1 Proof of Lemma 19

To prove this lemma, we use the one-step discretization technique. First note that, for any $\boldsymbol{X} \in \mathcal{S}_{k,\epsilon}$, there exists a matrix in $\boldsymbol{X}' \in \mathcal{S}_k$ such that $\|\boldsymbol{X} - \boldsymbol{X}'\|_{\mathrm{F}} \leq \epsilon$. Suppose that $\mathcal{N}_{k,\xi}$ is a ξ-net of $\mathcal{S}_k$ where $\xi \geq \epsilon$. Based on the definition of ξ-net, there exists $\boldsymbol{X}'' \in \mathcal{N}_{k,\xi}$ such that $\|\boldsymbol{X}' - \boldsymbol{X}''\|_{\mathrm{F}} \leq \xi$. This implies that $\|\boldsymbol{X} - \boldsymbol{X}''\|_{\mathrm{F}} \leq \|\boldsymbol{X} - \boldsymbol{X}'\|_{\mathrm{F}} + \|\boldsymbol{X}' - \boldsymbol{X}''\|_{\mathrm{F}} \leq 2\xi$, and hence, $\mathcal{N}_{k,\xi}$ is a 2ξ-net of $\mathcal{S}_{k,\epsilon}$. Given this fact, one can write the following chain of inequalities for every $\boldsymbol{Y} \in \mathbb{S}$:

$$\begin{aligned}
\mathbb{E}\left[\mathcal{G}_{\boldsymbol{Y}}\right] \leq & \underbrace{\mathbb{E}\left[\sup_{\boldsymbol{X}' \in \mathcal{N}_{k,\xi}} \frac{1}{m}\sum_{i=1}^{m} \operatorname{sign}\left(\langle \boldsymbol{A}_i, \boldsymbol{X}'\rangle - s_i\right)\langle \boldsymbol{A}_i, \boldsymbol{Y}\rangle - \varphi(\boldsymbol{X}')\left\langle \frac{\boldsymbol{X}'}{\|\boldsymbol{X}'\|_{\mathrm{F}}}, \boldsymbol{Y}\right\rangle\right]}_{(A)} \\
& + \underbrace{\mathbb{E}\left[\sup_{\|\boldsymbol{X}-\boldsymbol{X}'\|_{\mathrm{F}} \leq 2\xi} \frac{1}{m}\sum_{i=1}^{m} \left(\operatorname{sign}\left(\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle - s_i\right) - \operatorname{sign}\left(\langle \boldsymbol{A}_i, \boldsymbol{X}'\rangle - s_i\right)\right)\langle \boldsymbol{A}_i, \boldsymbol{Y}\rangle\right]}_{(B)} \\
& + \underbrace{\sup_{\|\boldsymbol{X}-\boldsymbol{X}'\|_{\mathrm{F}} \leq 2\xi} \left\langle \varphi(\boldsymbol{X})\frac{\boldsymbol{X}}{\|\boldsymbol{X}\|_{\mathrm{F}}} - \varphi(\boldsymbol{X}')\frac{\boldsymbol{X}'}{\|\boldsymbol{X}'\|_{\mathrm{F}}}, \boldsymbol{Y}\right\rangle}_{(C)}.
\end{aligned} \tag{2.112}$$

We control each term in the above inequality separately.

Bounding (A)**.** To control (A), note that $\frac{1}{m}\sum_{i=1}^{m}\operatorname{sign}\left(\langle \boldsymbol{A}_i, \boldsymbol{X}'\rangle - s_i\right)\langle \boldsymbol{A}_i, \boldsymbol{Y}\rangle - \varphi(\boldsymbol{X}')\left\langle \frac{\boldsymbol{X}'}{\|\boldsymbol{X}'\|_{\mathrm{F}}}, \boldsymbol{Y}\right\rangle$ is $\mathcal{O}\left(1/m\right)$-sub-Gaussian and (A) is the supremum of the sub-Gaussian random variable over a finite set $\mathcal{N}_{k,\xi}$. Hence, the Maximum Inequality implies that

$$(A) \lesssim \sqrt{\frac{dk}{m}\log\left(\frac{R}{\xi}\right)}. \tag{2.113}$$

Bounding (B)**.** Invoking Hölder's inequality, one can write

$$(B) \leq \mathbb{E}\left[\sup_{\|\boldsymbol{X}-\boldsymbol{X}'\|_{\mathrm{F}}\leq 2\xi}\left(\frac{1}{m}\sum_{i=1}^{m}\left|\operatorname{sign}\left(\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle - s_i\right) - \operatorname{sign}\left(\langle \boldsymbol{A}_i, \boldsymbol{X}'\rangle - s_i\right)\right|\right)\max_{1\leq i\leq m}\left|\langle \boldsymbol{A}_i, \boldsymbol{Y}\rangle\right|\right].$$

Note that if $\left|\langle \boldsymbol{A}_i, \boldsymbol{X}-\boldsymbol{X}'\rangle\right| \leq \left|\langle \boldsymbol{A}_i, \boldsymbol{X}'\rangle\right| - \lambda s_i$, then $\operatorname{sign}(\langle \boldsymbol{A}_i, \boldsymbol{X}'\rangle - \lambda s_i) = \operatorname{sign}(\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle - \lambda s_i)$. Therefore, the above term can be further bounded by

$$\begin{aligned}
&(B)\\
\leq&\mathbb{E}\left[\sup_{\|\boldsymbol{X}-X'\|_{\mathrm{F}}\leq 2\xi}\left(\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}\left(\left|\langle \boldsymbol{A}_i, \boldsymbol{X}-X'\rangle\right| \geq \left|\langle \boldsymbol{A}_i, \boldsymbol{X}'\rangle - s_i\right|\right)\right)\max_{1\leq i\leq m}\left|\langle \boldsymbol{A}_i, \boldsymbol{Y}\rangle\right|\right]\\
\leq&\mathbb{E}\left[\sup_{\|\boldsymbol{X}-\boldsymbol{X}'\|_{\mathrm{F}}\leq 2\xi}\left(\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}\left(\left|\langle \boldsymbol{A}_i, \boldsymbol{X}-\boldsymbol{X}'\rangle\right| \geq t\right) + \mathbb{1}\left(\left|\langle \boldsymbol{A}_i, \boldsymbol{X}'\rangle - s_i\right| \leq t\right)\right)\max_{1\leq i\leq m}\left|\langle \boldsymbol{A}_i, \boldsymbol{Y}\rangle\right|\right]\\
\leq&\underbrace{\mathbb{E}\left[\sup_{\|\boldsymbol{X}-\boldsymbol{X}'\|_{\mathrm{F}}\leq 2\xi}\left(\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}\left(\left|\langle \boldsymbol{A}_i, \boldsymbol{X}-\boldsymbol{X}'\rangle\right| \geq t\right)\right)\max_{1\leq i\leq m}\left|\langle \boldsymbol{A}_i, \boldsymbol{Y}\rangle\right|\right]}_{(B_1)}\\
&+\underbrace{\mathbb{E}\left[\sup_{\boldsymbol{X}'\in\mathcal{N}_{k,\xi}}\left(\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}\left(\left|\langle \boldsymbol{A}_i, \boldsymbol{X}'\rangle - s_i\right| \leq t\right)\right)\max_{1\leq i\leq m}\left|\langle \boldsymbol{A}_i, \boldsymbol{Y}\rangle\right|\right]}_{(B_2)}.
\end{aligned} \tag{2.114}$$

For (B_1), we have

$$\begin{aligned}(B_1) &\leq \mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}\left(\|\boldsymbol{A}_i\|_{\mathrm{F}} \geq \frac{t}{2\xi}\right)\right)\max_{1\leq i\leq m}|\langle \boldsymbol{A}_i, \boldsymbol{Y}\rangle|\right] \\ &\leq \mathbb{E}\left[\mathbb{1}\left(\|\boldsymbol{A}_i\|_{\mathrm{F}} \geq \frac{t}{2\xi}\right)\right]\mathbb{E}\left[\max_{j\neq i}|\langle A_j, \boldsymbol{Y}\rangle|\right] \\ &+ \mathbb{E}\left[\mathbb{1}\left(\|\boldsymbol{A}_i\|_{\mathrm{F}} \geq \frac{t}{2\xi}\right)|\langle \boldsymbol{A}_i, \boldsymbol{Y}\rangle|\right] \\ &\leq \mathcal{O}\left(e^{-C\frac{t^2}{\xi^2}}\sqrt{\log(m)}\right) + \mathbb{E}\left[\mathbb{1}\left(\|\boldsymbol{A}_i\|_{\mathrm{F}} \geq \frac{t}{2\xi}\right)|\langle \boldsymbol{A}_i, \boldsymbol{Y}\rangle|\right],\end{aligned}$$

where $C > 0$ is a universal constant and $t/\xi \geq \sqrt{d}$. Furthermore, applying Cauchy-Schwarz inequality, we have

$$\mathbb{E}\left[\mathbb{1}\left(\|\boldsymbol{A}_i\|_{\mathrm{F}} \geq \frac{t}{2\xi}\right)|\langle \boldsymbol{A}_i, \boldsymbol{Y}\rangle|\right] \leq \sqrt{\mathbb{P}\left(2\xi\|\boldsymbol{A}_i\|_{\mathrm{F}} \geq t\right)}\sqrt{\mathbb{E}\left[\langle \boldsymbol{A}_i, \boldsymbol{Y}\rangle^2\right]} \lesssim e^{-C\frac{t^2}{\xi^2}}.$$

Hence, we conclude that $(B_1) \lesssim e^{-C\frac{t^2}{\xi^2}}\sqrt{\log(m)}$.

Now we turn to bound (B_2). Since $\left(\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}\left(|\langle \boldsymbol{A}_i, \boldsymbol{X}'\rangle - s_i| \leq t\right)\right)\max_{1\leq i\leq m}|\langle \boldsymbol{A}_i, \boldsymbol{Y}\rangle|$ is $\mathcal{O}\left(\log(m)/m\right)$-sub-Gaussian, and $\mathcal{N}_{k,\xi}$ is a finite set, the Maximum Inequality yields

$$\begin{aligned}(B_2) \leq \sup_{\boldsymbol{X}'\in\mathcal{N}_{k,\xi}} &\mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}\left(|\langle \boldsymbol{A}_i, \boldsymbol{X}'\rangle - s_i| \leq t\right)\right)\max_{1\leq i\leq m}|\langle \boldsymbol{A}_i, \boldsymbol{Y}\rangle|\right] \\ &+ \mathcal{O}\left(\sqrt{\frac{dk\log(m)\log\left(R/\xi\right)}{m}}\right).\end{aligned} \tag{2.115}$$

For the first part, one can write

$$\begin{aligned}&\mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}\left(|\langle \boldsymbol{A}_i, \boldsymbol{X}'\rangle - s_i| \leq t\right)\right)\max_{1\leq i\leq m}|\langle \boldsymbol{A}_i, \boldsymbol{Y}\rangle|\right] \\ \leq &\mathbb{E}\left[\mathbb{1}\left(|\langle \boldsymbol{A}_i, \boldsymbol{X}'\rangle - s_i| \leq t\right)\right]\mathbb{E}\left[\max_{j\neq i}|\langle \boldsymbol{A}_i, \boldsymbol{Y}\rangle|\right] + \mathbb{E}\left[\mathbb{1}\left(|\langle \boldsymbol{A}_i, \boldsymbol{X}'\rangle - s_i| \leq t\right)|\langle \boldsymbol{A}_i, \boldsymbol{Y}\rangle|\right] \\ \leq &\sqrt{\log(m)}\frac{t}{\zeta}.\end{aligned}$$

Therefore, we conclude that $(B) \lesssim \sqrt{\log(m)}\frac{t}{\zeta} + e^{-C\frac{t^2}{\xi^2}}\sqrt{\log\left(m\right)} + \sqrt{\frac{dk\log(m)\log(R/\xi)}{m}}$. Finally, it remains to bound (C) in equation 2.112.

Bounding (C)**.** We have

$$\begin{aligned}(C) &= \sup_{\|\boldsymbol{X}-\boldsymbol{X}'\|_{\mathrm{F}}\le 2\xi}\left\{(\varphi(\boldsymbol{X})-\varphi(\boldsymbol{X}'))\left\langle\frac{\boldsymbol{X}}{\|\boldsymbol{X}\|_{\mathrm{F}}},\boldsymbol{Y}\right\rangle+\varphi(\boldsymbol{X}')\left\langle\frac{\boldsymbol{X}}{\|\boldsymbol{X}\|_{\mathrm{F}}}-\frac{\boldsymbol{X}'}{\|\boldsymbol{X}'\|_{\mathrm{F}}},\boldsymbol{Y}\right\rangle\right\}\\ &\le \sup_{\|\boldsymbol{X}-\boldsymbol{X}'\|_{\mathrm{F}}\le 2\xi}\{\varphi(\boldsymbol{X})-\varphi(\boldsymbol{X}')\}+\sup_{\|\boldsymbol{X}-\boldsymbol{X}'\|_{\mathrm{F}}\le 2\xi}\left\|\frac{\boldsymbol{X}}{\|\boldsymbol{X}\|_{\mathrm{F}}}-\frac{\boldsymbol{X}'}{\|\boldsymbol{X}'\|_{\mathrm{F}}}\right\|_{\mathrm{F}}.\end{aligned}$$

For the first part, we use the mean value theorem to write

$$|\varphi(\boldsymbol{X}')-\varphi(\boldsymbol{X})|\le\|\nabla\varphi(\boldsymbol{Z})\|_{\mathrm{F}}\|\boldsymbol{X}'-\boldsymbol{X}\|_{\mathrm{F}}\le 2\|\nabla\varphi(\boldsymbol{Z})\|_{\mathrm{F}}\xi,$$

where $\boldsymbol{Z}=\lambda\boldsymbol{X}+(1-\lambda)\boldsymbol{X}'$, $\lambda\in[0,1]$. Note that $\nabla\varphi(\boldsymbol{Z})=\sqrt{\frac{2}{\pi}}p\mathbb{E}\left[\frac{s^2\boldsymbol{Z}}{\|\boldsymbol{Z}\|_{\mathrm{F}}^4}e^{-\frac{s^2}{2\|\boldsymbol{Z}\|_{\mathrm{F}}^2}}\right]$. Hence, we have

$$\begin{aligned}\sup_{\|\boldsymbol{Z}\|_{\mathrm{F}}\ge\zeta}\|\nabla\varphi(\boldsymbol{Z})\|_{\mathrm{F}} &\lesssim \sup_{\|\boldsymbol{Z}\|_{\mathrm{F}}\ge\zeta}\mathbb{E}\left[\frac{s^2}{\|\boldsymbol{Z}\|_{\mathrm{F}}^3}e^{-\frac{s^2}{2\|\boldsymbol{Z}\|_{\mathrm{F}}^2}}\right]\\ &\le\frac{1}{\zeta}\sup_{\|\boldsymbol{Z}\|_{\mathrm{F}}\ge\zeta}\mathbb{E}\left[\frac{s^2}{\|\boldsymbol{Z}\|_{\mathrm{F}}^2}e^{-\frac{s^2}{2\|\boldsymbol{Z}\|_{\mathrm{F}}^2}}\right]\\ &\lesssim\frac{1}{\zeta}.\end{aligned}$$

For the second part, we have

$$\begin{aligned}&\sup_{\|\boldsymbol{X}-\boldsymbol{X}'\|_{\mathrm{F}}\le 2\xi}\left\|\frac{\boldsymbol{X}}{\|\boldsymbol{X}\|_{\mathrm{F}}}-\frac{\boldsymbol{X}'}{\|\boldsymbol{X}'\|_{\mathrm{F}}}\right\|_{\mathrm{F}}\\ &\le\sup_{\|\boldsymbol{X}-\boldsymbol{X}'\|_{\mathrm{F}}\le 2\xi}\left\|\frac{\boldsymbol{X}-\boldsymbol{X}'}{\|\boldsymbol{X}\|_{\mathrm{F}}}\right\|_{\mathrm{F}}+\left\|\frac{\boldsymbol{X}\left(\|\boldsymbol{X}'\|_{\mathrm{F}}-\|\boldsymbol{X}\|_{\mathrm{F}}\right)}{\|\boldsymbol{X}\|_{\mathrm{F}}\|\boldsymbol{X}'\|_{\mathrm{F}}}\right\|_{\mathrm{F}}\\ &\le\frac{4\xi}{\zeta}.\end{aligned}$$

Therefore, we conclude that $(C)\lesssim\frac{\xi}{\zeta}$.

Combining the derived upper bounds for (A), (B), and (C), we have

$$\begin{aligned}\mathbb{E}[\mathcal{G}_{\boldsymbol{Y}}] \le&(A)+(B)+(C)\\ \lesssim&\sqrt{\frac{dk}{m}\log\left(\frac{R}{\xi}\right)}+\sqrt{\log(m)}\frac{t}{\zeta}+e^{-C\frac{t^2}{\xi^2}}\sqrt{\log(m)}\\ &+\sqrt{\frac{dk\log(m)\log(R/\xi)}{m}}+\frac{\xi}{\zeta},\end{aligned}\tag{2.116}$$

provided that $\xi\ge\epsilon$ and $t/\xi\ge\sqrt{d}$. Let $\xi\asymp\zeta\sqrt{k/m}$ and $t\asymp\zeta\sqrt{dk\log(m)/m}$. Clearly,

these choices of parameters satisfy $t/\xi \geq \sqrt{d}$. Moreover, $\xi \asymp \zeta\sqrt{k/m}$ together with the assumption $\epsilon \lesssim \zeta\sqrt{k/m}$ implies that $\epsilon \leq \xi$. Finally, plugging these values in equation 2.116 leads to

$$\mathbb{E}[\mathcal{G}_{\boldsymbol{Y}}] \lesssim \sqrt{\frac{dk}{m}\log^2(m)\log\left(\frac{R}{\zeta}\right)},$$

for every $\boldsymbol{Y} \in \mathbb{S}$. This in turn implies

$$\sup_{\boldsymbol{Y}\in\mathbb{S}_k} \mathbb{E}[\mathcal{G}_{\boldsymbol{Y}}] \leq \sup_{\boldsymbol{Y}\in\mathbb{S}} \mathbb{E}[\mathcal{G}_{\boldsymbol{Y}}] \lesssim \sqrt{\frac{dk}{m}\log^2(m)\log\left(\frac{R}{\zeta}\right)},$$

which completes the proof. □

2.G.2 Proof of Lemma 23

To prove $0.5 \leq \left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\left(\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\right)^{-1}\right\| \leq 2$, it suffices to show that $0.5 \leq \lambda_{\min}\left(\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\right) \leq \left\|\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\right\| \leq 2$. For brevity, we only show $0.5 \leq \lambda_{\min}\left(\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\right)$, as the other part of the inequality can be proven in a similar fashion. One can write

$$\begin{aligned}\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top = {}& \boldsymbol{S}_t\boldsymbol{S}_t^\top - \bar{\eta}_t\boldsymbol{S}_t\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right) - \bar{\eta}_t\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right)\boldsymbol{S}_t\boldsymbol{S}_t^\top - 2\bar{\eta}_t\boldsymbol{S}_t\boldsymbol{E}_t\boldsymbol{E}_t^\top\boldsymbol{S}_t^\top \\ & + \bar{\eta}_t^2\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right)\boldsymbol{S}_t\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right) + \bar{\eta}_t^2\boldsymbol{S}_t\boldsymbol{E}_t^\top\boldsymbol{E}_t\boldsymbol{E}_t^\top\boldsymbol{E}_t\boldsymbol{S}_t^\top \\ & + \bar{\eta}_t^2\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right)\boldsymbol{S}_t\boldsymbol{E}_t\boldsymbol{E}_t^\top\boldsymbol{S}_t^\top + \bar{\eta}_t^2\boldsymbol{S}_t\boldsymbol{E}_t\boldsymbol{E}_t^\top\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right).\end{aligned} \tag{2.117}$$

Note that the eigenvalues of $\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}$ are real and nonnegative, due to its similarity to $\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1/2}\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1/2}$. One the other hand, one can write

$$\begin{aligned}&\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1} \\ =&\boldsymbol{I} + \bar{\eta}_t\boldsymbol{\Sigma} + \bar{\eta}_t\boldsymbol{S}_t\boldsymbol{S}_t^\top\boldsymbol{\Sigma}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1} - 2\bar{\eta}_t\boldsymbol{S}_t\boldsymbol{S}_t^\top - 2\bar{\eta}_t\boldsymbol{S}_t\boldsymbol{E}_t\boldsymbol{E}_t^\top\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1} \\ &+ \bar{\eta}_t^2\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right)\boldsymbol{S}_t\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right)\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1} + \bar{\eta}_t^2\boldsymbol{S}_t\boldsymbol{E}_t^\top\boldsymbol{E}_t\boldsymbol{E}_t^\top\boldsymbol{E}_t\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1} \\ &+ \bar{\eta}_t^2\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right)\boldsymbol{S}_t\boldsymbol{E}_t\boldsymbol{E}_t^\top\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1} + \bar{\eta}_t^2\boldsymbol{S}_t\boldsymbol{E}_t\boldsymbol{E}_t^\top\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{\Sigma}\right)\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}.\end{aligned}$$

We will show that every term in the above decomposition, except for the first term, is in the

order of $\mathcal{O}(\bar{\eta}_t\sigma_1)$. First, note that

$$\|\bar{\eta}_t\boldsymbol{\Sigma}\| = \left\|\bar{\eta}_t\boldsymbol{S}_t\boldsymbol{S}_t^\top\boldsymbol{\Sigma}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\right\| = \mathcal{O}(\bar{\eta}_t\sigma_1).$$

Similarly, we have

$$\begin{aligned}
&\left\|2\bar{\eta}_t\boldsymbol{S}_t\boldsymbol{S}_t^\top + 2\bar{\eta}_t\boldsymbol{S}_t\boldsymbol{E}_t\boldsymbol{E}_t^\top\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\right\| \\
&\leq 2\bar{\eta}_t\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| + 2\bar{\eta}_t\left\|\boldsymbol{S}_t\boldsymbol{E}_t\boldsymbol{E}_t^\top\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\right\| \\
&= 2\bar{\eta}_t\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| + 2\bar{\eta}_t\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|\left\|\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\boldsymbol{S}_t\right\| \\
&\leq 2\bar{\eta}_t\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| + 2\bar{\eta}_t\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| \\
&= \mathcal{O}(\bar{\eta}_t\sigma_1).
\end{aligned}$$

It is easy to see that all the remaining terms are in the order of $\mathcal{O}(\bar{\eta}_t\sigma_1)$; we omit their proofs for brevity. Combining the above bounds, we obtain

$$\lambda_{\min}\left(\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\right) = 1 - \mathcal{O}(\bar{\eta}_t\sigma_1) \geq 1/2,$$

where the last inequality is due to our choice of $\bar{\eta}_t$. This in turn implies that $\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\left(\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\right)^{-1}\right\| \leq 2$. Similarly, we can show that $\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\left(\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right)^{-1}\right\| \leq 3$, by proving $\lambda_{\min}\left(\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top(\boldsymbol{S}_t\boldsymbol{S}_t^\top)^{-1}\right) \geq 1/3$. This can be shown in an analogous fashion, after noting that

$$\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top = \tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top + \underbrace{\boldsymbol{V}^\top\boldsymbol{R}_t\boldsymbol{U}_t\tilde{\boldsymbol{S}}_{t+1}^\top + \tilde{\boldsymbol{S}}_{t+1}\boldsymbol{U}_t^\top\boldsymbol{R}_t^\top\boldsymbol{V} + \boldsymbol{V}^\top\boldsymbol{R}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{R}_t^\top\boldsymbol{V}}_{\text{perturbation}}.$$

Similar to the proof of Proposition 6, it can be shown that the norm of the perturbation term is upper bounded by $4\bar{\eta}_t\delta\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}} \leq 1/6$, due to our choice of δ and η, and our assumptions on $\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\|$ and $\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|$.

Finally, we prove the second statement by providing an upper bound for $\left\|\tilde{\boldsymbol{E}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\left(\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\right)^{-1}\right\|$. Due to the first part of the lemma, one can write

$$\begin{aligned}
\left\|\tilde{\boldsymbol{E}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\left(\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\right)^{-1}\right\| &\leq \left\|\tilde{\boldsymbol{E}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\right\|\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\left(\tilde{\boldsymbol{S}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\right)^{-1}\right\| \\
&\leq 2\left\|\tilde{\boldsymbol{E}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\right\|.
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}\tilde{\boldsymbol{E}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top &= \boldsymbol{E}_t\boldsymbol{S}_t^\top + \bar{\eta}_t\boldsymbol{E}_t\boldsymbol{S}_t^\top(\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^\top) - \bar{\eta}_t\boldsymbol{E}_t(\boldsymbol{S}_t^\top\boldsymbol{S}_t+\boldsymbol{E}_t^\top\boldsymbol{E}_t)\boldsymbol{S}_t^\top\\ &\quad+\bar{\eta}_t^2\boldsymbol{E}_t(\boldsymbol{S}_t^\top\boldsymbol{S}_t+\boldsymbol{E}_t^\top\boldsymbol{E}_t)\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top-\boldsymbol{\Sigma}\right)\\ &\quad-\bar{\eta}_t\boldsymbol{E}_t\boldsymbol{E}_t^\top\boldsymbol{E}_t\boldsymbol{S}_t^\top+\bar{\eta}_t^2\boldsymbol{E}_t\left(\boldsymbol{S}_t^\top\boldsymbol{S}_t+\boldsymbol{E}_t^\top\boldsymbol{E}_t\right)\boldsymbol{E}_t^\top\boldsymbol{E}_t\boldsymbol{S}_t^\top.\end{aligned}$$

Let us define $\boldsymbol{H}_t = \boldsymbol{E}_t\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}$. Based on this definition, one can verify that

$$\begin{aligned}&\tilde{\boldsymbol{E}}_{t+1}\tilde{\boldsymbol{S}}_{t+1}^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\\ =&\boldsymbol{H}_t+\bar{\eta}_t\boldsymbol{H}_t\boldsymbol{S}_t\boldsymbol{S}_t^\top(\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^\top)\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}-\bar{\eta}_t\boldsymbol{H}_t\boldsymbol{S}_t\boldsymbol{S}_t^\top-2\bar{\eta}_t\boldsymbol{E}_t\boldsymbol{E}_t^\top\boldsymbol{H}_t\\ &+\bar{\eta}_t^2H_t\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^2\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top-\boldsymbol{\Sigma}\right)\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}+\bar{\eta}_t^2\boldsymbol{E}_t\left(\boldsymbol{S}_t^\top\boldsymbol{S}_t+\boldsymbol{E}_t^\top\boldsymbol{E}_t\right)\boldsymbol{E}_t^\top\boldsymbol{H}_t\\ &+\bar{\eta}_t^2\boldsymbol{E}_t\boldsymbol{E}_t^\top\boldsymbol{H}_t\boldsymbol{S}_t\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top-\boldsymbol{\Sigma}\right)\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}.\end{aligned}$$

Now, note that

$$\begin{aligned}\left\|\bar{\eta}_t\boldsymbol{H}_t\boldsymbol{S}_t\boldsymbol{S}_t^\top(\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^\top)\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\right\| &\le \bar{\eta}_t\left\|\boldsymbol{H}_t\right\|\left\|\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|\\ &\lesssim \bar{\eta}_t\sigma_1\left\|\boldsymbol{H}_t\right\| \le \frac{1}{12}\left\|\boldsymbol{H}_t\right\|.\end{aligned}$$

Similarly, we have $\left\|\bar{\eta}_t\boldsymbol{H}_t\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| \le \frac{1}{12}\left\|\boldsymbol{H}_t\right\|$ and $\left\|\bar{\eta}_t\boldsymbol{E}_t\boldsymbol{E}_t^\top\boldsymbol{H}_t\right\| \le \frac{1}{12}\left\|\boldsymbol{H}_t\right\|$. Moreover, we have

$$\begin{aligned}\left\|\bar{\eta}_t^2H_t\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^2\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top-\boldsymbol{\Sigma}\right)\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\right\| &\le \bar{\eta}_t^2\left\|\boldsymbol{H}_t\right\|\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|\left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top-\boldsymbol{\Sigma}\right\|\\ &\lesssim (\bar{\eta}_t\sigma_1)^2\left\|\boldsymbol{H}_t\right\| \le \frac{1}{12}\left\|\boldsymbol{H}_t\right\|.\end{aligned}$$

In a similar fashion, it can be shown that $\left\|\bar{\eta}_t^2\boldsymbol{E}_t\left(\boldsymbol{S}_t^\top\boldsymbol{S}_t+\boldsymbol{E}_t^\top\boldsymbol{E}_t\right)\boldsymbol{E}_t^\top\boldsymbol{H}_t\right\| \le \frac{1}{12}\left\|\boldsymbol{H}_t\right\|$ and $\left\|\bar{\eta}_t^2\boldsymbol{E}_t\boldsymbol{E}_t^\top\boldsymbol{H}_t\boldsymbol{S}_t\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top-\boldsymbol{\Sigma}\right)\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\right\| \le \frac{1}{12}\left\|\boldsymbol{H}_t\right\|$. Combining the derived bounds completes the proof. □

2.G.3 Proof of Lemma 24

To prove this lemma, we show that

$$\boldsymbol{S}_{t+1}\boldsymbol{N}_t = (\boldsymbol{I}+\boldsymbol{\Xi})\boldsymbol{S}_t\boldsymbol{N}_t \tag{2.118}$$

for some matrix $\boldsymbol{\Xi}$ with $\|\boldsymbol{\Xi}\| < 1$. Before proceeding, we show that the above inequality is enough to prove the invertibility of $\boldsymbol{S}_{t+1}\boldsymbol{N}_t$. First, note that $\boldsymbol{S}_t\boldsymbol{N}_t = M_t\boldsymbol{D}_t$. Therefore, the matrix $\boldsymbol{M}_t\boldsymbol{D}_t$ is invertible due to the assumption $\boldsymbol{S}_t\boldsymbol{S}_t^\top \succ \boldsymbol{0}$. On the other hand, $\|\boldsymbol{\Xi}\| < 1$

implies that $\boldsymbol{I}+\boldsymbol{\Xi}$ is invertible, thereby completing the proof. To verify equation 2.118, it suffices to show that $\boldsymbol{\Xi}=\boldsymbol{S}_{t+1}\boldsymbol{N}_t\left(\boldsymbol{S}_t\boldsymbol{N}_t\right)^{-1}-\boldsymbol{I}$ has norm less than one. To this goal, we write

$$
\begin{aligned}
\boldsymbol{\Xi} =&\boldsymbol{S}_{t+1}\boldsymbol{N}_t\left(\boldsymbol{S}_t\boldsymbol{N}_t\right)^{-1}-\boldsymbol{I}\\
=&\boldsymbol{V}^\top\left(\boldsymbol{U}_t-\bar{\eta}_t\left(\boldsymbol{U}_t\boldsymbol{U}_t^\top-\boldsymbol{X}^\star\right)\boldsymbol{U}_t+\boldsymbol{R}_t\boldsymbol{U}_t\right)\boldsymbol{N}_t\left(\boldsymbol{V}^\top\boldsymbol{U}_t\boldsymbol{N}_t\right)^{-1}-\left(\boldsymbol{V}^\top\boldsymbol{U}_t\boldsymbol{N}_t\right)\left(\boldsymbol{V}^\top\boldsymbol{U}_t\boldsymbol{N}_t\right)^{-1}\\
=&\boldsymbol{V}^\top\left(-\bar{\eta}_t\left(\boldsymbol{U}_t\boldsymbol{U}_t^\top-\boldsymbol{X}^\star\right)+\boldsymbol{R}_t\right)\boldsymbol{U}_t\boldsymbol{N}_t\left(\boldsymbol{V}^\top\boldsymbol{U}_t\boldsymbol{N}_t\right)^{-1}\\
\overset{(a)}{=}&\boldsymbol{V}^\top\left(-\bar{\eta}_t\left(\boldsymbol{U}_t\boldsymbol{U}_t^\top-\boldsymbol{X}^\star\right)+\boldsymbol{R}_t\right)\boldsymbol{U}_t\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\\
=&\boldsymbol{V}^\top\left(-\bar{\eta}_t\left(\boldsymbol{U}_t\boldsymbol{U}_t^\top-\boldsymbol{X}^\star\right)+\boldsymbol{R}_t\right)\boldsymbol{V}+\boldsymbol{V}^\top\left(-\bar{\eta}_t\left(\boldsymbol{U}_t\boldsymbol{U}_t^\top-\boldsymbol{X}^\star\right)+\boldsymbol{R}_t\right)\boldsymbol{V}_\perp\boldsymbol{H}_t,
\end{aligned}
$$

where $\boldsymbol{H}_t$ is defined as $\boldsymbol{E}_t\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}$. Moreover, in (a), we used the following chain of equalities

$$
\boldsymbol{N}_t\left(\boldsymbol{V}^\top\boldsymbol{U}_t\boldsymbol{N}_t\right)^{-1}=\boldsymbol{N}_t\left(\boldsymbol{S}_t\boldsymbol{N}_t\right)^{-1}=\boldsymbol{N}_t\boldsymbol{D}_t^{-1}M_t^\top=\boldsymbol{N}_t\boldsymbol{D}_tM_t^\top M_t\boldsymbol{D}_t^{-2}M_t^\top=\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}
$$

Therefore, we have

$$
\begin{aligned}
\|\boldsymbol{\Xi}\| &\le \left\|-\bar{\eta}_t\left(\boldsymbol{U}_t\boldsymbol{U}_t^\top-\boldsymbol{X}^\star\right)+\boldsymbol{R}_t\right\|\left(1+\|\boldsymbol{H}_t\|\right)\\
&\le 2\bar{\eta}_t\|\boldsymbol{\Delta}_t\|+6\bar{\eta}_t\delta\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}\le 3\bar{\eta}_t\|\boldsymbol{\Delta}_t\|\le 1/2.
\end{aligned}
\tag{2.119}
$$

Here we used Lemma 26, $\|\boldsymbol{\Delta}_t\|\lesssim\sigma_1$, $\|\boldsymbol{R}_t\|\le 3\bar{\eta}_t\delta\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}$, $\|\boldsymbol{H}_t\|\le 1/3$, and our assumption on η. This completes the proof. □

2.G.4 Proof of Lemma 27

First, note that

$$
\begin{aligned}
\boldsymbol{E}_{s+1}\boldsymbol{S}_{s+1}^\top =& \boldsymbol{E}_s\boldsymbol{S}_s^\top+\bar{\eta}_t\boldsymbol{E}_s\boldsymbol{S}_s^\top(\boldsymbol{\Sigma}-\boldsymbol{S}_s\boldsymbol{S}_s^\top)-\bar{\eta}_t\boldsymbol{E}_s(\boldsymbol{S}_s^\top\boldsymbol{S}_s+\boldsymbol{E}_s^\top\boldsymbol{E}_s)\boldsymbol{S}_s^\top\\
&-\bar{\eta}_t\boldsymbol{E}_s\boldsymbol{E}_s^\top\boldsymbol{E}_s\boldsymbol{S}_s^\top+\bar{\eta}_t^2\boldsymbol{E}_s(\boldsymbol{S}_s^\top\boldsymbol{S}_s+\boldsymbol{E}_s^\top\boldsymbol{E}_s)\boldsymbol{S}_s^\top\left(\boldsymbol{S}_s\boldsymbol{S}_s^\top-\boldsymbol{\Sigma}\right)\\
&+\bar{\eta}_t^2\boldsymbol{E}_s\left(\boldsymbol{S}_s^\top\boldsymbol{S}_s+\boldsymbol{E}_s^\top\boldsymbol{E}_s\right)\boldsymbol{E}_s^\top\boldsymbol{E}_s\boldsymbol{S}_s^\top+\boldsymbol{V}_\perp^\top\boldsymbol{R}_s\boldsymbol{U}_s\boldsymbol{S}_{s+1}^\top\\
&+\boldsymbol{E}_{s+1}\boldsymbol{U}_s^\top\boldsymbol{R}_s^\top\boldsymbol{V}+\boldsymbol{V}_\perp^\top\boldsymbol{R}_s\boldsymbol{U}_s\boldsymbol{U}_s^\top\boldsymbol{R}_s^\top\boldsymbol{V}.
\end{aligned}
\tag{2.120}
$$

On the other hand, one can write

$$\begin{aligned}\boldsymbol{S}_{s+1}\boldsymbol{S}_{s+1}^{\top} &= \boldsymbol{S}_s\boldsymbol{S}_s^{\top} + \bar{\eta}_t\boldsymbol{S}_s\boldsymbol{S}_s^{\top}\left(\boldsymbol{\Sigma} - \boldsymbol{S}_s\boldsymbol{S}_s^{\top}\right) + \bar{\eta}_t\left(\boldsymbol{\Sigma} - \boldsymbol{S}_s\boldsymbol{S}_s^{\top}\right)\boldsymbol{S}_s\boldsymbol{S}_s^{\top} - 2\bar{\eta}_t\boldsymbol{S}_s\boldsymbol{E}_s^{\top}\boldsymbol{E}_s\boldsymbol{S}_s^{\top} \\ &+ \bar{\eta}_t^2\left(\boldsymbol{\Sigma} - \boldsymbol{S}_s\boldsymbol{S}_s^{\top}\right)\boldsymbol{S}_s\boldsymbol{S}_s^{\top}\left(\boldsymbol{\Sigma} - \boldsymbol{S}_s\boldsymbol{S}_s^{\top}\right) + \bar{\eta}_t^2\boldsymbol{S}_s\boldsymbol{E}_s^{\top}\boldsymbol{E}_s\boldsymbol{E}_s^{\top}\boldsymbol{E}_s\boldsymbol{S}_s^{\top} \\ &- \bar{\eta}_t^2\left(\boldsymbol{\Sigma} - \boldsymbol{S}_s\boldsymbol{S}_s^{\top}\right)\boldsymbol{S}_s\boldsymbol{E}_s^{\top}\boldsymbol{E}_s\boldsymbol{S}_s^{\top} - \bar{\eta}_t^2\boldsymbol{S}_s\boldsymbol{E}_s^{\top}\boldsymbol{E}_s\boldsymbol{S}_s^{\top}\left(\boldsymbol{\Sigma} - \boldsymbol{S}_s\boldsymbol{S}_s^{\top}\right) \\ &+ \boldsymbol{V}^{\top}\boldsymbol{R}_s\boldsymbol{U}_s\boldsymbol{S}_{s+1}^{\top} + \boldsymbol{S}_{s+1}\boldsymbol{U}_s^{\top}\boldsymbol{R}_s^{\top}\boldsymbol{V} + \boldsymbol{V}^{\top}\boldsymbol{R}_s\boldsymbol{U}_s\boldsymbol{U}_s^{\top}\boldsymbol{R}_s^{\top}\boldsymbol{V}.\end{aligned} \tag{2.121}$$

Pre-multiplying equation 2.121 with $\boldsymbol{H}_s$ leads to a relationship between $\boldsymbol{S}_{s+1}\boldsymbol{S}_{s+1}^{\top}$ and $\boldsymbol{E}_{s+1}\boldsymbol{S}_{s+1}^{\top}$:

$$\begin{aligned}\boldsymbol{E}_{s+1}\boldsymbol{S}_{s+1}^{T} &= \boldsymbol{H}_s\boldsymbol{S}_{s+1}\boldsymbol{S}_{s+1}^{\top} + \boldsymbol{T} \\ \Longrightarrow \boldsymbol{H}_{s+1} &= \boldsymbol{H}_s + \boldsymbol{T}(\boldsymbol{S}_{s+1}\boldsymbol{S}_{s+1})^{-1},\end{aligned}$$

where simple algebra reveals that

$$\begin{aligned}\boldsymbol{T} = &-H_s\left(\bar{\eta}_t\boldsymbol{\Sigma}\boldsymbol{S}_s\boldsymbol{S}_s^{\top} - 2\bar{\eta}_t\boldsymbol{S}_s\boldsymbol{E}_s^{\top}\boldsymbol{E}_s\boldsymbol{S}_s^{\top} + \bar{\eta}_t^2\boldsymbol{\Sigma}\boldsymbol{S}_s\boldsymbol{S}_s^{\top}\left(\boldsymbol{\Sigma} - \boldsymbol{S}_s\boldsymbol{S}_s^{\top}\right)\right. \\ &\left.+\bar{\eta}_t^2\boldsymbol{S}_s\boldsymbol{E}_s^{\top}\boldsymbol{E}_s\boldsymbol{E}_s^{\top}\boldsymbol{E}_s\boldsymbol{S}_s^{\top} - \bar{\eta}_t^2\boldsymbol{\Sigma}\boldsymbol{S}_s\boldsymbol{E}_s^{\top}\boldsymbol{E}_s\boldsymbol{S}_s^{\top} - \bar{\eta}_t^2\boldsymbol{S}_s\boldsymbol{E}_s^{\top}\boldsymbol{E}_s\boldsymbol{S}_s^{\top}\left(\boldsymbol{\Sigma} - \boldsymbol{S}_s\boldsymbol{S}_s^{\top}\right)\right) \\ &- 2\bar{\eta}_t\boldsymbol{E}_s\boldsymbol{E}_s^{\top}\boldsymbol{E}_s\boldsymbol{S}_s^{\top} + \bar{\eta}_t^2\boldsymbol{E}_s\boldsymbol{E}_s^{\top}\boldsymbol{E}_s\boldsymbol{S}_s^{\top}\left(\boldsymbol{S}_s\boldsymbol{S}_s^{\top} - \boldsymbol{\Sigma}\right) + \bar{\eta}_t^2\boldsymbol{E}_s\boldsymbol{E}_s^{\top}\boldsymbol{E}_s\boldsymbol{E}_s^{\top}\boldsymbol{E}_s\boldsymbol{S}_s^{\top} \\ &+ \boldsymbol{V}_{\perp}^{\top}\boldsymbol{R}_s\boldsymbol{U}_s\boldsymbol{S}_{s+1}^{\top} + \boldsymbol{E}_{s+1}\boldsymbol{U}_s^{\top}\boldsymbol{R}_s^{\top}\boldsymbol{V} + \boldsymbol{V}_{\perp}^{\top}\boldsymbol{R}_s\boldsymbol{U}_s\boldsymbol{U}_s^{\top}\boldsymbol{R}_s^{\top}\boldsymbol{V} \\ &- H_s\left(\boldsymbol{V}^{\top}\boldsymbol{R}_s\boldsymbol{U}_s\boldsymbol{S}_{s+1}^{\top} + \boldsymbol{S}_{s+1}\boldsymbol{U}_s^{\top}\boldsymbol{R}_s^{\top}\boldsymbol{V} + \boldsymbol{V}^{\top}\boldsymbol{R}_s\boldsymbol{U}_s\boldsymbol{U}_s^{\top}\boldsymbol{R}_s^{\top}\boldsymbol{V}\right).\end{aligned}$$

For simplicity, we define $\boldsymbol{D} = \boldsymbol{S}_s\boldsymbol{S}_s^{\top}\left(\boldsymbol{S}_{s+1}\boldsymbol{S}_{s+1}^{\top}\right)^{-1}$ in the sequel. Based on the definition of $\boldsymbol{T}$, one can write

$$\boldsymbol{H}_s + \boldsymbol{T}(\boldsymbol{S}_{s+1}\boldsymbol{S}_{s+1})^{-1} = \boldsymbol{H}_s\boldsymbol{A}_s + \boldsymbol{B}_s + \boldsymbol{C}_s, \tag{2.122}$$

where the matrices $\boldsymbol{A}_s$, $\boldsymbol{B}_s$, and $\boldsymbol{C}_s$ are defined as

$$\begin{aligned}\boldsymbol{A}_s =&\boldsymbol{I} - \bar{\eta}_t\boldsymbol{\Sigma}\boldsymbol{D} - \bar{\eta}_t^2\boldsymbol{\Sigma}\left(\boldsymbol{S}_s\boldsymbol{S}_s^{\top}\right)\boldsymbol{\Sigma}\left(\boldsymbol{S}_s\boldsymbol{S}_s^{\top}\right)^{-1}\boldsymbol{D} + \bar{\eta}_t^2\boldsymbol{\Sigma}\boldsymbol{S}_s\boldsymbol{S}_s^{\top}\boldsymbol{D}, \\ \boldsymbol{B}_s =& -2\bar{\eta}_t\boldsymbol{S}_s\boldsymbol{E}_s^{\top}\boldsymbol{H}_s\boldsymbol{D} - \bar{\eta}_t^2\boldsymbol{S}_s\boldsymbol{E}_s^{\top}\boldsymbol{E}_s\boldsymbol{E}_s^{\top}\boldsymbol{H}_s\boldsymbol{D} \\ &- \bar{\eta}_t^2\boldsymbol{\Sigma}\boldsymbol{S}_s\boldsymbol{E}_s^{\top}\boldsymbol{H}_s\boldsymbol{D} - \bar{\eta}_t^2\boldsymbol{S}_s\boldsymbol{E}_s^{\top}\boldsymbol{E}_s\boldsymbol{S}_s^{\top}\boldsymbol{\Sigma}\left(\boldsymbol{S}_s\boldsymbol{S}_s^{\top}\right)^{-1}\boldsymbol{D} - \bar{\eta}_t^2\boldsymbol{S}_s\boldsymbol{E}_s^{\top}\boldsymbol{E}_s\boldsymbol{S}_s^{\top}\boldsymbol{D} \\ &- \left(\boldsymbol{V}^{\top}\boldsymbol{R}_s\boldsymbol{U}_s\boldsymbol{S}_{s+1}^{\top} + \boldsymbol{S}_{s+1}\boldsymbol{U}_s^{\top}\boldsymbol{R}_s^{\top}\boldsymbol{V} + \boldsymbol{V}^{\top}\boldsymbol{R}_s\boldsymbol{U}_s\boldsymbol{U}_s^{\top}\boldsymbol{R}_s^{\top}\boldsymbol{V}\right)\left(\boldsymbol{S}_{s+1}\boldsymbol{S}_{s+1}^{\top}\right)^{-1}, \\ \boldsymbol{C}_s =& -2\bar{\eta}_t\boldsymbol{E}_s\boldsymbol{E}_s^{\top}\boldsymbol{H}_s\boldsymbol{D} + \bar{\eta}_t^2\boldsymbol{E}_s\boldsymbol{E}_s^{\top}\boldsymbol{H}_s\boldsymbol{S}_s\boldsymbol{S}_s^{\top}\left(\boldsymbol{S}_s\boldsymbol{S}_s^{\top} - \boldsymbol{\Sigma}\right)\left(\boldsymbol{S}_{s+1}\boldsymbol{S}_{s+1}^{\top}\right)^{-1} \\ &+ \left(\boldsymbol{V}_{\perp}^{\top}\boldsymbol{R}_s\boldsymbol{U}_s\boldsymbol{S}_{s+1}^{\top} + \boldsymbol{E}_{s+1}\boldsymbol{U}_s^{\top}\boldsymbol{R}_s^{\top}\boldsymbol{V} + \boldsymbol{V}_{\perp}^{\top}\boldsymbol{R}_s\boldsymbol{U}_s\boldsymbol{U}_s^{\top}\boldsymbol{R}_s^{\top}\boldsymbol{V}\right)\left(\boldsymbol{S}_{s+1}\boldsymbol{S}_{s+1}^{\top}\right)^{-1} \\ &+ \bar{\eta}_t^2\boldsymbol{E}_s\boldsymbol{E}_s^{\top}\boldsymbol{E}_s\boldsymbol{E}_s^{\top}\boldsymbol{H}_s\boldsymbol{D}.\end{aligned}$$

To provide an upper bound on $\|\boldsymbol{A}_s\|$, we define $\boldsymbol{P} = \left(\boldsymbol{I} + \bar{\eta}_t \boldsymbol{S}_s \boldsymbol{S}_s^\top \boldsymbol{\Sigma} \left(\boldsymbol{S}_s \boldsymbol{S}_s^\top\right)^{-1} - \bar{\eta}_t \boldsymbol{S}_s \boldsymbol{S}_s^\top\right) \boldsymbol{D}$ and $\boldsymbol{Q} = \boldsymbol{I} + \bar{\eta}_t \boldsymbol{S}_s \boldsymbol{S}_s^\top \boldsymbol{\Sigma} \left(\boldsymbol{S}_s \boldsymbol{S}_s^\top\right)^{-1} - \bar{\eta}_t \boldsymbol{S}_s \boldsymbol{S}_s^\top$. We have

$$\begin{aligned}
\|\boldsymbol{A}_s\|^2 = \|\boldsymbol{I} - \bar{\eta}_t \boldsymbol{\Sigma} \boldsymbol{P}\|^2 &= \lambda_{\max}\left(\left(\boldsymbol{I} - \bar{\eta}_t \boldsymbol{P} \boldsymbol{\Sigma}\right)^\top \left(\boldsymbol{I} - \bar{\eta}_t \boldsymbol{\Sigma} \boldsymbol{P}\right)\right) \\
&\leq 1 + \lambda_{\max}\left(-\bar{\eta}_t \boldsymbol{\Sigma} \boldsymbol{P} - \bar{\eta}_t \boldsymbol{P}^\top \boldsymbol{\Sigma} + \bar{\eta}_t^2 \boldsymbol{P}^\top \boldsymbol{\Sigma}^2 \boldsymbol{P}\right) \\
&= 1 - \bar{\eta}_t \lambda_{\min}\left(\boldsymbol{\Sigma} \boldsymbol{P} + \boldsymbol{P}^\top \boldsymbol{\Sigma} - \bar{\eta}_t \boldsymbol{P}^\top \boldsymbol{\Sigma}^2 \boldsymbol{P}\right) \\
&\leq 1 - \bar{\eta}_t \lambda_{\min}\left(\boldsymbol{\Sigma} \boldsymbol{P} + \boldsymbol{P}^\top \boldsymbol{\Sigma} - \boldsymbol{P}^\top \boldsymbol{\Sigma} \boldsymbol{P}\right),
\end{aligned}$$

where we used the fact that $\eta \lesssim \frac{1}{\sigma_1}$ in the last inequality. Now it suffices to provide a lower bound for $\lambda_{\min}\left(\boldsymbol{\Sigma} \boldsymbol{P} + \boldsymbol{P}^\top \boldsymbol{\Sigma} - \boldsymbol{P}^\top \boldsymbol{\Sigma} \boldsymbol{P}\right)$. To this goal, we use the following intermediate lemma.

Lemma 53 (Theorem 4.1. in Eisenstat and Ipsen [1998])**.** *Given a diagonal matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{d \times d}$ and its perturbed variant $\boldsymbol{\Lambda}' = \boldsymbol{\Lambda} \boldsymbol{R}$ for some $\boldsymbol{R} \in \mathbb{R}^{d \times d}$, we have*

$$\min_k |\lambda_i(\boldsymbol{\Lambda}') - \lambda_k(\boldsymbol{\Lambda})| \leq |\lambda_i(\boldsymbol{\Lambda}')| \left\|\boldsymbol{I} - \boldsymbol{R}^{-1}\right\|.$$

for every $i = 1, 2, \cdots, d$.

To apply this lemma, we choose $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}$ and $\boldsymbol{R} = \boldsymbol{P} + \boldsymbol{\Sigma}^{-1} \boldsymbol{P}^\top \boldsymbol{\Sigma} - \boldsymbol{\Sigma}^{-1} \boldsymbol{P}^\top \boldsymbol{\Sigma} P$, which leads to the equality $\boldsymbol{\Sigma} \boldsymbol{P} + \boldsymbol{P}^\top \boldsymbol{\Sigma} - \boldsymbol{P}^\top \boldsymbol{\Sigma} \boldsymbol{P} = \boldsymbol{\Sigma} \boldsymbol{R}$. Given this definition and Lemma 53, we have

$$\frac{\sigma_r}{1 + \|\boldsymbol{I} - \boldsymbol{R}^{-1}\|} \leq \lambda_{\min}(\boldsymbol{\Sigma} \boldsymbol{R}). \tag{2.123}$$

Now, we provide an upper bound for $\|\boldsymbol{I} - \boldsymbol{R}^{-1}\|$. First note that $\|\boldsymbol{I} - \boldsymbol{R}^{-1}\| \leq \|\boldsymbol{I} - \boldsymbol{R}\| \, \|\boldsymbol{R}^{-1}\|$. On the other hand

$$\|\boldsymbol{I} - \boldsymbol{R}\| = \left\|\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{P}^\top - \boldsymbol{I}\right) \boldsymbol{\Sigma} \left(\boldsymbol{I} - \boldsymbol{P}\right)\right\| \leq \|\boldsymbol{I} - \boldsymbol{P}\|^2 \leq \left(\|\boldsymbol{I} - \boldsymbol{D}\| + \|\left(\boldsymbol{I} - \boldsymbol{Q}\right) \boldsymbol{D}\|\right)^2.$$

Using a similar approach to the proof of Lemma 23, we have

$$\|\boldsymbol{I} - \boldsymbol{D}\| = \left\|\left(\boldsymbol{S}_{t+1} \boldsymbol{S}_{t+1}^\top - \boldsymbol{S}_t \boldsymbol{S}_t^\top\right)\left(\boldsymbol{S}_{t+1} \boldsymbol{S}_{t+1}^\top\right)^{-1}\right\| \leq 0.1.$$

On the other hand, one can write

$$\begin{aligned}\|(\boldsymbol{I}-\boldsymbol{Q})\boldsymbol{D}\| &\leq \|\boldsymbol{I}-\boldsymbol{Q}\|\,\|\boldsymbol{D}\| \\ &\overset{(a)}{\leq} 3\,\|\boldsymbol{I}-\boldsymbol{Q}\| \\ &= \bar{\eta}_t \left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\left(\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\right\| \\ &\leq \bar{\eta}_t \left\|\boldsymbol{\Sigma}-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| \\ &\leq 0.1,\end{aligned}$$

where, in (a), we used Lemma 23. Therefore, we have $\|\boldsymbol{I}-\boldsymbol{R}\| \leq 0.04$. Next, we provide an upper bound for $\|\boldsymbol{R}^{-1}\|$. Note that

$$\boldsymbol{R} = \boldsymbol{P} + \boldsymbol{\Sigma}^{-1}\boldsymbol{P}^\top\boldsymbol{\Sigma}\left(\boldsymbol{I}-\boldsymbol{P}\right).$$

By Weyl's inequality, we have

$$\begin{aligned}\sigma_{\min}(\boldsymbol{R}) &\geq \sigma_{\min}(\boldsymbol{P}) - \left\|\boldsymbol{\Sigma}^{-1}\boldsymbol{P}^\top\boldsymbol{\Sigma}\left(\boldsymbol{I}-\boldsymbol{P}\right)\right\| \\ &\geq \sigma_{\min}(\boldsymbol{P}) - \|\boldsymbol{P}\|\,\|\boldsymbol{I}-\boldsymbol{P}\| \\ &\geq 0.8 - 1.2\times 0.2 = 0.56.\end{aligned}$$

Here we used the fact that $\|\boldsymbol{I}-\boldsymbol{P}\| \leq 0.2$. The above inequality implies that $\|\boldsymbol{R}^{-1}\| = 1/\sigma_{\min}(\boldsymbol{R}) \leq 2$. Combining the above bounds, we have

$$\left\|\boldsymbol{I}-\boldsymbol{R}^{-1}\right\| \leq \|\boldsymbol{I}-\boldsymbol{R}\|\left\|\boldsymbol{R}^{-1}\right\| \leq 0.08.$$

This together with equation 2.123 implies that

$$\lambda_{\min}\left(\boldsymbol{\Sigma}\boldsymbol{R}\right) \geq 0.92\sigma_r.$$

Therefore, we have

$$\|\boldsymbol{A}_s\|^2 \leq 1 - 0.92\bar{\eta}_t\sigma_r \quad \Longrightarrow \quad \|\boldsymbol{A}_s\| \leq 1 - 0.46\bar{\eta}_t\sigma_r. \tag{2.124}$$

Next, we provide an upper bound for $\|\boldsymbol{B}_s\|$. Simple algebra reveals that

$$\begin{aligned}
&\Big\|2\bar{\eta}_t \boldsymbol{S}_s \boldsymbol{E}_s^\top \boldsymbol{H}_s \boldsymbol{D} + \bar{\eta}_t^2 \boldsymbol{S}_s \boldsymbol{E}_s^\top \boldsymbol{E}_s \boldsymbol{E}_s^\top \boldsymbol{H}_s \boldsymbol{D} + \bar{\eta}_t^2 \boldsymbol{\Sigma} \boldsymbol{S}_s \boldsymbol{E}_s^\top \boldsymbol{H}_s \boldsymbol{D} \\
&+\bar{\eta}_t^2 \boldsymbol{S}_s \boldsymbol{E}_s^\top \boldsymbol{E}_s \boldsymbol{S}_s^\top \boldsymbol{\Sigma} \left(\boldsymbol{S}_s \boldsymbol{S}_s^\top\right)^{-1} \boldsymbol{D} + \bar{\eta}_t^2 \boldsymbol{S}_s \boldsymbol{E}_s^\top \boldsymbol{E}_s \boldsymbol{S}_s^\top \boldsymbol{D}\Big\| \\
&\lesssim \bar{\eta}_t \left\|\boldsymbol{S}_s \boldsymbol{E}_s^\top\right\| \|\boldsymbol{H}_s\| .
\end{aligned}$$

Next, we provide a bound for the remaining terms in $\boldsymbol{B}_s$. We have

$$\begin{aligned}
&\left\|\boldsymbol{V}_\perp^\top \boldsymbol{R}_s \boldsymbol{U}_s \boldsymbol{S}_{s+1}^\top \left(\boldsymbol{S}_{s+1} \boldsymbol{S}_{s+1}^\top\right)^{-1}\right\| \\
&\le \|\boldsymbol{R}_s\| \left\|\left(\boldsymbol{I} - \bar{\eta}_t \left(\boldsymbol{U}_s \boldsymbol{U}_s^\top - \boldsymbol{X}^\star\right) + \boldsymbol{R}_s\right)^{-1} \boldsymbol{U}_{s+1} \boldsymbol{S}_{s+1}^\top \left(\boldsymbol{S}_{s+1} \boldsymbol{S}_{s+1}^\top\right)^{-1}\right\| \\
&\le \|\boldsymbol{R}_s\| \left\|\left(\boldsymbol{I} - \bar{\eta}_t \left(\boldsymbol{U}_s \boldsymbol{U}_s^\top - \boldsymbol{X}^\star\right) + \boldsymbol{R}_s\right)^{-1}\right\| \left\|\boldsymbol{U}_{s+1} \boldsymbol{S}_{s+1}^\top \left(\boldsymbol{S}_{s+1} \boldsymbol{S}_{s+1}^\top\right)^{-1}\right\| \\
&\lesssim \bar{\eta}_t \delta \|\boldsymbol{\Delta}_s\|_{\mathrm{F}} \left\|\boldsymbol{U}_{s+1} \boldsymbol{S}_{s+1}^\top \left(\boldsymbol{S}_{s+1} \boldsymbol{S}_{s+1}^\top\right)^{-1}\right\| .
\end{aligned}$$

To proceed, we provide an upper bound for $\left\|\boldsymbol{U}_{s+1} \boldsymbol{S}_{s+1}^\top \left(\boldsymbol{S}_{s+1} \boldsymbol{S}_{s+1}^\top\right)^{-1}\right\|$.

$$\begin{aligned}
&\left\|\boldsymbol{U}_{s+1} \boldsymbol{S}_{s+1}^\top \left(\boldsymbol{S}_{s+1} \boldsymbol{S}_{s+1}^\top\right)^{-1}\right\| \\
&\le \left\|\boldsymbol{V} \boldsymbol{S}_{s+1}^\top \boldsymbol{S}_{s+1}^\top \left(\boldsymbol{S}_{s+1} \boldsymbol{S}_{s+1}^\top\right)^{-1} + \boldsymbol{V}_\perp \boldsymbol{E}_{s+1} \boldsymbol{S}_{s+1}^\top \left(\boldsymbol{S}_{s+1} \boldsymbol{S}_{s+1}^\top\right)^{-1}\right\| \\
&\le 1 + \|\boldsymbol{H}_{s+1}\| .
\end{aligned}$$

Similarly, one can show that

$$\left\|\boldsymbol{V}_\perp^\top \boldsymbol{R}_s \boldsymbol{U}_s \boldsymbol{U}_s^\top \boldsymbol{R}_s^\top \boldsymbol{V} \left(\boldsymbol{S}_{s+1} \boldsymbol{S}_{s+1}^\top\right)^{-1}\right\| \lesssim \bar{\eta}_t^2 \delta^2 \bar{\varphi}^4 \|\boldsymbol{\Delta}_s\|_{\mathrm{F}}^2 (1 + \|\boldsymbol{H}_s\|) \lesssim \bar{\eta}_t \delta \|\boldsymbol{\Delta}_s\|_{\mathrm{F}} .$$

Combining the derived bounds leads to

$$\|\boldsymbol{B}_s\| \lesssim \bar{\eta}_t \left\|\boldsymbol{S}_s \boldsymbol{E}_s^\top\right\| \|\boldsymbol{H}_s\| + \bar{\eta}_t \delta \|\boldsymbol{\Delta}_s\|_{\mathrm{F}} (1 + \|\boldsymbol{H}_{s+1}\|). \tag{2.125}$$

In a similar way, one can show that

$$\|\boldsymbol{C}_s\| \lesssim \bar{\eta}_t \left\|\boldsymbol{E}_s \boldsymbol{E}_s^\top\right\| \|\boldsymbol{H}_s\| + \bar{\eta}_t \delta \|\boldsymbol{\Delta}_s\|_{\mathrm{F}} (1 + \|\boldsymbol{H}_{s+1}\|). \tag{2.126}$$

Substituting equation 2.124, equation 2.125, and equation 2.126 in equation 2.122 yields

$$
\begin{aligned}
(1 - c_1\bar{\eta}_t\delta \left\|\boldsymbol{\Delta}_s\right\|_{\mathrm{F}}) \left\|\boldsymbol{H}_{s+1}\right\| &\leq \big(1 - 0.46\bar{\eta}_t\sigma_r + c_2\bar{\eta}_t \left(\left\|\boldsymbol{S}_s\boldsymbol{E}_s^\top\right\| + \left\|\boldsymbol{E}_s\boldsymbol{E}_s^\top\right\|\right) \\
&\qquad + c_3\bar{\eta}_t\delta \left\|\boldsymbol{\Delta}_s\right\|_{\mathrm{F}}) \left\|\boldsymbol{H}_s\right\| + c_4\bar{\eta}_t\delta \left\|\boldsymbol{\Delta}_s\right\|_{\mathrm{F}} \\
\implies \left\|\boldsymbol{H}_{s+1}\right\| &\leq \left(\frac{1 - 0.46\bar{\eta}_t\sigma_r + c_2\bar{\eta}_t \left(\left\|\boldsymbol{S}_s\boldsymbol{E}_s^\top\right\| + \left\|\boldsymbol{E}_s\boldsymbol{E}_s^\top\right\|\right) + c_3\bar{\eta}_t\delta \left\|\boldsymbol{\Delta}_s\right\|_{\mathrm{F}}}{1 - c_1\bar{\eta}_t\delta \left\|\boldsymbol{\Delta}_s\right\|_{\mathrm{F}}}\right) \left\|\boldsymbol{H}_s\right\| \\
&\qquad + \frac{c_4\bar{\eta}_t\delta \left\|\boldsymbol{\Delta}_s\right\|_{\mathrm{F}}}{1 - c_1\bar{\eta}_t\delta \left\|\boldsymbol{\Delta}_s\right\|_{\mathrm{F}}} \\
\implies \left\|\boldsymbol{H}_{s+1}\right\| &\leq (1 - c_5\bar{\eta}_t\sigma_r) \left\|\boldsymbol{H}_s\right\| + c_6\sqrt{r}\sigma_1\bar{\eta}_t\delta,
\end{aligned}
$$

where the last inequality follows from the assumed upper bound on δ, as well as equation 2.38-equation 2.40. This completes the proof. □

2.G.5 Proof of Lemma 10

We first prove the upper bound. One can write

$$
\log\left(\prod_{t=0}^{T}\left(1 + \alpha\rho^t\right)\right) = \sum_{t=0}^{T}\log\left(1 + \alpha\rho^t\right) \leq \sum_{t=0}^{\infty}\alpha\rho^t = \frac{\alpha}{1-\rho}.
$$

Hence, we have $\prod_{t=0}^{T}\left(1 + \alpha\rho^t\right) \leq \exp\left(\frac{\alpha}{1-\rho}\right)$. For the lower bound, we have

$$
\begin{aligned}
\log\left(\prod_{t=0}^{\infty}\left(1 + \alpha\rho^t\right)\right) &= \sum_{t=0}^{T}\log\left(1 + \alpha\rho^t\right) \\
&\geq \sum_{t=0}^{T}\frac{\alpha\rho^t}{1 + \alpha\rho^t} \\
&\geq \sum_{t=0}^{T}\frac{\alpha\rho^t}{1 + \alpha} \\
&\geq \frac{\alpha}{1+\alpha}T\rho^T.
\end{aligned}
$$

Hence, we have $\prod_{t=0}^{T}\left(1 + \alpha\rho^t\right) \geq \exp\left(\frac{\alpha}{1+\alpha}T\rho^T\right)$. □

2.G.6 Proof of Lemma 11

We start with the proof of the first statement. We consider two cases:

- Suppose that $\left\|\boldsymbol{\Delta}_t\right\| \leq 0.01\sigma_r\rho^t$. Recall that $\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \frac{\eta_0\rho^t}{\left\|\boldsymbol{\Delta}_t\right\|}\boldsymbol{\Delta}_t\boldsymbol{U}_t + \boldsymbol{R}_t\boldsymbol{U}_t$. Hence, we

have

$$\boldsymbol{\Delta}_{t+1} = \left(\boldsymbol{U}_t - \frac{\eta_0\rho^t}{\|\boldsymbol{\Delta}_t\|}\boldsymbol{\Delta}_t\boldsymbol{U}_t + \boldsymbol{R}_t\boldsymbol{U}_t\right)\left(\boldsymbol{U}_t^\top - \frac{\eta_0\rho^t}{\|\boldsymbol{\Delta}_t\|}\boldsymbol{U}_t^\top\boldsymbol{\Delta}_t^\top + \boldsymbol{U}_t^\top\boldsymbol{R}_t^\top\right) - \boldsymbol{X}^\star.$$

The above equality leads to

$$\begin{aligned}
\|\boldsymbol{\Delta}_{t+1}\| &\leq \|\boldsymbol{\Delta}_t\| + \eta\rho^t \left\|\boldsymbol{U}_t\boldsymbol{U}_t^\top\right\| (2 + 2\|\boldsymbol{R}_t\|) \\
&\quad + 2\left\|\boldsymbol{U}_t\boldsymbol{U}_t^\top\right\| \|\boldsymbol{R}_t\| + \left\|\boldsymbol{U}_t\boldsymbol{U}_t^\top\right\| \|\boldsymbol{R}_t\|^2 + \eta^2\rho^{2t}\left\|\boldsymbol{U}_t\boldsymbol{U}_t^\top\right\| \\
&\overset{(a)}{\leq} \|\boldsymbol{\Delta}_t\| + 4\sigma_1\eta\rho^t + 4\sigma_1\eta\delta\rho^t\|\boldsymbol{\Delta}_t\|_{\mathrm{F}} + 2\sigma_1\eta^2\rho^{2t} \\
&\leq 0.02\sigma_r\rho^t.
\end{aligned}$$

On the other hand, we know that $\gamma_{t+1} \leq 5\|\boldsymbol{\Delta}_{t+1}\|$, which, together with the above inequality, implies that $\gamma_{t+1} \leq 0.1\sigma_r$.

- Suppose that $\|\boldsymbol{\Delta}_t\| \geq 0.01\sigma_r\rho^t$. Therefore, we have $\frac{\eta_0\rho^t}{\|\boldsymbol{\Delta}_t\|} \lesssim \frac{1}{\sigma_1}$. On the other hand, since $\gamma_t \leq 0.1\sigma_r$, we have

$$\lambda_{\min}\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right) \geq 0.9\sigma_r, \quad \left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| \leq 1.1\sigma_1,$$

and

$$\left\|\boldsymbol{E}_t\boldsymbol{E}_t^\top\right\| \leq 0.1\sigma_r, \quad \left\|\boldsymbol{E}_t\boldsymbol{S}_t^\top\left(\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)^{-1}\right\| \leq 0.2.$$

This implies that the assumptions of Propositions 8, 9, and 10 are satisfied at iteration t, and we have

$$\begin{aligned}
\gamma_{t+1} &\leq \left\|\boldsymbol{\Sigma} - \boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^\top\right\| + 2\left\|\boldsymbol{S}_{t+1}\boldsymbol{E}_{t+1}^\top\right\| + \|\boldsymbol{F}_{t+1}\|^2 + \|\boldsymbol{G}_{t+1}\|^2 \\
&\overset{(a)}{\leq} \left(1 - \Omega(1)\frac{\sigma_r\eta_0\rho^t}{\gamma_t}\right)\gamma_t + \mathcal{O}(1)\sqrt{r}\sigma_1\eta_0\delta\rho^t + \mathcal{O}\left(1\right)\frac{\sigma_r\eta_0\rho^t}{\|\boldsymbol{\Delta}_t\|}\|\boldsymbol{G}_t\|^2 \\
&\leq \gamma_t - \Omega\left(\sigma_r\eta_0\rho^t\right) + \mathcal{O}\left(1\right)\sigma_r\eta_0\rho^t\frac{\|\boldsymbol{G}_t\|^2}{\|\boldsymbol{\Delta}_t\|} \\
&\overset{(b)}{\leq} \gamma_t - \Omega\left(\sigma_r\eta_0\rho^t\right) + \mathcal{O}\left(\sigma_r\eta_0\rho^t/\sqrt{d}\right) \\
&\leq \gamma_t - \Omega\left(\sigma_r\eta_0\rho^t\right),
\end{aligned}$$

where (a) follows from the one-step dynamics of the signal, cross, and residual terms derived in Propositions 8, 9, and 10. Moreover, (b) follows from $\|\boldsymbol{\Delta}_t\| \geq \sqrt{d}\|\boldsymbol{G}_t\|^2$. Therefore, we have $\gamma_{t+1} \leq \gamma_t \leq 0.1\sigma_r$.

To complete the proof of this lemma, it suffices to show that if $\|\boldsymbol{\Delta}_{t+1}\| \leq \sqrt{d}\|\boldsymbol{G}_{t+1}\|^2$,

then $\|\boldsymbol{\Delta}_{t+1}\| \leq \sqrt{d}\alpha^{1-\mathcal{O}(\sqrt{r}\kappa\delta)}$. Note that based on our assumption and Phase 1 of the proof of Theorem 8, the one-step dynamic of $\boldsymbol{G}_s$ holds for every $0 \leq s \leq t+1$. Therefore, an analysis similar to Lemma equation 9 leads to $\|\boldsymbol{\Delta}_{t+1}\| \leq \sqrt{d}\alpha^{1-\mathcal{O}(\sqrt{r}\kappa\delta)}$. □

2.G.7 Proof of Lemma 12

Since $\|\boldsymbol{\Delta}_{t_0+T_3-1}\| \leq 0.02\sigma_r\rho^{t_0-1}$, an argument similar to the proof of Lemma 11 can be invoked to show that $\|\boldsymbol{\Delta}_{t_0+T_3}\| \leq 0.03\sigma_r\rho^{t_0}$ and $\gamma_{t_0+T_3} \leq 0.15\sigma_r\rho^{t_0}$. Let Δt be the first time that $\|\boldsymbol{\Delta}_{t_0+T_3+\Delta t}\| \leq 0.02\sigma_r\rho^{t_0+\Delta t}$. Note that since $\|\boldsymbol{\Delta}_{t_0+T_3}\| > 0.02\sigma_r\rho^{t_0}$, we have $\Delta t \geq 1$. This implies that, for every $0 \leq s \leq \Delta t - 1$, we have $\|\boldsymbol{\Delta}_{t_0+T_3+s}\| > 0.02\sigma_r\rho^{t_0+s}$. Therefore, equation 2.127 implies that $\gamma_{t_0+T_3+s+1} \leq \gamma_{t_0+T_3+s} - \Omega(\sigma_r\eta\rho^{t_0+s})$. This in turn leads to

$$\begin{aligned}
\gamma_{t_0+T_3+\Delta t} &\leq \gamma_{t_0+T_3} - \Omega\left(\sum_{s=0}^{\Delta t-1} \sigma_r\eta\rho^{t_0+s}\right) \\
&\leq 0.15\sigma_r\rho^{t_0} - \Omega\left(\sum_{s=0}^{\Delta t-1} \sigma_r\eta\rho^{t_0+s}\right) \\
&= \sigma_r\rho^{t_0}\left(0.15 - \Omega\left(\sum_{s=0}^{\Delta t-1} \eta\rho^{s}\right)\right).
\end{aligned}$$

Let us assume that $\Delta t \lesssim (\kappa/\eta)\log(1/\alpha)$. Under this assumption, we have $\rho^s = \Omega(1)$ for every $s \leq \Delta t$. This implies that $\Omega\left(\sum_{s=0}^{\Delta t-1} \eta\rho^s\right) = \Omega(\eta\Delta t)$. Therefore, upon choosing $\Delta t = \Omega(1/\eta)$, we have $\gamma_{t_0+T_3+\Delta t} \leq 0.15 - \Omega\left(\sum_{s=0}^{\Delta t-1} \eta\rho^s\right) \leq 0$. This implies that, there must exist $\tilde{t} \leq \Delta t$ such that $\gamma_{t_0+T_3+\tilde{t}} \leq 0.02\sigma_r\rho^{t_0+\tilde{t}}$. This completes the proof. □

CHAPTER 3

Robust Deep Linear Neural Network

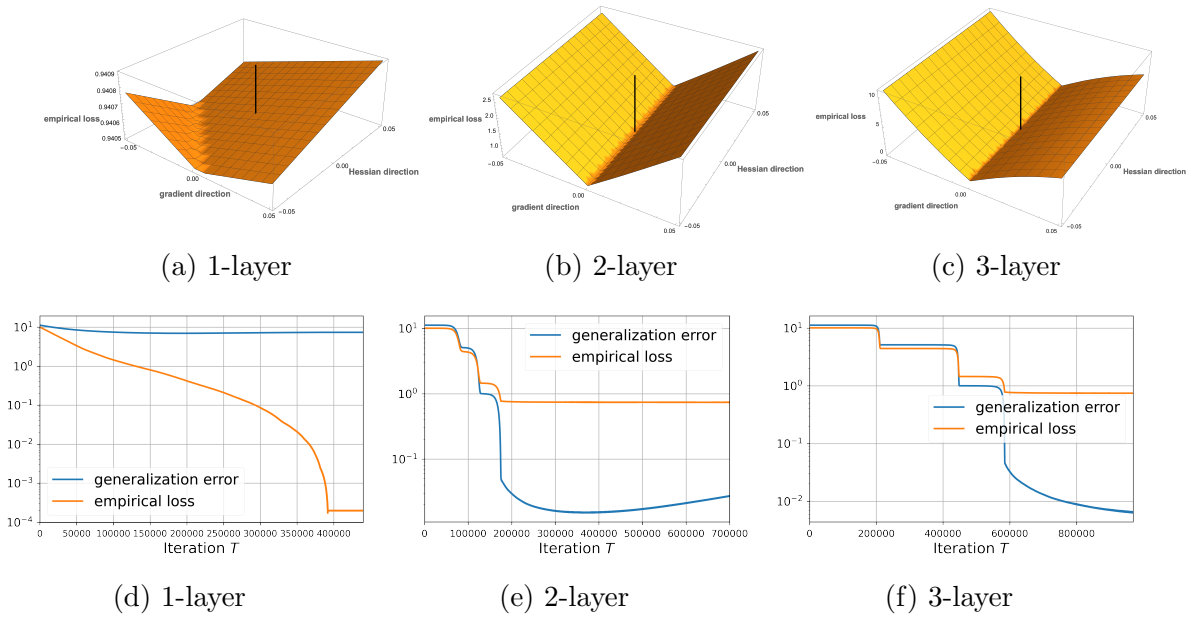
This chapter is based on the following paper:

- Jianhao Ma and Salar Fattahi. Blessing of depth in linear regression: Deeper models have flatter landscape around the true solution. In *Advances in Neural Information Processing Systems*, 2022c.

3.1 Introduction

Supported by the empirical success of deep models in contemporary learning tasks, it is by now a conventional wisdom that "deeper models generalize better" He et al. [2016], Novak, Bahri, Abolafia, Pennington, and Sohl-Dickstein [2018], Bommasani, Hudson, Adeli, Altman, Arora, von Arx, Bernstein, Bohg, Bosselut, Brunskill, et al. [2021]. Indeed, the flurry of recent attempts towards demystifying this phenomenon is a testament to the amount of research it has spawned: from simple linear regression to more complex and nonlinear models, it is shown that deeper models benefit from a range of desirable statistical properties, such as *depth separation* Safran and Lee [2021], Eldan and Shamir [2016], Telgarsky [2015, 2016], *implicit bias* Gunasekar et al. [2018], Arora et al. [2019a], Chou, Maly, and Rauhut [2021], and *hierarchical learning* Allen-Zhu and Li [2020], to name a few.

Despite the great promise of deeper models—both theoretically and empirically—the effect of depth on their optimization landscape has remained elusive to this day. A recent body of work attempts to characterize the effect of depth on the loss function through the notion of *benign landscape*. Roughly speaking, an optimization problem has a benign landscape if it is devoid of spurious local minima, and its true solutions—i.e., solutions corresponding to the ground truth—coincide with global minima. It has been shown that 2-layer Bhojanapalli et al. [2016] and multi-layer Kawaguchi [2016] linear neural networks with nearly-noiseless data have benign landscape. However, the notion of benign landscape is significantly stronger than what is needed in practice. For instance, the existence of spurious



(a) 1-layer (b) 2-layer (c) 3-layer

(d) 1-layer (e) 2-layer (f) 3-layer

Figure 3.1: **First row.** Local landscape around the balanced true solution $\boldsymbol{w}^\star = (\sqrt[N]{\boldsymbol{\theta}^\star}, \ldots, \sqrt[N]{\boldsymbol{\theta}^\star})$ for 1-, 2-, and 3-layer models. x and y axis correspond to the points of the form $\boldsymbol{w}^\star + \alpha \boldsymbol{d} + \beta \boldsymbol{h}$, for different values of α and β, where $\boldsymbol{d}$ and $\boldsymbol{h}$ are respectively the most descent sub-gradient direction and the most negatively curved direction of the Hessian after smoothing. **Second row.** Generalization error and the empirical loss of the solutions found by SubGM for 1-, 2-, and 3-layer models.

local minima may not pose any issue if an optimization algorithm can avoid them efficiently. Another line of research focuses on characterizing the solution trajectory of different local-search algorithms, showing that they enjoy an *implicit bias* that steers them away from undesirable solutions Li, Nguyen, Hegde, and Wong [2021], Vaskevicius et al. [2019], Chou et al. [2021], Gunasekar et al. [2018], Arora et al. [2019a], Gissin, Shalev-Shwartz, and Daniely [2019]. However, such guarantees only apply to specific trajectories of an algorithm, thereby falling short of any meaningful characterization of the optimization landscape around those trajectories.

3.1.1 Our Contributions

To shed light on the effect of depth on the optimization landscape of deep models, we consider a prototypical problem in machine learning, namely *robust linear regression*, where the goal is to recover a linear model from a limited number of grossly corrupted measurements. Given samples of the form $y_i = \langle \boldsymbol{\theta}^\star, \boldsymbol{x}_i \rangle + \epsilon_i$, we study the optimization landscape of ℓ_1-loss under an N-layer model defined as $y = f_{\boldsymbol{w}}(\boldsymbol{x}) := \langle \boldsymbol{w}_1 \odot \boldsymbol{w}_2 \odot \cdots \odot \boldsymbol{w}_N, \boldsymbol{x} \rangle$. Our results are summarized as follows:

- We prove that, for any $N \geq 1$, there exists at least one true solution that is neither

local nor global minimum of ℓ_1-loss, provided that at least a fraction $p > 0$ of the measurements are corrupted with noise.

- Despite the ubiquity of such "hidden" true solutions, we show that, for any N-layer model with $N \geq 2$, a simple sub-gradient method (SubGM) with small initialization converges to a small neighborhood of a balanced true solution. The radius of this neighborhood shrinks with the depth of the model, resulting in more accurate solutions. Moreover, the balancedness of the solution implies that each layer of the model inherits a similar sparsity pattern to the ground truth.

- We prove that deeper models take longer to train, but once trained, the algorithm will stay close to the ground truth for a longer time. This implies that early stopping of the algorithm becomes less crucial for deeper models.

- Finally, we prove that depth flattens the optimization landscape around the solution obtained by SubGM. In particular, we show that, within an γ-neighborhood of the true solution, the steepest descent direction can reduce the loss by at most $\mathcal{O}\left(\gamma^N\right)$, which decreases *exponentially* with N.

Motivating Example. To showcase our results, we consider an instance of robust linear regression in the over-parameterized setting, where the dimension of $\boldsymbol{\theta}^\star$ is 500 and the number of available samples is 300. Moreover, we assume that 10% of the measurements are corrupted with large noise. The first row of Figure 3.1 shows the landscape around the balanced ground truth $\boldsymbol{w}_1^\star = \sqrt[N]{\boldsymbol{\theta}^\star}, \boldsymbol{w}_2^\star = \sqrt[N]{\boldsymbol{\theta}^\star}, \ldots, \boldsymbol{w}_N^\star = \sqrt[N]{\boldsymbol{\theta}^\star}$.[1] In particular, x and y axis show the most descent sub-gradient direction and the most negatively curved direction of the Hessian after smoothing.[2] Evidently, there is a sharp transition in the landscape of N-layer models: for the 1-layer model, the true solution has strictly negative directions along both sub-gradient and negative curvature. However, these descent directions almost disappear in 2- and 3-layer models. The second row of Figure 3.1 shows the performance of SubGM on these models. It can be observed that a 1-layer model easily overfits to noise, leading to a vacuous generalization error. On the contrary, a 3-layer model can find a solution that generalizes progressively better than 1- and 2-layer models, demonstrating the algorithmic benefit of the depth.

Notations: For two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, their inner product is defined as $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = x^\top y$, and their Hadamard product is defined as $\boldsymbol{x} \odot \boldsymbol{y} = [x_1 y_1 \ \cdots \ x_d y_d]^\top$. For simplicity of notation,

[1] Later, we will show that a simple sub-gradient method converges to this balanced solution.

[2] To smoothen $|x|$, we replace it with $\sqrt{x^2 + \epsilon}$, for $\epsilon = 10^{-7}$.

we use $\prod_j \boldsymbol{w}_j$ to denote the Hadamard product of $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_N \in \mathbb{R}^d$. For a vector $\boldsymbol{x}$, $\|\boldsymbol{x}\|, \|\boldsymbol{x}\|_\infty$, and $\|\boldsymbol{x}\|_0$ refer to 2-norm, ∞-norm, and the number of nonzero elements, respectively. The symbols $a_t \lesssim b_t$ and $a_t = \mathcal{O}(b_t)$ are used to denote $a_t \leq Cb_t$, for a universal constant C. The notation $a_t = \Theta(b_t)$ is used to denote $a_t = O(b_t)$ and $b_t = \Omega(a_t)$. The sign$(\cdot)$ function is defined as sign$(x) = x/|x|$ if $x \neq 0$, and sign$(0) = [-1, 1]$. We denote $[n] := \{1, 2, \cdots, n\}$. For a vector $\boldsymbol{x} \in \mathbb{R}^d$, we define $\boldsymbol{x}^a = [x_1^a \ x_2^a \ \ldots \ x_d^a]^\top$, for any $a > 0$. In all of our probabilistic arguments, the randomness is only over the input data and noise.

3.2 Problem Formulation

We study the problem of robust and sparse linear regression, where the goal is to estimate a k-sparse vector $\boldsymbol{\theta}^\star \in \mathbb{R}^d$ ($k \ll d$) from a limited number of data points $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$, where $y_i = \langle \boldsymbol{\theta}^\star, \boldsymbol{x}_i \rangle + \epsilon_i$, $\boldsymbol{x}_i$ is i.i.d. standard Gaussian vector, and ϵ_i is noise. Moreover, for simplicity of our subsequent analysis, we assume that $\boldsymbol{\theta}^\star$ is a non-negative vector. We believe that this assumption can be relaxed without a significant change in our results.

Assumption 2 (Noise Model)**.** *Given a corruption probability p, the noise vector $\boldsymbol{\epsilon} = [\epsilon_1 \ \cdots \ \epsilon_m]^\top \in \mathbb{R}^m$ is generated as follows: first, a subset $\mathcal{S} \subset [m]$ with cardinality pm is chosen uniformly at random*[3]*. Then, for each entry $i \in \mathcal{S}$, the value of ϵ_i is drawn independently from a distribution P_o, and all the remaining entries are set to zero. Moreover, a random variable ζ under the distribution P_o satisfies $\mathbb{E}_{P_o}[\zeta] = 0$ and $\mathbb{P}(|\zeta| \geq t_0) \geq p_0$, for some strictly positive constants t_0 and p_0.*

Our considered noise model does not impose any assumption on the magnitude of the noise or the specific form of its distribution, which makes it particularly suitable for modeling outliers. Note that the assumption $\mathbb{P}(|\epsilon| \geq t_0) \geq p_0$ is very mild and satisfied for almost all common distributions. Roughly speaking, it implies that the noise takes a nonzero value with a nonzero probability.

To capture the input-output relationship, we consider a class of N-layer diagonal linear neural networks of the form $f_{\boldsymbol{w}}(\boldsymbol{x}) = \langle \boldsymbol{w}_1 \odot \cdots \odot \boldsymbol{w}_N, \boldsymbol{x} \rangle$, where $\boldsymbol{w} := (\boldsymbol{w}_1, \cdots, \boldsymbol{w}_N)$ collects the weights of the layers $\boldsymbol{w}_1, \cdots, \boldsymbol{w}_N \in \mathbb{R}^d$. Due to the sparse-and-large nature of the noise, it is natural to minimize the so-called empirical risk with ℓ_1-loss:

$$\min_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}) := \frac{1}{m} \sum_{i=1}^m |f_{\boldsymbol{w}}(\boldsymbol{x}_i) - y_i| = \frac{1}{m} \sum_{i=1}^m |\langle \boldsymbol{w}_1 \odot \cdots \odot \boldsymbol{w}_N, \boldsymbol{x}_i \rangle - y_i| \,. \tag{3.1}$$

[3]Here, for simplicity we assume pm is an integer.

Other variants of empirical risk minimization for linear regression have been studied in the literature. For instance, Chou et al. [2021] study the solution trajectory of gradient flow on ℓ_2-loss, showing that it converges to a solution with the smallest ℓ_1-norm. Similar analysis has also appeared in more general deep linear neural networks Du and Hu [2019], Du, Hu, and Lee [2018]. However, it is well-known that ℓ_2-loss is highly sensitive to outliers, and ℓ_1-loss is a better alternative to identify and reject large-and-sparse noise.

A solution $\bar{\boldsymbol{w}}$ is called *global* if it corresponds to a global minimizer of $\mathcal{L}(\boldsymbol{w})$. Moreover, a *local solution* $\bar{\boldsymbol{w}}$ corresponds to the minimum of $\mathcal{L}(\boldsymbol{w})$ within an open ball centered at $\bar{\boldsymbol{w}}$. Finally, a *true solution* $\bar{\boldsymbol{w}}$ satisfies $\bar{\boldsymbol{w}}_1 \odot \cdots \odot \bar{\boldsymbol{w}}_N = \boldsymbol{\theta}^\star$.

3.3 Main Results

3.3.1 Absence of Benign Landscape

We show that for any arbitrary corruption probability $0 < p < 1/2$ and any number of layers $N \geq 1$, there exists at least one true solution with a strictly negative descent direction, provided that the problem is *over-parameterized*, i.e., $m \lesssim d$.

Theorem 14 (unidentifiable true solutions)**.** *Define $\mathcal{W} = \{\boldsymbol{w} : \boldsymbol{w}_1 \odot \cdots \odot \boldsymbol{w}_N = \boldsymbol{\theta}^\star\}$ as the set of all true solutions of an N-layer model. For any $N \geq 1$ and $0 < p < 1/2$, the following statements hold:*

- (Over-parameterized regime) If $m \leq 0.1d$, with probability of at least $1/16$, we have

$$\inf_{\boldsymbol{w}^\star \in \mathcal{W}} \quad \inf_{\boldsymbol{w}: \|\boldsymbol{w} - \boldsymbol{w}^\star\|_\infty \leq \gamma} \{\mathcal{L}(\boldsymbol{w}) - \mathcal{L}(\boldsymbol{w}^\star)\} \lesssim -p_0 p \gamma, \tag{3.2}$$

for any $\gamma \lesssim t_0$.

- (Under-parameterized regime) If $m \gtrsim \frac{d}{(1-2p)^2}$, with probability of at least $1 - e^{-\Omega(d)}$, we have

$$\inf_{\boldsymbol{w}^\star \in \mathcal{W}} \quad \inf_{\boldsymbol{w}} \{\mathcal{L}(\boldsymbol{w}) - \mathcal{L}(\boldsymbol{w}^\star)\} \geq 0. \tag{3.3}$$

The above proposition unravels a sharp transition in the landscape of robust linear regression with an N-layer model: when $m \lesssim d$, some of the true solutions are likely to be non-critical points, and hence, cannot be recovered via any first-order algorithm. As soon as $m \gtrsim d$, all true solutions become global. This is in stark contrast with the recent results on the benign landscape of robust low-rank matrix recovery with ℓ_1-loss, which show that, under the so-called *restricted isometry property* (RIP), all the true solutions are global and vice versa Li et al. [2020b], Ding et al. [2021], Fattahi and Sojoudi [2020]. The vector version of RIP is known to hold with $m = \tilde{\Omega}(k)$ samples (see e.g. Baraniuk, Davenport, DeVore, and

Wakin [2008] for a simple proof). Theorem 14 shows that, unlike the low-rank matrix recovery, RIP is *not enough* to guarantee the equivalence between the true and global solutions in deep linear models.

The detailed proof of this theorem can be found in Appendix 3.C.1. Here, we provide a proof sketch for $N = 1$ and $N = 2$ to elucidate the key ideas.

Proof sketch of Theorem 14. For 1-layer model, the set of true solutions $\mathcal{W}$ reduces to a singleton $\{\boldsymbol{\theta}^\star\}$. Upon choosing $\boldsymbol{w} = \boldsymbol{\theta}^\star$, we prove the existence of a perturbation $\|\boldsymbol{\Delta\theta}\| \leq \gamma$ that satisfies $\mathcal{L}(\boldsymbol{\theta}^\star) - \mathcal{L}(\boldsymbol{\theta}^\star + \boldsymbol{\Delta\theta}) < 0$. The perturbed loss takes the following form

$$\mathcal{L}\left(\boldsymbol{\theta}^\star + \boldsymbol{\Delta\theta}\right) = \frac{1}{m}\sum_{i\in\bar{\mathcal{S}}} \left|\langle \boldsymbol{\Delta\theta}, \boldsymbol{x}_i\rangle\right| + \frac{1}{m}\sum_{i\in\mathcal{S}} \left|\langle \boldsymbol{\Delta\theta}, \boldsymbol{x}_i\rangle - \epsilon_i\right|,$$

Consider the following feasibility problem:

$$\text{find} \quad \boldsymbol{\alpha} \quad \text{s.t.} \quad \langle \boldsymbol{\alpha}, \boldsymbol{x}_i\rangle = 0, \forall i \in \bar{\mathcal{S}}, \quad \langle \boldsymbol{\alpha}, \boldsymbol{x}_i\rangle = \epsilon_i, \forall i \in \mathcal{S}.$$

Since $m \leq d$ and $\{\boldsymbol{x}_i\}_{i=1}^m$ are i.i.d. standard Gaussian vectors, they are linearly independent almost surely. Moreover, with high probability, at least one of ϵ_i's will be nonzero. Therefore, with high probability, the above system of linear equations has at least one nonzero feasible solution. Suppose that $\bar{\boldsymbol{\alpha}} \neq 0$ is one such solution. Define $\boldsymbol{\Delta\theta} = \gamma\bar{\boldsymbol{\alpha}}/\left\|\bar{\boldsymbol{\alpha}}\right\|$ for some $0 < \gamma < \|\bar{\boldsymbol{\alpha}}\|$. One can write

$$\mathcal{L}\left(\boldsymbol{\theta}^\star + \boldsymbol{\Delta\theta}\right) = \frac{1}{m}\left(1 - \frac{\gamma}{\|\bar{\boldsymbol{\alpha}}\|}\right)\sum_{i\in\mathcal{S}} |\epsilon_i| < \mathcal{L}\left(\boldsymbol{\theta}^\star\right),$$

implying that $\boldsymbol{\Delta\theta}$ is indeed a descent direction. Now, consider a 2-layer model. It is easy to verify the existence of a true solution $\boldsymbol{w} = (\boldsymbol{w}_1, \boldsymbol{w}_2)$ such that $\boldsymbol{w}_1 \odot \boldsymbol{w}_2 = \boldsymbol{\theta}^\star$ and $\|\boldsymbol{w}_1\|_0 = d$. Consider a perturbation of the form $\boldsymbol{\Delta w} = (0, \boldsymbol{\Delta w}_2)$. One can write

$$\mathcal{L}\left(\boldsymbol{w} + \boldsymbol{\Delta w}\right) = \frac{1}{m}\sum_{i\in\bar{\mathcal{S}}} \left|\langle \boldsymbol{w}_1 \odot \boldsymbol{\Delta w}_2, \boldsymbol{x}_i\rangle\right| + \frac{1}{m}\sum_{i\in\mathcal{S}} \left|\langle \boldsymbol{w}_1 \odot \boldsymbol{\Delta w}_2, \boldsymbol{x}_i\rangle - \epsilon_i\right|.$$

Since $\boldsymbol{w}_1$ is devoid of zero elements, there exists a nonzero $\boldsymbol{\Delta w}_2$ such that $\boldsymbol{w}_1 \odot \boldsymbol{\Delta w}_2 = \gamma\bar{\boldsymbol{\alpha}}/\left\|\bar{\boldsymbol{\alpha}}\right\|$ for $\gamma < \|\bar{\boldsymbol{\alpha}}\|$. Arguments analogous to 1-layer model can then be invoked to show that the constructed perturbation is indeed a descent direction. A similar idea can be naturally extended to $N \geq 3$. □

Theorem 14 implies that, despite their convexity, 1-layer models are *not* suitable for the robust linear regression since the set of true solutions (which is a singleton $\mathcal{W} = \{\boldsymbol{\theta}^\star\}$) is unidentifiable. However, despite the existence of unidentifiable true solutions in N-layer

models with $N \geq 2$, we will show that a simple SubGM converges to a *balanced* true solution, even if an arbitrarily large fraction of the measurements are corrupted with arbitrarily large noise values. This further sheds light on the desirable landscape of deeper models in the context of linear regression.

Algorithm 3 Sub-gradient Method

Input: Data points $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$, number of iterations T, the initial point $\boldsymbol{w}_0$, and the step-size $\{\eta^{(t)}\}_{t=0}^T$;
Output: Solution $\boldsymbol{w}^{(T)}$ to equation 3.1;
for $t \leq T$ **do**
Select a direction $\boldsymbol{d}^{(t)}$ from the sub-differential $\partial \mathcal{L}(\boldsymbol{w}^{(t)})$ defined as:

$$\partial_{\boldsymbol{w}_i} \mathcal{L}(\boldsymbol{w}) = \frac{1}{m} \sum_{j=1}^{m} \operatorname{sign}\left(y_j - \left\langle \prod_k \boldsymbol{w}_k, \boldsymbol{x}_j \right\rangle \right) \boldsymbol{x}_j \odot \prod_{k \neq i} \boldsymbol{w}_k; \tag{3.4}$$

Set $\boldsymbol{w}^{(t+1)} \leftarrow \boldsymbol{w}^{(t)} - \eta^{(t)} \boldsymbol{d}^{(t)}$;
end for

3.3.2 Convergence of Sub-gradient Method

At every iteration t, SubGM selects a direction $\boldsymbol{d}^{(t)}$ from the sub-differential of the ℓ_1-loss (defined as equation 3.4), and updates the solution as $\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \eta^{(t)} \boldsymbol{d}^{(t)}$; see Algorithm 3 for details. Our next two theorems characterize the performance of SubGM with small initialization on N-layer models. We consider the cases $N = 2$ and $N \geq 3$ separately, as SubGM behaves differently on these models. We define $\kappa = \theta^\star_{\max}/\theta^\star_{\min}$ as the condition number, where $\theta^\star_{\max}$ and $\theta^\star_{\min}$ are the maximum and minimum nonzero elements of $\boldsymbol{\theta}^\star$, respectively.

Theorem 15 (2-layer model)**.** *Consider the iterations of SubGM $\{\boldsymbol{w}^{(t)}\}_{t=0}^T$ applied to $\mathcal{L}(\boldsymbol{w})$ with $N = 2$ and step-size $\eta \lesssim 1$. Suppose that the initial point satisfies $\boldsymbol{w}_j^{(0)} = \Theta(\sqrt{\alpha}\mathbf{1})$, $j = 1, 2$, where $0 < \alpha \lesssim d^2 m/k$. Moreover, suppose that $m \gtrsim \frac{k^2 \kappa^2 \log^2(m) \log(d) \log(\|\boldsymbol{\theta}^\star\|/\alpha)}{(1-p)^2}$. Then, the following statements hold with probability of $1 - Ce^{-\tilde{\Omega}(k)}$:*

- ***Convergence guarantee:*** *After $\frac{1}{\eta} \log\left(\frac{1}{\alpha}\right) \lesssim \bar{T} \lesssim \frac{k^{3/2}}{\eta} \log\left(\frac{1}{\alpha}\right)$ iterations, we have*

$$\left\| \boldsymbol{w}_1^{(\bar{T})} \odot \boldsymbol{w}_2^{(\bar{T})} - \boldsymbol{\theta}^\star \right\| \lesssim \eta \theta^\star_{\max} \vee \sqrt{d^2 m} \alpha^{1 - \tilde{\Theta}\left(\frac{k^2}{\sqrt{(1-p)^2 m}}\right)}.$$

- ***Balanced property:*** *For every* $0 \le t \le \bar{T}$*, we have*

$$\left\| \boldsymbol{w}_1^{(t)} - \boldsymbol{w}_2^{(t)} \right\|_\infty \lesssim \alpha^{0.5-\tilde{\Theta}\left(\frac{k^2}{\sqrt{(1-p)^2 m}}\right)}.$$

- ***Long escape time:*** *For every* $\bar{T} \le t \le \sqrt{\frac{m(1-p)^2}{k}}\bar{T}$*, we have*

$$\left\| \boldsymbol{w}_1^{(t)} \odot \boldsymbol{w}_2^{(t)} - \boldsymbol{\theta}^\star \right\| \lesssim \eta\theta^\star_{\max} \vee \sqrt{d^2 m}\alpha^{0.5-\tilde{\Theta}\left(\frac{k^2}{\sqrt{(1-p)^2 m}}\right)}.$$

Furthermore, if $m \gtrsim d\log(m)/(1-p)^2$*, with probability of* $1 - Ce^{-\tilde{\Omega}(k)}$ *and for every* $t \ge \bar{T}$*, we have*

$$\left\| \boldsymbol{w}_1^{(t)} \odot \boldsymbol{w}_2^{(t)} - \boldsymbol{\theta}^\star \right\| \lesssim \eta\theta^\star_{\max} \vee \sqrt{d^2 m}\alpha^{1-\tilde{\Theta}\left(\frac{k^2}{\sqrt{m(1-p)^2}}\right)} \left(1 - \Omega\left(\eta/\sqrt{d}\right)\right)^{t-\bar{T}}.$$

We provide the main idea behind the proof of Theorem 15 in Section 3.4. The formal proof can be found in the appendix. A few observations are in order based on Theorem 15. First, for any $\epsilon > 0$, SubGM is guaranteed to satisfy $\left\| \boldsymbol{w}_1^{(t)} \odot \boldsymbol{w}_2^{(t)} - \boldsymbol{\theta}^\star \right\| \lesssim \epsilon$ after $\mathcal{O}((1/\epsilon)\log(d/\epsilon))$ iterations, provided that $\eta = \Theta(\epsilon)$ and $\alpha = \epsilon^2/(d^2 m)$. Based on our numerical results (provided in the appendix), we believe that it is possible to establish a linear convergence for SubGM with a geometric step-size; a rigorous verification of this conjecture is considered future work. Second, although SubGM converges to a vicinity of a true solution quickly, it will stay there for a significantly longer time—in particular, $\sqrt{m(1-p)^2/k}$ times longer than its initial convergence time. Such behavior is also exemplified in our simulations (see Figure 3.1e). After this escape time, the algorithm may slowly converge to an *overfitted* solution with a better training loss. Moreover, if $m \gtrsim d$, SubGM will continuously converge to a true solution at an exponential rate, and it will never diverge. Finally, Theorem 15 shows that SubGM implicitly favors balanced solutions, i.e. solutions whose factors have similar magnitudes. Combined with the convergence result of SubGM, we immediately conclude that SubGM converges to a particular solution of the form $(\sqrt{\boldsymbol{\theta}^\star}, \sqrt{\boldsymbol{\theta}^\star})$. Therefore, the solution found by SubGM will enjoy the same (approximate) sparsity pattern as $\boldsymbol{\theta}^\star$.

Theorem 16 (N-layer models)**.** *Consider the iterations of SubGM* $\{\boldsymbol{w}^{(t)}\}_{t=0}^T$ *applied to* $\mathcal{L}(\boldsymbol{w})$ *with* $N \ge 3$ *and step-size* $\eta \lesssim N^{-1}\kappa^{-\frac{N-2}{N}}$*. Suppose that the initial point satisfies* $\boldsymbol{w}_j^{(0)} = \Theta(\alpha^{1/N}\mathbf{1})$*, where* $0 < \alpha \lesssim d^2 m/k$*. Moreover, suppose that* $m \gtrsim \frac{k^2\kappa^4 \log^2(m)\log(d)\log(\|\boldsymbol{\theta}^\star\|/\alpha)}{(1-p)^2}$*. Then, the following statements hold with probability of* $1 - Ce^{-\tilde{\Omega}(k)}$*:*

- ***Convergence guarantee:*** *After* $\frac{1}{N\eta}\alpha^{-\frac{N-2}{N}} \lesssim \bar{T} \lesssim \frac{k^{3/2}}{N\eta}\alpha^{-\frac{N-2}{N}}$ *iterations, we have*

$$\left\|\prod \boldsymbol{w}_i^{(\bar{T})} - \boldsymbol{\theta}^\star\right\| \lesssim N\eta\theta^\star_{\max} \vee \sqrt{d^2 m}\alpha.$$

- ***Balanced property:*** *For every* $0 \le t \le \bar{T}$*, we have*

$$\begin{aligned}
\left|w_{i,l}^{(t)} - w_{j,l}^{(t)}\right| &= \mathcal{O}\left(\alpha^{1/N}\right), && \text{for } 1 \le i < j \le N, l : \boldsymbol{\theta}_l^\star = 0,\\
\left|w_{i,l}^{(t)} - w_{j,l}^{(t)}\right| &= \tilde{\mathcal{O}}\left(\sqrt[N]{\theta_l^\star}\sqrt{k^3/m}\right), && \text{for } 1 \le i < j \le N, l : \boldsymbol{\theta}_l^\star \neq 0.
\end{aligned}$$

- ***Long escape time:*** *For every* $\bar{T} \le t \le \sqrt{\frac{m(1-p)^2}{k}}\bar{T}$*, we have*

$$\left\|\prod \boldsymbol{w}_i^{(t)} - \boldsymbol{\theta}^\star\right\| \lesssim N\eta\theta^\star_{\max} \vee \sqrt{d^2 m}\alpha,$$

Furthermore, if $m \gtrsim d^{\frac{2N-2}{N}}\log(m)/(1-p)^2$*, with probability of at least* $1 - Ce^{-\tilde{\Omega}(k)}$ *and for every* $t > \bar{T}$*, we have*

$$\left\|\prod \boldsymbol{w}_i^{(t)} - \boldsymbol{\theta}^\star\right\| \lesssim N\eta\theta^\star_{\max} \vee \left(\frac{\sqrt{d^2 m}\alpha}{\sqrt{d^2 m}\alpha N\eta d^{-\frac{N-1}{N}}(t-\bar{T})+1}\right)^{\frac{N}{N-2}}.$$

The proof of this theorem can be found in the appendix. Theorem 16 sheds light on an important benefit of N-layer models with $N \ge 3$ compared to 2-layer models: for sufficiently small step-size, deeper models improve the generalization error by a factor of $(1/\alpha)^{\tilde{\Theta}\left(k^2/\sqrt{((1-p)^2 m)}\right)}$. This improvement is particularly significant when both α and m are small. However, such improvement comes at the expense of a slower convergence rate. In particular, after setting $\eta = \Theta(\epsilon/N)$, and $\alpha = \epsilon/\sqrt{d^2 m}$, SubGM needs $\mathcal{O}\left((1/\epsilon)^{1+\frac{N-2}{N}}\right)$ iterations to obtain an ϵ-accurate solution. Evidently, the convergence rate deteriorates with N, ultimately approaching $\mathcal{O}\left(1/\epsilon^2\right)$ for infinitely deep models. This can be observed in practice: Figures 3.1e and 3.1f show that 3-layer model enjoys a better generalization error compared to 2-layer model, but suffers from a slower convergence rate. This slower convergence rate also manifests itself in a more stable behavior of the algorithm: for deeper models, SubGM stays close to the ground truth for a longer time. Finally, the balanced property of the solution obtained via SubGM extends to N-layer models. In particular, SubGM converges to a particular solution of the form $(\sqrt[N]{\boldsymbol{\theta}^\star}, \ldots, \sqrt[N]{\boldsymbol{\theta}^\star})$, thereby inheriting the same sparsity pattern as $\boldsymbol{\theta}^\star$.

3.3.3 Local Landscape Around Balanced Solution

In the previous section, we showed that SubGM converges to a balanced solution. In this section, we characterize the local landscape around this balanced solution, proving that it becomes flatter for deeper models.

Theorem 17 (flatness around balanced solution)**.** *Suppose that* $k\log(d)/(1-2p)^2 \lesssim m \leq 0.1d$ *and* $p < 1/2$*. Let* $\boldsymbol{w}^\star = (\sqrt[N]{\boldsymbol{\theta}^\star}, \ldots, \sqrt[N]{\boldsymbol{\theta}^\star})$*. Then, for any* $N \geq 2$ *and* $\gamma \leq t_0/\sqrt{d} \wedge 1$*, the following statements hold:*

- *With probability at least* $1 - e^{-\Omega(k)}$*, we have*

$$\inf_{\boldsymbol{w}:\|\boldsymbol{w}-\boldsymbol{w}^\star\|_\infty \leq \gamma} \{\mathcal{L}(\boldsymbol{w}^\star) - \mathcal{L}(\boldsymbol{w})\} \gtrsim -\frac{d}{\sqrt{m}}\gamma^N.$$

- *With probability at least* $1/16$*, we have*

$$\inf_{\boldsymbol{w}:\|\boldsymbol{w}-\boldsymbol{w}^\star\|_\infty \leq \gamma} \{\mathcal{L}(\boldsymbol{w}^\star) - \mathcal{L}(\boldsymbol{w})\} \lesssim -\sqrt{p_0 p}\frac{d}{\sqrt{m}}\gamma^N.$$

Theorem 17 shows that, within a γ-neighborhood of $\boldsymbol{w}^\star$, the most descent direction from $\boldsymbol{w}^\star$ can reduce the loss by at most $\mathcal{O}\left(d/\sqrt{m}\cdot\gamma^N\right)$, which decreases exponentially with N. Moreover, in the noisy setting, the above theorem implies that $\boldsymbol{w}^\star$ is likely to be neither local nor global minimum since it has a descent direction. However, the flatness of the landscape around $\boldsymbol{w}^\star$ enables SubGM to stay close to the balanced solution for a long time.

Remark 3. *Note that the choice of* ℓ_∞*-ball for the perturbation set is to ensure that the size of the possible perturbations per layer remains independent of the depth of the model. This is indeed crucial to ensure a fair comparison between models with different depths: alternative choices of the perturbation set, such as* ℓ_q*-ball with* $1 \leq q < \infty$ *(e.g.* ℓ_2*-ball) would shrink the size of the feasible per-layer perturbations with* N*, thereby leading to an unfair advantage to deeper models.*

3.4 Proof Techniques

At the crux of our proof technique for Theorems 15 and 16 lies the following decomposition of the sub-differential:

$$\partial\mathcal{L}(\boldsymbol{w}) = \underbrace{\xi\cdot\partial\bar{\mathcal{L}}(\boldsymbol{w})}_{\text{expected subdiff.}} + \underbrace{\left(\partial\mathcal{L}(\boldsymbol{w}) - \xi\cdot\partial\bar{\mathcal{L}}(\boldsymbol{w})\right)}_{\text{subdiff. deviation}}, \quad \text{for some strictly positive } \xi.$$

In the above decomposition, $\bar{\mathcal{L}}(\boldsymbol{w})$ is called *expected loss*, and is defined as $\bar{\mathcal{L}}(\boldsymbol{w}) = \|\boldsymbol{w}_1 \odot \cdots \odot \boldsymbol{w}_N - \boldsymbol{\theta}^\star\|$. As will be shown later, $\bar{\mathcal{L}}(\boldsymbol{w})$ captures the expected behavior of the empirical loss $\mathcal{L}(\boldsymbol{w})$. To analyze the behavior of SubGM on $\mathcal{L}(\boldsymbol{w})$, we first consider the ideal scenario, where $\mathcal{L}(\boldsymbol{w})$ coincides with its expectation. Then, we extend our analysis to the general case by controlling the sub-differential deviation. In particular, we show that the desirable convergence properties of SubGM extends to $\mathcal{L}(\boldsymbol{w})$, provided that its sub-differentials are "direction-preserving", i.e., $\boldsymbol{d} \approx \xi \bar{\boldsymbol{d}}$, for every $\boldsymbol{d} \in \partial \mathcal{L}(\boldsymbol{w}), \bar{\boldsymbol{d}} \in \partial \bar{\mathcal{L}}(\boldsymbol{w})$ and some $\xi > 0$. To formalize this idea, we first provide a more concise characterization of $\partial \mathcal{L}(\boldsymbol{w})$:

$$\partial_{\boldsymbol{w}_i} \mathcal{L}(\boldsymbol{w}) = \left\{ \boldsymbol{q} \odot \prod_{k \neq i} \boldsymbol{w}_k : \boldsymbol{q} \in \mathcal{Q}\left(\boldsymbol{\theta}^\star - \prod_k \boldsymbol{w}_k\right) \right\}, \text{ where } \mathcal{Q}(\boldsymbol{z}) = \frac{1}{m} \sum_{i=1}^{m} \operatorname{sign}\left(\langle \boldsymbol{x}_i, \boldsymbol{z} \rangle + \epsilon_i\right) \boldsymbol{x}_i.$$

Definition 15 (approximately sparse vectors)**.** *We say a vector $\boldsymbol{v} \in \mathbb{R}^d$ is (k, ϑ)-approximately sparse if there exists a vector $\boldsymbol{u}$, such that $\|\boldsymbol{u}\|_0 \leq k$ and $\|\boldsymbol{u} - \boldsymbol{v}\| \leq \vartheta$.*

Proposition 11 (direction-preserving property)**.** *Suppose that $m \gtrsim \frac{k \log^2(m) \log(d) \log(R/\vartheta)}{(1-p)^2 \delta^2}$ for some $R, \vartheta, \delta > 0$. Then, with probability of at least $1 - Ce^{-\Omega(m\delta^2)}$, the following inequality holds for any $\boldsymbol{q} \in \mathcal{Q}(\boldsymbol{z})$ and any (k, ϑ)-approximately sparse vector $\boldsymbol{z}$ that satisfies $\sqrt{dm/k}\vartheta \log(1/\vartheta) \lesssim \|\boldsymbol{z}\| \leq R$:*

$$\left\| \boldsymbol{q} - \sqrt{\frac{2}{\pi}} \left(1 - p + p\mathbb{E}\left[e^{-\epsilon^2/(2\|z\|)}\right]\right) \frac{\boldsymbol{z}}{\|\boldsymbol{z}\|} \right\|_\infty \leq \delta. \tag{3.5}$$

Moreover, if $m \gtrsim \frac{d \log(m)}{(1-p)^2 \delta^2}$, with probability of $1 - Ce^{-\Omega(m\delta^2)}$, equation 3.5 holds for every $\boldsymbol{z} \in \mathbb{R}^d$.

Proposition 11 is analogous to *Sign-RIP* condition introduced in Ma and Fattahi [2023a, 2021b] for the robust low-rank matrix recovery, and is at the heart of our proofs for Theorems 15 and 16. Suppose that $\boldsymbol{\theta}^\star - \prod_k \boldsymbol{w}_k$ is a (k, ϑ)-approximately sparse and satisfies equation 3.5. Then, we have $\left\|\boldsymbol{d} - \bar{\boldsymbol{d}}\right\|_\infty \leq \left(\max_i \left\{\prod_{k \neq i} \boldsymbol{w}_k\right\}\right) \delta$, which in turn provides an upper bound on the sub-differential deviation.

3.4.1 Proof Sketch of Theorem 15

To streamline the presentation, here we only provide simplified versions of our key ideas, which inevitably lead to looser guarantees. To streamline the proof, we assume that $\theta_1^\star \geq \cdots \geq \theta_k^\star > \theta_{k+1}^\star = \cdots = \theta_d^\star = 0$. Moreover, for simplicity of notation, we denote $\boldsymbol{u} = \boldsymbol{w}_1$ and

$\boldsymbol{v} = \boldsymbol{w}_2$. Consider the following decomposition:

$$\boldsymbol{u} \odot \boldsymbol{v} = [\underbrace{u_1 v_1 \dots u_k v_k}_{\boldsymbol{S}} \; \underbrace{u_{k+1} v_{k+1} \dots u_d v_d}_{\boldsymbol{E}}]^\top. \tag{3.6}$$

The vectors $\boldsymbol{S}$ and $\boldsymbol{E}$ are called *signal* and *residual terms*, respectively. Evidently, we have $\boldsymbol{u} \odot \boldsymbol{v} = \boldsymbol{\theta}^\star$ if and only if $\boldsymbol{S} = [\boldsymbol{\theta}_1^\star, \dots, \boldsymbol{\theta}_k^\star]^\top$ and $\boldsymbol{E} = 0$. Based on this observation, our goal is to show that the signal term converges to $[\boldsymbol{\theta}_1^\star, \dots, \boldsymbol{\theta}_k^\star]^\top$ exponentially fast, while the error term remains small throughout the solution trajectory.

Lemma 54 (signal dynamic; informal)**.** *Suppose that equation 3.5 holds for $\boldsymbol{z} = \boldsymbol{\theta}^\star - u^{(t)} \odot v^{(t)}$, and $\left\|\boldsymbol{\theta}^\star - u^{(t)} \odot v^{(t)}\right\| \gtrsim \eta \left\|\boldsymbol{\theta}^\star\right\|$. Then, we have*

$$u_i^{(t+1)} v_i^{(t+1)} \geq \left(1 + 2\eta \left(\frac{\theta_i^\star - u_i^{(t)} v_i^{(t)}}{\left\|u^{(t)} \odot v^{(t)} - \boldsymbol{\theta}^\star\right\|} + \delta_i\right)\right) u_i^{(t)} v_i^{(t)}, \tag{3.7}$$

for some $|\delta_i| \leq \delta$ and every $i = 1, \dots, k$.

Lemma 55 (residual dynamic; informal)**.** *Suppose that equation 3.5 holds for $\boldsymbol{z} = \boldsymbol{\theta}^\star - u^{(t)} \odot v^{(t)}$, and $\left\|\boldsymbol{\theta}^\star - u^{(t)} \odot v^{(t)}\right\| \gtrsim \eta \left\|\boldsymbol{\theta}^\star\right\|$. Then, we have*

$$\left(u_i^{(t+1)}\right)^2 + \left(v_i^{(t+1)}\right)^2 \leq (1 + \mathcal{O}(\eta\delta)) \left(\left(u_i^{(t)}\right)^2 + \left(v_i^{(t)}\right)^2\right), \tag{3.8}$$

for every $i = k+1, \dots, d$.

Lemma 56 (difference dynamic; informal)**.** *Suppose that equation 3.5 holds for $\boldsymbol{z} = \boldsymbol{\theta}^\star - u^{(t)} \odot v^{(t)}$, and $\left\|\boldsymbol{\theta}^\star - u^{(t)} \odot v^{(t)}\right\| \gtrsim \eta \left\|\boldsymbol{\theta}^\star\right\|$. Then, we have*

$$u_i^{(t+1)} - v_i^{(t+1)} = \left(u_i^{(t)} - v_i^{(t)}\right) \left(1 - \eta \frac{\theta_i^\star - u_i^{(t)} v_i^{(t)}}{\left\|u^{(t)} \odot v^{(t)} - \boldsymbol{\theta}^\star\right\|} + \eta\delta_i\right), \tag{3.9}$$

for some $|\delta_i| \leq \delta$ and every $i = 1, \dots, d$.

Convergence guarantee. For any fixed $i = 1, \dots, k$, we show that $u_i^{(t)} v_i^{(t)} = \boldsymbol{\theta}_i^\star \pm \mathcal{O}(\delta) \left\|\boldsymbol{\theta}^\star\right\|$ after $\mathcal{O}(\|\boldsymbol{\theta}^\star\|/(\eta \boldsymbol{\theta}_i^\star) \log(1/\alpha))$ iterations. To see this, suppose that T_i is the largest iteration such that $u_i^{(t)} v_i^{(t)} \leq \boldsymbol{\theta}_i^\star$ for every $t \leq T_i$. Moreover, suppose that $\left\|u^{(t)} \odot v^{(t)}\right\| \leq C \left\|\boldsymbol{\theta}^\star\right\|$, for sufficiently large C (this is proven in the appendix). Under

these assumptions, equation 3.7 reduces to

$$u_i^{(t+1)} v_i^{(t+1)} \geq \left(1 + \Omega(1) \frac{\eta \theta_i^\star}{\|\boldsymbol{\theta}^\star\|}\right) u_i^{(t)} v_i^{(t)}. \tag{3.10}$$

which implies that $T_i \lesssim \|\boldsymbol{\theta}^\star\|/(\eta \boldsymbol{\theta}_i^\star) \log(1/\alpha)$. For any $t > T_i$, define $y_i^{(t)} = \boldsymbol{\theta}_i^\star - u_i^{(t)} v_i^{(t)}$. One can write

$$y_i^{(t+1)} \leq \left(1 - \Omega(1) \frac{\eta \theta_i^\star}{\|\boldsymbol{\theta}^\star\|}\right) y_i^{(t)} + \eta \delta \theta_i^\star. \tag{3.11}$$

Hence, with additional $\mathcal{O}\left(\|\boldsymbol{\theta}^\star\| / (\eta \theta_1^\star)\right)$ iterations, we have $u_i^{(t)} v_i^{(t)} = \theta_i^\star \pm \mathcal{O}(\delta) \|\boldsymbol{\theta}^\star\|$. On the other hand, Lemma 55 implies that, for any $i = k+1, \ldots, d$ and $t \lesssim \|\boldsymbol{\theta}^\star\|/(\eta \boldsymbol{\theta}_k^\star) \log(1/\alpha)$, we have

$$\left(u_i^{(t)}\right)^2 + \left(v_i^{(t)}\right)^2 \lesssim \alpha \left(1 + \mathcal{O}(\eta\delta)\right)^{\mathcal{O}\left(\frac{\|\boldsymbol{\theta}^\star\|}{\eta \theta_k^\star} \log\left(\frac{1}{\alpha}\right)\right)} \lesssim \alpha^{1 - \mathcal{O}\left(\sqrt{k}\kappa\delta\right)},$$

where $\kappa = \boldsymbol{\theta}_1^\star / \boldsymbol{\theta}_k^\star$ is the condition number of $\boldsymbol{\theta}^\star$. Combining the above dynamics, we have

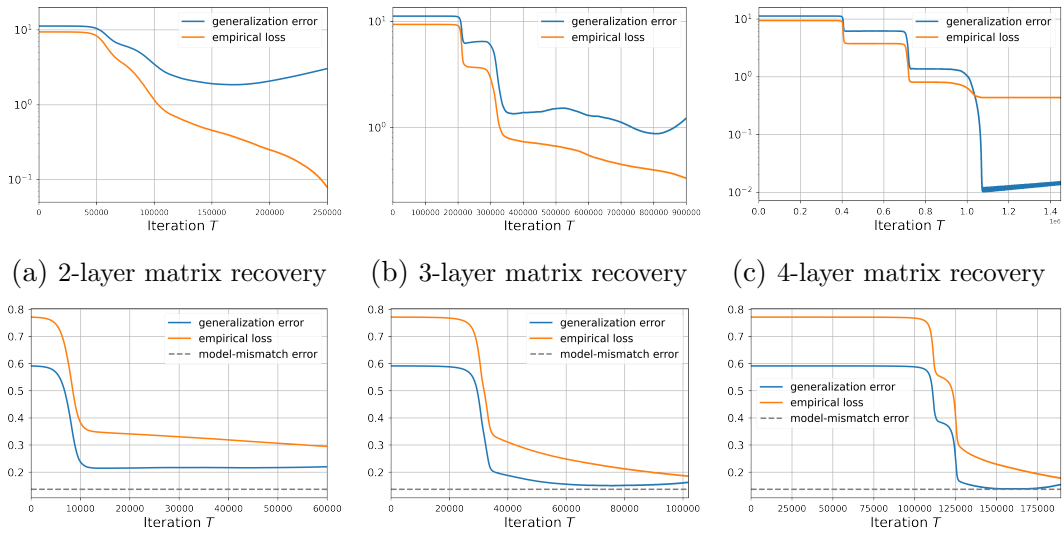
$$\left\| u^{(t)} \odot v^{(t)} - \boldsymbol{\theta}^\star \right\| \lesssim \eta \|\boldsymbol{\theta}^\star\| \vee \sqrt{k} \|\boldsymbol{\theta}^\star\| \delta \vee \sqrt{d} \alpha^{1 - \mathcal{O}\left(\sqrt{k}\kappa\delta\right)}.$$

In the appendix, we provide a more refined analysis that relaxes the dependency of the final error on δ and κ.

Long escape time. We show in the appendix that after the first stage, the residual becomes the dominant term in the final error. This together with Lemma 55 implies that, for every $t \lesssim \frac{\|\boldsymbol{\theta}^\star\|}{\eta \theta_k^\star \sqrt{\delta}} \log(1/\alpha)$, we have $\|E\| \lesssim \sqrt{d} \alpha^{1 - \sqrt{k}\kappa\sqrt{\delta}}$.

Balanced property. We have $u_i^{(t)} v_i^{(t)} \leq \boldsymbol{\theta}_i^\star$ for every $i \in [k]$, and $|u_i^{(t)} v_i^{(t)}| \lesssim \alpha^{1 - \mathcal{O}(\sqrt{k}\kappa\delta)}$ for every $i = k+1, \ldots, d$. Therefore, Lemma 56 can be invoked to verify $\left|u_i^{(t+1)} - v_i^{(t+1)}\right| \leq (1 + \mathcal{O}(\eta\delta)) \left|u_i^{(t)} - v_i^{(t)}\right|$. This in turn leads to

$$\left|u_i^{(t)} - v_i^{(t)}\right| \lesssim \sqrt{\alpha} \left(1 + \mathcal{O}(\eta\delta)\right)^{\mathcal{O}\left(\frac{\|\boldsymbol{\theta}^\star\|}{\eta \theta_k^\star} \log\left(\frac{1}{\alpha}\right)\right)} \lesssim \alpha^{0.5 - \mathcal{O}(\sqrt{k}\kappa\delta)}.$$



(a) 2-layer matrix recovery (b) 3-layer matrix recovery (c) 4-layer matrix recovery

(d) 2-layer ReLU network (e) 3-layer ReLU network (f) 4-layer ReLU network

Figure 3.2: **Deep matrix recovery (first row).** The ground truth $\boldsymbol{X}^\star \in \mathbb{R}^{20\times 20}$ with $\text{rank}(\boldsymbol{X}^\star) = 3$ is chosen randomly. The elements of the measurement matrices are selected from $\mathcal{N}(0,1)$, and the sample size is set to $m = 180$. The corruption probability is set to $p = 0.05$ with distribution $\mathcal{N}(0, 100)$. We use SubGM with step-size $\eta = 0.001$ and Gaussian initialization with an initialization scale $\alpha = 1\times 10^{-3}$. **ReLU models (second row).** The samples are chosen from $y_i = \sin(\theta^{\star\top}\boldsymbol{x}_i) + \epsilon_i$ where $\boldsymbol{\theta}^\star \in \mathbb{R}^{50}$ is randomly generated with $\|\boldsymbol{\theta}^\star\|_0 = 2$, $\boldsymbol{x}_i \sim \mathcal{N}(0, I_{50})$, and $\epsilon_i \sim \mathcal{N}(0, 25)$ with corruption probability $p = 0.05$. The sample size is set to be $m = 1500$. We use SubGM with step-size $\eta = 0.001$ and Gaussian initialization $\mathcal{N}(0, \alpha^{2/N}/d)$.

3.5 Numerical Experiments: Beyond Linear Regression

In this section, we empirically verify that the benefits of depth extend to the robust variants of deep matrix recovery and ReLU networks with ℓ_1-loss.

Deep Matrix Recovery. In low-rank matrix recovery, the goal is to recover a low-rank matrix $X^\star \in \mathbb{R}^{d\times d}$, from a limited number of noisy measurements of the form $y_i = \langle A_i, X^\star\rangle + \epsilon_i$. To recover $X^\star$, we consider a deep factorized model of the form $W_1 W_2 \dots W_N$, where $W_i \in \mathbb{R}^{d\times d}$ for $i = 1, \dots, N$, and minimize the ℓ_1-loss $(1/m)\sum_{i=1}^m |y_i - \langle A_i, W_1 W_2 \cdots W_N\rangle|$ via SubGM. When $N = 2$, the above model reduces to the famous Burer-Monteiro approach Burer and Monteiro [2003]. We assume that 5% of the measurements are grossly corrupted with noise. The first row of Figure 3.2 shows the performance of SubGM on 2-, 3-, and 4-layer models. It can be seen that the 4-layer model outperforms shallower models, achieving a generalization error that is proportional to the step-size.

Deep ReLU Network on Synthetic Dataset. As another experiment, we analyze the effect of depth on the performance of SubGM with ReLU networks and ℓ_1-loss. Given an input $\boldsymbol{x} \in \mathbb{R}^d$, the output of an N-layer ReLU network is defined as $f_{\boldsymbol{W}}(\boldsymbol{x}) = \boldsymbol{W}_N \sigma\left(\boldsymbol{W}_{N-1} \cdots \sigma\left(\boldsymbol{W}_1 x\right) \cdots\right)$, where $\boldsymbol{W}_1 \in \mathbb{R}^{m \times d}, \boldsymbol{W}_2, \cdots, \boldsymbol{W}_{N-1} \in \mathbb{R}^{m \times m}$, and $\boldsymbol{W}_N \in \mathbb{R}^{1 \times m}$. Moreover, $\sigma(x) = \max\{0, x\}$ is the ReLU activation function. Given the true function $f^\star(\boldsymbol{x}) = \sin(\boldsymbol{\theta}^{\star\top} \boldsymbol{x})$, our goal is to train a ReLU model to approximate $f^\star$ as accurately as possible. To this goal, we minimize the ℓ_1-loss $(1/m) \sum_{i=1}^m |y_i - f_{\boldsymbol{W}}(\boldsymbol{x}_i)|$. The second row of Figure 3.2 illustrates the performance of SubGM. It is worth noting that, unlike robust linear regression and deep matrix recovery, there always exists a non-diminishing model-mismatch error between the true and considered ReLU model (shown as a dashed line). Nonetheless, SubGM can achieve this model-mismatch error on a 4-layer ReLU model with only 1500 samples, even if 5% of the measurements are corrupted with large noise.

Deep ReLU Network on CIFAR Dataset. We verify that the desirable performance of SubGM with ℓ_1-loss can be extended to its stochastic variant with mini-batches on CIFAR-10 and CIFAR-100 Krizhevsky, Hinton, et al. [2009], outperforming cross-entropy (CE) loss, which is considered as one of the most suitable loss functions for CIFAR datasets. To show this, we use standard ResNet architectures He et al. [2016] with ℓ_1-loss and compare it with the cross-entropy loss on noisy CIFAR datasets, where we randomize the labels of 10% of the training dataset. For CIFAR-100 experiment, we use the "loss scaling" trick introduced in Hui and Belkin [2020]. The training details are deferred to Section 3.B.3. The best test accuracy for both CIFAR-10 and CIFAR-100 is reported in Table 5.1. One can see that ℓ_1-loss outperforms cross-entropy loss significantly, demonstrating that our framework may be extended to more realistic settings. Moreover, we do observe that the deeper model performs better on CIFAR-100, which aligns with our theoretical result. Based on our simulations, an interesting and important future direction would be to study the optimization landscape of ℓ_1-loss with more general neural network architectures.

	CIFAR-10			CIFAR-100		
	ResNet-18	ResNet-34	ResNet-50	ResNet-18	ResNet-34	ResNet-50
CE loss	91.52%	91.53%	90.87%	70.17%	71.22%	71.30%
ℓ_1-loss	**94.16**%	**93.13**%	**92.68**%	**73.14**%	**74.86**%	**76.46**%

Table 3.1: Test accuracy for ResNets on CIFAR-10 and CIFAR-100 datasets with 10% label noise.

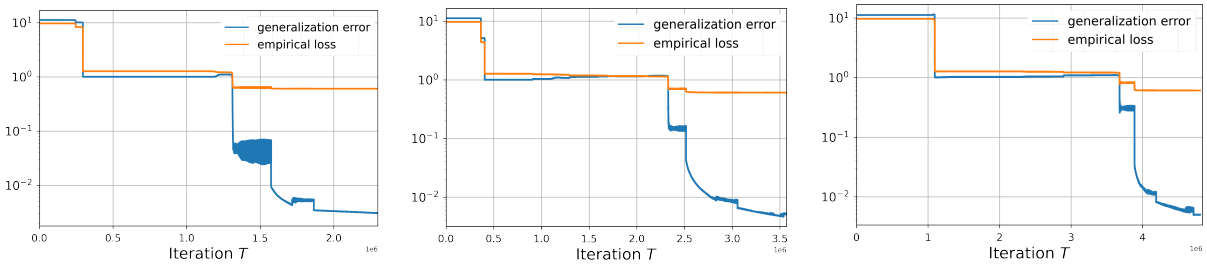
3.6 Conclusion

Modern problems in machine learning are naturally nonconvex but can be solved reasonably well in practice. To explain this, a recent body of work has postulated that many optimization problems in machine learning are "convex-like", i.e., they are devoid of spurious local minima. Our work shows that such global property is too restrictive to hold even in the context of linear regression, and instead propose a more refined *trajectory analysis* to better capture the landscape of the problem around the solution trajectory. We show that convex models may be fundamentally ill-suited for linear models, and deeper models–despite their nonconvexity–have provably better optimization landscape around the solution trajectory. Empirically, we show that our analysis may extend beyond linear regression; formal verification of this conjecture is considered an enticing challenge for future research.

3.A Related Works

Deep models: It is known that deeper models enjoy better *approximation power*. For instance, [Eldan and Shamir, 2016, Telgarsky, 2015, 2016] introduce several functions that are expressible by deep models of moderate size; yet, they cannot be approximated via any shallow network of sub-exponential size. A recent work [Safran and Lee, 2021] shows that depth separation may lead to optimization separation. In other words, functions that can be expressed by deeper models can also be efficiently learned via gradient descent. Another line of work shows that deep linear models have a strong implicit bias towards the true solution [Gunasekar et al., 2018, Arora et al., 2019a, Chou et al., 2021], and that they benefit from *incremental learning* [Gissin et al., 2019, Li et al., 2020c]. More generally, [Allen-Zhu and Li, 2020] show that stochastic gradient descent on deep nonlinear models can provably learn certain complex functions by automatically decomposing them into a series of simpler ones.

Robust and sparse linear regression: Robust and sparse linear regression is a classical problem in statistics, with a wide range of applications in signal and image processing. Regularized methods, including Lasso [Van de Geer, 2007, Candes and Tao, 2007, Bickel, Ritov, and Tsybakov, 2009, Raskutti, Wainwright, and Yu, 2011], Best Subset Selection [Mazumder, Radchenko, and Dedieu, 2017, Bertsimas, King, and Mazumder, 2016], and Forward and Backward Step-wise Regression [Friedman, Hastie, Tibshirani, et al., 2001], are considered as most widely used methods for solving robust and sparse linear regression that come equipped with strong statistical and computational guarantees. Recently, sparse



(a) 4-layer (b) 5-layer (c) 6-layer

Figure 3.3: The optimization trajectories of deep models ($N = 4, 5, 6$).

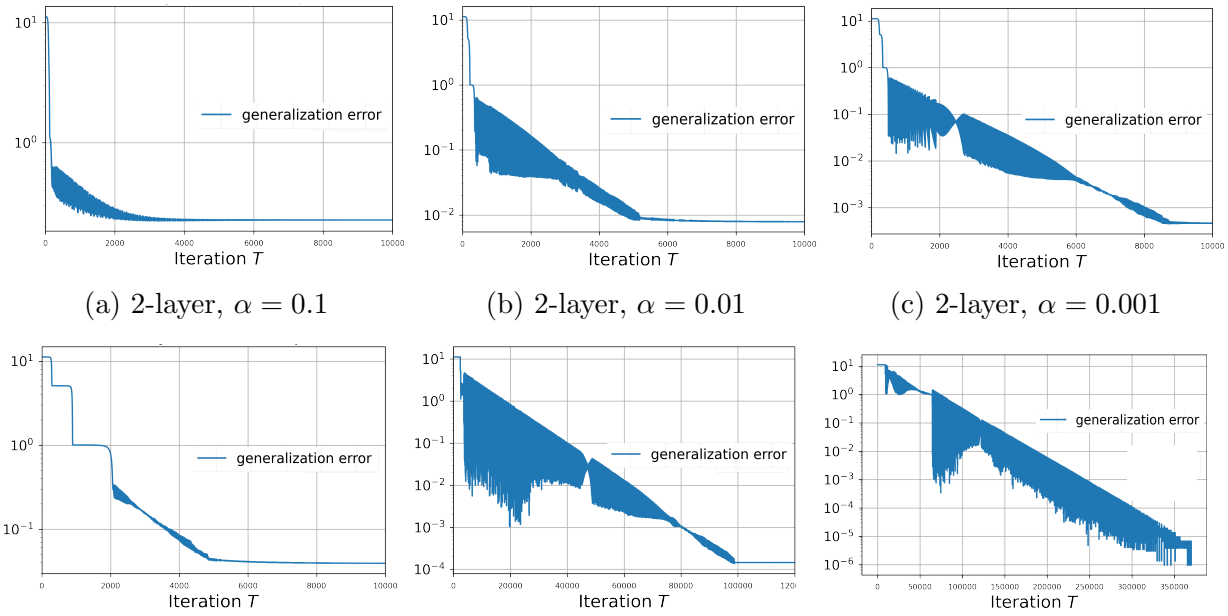
linear regression has been used to explore the implicit bias of different optimization algorithms and initialization regimes. [Vaskevicius et al., 2019, Zhao et al., 2019] show that, for some unregularized overparameterized models, gradient descent (with early stopping) for sparse linear regression achieves minimax error rate. Moreover, [Woodworth, Gunasekar, Lee, Moroshko, Savarese, Golan, Soudry, and Srebro, 2020] study how the scale of the initial point controls the transition between the "kernel" (lazy training) and "rich" regimes, and their corresponding generalization performance. [HaoChen, Wei, Lee, and Ma, 2021] use this problem setting to explore the role of label noise in stochastic gradient descent.

3.B Additional Experiments

In this section, we provide additional experiments on the performance of SubGM on deep models. Our goal is to verify our theoretical results and show the benefits of both small initialization and geometric step-size. Moreover, we show that the desirable performance of SubGM can be observed in its stochastic variant, as well as for different architectures of ResNets with ℓ_1-loss, and more realistic CIFAR-10 dataset. All simulations are run on a desktop computer with an Intel Core i9 3.50 GHz CPU and 128GB RAM. The reported results are for implementation in Python.

3.B.1 Deeper Linear Models

In this experiment, we study the performance of SubGM on deeper models ($N = 4, 5, 6$). To accelerate the training process, we first use a large step-size $\eta_1 = 1 \times 10^{-3}$, and then progressively apply smaller step-sizes $\eta_2 = 1 \times 10^{-4}$ and $\eta_3 = 1 \times 10^{-5}$ as the training loss continues to decay. As shown in Figure 3.3, deeper models share similar generalization error, outperforming 1- and 2-layer models presented in Figure 3.1.



(a) 2-layer, $\alpha = 0.1$ (b) 2-layer, $\alpha = 0.01$ (c) 2-layer, $\alpha = 0.001$

(d) 3-layer, $\alpha = 0.1$ (e) 3-layer, $\alpha = 0.01$ (f) 3-layer, $\alpha = 0.001$

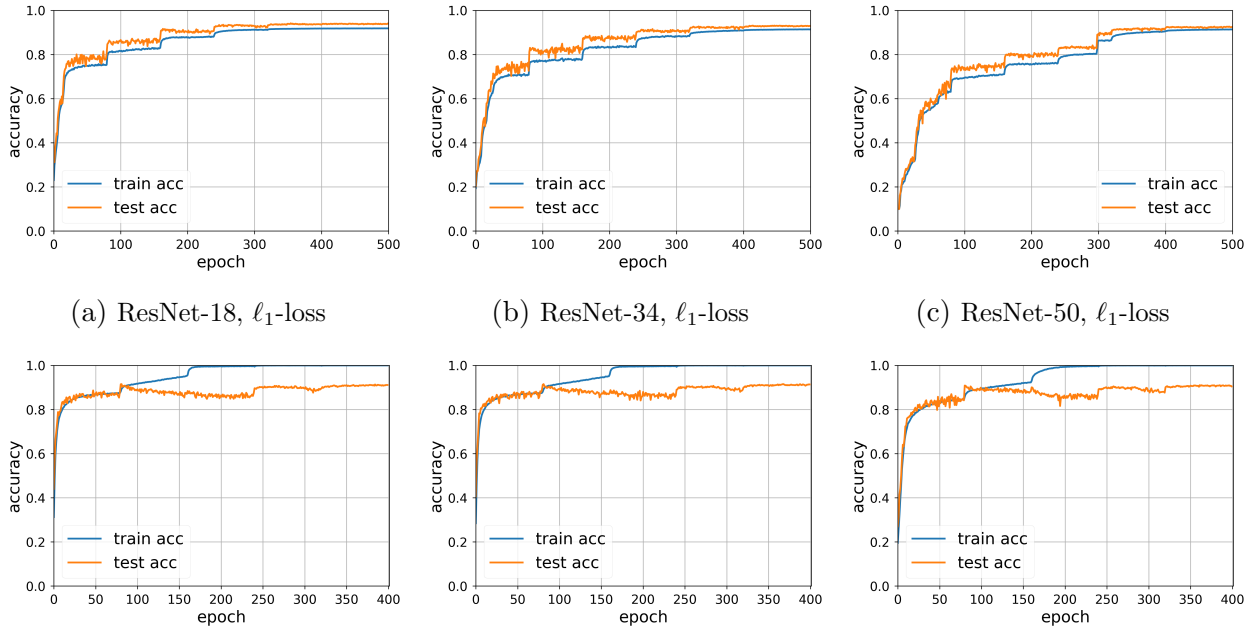
Figure 3.4: The optimization trajectories of 2 and 3-layer models with different initialization size $\alpha = 0.1, 0.01, 0.001$ using exponentially decayed step-size.

3.B.2 Geometric Step-size

As shown in our simulations, training N-layer models may require millions of iterations even on a small synthetic dataset. As proven in Theorem 16, the training process may become even slower for deeper models. In this experiment, we explore the performance of geometric (i.e., exponentially decaying) step-size on the same dataset. SubGM with a geometric step-size has been widely used for the optimization of ℓ_1-loss [Li et al., 2020b, Ma and Fattahi, 2021b], and more general sharp weakly convex functions [Davis and Drusvyatskiy, 2018]. Figure 3.4 shows that a geometric steps-size can lead to a 1000-fold reduction in the required number of iterations. Moreover, a geometric step-size improves the convergence rate to linear. The theoretical justification of this improvement is left as an enticing challenge for future research. Finally, it can be observed that SubGM with geometric step-size performs surprisingly well on deeper models, achieving a generalization error in the order of 10^{-6}. This further supports the benefits of depth.

3.B.3 Experiments on CIFAR Dataset

In this section, we provide the training details for the experiments on both CIFAR-10 and CIFAR-100 where 10% of the training data points are randomly labeled. For CIFAR-100



(a) ResNet-18, ℓ_1-loss (b) ResNet-34, ℓ_1-loss (c) ResNet-50, ℓ_1-loss

(d) ResNet-18, CE (e) ResNet-34, CE (f) ResNet-50, CE

Figure 3.5: We apply ResNet-18, 34, 50 on noisy CIFAR-10 with both ℓ_1-loss and cross-entropy loss (CE). For the training dataset, we randomly choose 10% samples and replace their labels with uniform random labels. We use SGD with initial learning rate $\eta = 0.1$, momentum 0.9, batch size $B = 32$. For every 80 epochs, we decay the learning rate by a factor 0.33. We use standard data augmentation. The initialization is set by default in PyTorch.

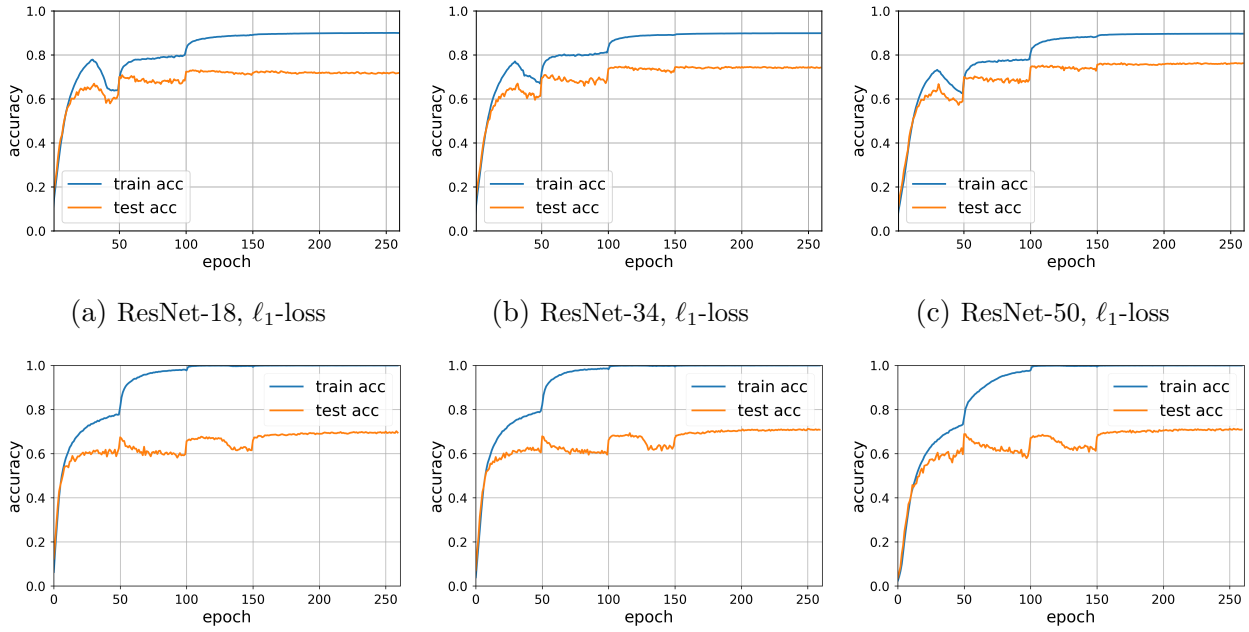
experiment, we use the "loss scaling" trick introduced in [Hui and Belkin, 2020]. In particular, we denote the neural network by $f_\theta : \mathbb{R}^d \to \mathbb{R}^C$, where d is the input dimension and C is the number of class. The standard ℓ_1-loss for the one-hot encoded label vector can be written (at a single point) as

$$\ell = |f_\theta(\boldsymbol{x})[c] - 1| + \sum_{c' \neq c} |f_\theta(\boldsymbol{x})[c']| \,. \tag{3.12}$$

Here c is the position of the label and $f_\theta(\boldsymbol{x})[i]$ is the i-th coordinate of the prediction. The rescaled ℓ_1-loss is defined by two parameters k and M as follows:

$$\ell_{\text{scaling}} = k \cdot |f_\theta(\boldsymbol{x})[c] - M| + \sum_{c' \neq c} |f_\theta(\boldsymbol{x})[c']| \,. \tag{3.13}$$

In our simulation, we choose $k = 5$, and $M = 2$. Moreover, in the CIFAR-100 experiment, we use ℓ_2-loss with the same k, M in the first 30 epochs as the warm-up phase. The evolution of the training and test accuracy for CIFAR-10 and CIFAR-100 with both ℓ_1 and CE losses are shown in Figures 3.5 and 3.6.



(a) ResNet-18, ℓ_1-loss (b) ResNet-34, ℓ_1-loss (c) ResNet-50, ℓ_1-loss

(d) ResNet-18, CE (e) ResNet-34, CE (f) ResNet-50, CE

Figure 3.6: We apply ResNet-18, 34, 50 on noisy CIFAR-100 with both ℓ_1-loss and cross-entropy loss (CE). The data generator and optimizer are the same as those in CIFAR-10 experiment. For each trial, we set the initial learning rate to be $\eta = 0.05$ and run 260 epochs and decay the learning rate with factor 0.33 for each 50 epochs.

3.C Proofs of Landscape Analysis

3.C.1 Proof of Theorem 14

Over-parameterized Regime

We first provide the proof for the 1-layer model. The values of $\mathcal{L}(\boldsymbol{\theta}^\star), \mathcal{L}(\boldsymbol{\theta}^\star + \boldsymbol{\Delta\theta})$ are provided as

$$\mathcal{L}(\boldsymbol{\theta}^\star) = \frac{1}{m}\sum_{i\in\mathcal{S}} |\epsilon_i|, \quad \mathcal{L}(\boldsymbol{\theta}^\star + \boldsymbol{\Delta\theta}) = \frac{1}{m}\sum_{i\in\bar{\mathcal{S}}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}\rangle| + \frac{1}{m}\sum_{i\in\mathcal{S}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}\rangle - \epsilon_i|. \tag{3.14}$$

Here $\mathcal{S} \in [m] : \{1, 2, \ldots, m\}$ is the support of the noise vector $\boldsymbol{\epsilon} = [\epsilon_1, \cdots, \epsilon_m]^\top$, and $\bar{\mathcal{S}} = [m] - \mathcal{S}$. Hence, we have

$$\mathcal{L}(\boldsymbol{\theta}^\star + \boldsymbol{\Delta\theta}) - \mathcal{L}(\boldsymbol{\theta}^\star) = \frac{1}{m}\sum_{i\in\mathcal{S}} (|\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}\rangle - \epsilon_i| - |\epsilon_i|) + \frac{1}{m}\sum_{i\in\bar{\mathcal{S}}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}\rangle|. \tag{3.15}$$

For simplicity, we denote $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m]^\top \in \mathbb{R}^{m\times d}$ by the design matrix. To upper bound the infimum of $\mathcal{L}(\boldsymbol{\theta}^\star + \boldsymbol{\Delta\theta}) - \mathcal{L}(\boldsymbol{\theta}^\star)$ over the ℓ_∞-norm ball $\{\boldsymbol{\Delta\theta} : \|\boldsymbol{\Delta\theta}\|_\infty \le \gamma\}$, we choose a specific $\overline{\boldsymbol{\Delta\theta}}$ for each realization $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$ and show that $\overline{\boldsymbol{\Delta\theta}}$ lies in the ℓ_∞-norm ball with high probability. To this goal, we pick $\overline{\boldsymbol{\Delta\theta}}$ by solving the following linear equation

$$\boldsymbol{X}\boldsymbol{\Delta\theta} = \boldsymbol{b}, \tag{3.16}$$

where $b[i] = \operatorname{sign}(\epsilon_i)(\xi \wedge |\epsilon_i|)$ for $i \in \mathcal{S}$, and $b[i] = 0$ otherwise. Here $\xi > 0$ is a hyperparameter to be tuned later. Before showing that such $\overline{\boldsymbol{\Delta\theta}}$ lies in the ℓ_∞-norm ball with high probability, we first further upper bound $\mathcal{L}(\boldsymbol{\theta}^\star + \boldsymbol{\Delta\theta}) - \mathcal{L}(\boldsymbol{\theta}^\star)$ for this specific choice of $\overline{\boldsymbol{\Delta\theta}}$. By our construction of $\overline{\boldsymbol{\Delta\theta}}$, we first have

$$\begin{aligned}\mathcal{L}\left(\boldsymbol{\theta}^\star + \overline{\boldsymbol{\Delta\theta}}\right) - \mathcal{L}(\boldsymbol{\theta}^\star) &= \frac{1}{m}\sum_{i\in\mathcal{S}}\left(\left|\left\langle \boldsymbol{x}_i, \overline{\boldsymbol{\Delta\theta}}\right\rangle - \epsilon_i\right| - |\epsilon_i|\right) + \frac{1}{m}\sum_{i\in\bar{\mathcal{S}}}\left|\left\langle \boldsymbol{x}_i, \overline{\boldsymbol{\Delta\theta}}\right\rangle\right| \\ &= -\frac{1}{m}\sum_{i\in\mathcal{S}}\xi \wedge |\epsilon_i|.\end{aligned} \tag{3.17}$$

On the other hand, based on Assumption 2, each $\epsilon_i, i \in \mathcal{S}$ has nonzero probability to be away from zero. Hence, we can define the event $\mathcal{E} := \{\text{There are at least } p_0 pm \text{ elements of } |\epsilon| \text{ larger than } t_0\}$. Define $\mathcal{M}_0 = \{i : |\epsilon_i| \ge t_0\}$. Using the tail bound of binomial distribution, we have $\mathbb{P}(\mathcal{E}) \ge \frac{1}{4}$. Conditioned on the event $\mathcal{E}$, we have $m_0 = |\mathcal{M}_0| \ge p_0 pm$ and

$$\mathcal{L}\left(\boldsymbol{\theta}^\star + \overline{\boldsymbol{\Delta\theta}}\right) - \mathcal{L}(\boldsymbol{\theta}^\star) = -\frac{1}{m}\sum_{i\in\mathcal{S}}\xi \wedge |\epsilon_i| \le -p_0 p \cdot (\xi \wedge t_0). \tag{3.18}$$

Now it suffices to show that such $\overline{\boldsymbol{\Delta\theta}}$ belongs to the ℓ_∞-norm ball with high probability. To this goal, we first choose a specific $\overline{\boldsymbol{\Delta\theta}}$ satisfying $\boldsymbol{X}\overline{\boldsymbol{\Delta\theta}} = b$. We divide $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2]$ where $\boldsymbol{X}_1 \in \mathbb{R}^{m\times m}$ and $\boldsymbol{X}_2 \in \mathbb{R}^{m\times(d-m)}$. Note that $\boldsymbol{X}_1$ is non-singular almost surely. Hence, we can pick $\overline{\boldsymbol{\Delta\theta}} = [b^\top \boldsymbol{X}_1(\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1}, \boldsymbol{0}]^\top$. Next, it suffices to verify that with high probability, $\left\|\overline{\boldsymbol{\Delta\theta}}\right\|_\infty \le \gamma$, which is equivalent to

$$\left\|(\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top b\right\|_\infty \le \gamma. \tag{3.19}$$

To this goal, note that with probability at least $1 - e^{-\Omega(m)}$

$$\begin{aligned}\left\|(\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top b\right\|_\infty &\le \left\|(\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top b\right\| \\ &\le \left\|\left(\frac{1}{m}\boldsymbol{X}_1^\top \boldsymbol{X}_1\right)^{-1}\right\| \left\|\frac{1}{m}\boldsymbol{X}_1\right\| \|\boldsymbol{b}\| \\ &\overset{(a)}{\lesssim} \frac{1}{\sqrt{m}}\|\boldsymbol{b}\| \\ &\lesssim \xi.\end{aligned} \tag{3.20}$$

Here in (a) we used the fact that $\left\|\left(\frac{1}{m}\boldsymbol{X}_1^\top \boldsymbol{X}_1\right)^{-1}\right\| \lesssim 1$ and $\left\|\frac{1}{m}\boldsymbol{X}_1\right\| \lesssim \frac{1}{\sqrt{m}}$ with probability at least $1 - e^{-\Omega(m)}$. Hence, we only need to set $\xi \lesssim \gamma$, so that the chosen $\overline{\boldsymbol{\Delta\theta}}$ belongs to the ℓ_∞-norm ball with probability at least $1 - e^{-\Omega(m)}$.

Combining the above steps and choosing $\xi \lesssim \gamma$, we know that with probability at least $\frac{1}{16}$, the below upper bound holds

$$\begin{aligned}\inf_{\|\boldsymbol{\Delta\theta}\|_\infty \le \gamma} \{\mathcal{L}(\boldsymbol{\theta}^\star + \boldsymbol{\Delta\theta}) - \mathcal{L}(\boldsymbol{\theta}^\star)\} &\overset{\text{w.p. } 1-e^{-\Omega(m)}}{\le} \mathcal{L}(\boldsymbol{\theta}^\star + \overline{\boldsymbol{\Delta\theta}}) - \mathcal{L}(\boldsymbol{\theta}^\star) \\ &\overset{\text{w.p. } \frac{1}{4}}{\le} -p_0 p\xi \\ &\lesssim -p_0 p\gamma.\end{aligned} \tag{3.21}$$

This completes the proof for the 1-layer model. For general N-layer models, we consider the true solution $\boldsymbol{w} \in \mathcal{W}$ with $\boldsymbol{w}_1 = \boldsymbol{\theta}^\star$, and $\boldsymbol{w}_2 = \cdots = \boldsymbol{w}_N = \mathbf{1}$. Moreover, for $\boldsymbol{\Delta w}$ we choose $\boldsymbol{\Delta w}_2 = \cdots \boldsymbol{\Delta w}_N = \mathbf{0}$. It is easy to verify that

$$\mathcal{L}(\boldsymbol{w}) - \mathcal{L}(\boldsymbol{w} + \boldsymbol{\Delta w}) = \frac{1}{m}\sum_{i\in\mathcal{S}} \left(\left|\langle \boldsymbol{x}_i, \boldsymbol{\Delta w}_1\rangle - \epsilon_i\right| - |\epsilon_i|\right) + \frac{1}{m}\sum_{i\in\bar{\mathcal{S}}} \left|\langle \boldsymbol{x}_i, \boldsymbol{\Delta w}_1\rangle\right|. \tag{3.22}$$

Therefore, an argument similar to the 1-layer model can be used to write

$$\inf_{\boldsymbol{w}\in\mathcal{W}} \inf_{\boldsymbol{w}':\|\boldsymbol{w}-\boldsymbol{w}'\|_\infty \le \gamma} \{\mathcal{L}(\boldsymbol{w}) - \mathcal{L}(\boldsymbol{w}')\} \le \inf_{\|\boldsymbol{\Delta w}_1\|_\infty \le \gamma} \{\mathcal{L}(\boldsymbol{w}) - \mathcal{L}(\boldsymbol{w} + \boldsymbol{\Delta w})\} \lesssim -p_0 p\gamma, \tag{3.23}$$

with probability of $\frac{1}{16}$.

Under-parameterized Regime

Given $\boldsymbol{w} \in \mathcal{W}$ and any $\boldsymbol{\Delta w}$, consider $\boldsymbol{w}' = \boldsymbol{w} + \boldsymbol{\Delta w}$ and define

$$\boldsymbol{\Delta\theta} = \sum_{i=1}^{N} (\boldsymbol{\theta}^\star)^{\frac{N-i}{N}} \odot \sum_{j_1,\cdots,j_i} \boldsymbol{\Delta w}_{j_1} \odot \cdots \odot \boldsymbol{\Delta w}_{j_i}. \tag{3.24}$$

We have

$$\begin{aligned} \mathcal{L}(\boldsymbol{w}) - \mathcal{L}(\boldsymbol{w} + \boldsymbol{\Delta w}) &= \frac{1}{m} \sum_{i \in \mathcal{S}} \left(\left| \langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta} \rangle - \epsilon_i \right| - |\epsilon_i| \right) + \frac{1}{m} \sum_{i \in \bar{\mathcal{S}}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta} \rangle| \\ &\geq \frac{1}{m} \sum_{i \in \bar{\mathcal{S}}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta} \rangle| - \frac{1}{m} \sum_{i \in \mathcal{S}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta} \rangle| . \end{aligned} \tag{3.25}$$

Hence, it suffices to show that, with probability of $1 - e^{-\Omega(d)}$,

$$\inf_{\boldsymbol{\Delta\theta} \in \mathbb{R}^d} \left\{ \frac{1}{m} \sum_{i \in \bar{\mathcal{S}}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta} \rangle| - \frac{1}{m} \sum_{i \in \mathcal{S}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta} \rangle| \right\} \geq 0. \tag{3.26}$$

Note that the above inequality is invariant with respect to scaling. Hence, it suffices to show that it holds for arbitrary $\boldsymbol{\Delta\theta} \in \mathbb{S}^{d-1}$ where $\mathbb{S}^{d-1} := \{x \in \mathbb{R}^d : \|\boldsymbol{x}\| = 1\}$ is the standard sphere. Hence, it suffices to show

$$\begin{aligned} &\inf_{\boldsymbol{\Delta\theta} \in \mathbb{S}^{d-1}} \left\{ \frac{1}{m} \sum_{i \in \bar{\mathcal{S}}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta} \rangle| - \frac{1}{m} \sum_{i \in \mathcal{S}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta} \rangle| \right\} \\ &\geq \inf_{\boldsymbol{\Delta\theta} \in \mathbb{S}^{d-1}} \left\{ \frac{1}{m} \sum_{i \in \bar{\mathcal{S}}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta} \rangle| \right\} - \sup_{\boldsymbol{\Delta\theta} \in \mathbb{S}^{d-1}} \left\{ \frac{1}{m} \sum_{i \in \mathcal{S}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta} \rangle| \right\} \geq 0. \end{aligned} \tag{3.27}$$

For the first term, applying Lemma 59, we have that with probability at least $1 - e^{-\Omega(d)}$

$$\inf_{\boldsymbol{\Delta\theta} \in \mathbb{S}^{d-1}} \frac{1}{m} \sum_{i \in \bar{\mathcal{S}}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta} \rangle| \geq \sqrt{\frac{2}{\pi}}(1-p) - \sqrt{\frac{(1-p)d}{m}}. \tag{3.28}$$

Similarly, for the second part, with probability of at least $1 - e^{-\Omega(d)}$, we have

$$\sup_{\boldsymbol{\Delta\theta} \in \mathbb{S}^{d-1}} \frac{1}{m} \sum_{i \in \mathcal{S}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta} \rangle| \leq \sqrt{\frac{2}{\pi}} p + \sqrt{\frac{pd}{m}}. \tag{3.29}$$

Combining both parts, we have that with probability at least $1 - e^{-\Omega(d)}$

$$\inf_{\boldsymbol{\Delta\theta}\in\mathbb{S}^{d-1}} \left\{ \frac{1}{m} \sum_{i\in\bar{\mathcal{S}}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}\rangle| \right\} - \sup_{\boldsymbol{\Delta\theta}\in\mathbb{S}^{d-1}} \left\{ \frac{1}{m} \sum_{i\in\mathcal{S}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}\rangle| \right\} \geq \sqrt{\frac{2}{\pi}}(1-2p) - 2\sqrt{\frac{d}{m}} \geq 0. \tag{3.30}$$

The last inequality follows from the fact that $m \gtrsim \frac{d}{(1-2p)^2}$. This completes the proof. □

3.C.2 Proof of Theorem 17

Given any $\boldsymbol{\Delta w} = [\boldsymbol{\Delta w}_1, \cdots, \boldsymbol{\Delta w}_N]^\top$, the following equality holds for any point $\boldsymbol{w} = \boldsymbol{w}^\star + \boldsymbol{\Delta w}$ where $\boldsymbol{w}^\star = [\sqrt[N]{\boldsymbol{\theta}^\star}, \cdots, \sqrt[N]{\boldsymbol{\theta}^\star}]^\top$:

$$\begin{aligned}(\sqrt[N]{\boldsymbol{\theta}^\star} + \boldsymbol{\Delta w}_1) \odot \cdots \odot (\sqrt[N]{\boldsymbol{\theta}^\star} + \boldsymbol{\Delta w}_N) - \boldsymbol{\theta}^\star = & \underbrace{\sum_{i=1}^{N-1} (\boldsymbol{\theta}^\star)^{\frac{N-i}{N}} \odot \sum_{j_1,\cdots,j_i} \boldsymbol{\Delta w}_{j_1} \odot \cdots \odot \boldsymbol{\Delta w}_{j_i}}_{:=\boldsymbol{\Delta\theta}_1} \\ & + \underbrace{\boldsymbol{\Delta w}_1 \odot \cdots \odot \boldsymbol{\Delta w}_N}_{:=\boldsymbol{\Delta\theta}_2}.\end{aligned} \tag{3.31}$$

Hence, we have

$$\mathcal{L}(\boldsymbol{w}^\star) - \mathcal{L}(\boldsymbol{w}^\star + \boldsymbol{\Delta w}) = \frac{1}{m} \sum_{i\in\mathcal{S}} \left(|\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}_1 + \boldsymbol{\Delta\theta}_2\rangle - \epsilon_i| - |\epsilon_i| \right) + \frac{1}{m} \sum_{i\in\bar{\mathcal{S}}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}_1 + \boldsymbol{\Delta\theta}_2\rangle|. \tag{3.32}$$

For simplicity, we denote Θ_1 as the set of $\boldsymbol{\Delta\theta}_1$ defined in equation 3.31 with $\|\boldsymbol{\Delta w}\|_\infty \leq \gamma$. Similarly, Θ_2 is the set of $\boldsymbol{\Delta\theta}_2$ defined in equation 3.31 with $\|\boldsymbol{\Delta w}\|_\infty \leq \gamma$.

Lower bound. To prove the lower bound, one can write

$$\begin{aligned}\mathcal{L}(\boldsymbol{w}^\star) - \mathcal{L}(\boldsymbol{w}^\star + \boldsymbol{\Delta w}) &\geq \frac{1}{m} \sum_{i\in\bar{\mathcal{S}}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}_1 + \boldsymbol{\Delta\theta}_2\rangle| - \frac{1}{m} \sum_{i\in\mathcal{S}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}_1 + \boldsymbol{\Delta\theta}_2\rangle| \\ &\geq \frac{1}{m} \sum_{i\in\bar{\mathcal{S}}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}_1\rangle| - \frac{1}{m} \sum_{i\in\mathcal{S}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}_1\rangle| - \frac{1}{m} \sum_{i=1}^{m} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}_2\rangle|\end{aligned} \tag{3.33}$$

Hence, we have

$$\inf_{\|\boldsymbol{\Delta w}\|_\infty \leq \gamma} \{\mathcal{L}(\boldsymbol{w}^\star) - \mathcal{L}(\boldsymbol{w}^\star + \boldsymbol{\Delta w})\} \geq \inf_{\boldsymbol{\Delta\theta}_1 \in \Theta_1} \left\{ \frac{1}{m} \sum_{i \in \bar{\mathcal{S}}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}_1 \rangle| - \frac{1}{m} \sum_{i \in \mathcal{S}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}_1 \rangle| \right\} - \sup_{\boldsymbol{\Delta\theta}_2 \in \Theta_2} \left\{ \frac{1}{m} \sum_{i=1}^{m} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}_2 \rangle| \right\}. \tag{3.34}$$

First, we bound the term $\inf_{\boldsymbol{\Delta\theta}_1 \in \Theta_1} \left\{ \frac{1}{m} \sum_{i \in \bar{\mathcal{S}}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}_1 \rangle| - \frac{1}{m} \sum_{i \in \mathcal{S}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}_1 \rangle| \right\}$. It is easy to see that the vector $\boldsymbol{\Delta\theta}_1$ is k-sparse and has the same sparsity pattern as $\boldsymbol{\theta}^\star$. Therefore, according to Lemma 59, there exist universal constants $C, c > 0$ such that the following hold

$$\mathbb{P}\left(\sup_{\boldsymbol{\Delta\theta}_1 \in \Theta_1} \left| \frac{1}{m' \|\boldsymbol{\Delta\theta}_1\|} \sum_{i \in \bar{\mathcal{S}}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}_1 \rangle| - \sqrt{\frac{2}{\pi}} \right| \geq C\sqrt{\frac{k}{m'}} + \delta \right) \leq e^{-cm'\delta^2}, \tag{3.35}$$

and

$$\mathbb{P}\left(\sup_{\boldsymbol{\Delta\theta}_1 \in \Theta_1} \left| \frac{1}{m'' \|\boldsymbol{\Delta\theta}_1\|} \sum_{i \in \mathcal{S}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}_1 \rangle| - \sqrt{\frac{2}{\pi}} \right| \geq C\sqrt{\frac{k}{m''}} + \delta \right) \leq e^{-cm''\delta^2}. \tag{3.36}$$

Here $m' = (1-p)m$, and $m'' = pm$. The inequality equation 3.35 implies that

$$\begin{aligned} &\mathbb{P}\left(\frac{1}{m} \sum_{i \in \bar{\mathcal{S}}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}_1 \rangle| \geq \|\boldsymbol{\Delta\theta}_1\| \left(\sqrt{\frac{2}{\pi}}(1-p) - C\sqrt{\frac{(1-p)k}{m}} - \delta_1 \right), \forall \boldsymbol{\Delta\theta}_1 \in \Theta_1 \right) \\ &\geq 1 - e^{-\frac{cm\delta_1^2}{1-p}}. \end{aligned} \tag{3.37}$$

Similarly, the inequality equation 3.36 leads to

$$\mathbb{P}\left(\frac{1}{m} \sum_{i \in \mathcal{S}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}_1 \rangle| \leq \|\boldsymbol{\Delta\theta}_1\| \left(\sqrt{\frac{2}{\pi}}p + C\sqrt{\frac{pk}{m}} + \delta_2 \right), \forall \boldsymbol{\Delta\theta}_1 \in \Theta_1 \right) \geq 1 - e^{-\frac{cm\delta_2^2}{p}}. \tag{3.38}$$

Upon setting $\delta_1 = \sqrt{\frac{(1-p)k}{m}}$ and $\delta_2 = \sqrt{\frac{pk}{m}}$, with probability of $1 - e^{-\Omega(k)}$, for all $\boldsymbol{\Delta\theta}_1 \in \Theta_1$, we have

$$\frac{1}{m} \sum_{i \in \bar{\mathcal{S}}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}_1 \rangle| - \frac{1}{m} \sum_{i \in \mathcal{S}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}_1 \rangle| \geq \sqrt{\frac{2}{\pi}}(1-2p) - C'\sqrt{\frac{k}{m}} \geq 0. \tag{3.39}$$

In the last inequality we used the assumption $m \gtrsim \frac{k}{(1-2p)^2}$. The above argument implies

$$\inf_{\boldsymbol{\Delta\theta}_1 \in \Theta_1} \left\{ \frac{1}{m} \sum_{i \in \bar{\mathcal{S}}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}_1 \rangle| - \frac{1}{m} \sum_{i \in \mathcal{S}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}_1 \rangle| \right\} \geq 0, \tag{3.40}$$

with probability of at least $1 - e^{-\Omega(k)}$. Now we turn to bound the second part $\sup_{\boldsymbol{\Delta\theta}_2 \in \Theta_2} \left\{ \frac{1}{m} \sum_{i=1}^m |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}_2 \rangle| \right\}$. To this goal, we first apply Lemma 59, which leads to

$$\mathbb{P}\left(\sup_{\boldsymbol{\Delta\theta}_2 \in \Theta_2} \left| \frac{1}{m \|\boldsymbol{\Delta\theta}_2\|} \sum_{i=1}^m |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}_2 \rangle| - \sqrt{\frac{2}{\pi}} \right| \geq C\sqrt{\frac{d}{m}} + \delta \right) \leq e^{-cm\delta^2}. \tag{3.41}$$

Therefore, upon setting $\delta = \sqrt{\frac{d}{m}}$, with probability of $1 - e^{-\Omega(d)}$, we have

$$\begin{aligned} \sup_{\boldsymbol{\Delta\theta}_2 \in \Theta_2} \left\{ \frac{1}{m} \sum_{i=1}^m |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}_2 \rangle| \right\} &\leq \sup_{\boldsymbol{\Delta\theta}_2 \in \Theta_2} \|\boldsymbol{\Delta\theta}_2\| \left(\sqrt{\frac{2}{\pi}} + (C+1)\sqrt{\frac{d}{m}} \right) \\ &\lesssim \sqrt{\frac{d}{m}} \sup_{\|\boldsymbol{\Delta w}\|_\infty \leq \gamma} \|\boldsymbol{\Delta w}_1 \odot \cdots \odot \boldsymbol{\Delta w}_N\| \\ &= \sqrt{\frac{d}{m}} \sup_{|\boldsymbol{\Delta w}_i[j]| \leq \gamma} \sqrt{\sum_{j \in [d]} \left(\prod_{i \in [N]} \boldsymbol{\Delta w}_i[j] \right)^2} \\ &= \frac{d}{\sqrt{m}} \gamma^N. \end{aligned} \tag{3.42}$$

Here $\boldsymbol{\Delta w}_i[j]$ is the j-th element of $\boldsymbol{\Delta w}_i$. Therefore, we conclude that

$$\inf_{\|\boldsymbol{\Delta w}\|_\infty \leq \gamma} \{\mathcal{L}(\boldsymbol{w}^\star) - \mathcal{L}(\boldsymbol{w}^\star + \boldsymbol{\Delta w})\} \gtrsim -\frac{d}{\sqrt{m}} \gamma^N, \tag{3.43}$$

with probability of $1 - e^{-\Omega(k)}$, thereby completing the proof of the lower bound.

Upper bound. To this goal, we first define a restricted set of perturbation vectors

$$\mathcal{V} := \{\mathbf{v} : \|\mathbf{v}\|_\infty \leq \gamma, v_i[j] = 0, \forall i \in [N], j \in \mathrm{supp}(\boldsymbol{\theta}^\star)\}, \tag{3.44}$$

where $\text{supp}(\boldsymbol{\theta}^\star)$ is the support of $\boldsymbol{\theta}^\star$. Based on this definition, we have

$$\begin{aligned}
&\inf_{\|\boldsymbol{\Delta w}\|_\infty \le \gamma} \{\mathcal{L}(\boldsymbol{w}^\star) - \mathcal{L}(\boldsymbol{w}^\star + \boldsymbol{\Delta w})\} \\
&\le \inf_{\boldsymbol{\Delta w} \in \mathcal{V}} \{\mathcal{L}(\boldsymbol{w}^\star) - \mathcal{L}(\boldsymbol{w}^\star + \boldsymbol{\Delta w})\} \\
&= \inf_{\boldsymbol{\Delta w} \in \mathcal{V}} \left\{ \frac{1}{m} \sum_{i \in \mathcal{S}} (|\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}_2 \rangle - \epsilon_i| - |\epsilon_i|) + \frac{1}{m} \sum_{i \in \bar{\mathcal{S}}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}_2 \rangle| \right\}.
\end{aligned} \tag{3.45}$$

where $\boldsymbol{\Delta\theta}_2 = \boldsymbol{\Delta w}_1 \odot \cdots \odot \boldsymbol{\Delta w}_N$ is the same as before. Note that for any $\|\boldsymbol{\Delta w}\|_\infty \le \gamma$, we have $\|\boldsymbol{\Delta\theta}_2\| \le \sqrt{d}\gamma^N$. Moreover, this bound is attainable when $\boldsymbol{\Delta w}_i \equiv [\pm\gamma, \cdots, \pm\gamma]^\top$. Hence, we have

$$\begin{aligned}
&\inf_{\boldsymbol{\Delta w} \in \mathcal{V}} \left\{ \frac{1}{m} \sum_{i \in \mathcal{S}} (|\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}_2 \rangle - \epsilon_i| - |\epsilon_i|) + \frac{1}{m} \sum_{i \in \bar{\mathcal{S}}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}_2 \rangle| \right\} \\
&\le \inf_{\|\boldsymbol{\Delta\theta}_2\| \le \sqrt{d}\gamma^N} \left\{ \frac{1}{m} \sum_{i \in \mathcal{S}} (|\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}_2 \rangle - \epsilon_i| - |\epsilon_i|) + \frac{1}{m} \sum_{i \in \bar{\mathcal{S}}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}_2 \rangle| \right\}.
\end{aligned} \tag{3.46}$$

An argument similar to the proof of Theorem 14 can be used to show that, with probability of at least $1/16$, we have

$$\inf_{\|\boldsymbol{\Delta\theta}_2\| \le \sqrt{d}\gamma^N} \left\{ \frac{1}{m} \sum_{i \in S} (|\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}_2 \rangle - \epsilon_i| - |\epsilon_i|) + \frac{1}{m} \sum_{i \in \bar{S}} |\langle \boldsymbol{x}_i, \boldsymbol{\Delta\theta}_2 \rangle| \right\} \lesssim -\sqrt{\frac{p_0 p(d-k)}{m}} \sqrt{d}\gamma^N. \tag{3.47}$$

Recalling that $k \ll d$, we have with probability of $1/16$

$$\inf_{\|\boldsymbol{\Delta w}\|_\infty \le \gamma} \{\mathcal{L}(\boldsymbol{w}^\star) - \mathcal{L}(\boldsymbol{w}^\star + \boldsymbol{\Delta w})\} \lesssim -\sqrt{p_0 p} \frac{d}{\sqrt{m}} \gamma^N. \tag{3.48}$$

This completes the proof. □

3.D Proofs of Convergence Analysis

3.D.1 Proof of Theorem 15

For simplicity of notation, we denote $\boldsymbol{u} = \boldsymbol{w}_1$ and $\boldsymbol{v} = \boldsymbol{w}_2$. Moreover, without loss of generality, we assume that the elements of $\boldsymbol{\theta}^\star$ are arranged in descending order, i.e., $\theta_1^\star \ge \cdots \ge \theta_k^\star > \theta_{k+1}^\star = \cdots = \theta_d^\star = 0$, and the initial point satisfies $u_i, v_i = \Theta(\sqrt{\alpha}), \forall i \in [d]$. Moreover, for $v \in \mathbb{R}^d$, we define $\boldsymbol{v}_{:i} = [v_1, \cdots, v_i]^\top$ and $\boldsymbol{v}_{i:} = [v_i, \cdots, v_d]^\top$. For short, we

denote $\boldsymbol{v}_{-i} = \boldsymbol{v}_{i+1:}$. Finally, we define $\kappa = \theta_1^\star/\theta_k^\star$ as the condition number. Moreover, without loss of generality, we assume that $\theta_1^\star \geq 1 \geq \theta_k^\star$.

First, according to Proposition 11, the sub-differential of $\mathcal{L}(\boldsymbol{u}, \boldsymbol{v})$ is uniformly concentrated around its population gradient. In particular, with probability at least $1 - Ce^{-cm\delta^2}$, we have

$$u_{i,t+1} = u_{i,t} + \eta \frac{\theta_i^\star - u_{i,t}v_{i,t}}{\|\boldsymbol{u}_t \odot \boldsymbol{v}_t - \boldsymbol{\theta}^\star\|} v_{i,t} + \eta\delta_i v_{i,t}, \text{ and } |\delta_i| \leq \delta, \forall i \in [d], \tag{3.49}$$

$$v_{i,t+1} = v_{i,t} + \eta \frac{\theta_i^\star - u_{i,t}v_{i,t}}{\|\boldsymbol{u}_t \odot \boldsymbol{v}_t - \boldsymbol{\theta}^\star\|} u_{i,t} + \eta\delta_i u_{i,t}, \text{ and } |\delta_i| \leq \delta, \forall i \in [d]. \tag{3.50}$$

Hence, we have

$$u_{i,t+1}v_{i,t+1} = u_{i,t}v_{i,t} + \eta \left(\frac{\theta_i^\star - u_{i,t}v_{i,t}}{\|\boldsymbol{u}_t \odot \boldsymbol{v}_t - \boldsymbol{\theta}^\star\|} + \delta_i \right) \left(u_{i,t}^2 + v_{i,t}^2\right) + \eta^2 \left(\frac{\theta_i^\star - u_{i,t}v_{i,t}}{\|\boldsymbol{u}_t \odot \boldsymbol{v}_t - \boldsymbol{\theta}^\star\|} + \delta_i \right)^2 u_{i,t}v_{i,t}. \tag{3.51}$$

Moreover, we have

$$\begin{aligned} &u_{i,t+1}^2 + v_{i,t+1}^2 \\ &= \left(u_{i,t}^2 + v_{i,t}^2\right) \left(1 + \eta^2 \left(\frac{(\theta_i^\star - u_{i,t}v_{i,t})}{\|\boldsymbol{u}_t \odot \boldsymbol{v}_t - \boldsymbol{\theta}^\star\|} + \delta_i \right)^2 \right) + 4\eta \left(\frac{\theta_i^\star - u_{i,t}v_{i,t}}{\|\boldsymbol{u}_t \odot \boldsymbol{v}_t - \boldsymbol{\theta}^\star\|} + \delta_i \right) u_{i,t}v_{i,t}. \end{aligned} \tag{3.52}$$

Signal Dynamics

We first study the behavior of the signal term $\boldsymbol{S}_t = \boldsymbol{u}_{:k,t}\boldsymbol{v}_{:k,t}$ for the first k components of the model $\boldsymbol{u}_t \odot \boldsymbol{v}_t$. We divide the dynamics into $k+1$ stages. In the first k stages, each component $u_{i,t}v_{i,t}$ converges to $\theta_i^\star$ sequentially. Once all the components are close to the ground truth, the distance between signal term and the ground truth $\|\boldsymbol{S}_t - \boldsymbol{\theta}_{:k}^\star\|$ will further decrease to $\mathcal{O}\left(\sqrt{d^2m}\alpha^2 \vee \eta\theta_1^\star\right)$.

Stage 1: In this stage, the first component u_1v_1 grows to $\theta_1 - \delta\|\boldsymbol{\theta}^\star\|$ within $\Theta\left(\frac{\|\boldsymbol{\theta}^\star\|}{\eta\theta_1^\star}\log\left(\frac{1}{\alpha}\right)\right)$ iterations. At the initial point, we have $u_{1,0}v_{1,0} = \Theta(\alpha)$. For iteration $t+1$, according to equation 3.51, we have

$$\begin{aligned} u_{1,t+1}v_{1,t+1} &\geq u_{1,t}v_{1,t} + \eta \left(\frac{\theta_1^\star - u_{1,t}v_{1,t}}{\|\boldsymbol{u}_t \odot \boldsymbol{v}_t - \boldsymbol{\theta}^\star\|} + \delta_1 \right) \left(u_{1,t}^2 + v_{1,t}^2\right) \\ &\geq u_{1,t}v_{1,t} + 2\eta \left(\frac{\theta_1^\star - u_{1,t}v_{1,t}}{\|\boldsymbol{u}_t \odot \boldsymbol{v}_t - \boldsymbol{\theta}^\star\|} + \delta_1 \right) u_{1,t}v_{1,t}. \end{aligned} \tag{3.53}$$

We further divide our analysis into two substages. In the first substage, we have $u_1 v_1 \leq \theta_1^\star/2$. Note that $\|\boldsymbol{u}_t \odot \boldsymbol{v}_t - \boldsymbol{\theta}^\star\| = \mathcal{O}(\|\boldsymbol{\theta}^\star\|)$ and $|\delta_1| \leq \delta \lesssim \frac{1}{\kappa}$. Hence, equation 3.53 can be further simplified as

$$u_{1,t+1} v_{1,t+1} \geq \left(1 + \frac{1}{4} \frac{\eta \theta_1^\star}{\|\boldsymbol{\theta}^\star\|}\right) u_{1,t} v_{1,t}. \tag{3.54}$$

Therefore, this stage ends within $\mathcal{O}\left(\frac{\|\boldsymbol{\theta}^\star\|}{\eta \theta_1^\star} \log\left(\frac{1}{\alpha}\right)\right)$ iterations. In the second substage, we have $u_1 v_1 \geq \theta_1^\star/2$. Upon defining $x_t = \theta_1^\star - u_{1,t} v_{1,t}$, one can write

$$\begin{aligned} x_{t+1} &\leq \left(1 - 2\eta \frac{u_{1,t} v_{1,t}}{\|\boldsymbol{u}_t \odot \boldsymbol{v}_t - \boldsymbol{\theta}^\star\|}\right) x_t - 2\eta \delta_1 \boldsymbol{u}_t \odot \boldsymbol{v}_t \\ &\leq \left(1 - \frac{\eta \theta_1^\star}{\|\boldsymbol{\theta}^\star\|}\right) x_t + \eta \delta \theta_1^\star. \end{aligned} \tag{3.55}$$

Hence, within additional $\mathcal{O}\left(\|\boldsymbol{\theta}^\star\| / (\eta \theta_1^\star)\right)$ iterations, we have $u_{1,T_1} v_{1,T_1} = \theta_1^\star \pm \delta \|\boldsymbol{\theta}^\star\|$. Overall, within $T_1 = \mathcal{O}\left(\frac{\|\boldsymbol{\theta}^\star\|}{\eta \theta_1^\star} \log\left(\frac{1}{\alpha}\right)\right)$ iterations, we have $u_{1,T_1} v_{1,T_1} = \theta_1^\star \pm \delta \|\boldsymbol{\theta}^\star\|$. Now, we turn to show the lower bound on T_1 by analyzing the trajectory of $u_{1,t}^2 + v_{1,t}^2$. Due to equation 3.52, when $u_{1,t}^2 + v_{1,t}^2 \leq \frac{\theta_1^\star}{2}$, we have

$$u_{1,t+1}^2 + v_{1,t+1}^2 \leq \left(u_{1,t}^2 + v_{1,t}^2\right)\left(1 + 10 \frac{\eta \theta_1^\star}{\|\boldsymbol{\theta}^\star\|}\right). \tag{3.56}$$

Hence, at least $\Omega\left(\frac{\|\boldsymbol{\theta}^\star\|}{\eta \theta_1^\star} \log\left(\frac{1}{\alpha}\right)\right)$ iterations are needed for $u_{1,t}^2 + v_{1,t}^2$ to be larger than $\frac{\theta_1^\star}{2}$. Since $u_{1,t}^2 + v_{1,t}^2 \geq u_{1,t} v_{1,t}$, we immediately obtain $T_1 = \Omega\left(\frac{\|\boldsymbol{\theta}^\star\|}{\eta \theta_1^\star} \log\left(\frac{1}{\alpha}\right)\right)$.

Stage 2 to Stage k: In the next $k-1$ stages, each component $u_i v_i$ will converge to $\theta_i^\star \pm \delta \|\boldsymbol{\theta}^\star\|$ sequentially. To show this, we use an inductive argument. In each stage i, we assume that the first $i-1$ components have already converged close to $\boldsymbol{\theta}_j^\star, \forall j \in [i-1]$. Hence, we have $\|\boldsymbol{u}_t \odot \boldsymbol{v}_t - \boldsymbol{\theta}^\star\| = \Theta\left(\left\|\boldsymbol{\theta}^\star_{-(i-1)}\right\|\right)$. Repeating the procedure in Stage 1, we can show that, at stage i, $T_i = \mathcal{O}\left(\frac{\left\|\boldsymbol{\theta}^\star_{-(i-1)}\right\|}{\eta \theta_i^\star} \log\left(\frac{1}{\alpha}\right)\right)$ iterations are needed for $u_i v_i$ to converge to $\theta_i^\star \pm \delta \|\boldsymbol{\theta}^\star\|$. Overall, after $T = T_1 + \cdots T_k = \mathcal{O}\left(\frac{k^{3/2}}{\eta} \log\left(\frac{1}{\alpha}\right)\right)$ iterations, we have

$$\left\|\boldsymbol{u}_{:k,T} \boldsymbol{v}_{:k,T} - \boldsymbol{\theta}^\star_{:k,T}\right\| \lesssim \sqrt{k} \delta \|\boldsymbol{\theta}^\star\|. \tag{3.57}$$

Stage $k+1$**:** In the final stage, the signal term will quickly decrease to $\mathcal{O}\left(\sqrt{d^2m}\alpha^{1-\Theta(\delta)} \vee \eta\theta_1^\star\right)$ within $T_{k+1} = \mathcal{O}\left(\frac{\delta\|\boldsymbol{\theta}^\star\|}{\eta\theta_k^\star}\right)$ iterations. To show this, we write

$$\begin{aligned}
&\boldsymbol{\theta}_{:k}^\star - \boldsymbol{S}_{t+1} \\
&= \boldsymbol{\theta}_{:k}^\star - \boldsymbol{S}_t - \eta\frac{\boldsymbol{u}_{:k,t}^2 + \boldsymbol{v}_{:k,t}^2}{\|\boldsymbol{u}_t \odot \boldsymbol{v}_t - \boldsymbol{\theta}^\star\|} \odot (\boldsymbol{\theta}_{:k}^\star - \boldsymbol{S}_t) - \eta\boldsymbol{\delta}\left(\boldsymbol{u}_t^2 + \boldsymbol{v}_t^2\right) + \eta^2\left(\frac{\boldsymbol{\theta}_{:k}^\star - \boldsymbol{S}_t}{\|\boldsymbol{u}_t \odot \boldsymbol{v}_t - \boldsymbol{\theta}^\star\|} + \boldsymbol{\delta}\right)^2 \odot \boldsymbol{S}_t.
\end{aligned} \tag{3.58}$$

Here we denote $\boldsymbol{\delta} = [\delta_1, \cdots, \delta_k]^\top$. Note that $\|\boldsymbol{u}_t \odot \boldsymbol{v}_t - \boldsymbol{\theta}^\star\| \le \|\boldsymbol{\theta}_{:k}^\star - \boldsymbol{S}_t\| + \|\boldsymbol{E}_t\|$. Moreover, based on our assumption, we have $\|\boldsymbol{E}_t\| \lesssim \sqrt{d}\alpha^{1-\Theta(\delta)}$. Finally, the balanced property implies that $|u_{i,t} - v_{i,t}| = \mathcal{O}\left(\alpha^{1-\Theta(\delta)}\right)$ (this will be proven later). Hence, we have

$$\begin{aligned}
\|\boldsymbol{\theta}_{:k}^\star - \boldsymbol{S}_{t+1}\| &\le \left\|(\boldsymbol{\theta}_{:k}^\star - \boldsymbol{S}_t) \odot \left(\boldsymbol{1} - \eta\frac{\boldsymbol{u}_t^2 + \boldsymbol{v}_t^2}{\|\boldsymbol{u}_t \odot \boldsymbol{v}_t - \boldsymbol{\theta}^\star\|} + \eta^2\frac{(\boldsymbol{\theta}_{:k}^\star - \boldsymbol{S}_t) \odot \boldsymbol{S}_t}{\|\boldsymbol{u}_t \odot \boldsymbol{v}_t - \boldsymbol{\theta}^\star\|^2}\right)\right\| + 4\eta\delta\|\boldsymbol{\theta}^\star\| \\
&\le \|\boldsymbol{\theta}_{:k}^\star - \boldsymbol{S}_t\|\left(1 - \frac{\eta}{2}\frac{\theta_k^\star}{\|\boldsymbol{\theta}_{:k}^\star - \boldsymbol{S}_t\| + \|\boldsymbol{E}_t\|}\right) + 4\eta\delta\|\boldsymbol{\theta}^\star\| \\
&\le \|\boldsymbol{\theta}_{:k}^\star - \boldsymbol{S}_t\| - 0.25\eta\theta_k^\star + 4\eta\delta\|\boldsymbol{\theta}^\star\| \\
&\le \|\boldsymbol{\theta}_{:k}^\star - \boldsymbol{S}_t\| - 0.1\eta\theta_k^\star.
\end{aligned} \tag{3.59}$$

Here, we used the fact that $\delta \lesssim \frac{1}{\kappa}$. On the other hand, we have

$$\begin{aligned}
\|\boldsymbol{S}_{t+1} - \boldsymbol{S}_t\| &\le \frac{\eta}{\|\boldsymbol{u}_t \odot \boldsymbol{v}_t - \boldsymbol{\theta}^\star\|}\left\|\left(\boldsymbol{u}_t^2 + \boldsymbol{v}_t^2\right) \odot (\boldsymbol{\theta}_{:k}^\star - \boldsymbol{S}_t)\right\| + 4\eta\delta\|\boldsymbol{\theta}^\star\| \\
&\overset{(a)}{\le} 2\eta\theta_1^\star + 4\eta\delta\|\boldsymbol{\theta}^\star\| \\
&\le 3\eta\theta_1^\star,
\end{aligned} \tag{3.60}$$

where in (a) we used Lemma 60. The above inequality indicates that the signal propagation in each step is upper bounded by $\mathcal{O}\left(\eta\theta_1^\star\right)$. Hence, we conclude that within $T_{k+1} = \mathcal{O}\left(\frac{\delta\|\boldsymbol{\theta}^\star\|}{\eta\theta_k^\star}\right)$ iterations, we have $\|\boldsymbol{\theta}_{:k}^\star - \boldsymbol{S}_t\| \lesssim \sqrt{d^2m}\alpha^{1-\Theta(\delta)} \vee \eta\theta_1^\star$. Since $\delta \lesssim \frac{1}{\kappa}$, the total iteration complexity is upper bounded by $T' = T_1 + \cdots T_{k+1} = \mathcal{O}\left(\frac{k^{3/2}}{\eta}\log\left(\frac{1}{\alpha}\right)\right)$.

Residual Dynamics

Now, we analyze the residual dynamics. Instead of analyzing the dynamics of $u_{i,t}v_{i,t}$, we analyze its surrogate $u_{i,t}^2 + v_{i,t}^2$. Based on equation 3.52, we can naturally bound it as follows

$$u_{i,t+1}^2 + v_{i,t+1}^2 \le u_{i,t}^2 + v_{i,t}^2 + 6\eta\delta u_{i,t}v_{i,t} \le (1 + 3\eta\delta)\left(u_{i,t}^2 + v_{i,t}^2\right). \tag{3.61}$$

Therefore, during the training process, we can bound the residual term as

$$u_{i,t}^2 + v_{i,t}^2 \lesssim \alpha \left(1+\eta\delta\right)^{\mathcal{O}\left(\frac{k^{3/2}}{\eta}\log\left(\frac{1}{\alpha}\right)\right)} \lesssim \alpha^{1-\mathcal{O}\left(k^{3/2}\delta\right)}. \tag{3.62}$$

Hence, we have

$$\|\boldsymbol{E}_t\| \leq \left(\sum_{i=k+1}^{d} u_{i,t}^2 + v_{i,t}^2\right)^{1/2} \lesssim \sqrt{d}\alpha^{1-\mathcal{O}\left(k^{3/2}\delta\right)}. \tag{3.63}$$

Therefore, we conclude that within $\bar{T} = \mathcal{O}\left(\frac{k^{3/2}}{\eta}\log\left(\frac{1}{\alpha}\right)\right)$ iterations, we have

$$\|\boldsymbol{u}_{\bar{T}} \odot \boldsymbol{v}_{\bar{T}} - \boldsymbol{\theta}^\star\| \leq \|\boldsymbol{\theta}^\star_{:k} - \boldsymbol{S}_{\bar{T}}\| + \|\boldsymbol{E}_{\bar{T}}\| \lesssim \sqrt{d^2 m}\alpha^{1-\mathcal{O}(\delta)} \vee \eta\theta_1^\star. \tag{3.64}$$

Therefore, with probability at least $1 - e^{-C\log^2(m)\log(d)\log(\|\boldsymbol{\theta}^\star\|/\alpha)}$, we have

$$\|\boldsymbol{u}_{\bar{T}} \odot \boldsymbol{v}_{\bar{T}} - \boldsymbol{\theta}^\star\| \lesssim \sqrt{d^2 m}\alpha^{1-\tilde{\Theta}\left(\frac{k^2}{\sqrt{m(1-p)^2}}\right)} \vee \eta\theta_1^\star. \tag{3.65}$$

Long Escape Time

Based on the above analysis, it can be seen that, after $\bar{T}$ iterations, the residual term is the dominant term, which may cause the algorithm to diverge, as captured by Figure 3.1. We now show that the residual term will not diverge within $T' = \sqrt{\frac{m(1-p)^2}{k}}\bar{T}$. To this goal, recall that for $\forall k+1 \leq i \leq d$, we have

$$u_{i,t+1}^2 + v_{i,t+1}^2 \leq \left(1+3\eta\delta\right)\left(u_{i,t}^2 + v_{i,t}^2\right). \tag{3.66}$$

Hence, for $t \leq \frac{1}{6\eta\delta}\log\left(\frac{1}{\alpha}\right)$, we have

$$u_{i,t}^2 + v_{i,t}^2 \lesssim \alpha\left(1+3\eta\delta\right)^{\frac{1}{6\eta\delta}\log\left(\frac{1}{\alpha}\right)} \leq \alpha e^{0.5\log\left(\frac{1}{\alpha}\right)} = \sqrt{\alpha}. \tag{3.67}$$

The proof is completed by noticing that $m = \tilde{\Omega}\left(\frac{k}{(1-p)^2\delta^2}\right)$.

Balanced Property

To prove the balanced property, we directly calculate the dynamic of the difference $u_{i,t} - v_{i,t}$. One can write

$$u_{i,t+1} - v_{i,t+1} = \left(u_{i,t} - v_{i,t}\right)\left(1 - \eta\frac{\theta_i^\star - u_{i,t}v_{i,t}}{\|\boldsymbol{u}_t \odot \boldsymbol{v}_t - \boldsymbol{\theta}^\star\|} - \eta\delta_i\right). \tag{3.68}$$

Since $\theta_i^\star - u_{i,t}v_{i,t} \geq 0$, we have

$$|u_{i,t+1} - v_{i,t+1}| \leq |u_{i,t} - v_{i,t}| (1 + \eta\delta), \tag{3.69}$$

for $0 \leq i \leq k$. We conclude that

$$|u_{i,t} - v_{i,t}| \leq \sqrt{\alpha} (1 + \eta\delta)^t \lesssim \alpha^{0.5 - \Theta(k^{3/2}\delta)}. \tag{3.70}$$

for $\forall t \lesssim \frac{k^{3/2}}{\eta} \log\left(\frac{1}{\alpha}\right)$. On the other hand, for $i \geq k$, we can write

$$|u_{i,t} - v_{i,t}| \leq \sqrt{u_{i,t}^2 + v_{i,t}^2} \lesssim \alpha^{0.5 - \Theta(k^{3/2}\delta)}. \tag{3.71}$$

The proof is completed by noticing that $m = \tilde{\Omega}\left(\frac{k}{(1-p)^2\delta^2}\right)$.

Convergence in Under-parameterized Regime

In this section, we study the under-parameterized regime, where we assume that $m = \tilde{\Omega}(\frac{d}{(1-p)^2})$. The analysis of the signal term is the same as the over-parameterized regime and hence omitted for brevity. Here we only analyze the residual dynamic. One can write

$$\begin{aligned} \boldsymbol{E}_{t+1} = \boldsymbol{E}_t \left(\mathbf{1} - \eta \frac{\boldsymbol{u}_{k+1:,t}^2 + \boldsymbol{v}_{k+1:,t}^2}{\|\boldsymbol{u}_t \odot \boldsymbol{v}_t - \boldsymbol{\theta}^\star\|} \right) + \eta \boldsymbol{\delta}_{k+1:} \left(\boldsymbol{u}_{k+1:,t}^2 + \boldsymbol{v}_{k+1:,t}^2 \right) \\ + \eta^2 \left(\frac{-\boldsymbol{E}_t}{\|\boldsymbol{u}_t \odot \boldsymbol{v}_t - \boldsymbol{\theta}^\star\|} + \boldsymbol{\delta}_{k+1:} \right)^2 \boldsymbol{E}_t. \end{aligned} \tag{3.72}$$

When the residual term becomes the dominant term, i.e., $\|\boldsymbol{\theta}_{:k}^\star - \boldsymbol{S}_t\| \leq \|\boldsymbol{E}_t\|$, we have the simplified dynamic

$$\begin{aligned} \|\boldsymbol{E}_{t+1}\| &\leq \left\| \boldsymbol{E}_t \left(\mathbf{1} - 0.5\eta \frac{\boldsymbol{u}_{k+1:,t}^2 + \boldsymbol{v}_{k+1:,t}^2}{\|\boldsymbol{u}_t \odot \boldsymbol{v}_t - \boldsymbol{\theta}^\star\|} \right) + \eta \boldsymbol{\delta}_{k+1:} \left(\boldsymbol{u}_{k+1:,t}^2 + \boldsymbol{v}_{k+1:,t}^2 \right) \right\| + 2\eta^2 \left(\|\boldsymbol{E}_t\|^3 + \delta^2 \|\boldsymbol{E}_t\| \right) \\ &\overset{(a)}{\leq} \left\| \boldsymbol{E}_t \left(\mathbf{1} - \frac{\eta \boldsymbol{E}_t}{\|\boldsymbol{E}_t\|} \right) \right\| + 4\eta\delta \|\boldsymbol{E}_t\| \\ &\leq \left(1 - \frac{\eta}{\sqrt{d}} \right) \|\boldsymbol{E}_t\| + 4\eta\delta \|\boldsymbol{E}_t\| \\ &\overset{(b)}{\leq} \left(1 - 0.5\frac{\eta}{\sqrt{d}} \right) \|\boldsymbol{E}_t\|. \end{aligned} \tag{3.73}$$

Here in (a) we used the balanced property, which results in $\boldsymbol{u}_{k+1:,t}^2+\boldsymbol{v}_{k+1:,t}^2 \asymp 2\boldsymbol{E}_t$. Moreover, (b) is implied by the fact that $m \gtrsim \frac{d}{(1-p)^2}$, which in turn implies $\delta \lesssim \frac{1}{\sqrt{d}}$. Hence, we have

$$\|\boldsymbol{u}_{t+1} \odot \boldsymbol{v}_{t+1} - \boldsymbol{\theta}^\star\| \leq \left(1 - \Omega\left(\frac{\eta}{\sqrt{d}}\right)\right) \|\boldsymbol{u}_t \odot \boldsymbol{v}_t - \boldsymbol{\theta}^\star\|. \tag{3.74}$$

Then, for $t \geq \bar{T}$, we have

$$\|\boldsymbol{u}_t \odot \boldsymbol{v}_t - \boldsymbol{\theta}^\star\| \lesssim \sqrt{d^2 m}\alpha^{1-\tilde{\Theta}\left(\frac{k^2}{\sqrt{m(1-p)^2}}\right)} \left(1 - \Omega\left(\frac{\eta}{\sqrt{d}}\right)\right)^{t-\bar{T}} \vee \eta\theta_1^\star. \tag{3.75}$$

3.D.2 Proof of Theorem 16

The proof of N-layer model is similar to that of 2-layer model. First, we study the signal dynamics, showing that the first k components $\prod w_{i,j}^{(t)}$ converge to $\theta_j^\star$ sequentially for $1 \leq j \leq k$. We also prove that the residual term remains small along the optimization trajectory. Based on the SubGM update rule, one can write

$$\boldsymbol{w}_i^{(t+1)} = \boldsymbol{w}_i^{(t)} + \eta \frac{\boldsymbol{\theta}^\star - \prod \boldsymbol{w}_j^{(t)}}{\left\|\boldsymbol{\theta}^\star - \prod \boldsymbol{w}_j^{(t)}\right\|} \prod_{j\neq i} \boldsymbol{w}_j^{(t)} + \eta\boldsymbol{\delta} \prod_{j\neq i} \boldsymbol{w}_j^{(t)}, \text{ and } \|\boldsymbol{\delta}\|_\infty \leq \delta, \forall i \in [N]. \tag{3.76}$$

Signal Dynamics

Stage 1: In this stage, we show that $\prod w_{i,1}$ will converge to $\theta_1^\star$ within $T_1 = \Theta\left(\frac{\|\boldsymbol{\theta}^\star\|}{N\eta\theta_1^\star}\alpha^{-\frac{N-2}{N}}\right)$ iterations. We first prove the upper bound. According to the update rule, we have

$$\begin{aligned}
\prod w_{i,1}^{(t+1)} &= \prod w_{i,1}^{(t)} + \sum_{i=1}^N \eta \left(\frac{\theta_1^\star - \prod w_{j,1}^{(t)}}{\left\|\boldsymbol{\theta}^\star - \prod \boldsymbol{w}_j^{(t)}\right\|} + \delta_1\right) \left(\prod_{j\neq i} w_{j,1}^{(t)}\right)^2 + \text{higher order terms of } \eta \\
&\geq \prod w_{i,1}^{(t)} + \sum_{i=1}^N \eta \left(\frac{\theta_1^\star - \prod w_{j,1}^{(t)}}{\left\|\boldsymbol{\theta}^\star - \prod \boldsymbol{w}_j^{(t)}\right\|} + \delta_1\right) \left(\prod_{j\neq i} w_{j,1}^{(t)}\right)^2 \\
&\geq \prod w_{i,1}^{(t)} + N\eta \left(\frac{\theta_1^\star - \prod w_{j,1}^{(t)}}{\left\|\boldsymbol{\theta}^\star - \prod \boldsymbol{w}_j^{(t)}\right\|} - \delta\right) \left(\prod w_{i,1}^{(t)}\right)^{\frac{2(N-1)}{N}}.
\end{aligned} \tag{3.77}$$

Here we use the fact that $\prod w_{i,1}^{(t)} \leq \theta_1^\star$ and $\eta \lesssim \frac{1}{N}\frac{1}{\kappa}^{\frac{N-2}{N}}$ so that we can drop the higher order terms. For brevity, we only show how we can drop the 2-th order term of η. The proof of

the higher order terms is similar. One can write

$$\begin{aligned}\eta^2 \sum_{i\neq j}\left(\prod_{k\neq i} w_{k,1}^{(t)} \prod_{k\neq j} w_{k,1}^{(t)} \prod_{k\neq i,j} w_{k,1}^{(t)}\right) &\overset{(a)}{\leq} \eta^2 \left(\theta_1^\star\right)^{\frac{N-2}{N}} \sum_{i\neq j}\left(\prod_{k\neq i} w_{k,1}^{(t)} \prod_{k\neq j} w_{k,1}^{(t)}\right) \\ &\overset{(b)}{\leq} (N-1)\eta \left(\theta_1^\star\right)^{\frac{N-2}{N}} \eta \sum_{i=1}^{m}\left(\prod_{j\neq i} w_{j,1}^{(t)}\right)^2 \\ &\overset{(c)}{\lesssim} \eta \sum_{i=1}^{m}\left(\prod_{j\neq i} w_{j,1}^{(t)}\right)^2.\end{aligned} \tag{3.78}$$

Here in (a) we use the balanced property and the fact that $\prod w_{i,1}^{(t)} \leq \theta_1^\star$. Moreover, in (b), we use the rearrangement Inequality. Finally in (c) we use the assumption that $\eta \lesssim N^{-1}\kappa^{-\frac{N-2}{N}}$. For simplicity, we denote $x_t = \prod w_{i,1}^{(t)}$. Note that $\left\|\boldsymbol{\theta}^\star - \prod \boldsymbol{w}_j^{(t)}\right\| \leq \|\boldsymbol{\theta}^\star\|$. Hence, the dynamic can be simplified as

$$x_{t+1} \geq x_t + \frac{N\eta}{\|\boldsymbol{\theta}^\star\|}\left(\theta_1^\star - \delta\|\boldsymbol{\theta}^\star\| - x_t\right) x_t^{\frac{2(N-1)}{N}}. \tag{3.79}$$

We next show that $x_T \geq \theta_1^\star - 2\delta\|\boldsymbol{\theta}^\star\|$ within $T_1 = \Theta\left(\frac{\|\boldsymbol{\theta}^\star\|}{N\eta\theta_1^\star}\alpha^{-\frac{N-2}{N}}\right)$ iterations provided that $x_0 = \Theta(\alpha)$. To this goal, we divide our analysis into two substages.

- $x_t \leq \frac{\theta_1^\star}{2}$: In this substage, we assume that $x_t \leq \frac{\theta_1^\star}{2}$. Hence, we can further simplify the dynamic as

$$x_{t+1} \geq x_t + 0.5N\eta\frac{\theta_1^\star}{\|\boldsymbol{\theta}^\star\|}x_t^{\frac{2(N-1)}{N}}. \tag{3.80}$$

  Without loss of generality, we assume that $x_0 = \alpha$. Now we divide the interval $[\alpha, 0.5\theta_1^\star]$ into a series of sub-intervals $\{\mathcal{I}_k\}$, where $\mathcal{I}_k = [2^k\alpha, 2^{k+1}\alpha)$. In each $\mathcal{I}_k$, the dynamic can be further simplified as

$$x_{t+1} \geq \left(1 + 0.5N\eta\frac{\theta_1^\star}{\|\boldsymbol{\theta}^\star\|}\left(2^k\alpha\right)^{\frac{N-2}{N}}\right)x_t. \tag{3.81}$$

  Therefore, the number of iterations that x_t spends in each interval $\mathcal{I}_k$ is $\mathcal{O}\left(\frac{\|\boldsymbol{\theta}^\star\|}{N\eta\theta_1^\star}\left(2^k\alpha\right)^{-\frac{N-2}{N}}\right)$. Hence, the total number of iterations is upper bounded by $\mathcal{O}\left(\sum_{k=0}^{\infty}\frac{\|\boldsymbol{\theta}^\star\|}{N\eta\theta_1^\star}\left(2^k\alpha\right)^{-\frac{N-2}{N}}\right) = \mathcal{O}\left(\frac{\|\boldsymbol{\theta}^\star\|}{N\eta\theta_1^\star}\alpha^{-\frac{N-2}{N}}\right)$.

- $x_t \geq \frac{\theta_1^\star}{2}$: In this substage, we define $y_t = \theta_1^\star - \delta\|\boldsymbol{\theta}^\star\| - x_t$. Via a similar trick, we can show that within additional $\mathcal{O}\left(\frac{\|\boldsymbol{\theta}^\star\|}{N\eta}(\theta_1^\star)^{-\frac{2N-2}{N}}\right)$ iterations, we have $x_t \geq \theta_1^\star - 2\delta\|\boldsymbol{\theta}^\star\|$. Overall, after $T_1 = \Theta\left(\frac{\|\boldsymbol{\theta}^\star\|}{N\eta\theta_1^\star}\alpha^{-\frac{N-2}{N}}\right)$ iterations, we have $\theta_1^\star - 2\delta\|\boldsymbol{\theta}^\star\| \leq \prod w_{i,1}^{(T_1)} \leq \theta_1^\star$.

Stages 2 to k: Similarly, for component $\prod w_{j,i}^{(t)}$, it takes $\mathcal{O}\left(\frac{\left\|\boldsymbol{\theta}^\star_{-(i-1)}\right\|}{N\eta\theta_i^\star}\alpha^{-\frac{N-2}{N}}\right)$ iterations to attain $\theta_i^\star - 2\delta\left\|\boldsymbol{\theta}^\star\right\|$. Overall, Stages 2 to k take $\Theta\left(\frac{k^{\frac{3}{2}}}{N\eta}\alpha^{-\frac{N-2}{N}}\right)$ iterations to terminate.

Stage $k+1$: In this stage, we take $\boldsymbol{S}_t = \prod \boldsymbol{w}_{j,:k}^{(t)}$. Hence, we have

$$\begin{aligned}
\left\|\boldsymbol{\theta}_{:k}^\star - \boldsymbol{S}_{t+1}\right\| &\leq \left\|(\boldsymbol{\theta}_{:k}^\star - \boldsymbol{S}_t)\left(1 - \eta\frac{\sum_{i=1}^N\left(\prod_{j\neq i} w_{j,1}^{(t)}\right)^2}{\left\|\boldsymbol{\theta}^\star - \prod \boldsymbol{w}_j^{(t)}\right\|}\right)\right\| + 4N\eta\delta\sqrt{k}(\theta_1^\star)^{\frac{2(N-1)}{N}} \\
&\leq \left\|\boldsymbol{\theta}_{:k}^\star - \boldsymbol{S}_t\right\|\left(1 - N\eta\frac{(\theta_k^\star)^{\frac{2(N-1)}{N}}}{\left\|\boldsymbol{\theta}_{:k}^\star - \boldsymbol{S}_t\right\| + \left\|\boldsymbol{E}_t\right\|}\right) + 4N\eta\delta\sqrt{k}(\theta_1^\star)^{\frac{2(N-1)}{N}} \qquad (3.82) \\
&\leq \left\|\boldsymbol{\theta}_{:k}^\star - \boldsymbol{S}_t\right\| - 0.5N\eta(\theta_k^\star)^{\frac{2(N-1)}{N}} + 4N\eta\delta\sqrt{k}(\theta_1^\star)^{\frac{2(N-1)}{N}} \\
&\leq \left\|\boldsymbol{\theta}_{:k}^\star - \boldsymbol{S}_t\right\| - 0.1N\eta(\theta_k^\star)^{\frac{2(N-1)}{N}}.
\end{aligned}$$

Here we used the fact that $\left\|\boldsymbol{\theta}^\star - \prod \boldsymbol{w}_j^{(t)}\right\| \leq \left\|\boldsymbol{\theta}_{:k}^\star - \boldsymbol{S}_t\right\| + \left\|\boldsymbol{E}_t\right\| \leq 2\left\|\boldsymbol{\theta}_{:k}^\star - \boldsymbol{S}_t\right\|$, and the assumption that $\delta \lesssim \frac{1}{N}\frac{1}{\kappa}^{\frac{2N-2}{N}}$. On the other hand, we have

$$\begin{aligned}
\left\|\boldsymbol{S}_{t+1} - \boldsymbol{S}_t\right\| &\leq \left\|\sum_{i=1}^N \eta\left(\frac{\boldsymbol{\theta}_{:k}^\star - \boldsymbol{S}_t}{\left\|\boldsymbol{\theta}^\star - \prod \boldsymbol{w}_j^{(t)}\right\|} + \boldsymbol{\delta}_{i,:k}\right)\left(\prod_{j\neq i} \boldsymbol{w}_{j,:k}^{(t)}\right)^2\right\| \qquad (3.83) \\
&\leq 2N\eta\sqrt{k}(\theta_1^\star)^{\frac{2(N-1)}{N}}.
\end{aligned}$$

Hence, we conclude that within $\mathcal{O}\left(\frac{\sqrt{k}\delta}{N\eta}\frac{1}{\kappa}^{\frac{2(N-1)}{N}}\right)$ iterations, we have $\left\|\boldsymbol{\theta}_{:k}^\star - \boldsymbol{S}_t\right\| \lesssim \sqrt{d^2m}\alpha \vee N\eta(\theta_1^\star)^{\frac{2(N-1)}{N}}$. Overall, the total iteration complexity is bounded by $\mathcal{O}\left(\frac{k^{\frac{3}{2}}}{N\eta}\alpha^{-\frac{N-2}{N}}\right)$.

Residual Dynamics

Similar to the 2-layer model, here we study the surrogate of the residual term $\sum_{i=1}^{N}\left(w_{i,l}^{(t)}\right)^2$ for $l \geq k+1$. To this goal, we first notice that

$$\begin{aligned}\sum_{i=1}^{N}\left(w_{i,l}^{(t+1)}\right)^2 &= \sum_{i=1}^{N}\left(w_{i,l}^{(t)}\right)^2 + 2N\eta\left(\frac{-\prod w_{j,l}^{(t)}}{\left\|\boldsymbol{\theta}^\star - \prod \boldsymbol{w}_j^{(t)}\right\|} + \delta_{i,l}\right)\prod w_{j,l}^{(t)} \\ &\quad + \eta^2\sum_{i=1}^{N}\left(\frac{-\prod w_{j,l}^{(t)}}{\left\|\boldsymbol{\theta}^\star - \prod \boldsymbol{w}_j^{(t)}\right\|} + \delta_{i,l}\right)^2\left(\prod_{j\neq i} w_{j,l}^{(t)}\right)^2 \\ &\leq \sum_{i=1}^{N}\left(w_{i,l}^{(t)}\right)^2 + 4N\eta\delta\prod w_{j,l}^{(t)} \\ &\leq \sum_{i=1}^{N}\left(w_{i,l}^{(t)}\right)^2 + 4N\eta\delta\left(\frac{\sum_{i=1}^{N}\left(w_{i,l}^{(t)}\right)^2}{N}\right)^{\frac{N}{2}}.\end{aligned} \tag{3.84}$$

Hence, once we set $z_t = \sum_{i=1}^{N}\left(w_{i,l}^{(t)}\right)^2$, we have the following simplified dynamic

$$z_{t+1} \leq z_t + 4N\eta\delta\left(\frac{z_t}{N}\right)^{\frac{N}{2}}, \tag{3.85}$$

with $z_0 = \Theta\left(N\alpha^{\frac{2}{N}}\right)$. We claim that within $\mathcal{O}\left(\frac{1}{N\eta\alpha}\right)$ iterations, we still have $z_t = \Theta\left(N\alpha^{\frac{2}{N}}\right)$. To show this, we suppose without loss of generality that $z_0 = N\alpha^{\frac{2}{N}}$, and define T as the first time that $z_T \geq 2N\alpha^{\frac{2}{N}}$. For any $0 \leq t \leq T-1$, we have

$$z_{t+1} \leq z_t + 4N\eta\delta 2^{\frac{N}{2}}\alpha. \tag{3.86}$$

We conclude that

$$T \geq \frac{N\alpha^{\frac{2}{N}}}{4N\eta\delta 2^{\frac{N}{2}}\alpha} = \frac{1}{4\delta 2^{\frac{N}{2}}}\frac{1}{\eta}\alpha^{-\frac{N-2}{N}} \gtrsim \frac{1}{N\eta}\alpha^{-\frac{N-2}{N}}. \tag{3.87}$$

Therefore, via a basic inequality, we have

$$\prod w_{i,l}^{(t)} \leq \left(\frac{\sum_{i=1}^{N}\left(w_{i,l}^{(t)}\right)^2}{N}\right)^{\frac{N}{2}} \lesssim \alpha. \tag{3.88}$$

Combining the analysis of both signal and residual terms, we conclude that within $\Theta\left(\frac{1}{N\eta}\alpha^{-\frac{N-2}{N}}\right)$ iterations, we have

$$\left\|\prod \boldsymbol{w}_i^{(t)} - \boldsymbol{\theta}^\star\right\| \lesssim \sqrt{d^2 m}\alpha \vee (\eta\theta_1^\star)^{\frac{2(N-1)}{N}}. \tag{3.89}$$

Long Time Guarantee

Similar to the proof of the 2-layer model, one can show that the residual term becomes the dominant term in the generalization error, and it stays in the order of α within $\Omega\left(\frac{1}{N\eta\delta}\alpha^{-\frac{N-2}{N}}\right)$ iterations. The details are omitted for brevity.

Balanced Property

To prove the balanced property, we first study the dynamic of $w_{i,l}^{(t)} - w_{j,l}^{(t)}$, $\forall l \in [k], i, j \in [N]$. To this goal, we have

$$w_{i,l}^{(t+1)} - w_{j,l}^{(t+1)} = \left(w_{i,l}^{(t)} - w_{j,l}^{(t)}\right)\left(1 - \eta\left(\frac{\theta_l^\star - \prod w_{j,l}^{(t)}}{\left\|\boldsymbol{\theta}^\star - \prod \boldsymbol{w}_j^{(t)}\right\|} + \delta_l\right)\prod_{f\neq i,j} w_{f,l}^{(t)}\right), \tag{3.90}$$

which in turn implies

$$\left|w_{i,l}^{(t+1)} - w_{j,l}^{(t+1)}\right| \leq \left|w_{i,l}^{(t)} - w_{j,l}^{(t)}\right|\left(1 - \eta\left(\frac{\theta_l^\star - \prod w_{j,l}^{(t)}}{\left\|\boldsymbol{\theta}^\star - \prod \boldsymbol{w}_j^{(t)}\right\|} + \delta_l\right)\prod_{f\neq i,j} w_{f,l}^{(t)}\right). \tag{3.91}$$

If $\prod w_{j,l}^{(t)} \leq \theta_l^\star - \delta\left\|\boldsymbol{\theta}^\star\right\|$, the above inequality can be simplified as

$$\left|w_{i,l}^{(t)} - w_{j,l}^{(t)}\right| \leq \left|w_{i,l}^{(0)} - w_{j,l}^{(0)}\right| \lesssim \alpha^{\frac{1}{N}}, \forall i, j \in [N], l \in [k]. \tag{3.92}$$

Once $\prod w_{j,l}^{(t)} \geq \theta_l^\star - \delta\left\|\boldsymbol{\theta}^\star\right\|$, we immediately have $w_{j,l}^{(t)} = \sqrt[N]{\theta_l^\star} \pm \mathcal{O}(\sqrt[N]{\alpha})$. Then, we show that $w_{j,l}^{(t)}$ will stay close to $\sqrt[N]{\theta_l^\star}$. To this goal, we first observe that $\left(w_{i,l}^{(t+1)} - w_{i,l}^{(t)}\right) w_{i,l}^{(t)} \equiv \left(w_{j,l}^{(t+1)} - w_{j,l}^{(t)}\right) w_{j,l}^{(t)}$, $\forall i, j \in [N]$, which indicates that $w_{i,l}^{(t)}$ increases or decreases simultaneously. Hence, we conclude that $\left|w_{i,l}^{(t)} - w_{j,l}^{(t)}\right| \lesssim \delta\sqrt[N]{\theta_l^\star}$.

For the residual term, we can derive a tighter bound. First, we have

$$\begin{aligned}\left(w_{i,l}^{(t+1)}\right)^2 = \left(w_{i,l}^{(t)}\right)^2 &+ 2\eta\left(\frac{-\prod w_{j,l}^{(t)}}{\left\|\boldsymbol{\theta}^\star - \prod \boldsymbol{w}_j^{(t)}\right\|} + \delta_l\right)\prod w_{j,l}^{(t)} \\ &+ \eta^2\left(\frac{-\prod w_{j,l}^{(t)}}{\left\|\boldsymbol{\theta}^\star - \prod \boldsymbol{w}_j^{(t)}\right\|} + \delta_l\right)^2\left(\prod_{j\neq i} w_{j,l}^{(t)}\right)^2.\end{aligned} \tag{3.93}$$

Since we have already shown that $\prod w_{i,l}^{(t)} \lesssim \alpha$, we further have

$$\left(w_{i,l}^{(t+1)}\right)^2 \leq \left(w_{i,l}^{(t)}\right)^2 + 4\eta\delta\alpha. \tag{3.94}$$

Therefore, one can write $\left(w_{i,l}^{(t)}\right)^2 \lesssim \left(w_{i,l}^{(0)}\right)^2 + 4\eta\delta\alpha\frac{1}{\eta}\alpha^{-\frac{N-2}{N}} \lesssim \alpha^{\frac{2}{N}}$, which in turn implies $\left|w_{i,l}^{(t)} - w_{j,l}^{(t)}\right| \leq \left|w_{i,l}^{(t)}\right| + \left|w_{j,l}^{(t)}\right| \lesssim \alpha^{1/N}$.

Convergence in Under-parameterized Regime

Similar to the 2-layer model, we consider the dynamic of $\boldsymbol{E}_t = \prod \boldsymbol{w}_{i,k+1:}^{(t)}$, which is characterized as follows

$$\begin{aligned}\|\boldsymbol{E}_{t+1}\| &\leq \left\|\boldsymbol{E}_t + \sum_{i=1}^N \eta\left(\frac{-\boldsymbol{E}_t}{\left\|\boldsymbol{\theta}^\star - \prod \boldsymbol{w}_j^{(t)}\right\|} + \boldsymbol{\delta}_{k+1:}\right)\left(\prod_{j\neq i} \boldsymbol{w}_{j,k+1:}^{(t)}\right)^2\right\| \\ &\leq \left\|\boldsymbol{E}_t + N\eta\left(\frac{-\boldsymbol{E}_t}{\left\|\boldsymbol{\theta}^\star - \prod \boldsymbol{w}_j^{(t)}\right\|} + \delta\mathbf{1}\right)\boldsymbol{E}_t^{\frac{2(N-1)}{N}}\right\| \\ &\leq \left\|\boldsymbol{E}_t + N\eta\left(\frac{-\boldsymbol{E}_t}{\|\boldsymbol{E}_t\|} + \delta\mathbf{1}\right)\boldsymbol{E}_t^{\frac{2(N-1)}{N}}\right\| \\ &\leq \|\boldsymbol{E}_t\| - N\eta d^{-\frac{N-1}{N}}\|\boldsymbol{E}_t\|^{\frac{2N-2}{N}} + N\eta\delta\left\|\boldsymbol{E}_t^{\frac{2(N-1)}{N}}\right\| \\ &\leq \|\boldsymbol{E}_t\| - N\eta d^{-\frac{N-1}{N}}\|\boldsymbol{E}_t\|^{\frac{2N-2}{N}} + N\eta\delta\|\boldsymbol{E}_t\|^{\frac{2N-2}{N}} \\ &\leq \|\boldsymbol{E}_t\| - N\eta d^{-\frac{N-1}{N}}\|\boldsymbol{E}_t\|^{\frac{2N-2}{N}}.\end{aligned} \tag{3.95}$$

The last inequality comes from the fact that $\delta \lesssim d^{-\frac{N-1}{N}}$ since we assume $m \gtrsim \frac{d^{\frac{2N-2}{N}}}{(1-p)^2}$. Hence, we have

$$\|\boldsymbol{E}_t\| \lesssim \left(\frac{1}{N\eta d^{-(N-1)/N}(t-\bar{T}) + 1/\|\boldsymbol{E}_{\bar{T}}\|}\right)^{\frac{N}{N-2}}. \tag{3.96}$$

Since the residual term is the dominant term in the generalization error, we have

$$\left\|\prod \boldsymbol{w}_i^{(t)} - \boldsymbol{\theta}^\star\right\| \lesssim \left(\frac{\left\|\prod \boldsymbol{w}_i^{(\bar{T})} - \boldsymbol{\theta}^\star\right\|}{\left\|\prod \boldsymbol{w}_i^{(\bar{T})} - \boldsymbol{\theta}^\star\right\| N\eta d^{-(N-1)/N}(t-\bar{T}) + 1}\right)^{\frac{N}{N-2}}, \tag{3.97}$$

which completes the proof.

3.E Proof of Proposition 11

First, we provide an upper bound of the covering number for the (k,ϑ)-approximate sparse unit ball. We defer a preliminary discussion on covering number to Appendix 3.H .

Lemma 57. *Let $\mathcal{T}_{k,\vartheta} := \{\boldsymbol{u} \in \mathbb{R}^d : \boldsymbol{u}$ is (k,ϑ)-approximate sparse, $\|\boldsymbol{u}\| \le 1\}$. Then its covering number $N(\mathcal{T}_{k,\vartheta}, \epsilon, \|\cdot\|)$ is upper bounded by*

$$N(\mathcal{T}_{k,\vartheta}, \epsilon, \|\cdot\|) \le \left(\frac{ed}{k}\right)^k \left(1 + \frac{4}{\epsilon}\right)^k, \tag{3.98}$$

provided that $\epsilon \ge \vartheta$.

The next lemma will play a crucial role in proving Proposition 11.

Lemma 58. *Suppose $\boldsymbol{x} \in \mathbb{R}^d$ is a standard Gaussian vector, i.e., $x_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, and the noise ϵ satisfies Assumption 2, then we have*

$$\varphi(\boldsymbol{u}) = \frac{\mathbb{E}\left[\operatorname{sign}\left(\langle \boldsymbol{x}, \boldsymbol{u}\rangle + \epsilon\right)\langle \boldsymbol{x}, \boldsymbol{v}\rangle\right]}{\left\langle \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|}, \boldsymbol{v}\right\rangle} = \sqrt{\frac{2}{\pi}}(1-p) + \sqrt{\frac{2}{\pi}} p \mathbb{E}\left[e^{-\epsilon^2/\left(2\|\boldsymbol{u}\|^2\right)}\right].$$

The proof of this lemma can be found in Appendix 3.G.1. Now, we are ready to prove Proposition 11. Our goal is to show that for arbitrary $u \in \mathcal{A}$, the following inequality holds

$$\left\|\frac{1}{m}\sum_{i=1}^m \operatorname{sign}\left(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle + \epsilon_i\right)\boldsymbol{x}_i - \varphi(\boldsymbol{u})\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|}\right\|_\infty \le \delta \tag{3.99}$$

with probability at least $1 - Ce^{-cm\delta^2}$ provided that $m \gtrsim \frac{k\log(d)\log(R)\log(\frac{1}{\vartheta})}{(1-p)^2}$. Here we define $\mathcal{A} := \{u : r \le \|\boldsymbol{u}\| \le R, u$ is (k,ϑ)-approximate sparse$\}$, where $r \gtrsim \sqrt{dm/k}\vartheta\log(1/\vartheta)$. Moreover, we define $\mathcal{B} := \{u : r \le \|\boldsymbol{u}\| \le R, \|\boldsymbol{u}\|_0 \le k\}$ and $\mathcal{C} := \{(\boldsymbol{u}, \boldsymbol{u}') : u \in \mathcal{A}, \boldsymbol{v} \in \mathcal{B}_\zeta, \|\boldsymbol{u} - \boldsymbol{u}'\| \le \zeta\}$. Here $\mathcal{B}_\zeta$ is the ζ-net of $\mathcal{B}$ with $\zeta \gtrsim r$. Finally, we define $\mathcal{D} := \{\pm\mathbf{e}_j\}_{j\in[d]}$,

where $\mathbf{e}_j$ forms the standard basis of $\mathbb{R}^d$. Based on these definitions, we have

$$\begin{aligned}
&\sup_{\boldsymbol{u}\in\mathcal{A}}\left\|\frac{1}{m}\sum_{i=1}^{m}\operatorname{sign}\left(\langle\boldsymbol{x}_i,\boldsymbol{u}\rangle+\epsilon_i\right)\boldsymbol{x}_i-\varphi(\boldsymbol{u})\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|}\right\|_\infty \\
&=\sup_{\boldsymbol{u}\in\mathcal{A},\boldsymbol{v}\in\mathcal{D}}\frac{1}{m}\sum_{i=1}^{m}\operatorname{sign}\left(\langle\boldsymbol{x}_i,\boldsymbol{u}\rangle+\epsilon_i\right)\langle\boldsymbol{x}_i,\boldsymbol{v}\rangle-\frac{\varphi(\boldsymbol{u})}{\|\boldsymbol{u}\|}\langle\boldsymbol{u},\boldsymbol{v}\rangle \\
&=\sup_{v\in\mathcal{D}}\left\{\sup_{\boldsymbol{u}\in\mathcal{A}}\frac{1}{m}\sum_{i=1}^{m}\operatorname{sign}\left(\langle\boldsymbol{x}_i,\boldsymbol{u}\rangle+\epsilon_i\right)\langle\boldsymbol{x}_i,\boldsymbol{v}\rangle-\frac{\varphi(\boldsymbol{u})}{\|\boldsymbol{u}\|}\langle\boldsymbol{u},\boldsymbol{v}\rangle\right\}.
\end{aligned}\tag{3.100}$$

We then show that for each element $y\in\mathcal{D}$, $\sup_{\boldsymbol{u}\in\mathcal{A}}\frac{1}{m}\sum_{i=1}^{m}\operatorname{sign}\left(\langle\boldsymbol{x}_i,\boldsymbol{u}\rangle+\epsilon_i\right)\langle\boldsymbol{x}_i,\boldsymbol{y}\rangle-\varphi(\boldsymbol{u})\frac{\langle\boldsymbol{u},\boldsymbol{y}\rangle}{\|\boldsymbol{u}\|}$, $j\in[d]$ is $\mathcal{O}\left(\frac{1}{m}\right)$-sub-Gaussian random variable. To see this, note that

$$\begin{aligned}
\left\|\operatorname{sign}\left(\langle\boldsymbol{x}_i,\boldsymbol{u}\rangle+\epsilon_i\right)x_{i,j}-\varphi(\boldsymbol{u})\frac{u_j}{\|\boldsymbol{u}\|}\right\|_{\psi_2}&\le\left\|\operatorname{sign}\left(\langle\boldsymbol{x}_i,\boldsymbol{u}\rangle+\epsilon_i\right)x_{i,j}\right\|_{\psi_2}+\sqrt{\frac{2}{\pi}} \\
&\le\left\|x_{i,j}\right\|_{\psi_2}+\sqrt{\frac{2}{\pi}}=\mathcal{O}(1).
\end{aligned}\tag{3.101}$$

Here we use the property of sub-Gaussian norm. This implies that $\frac{1}{m}\sum_{i=1}^{m}\operatorname{sign}\left(\langle\boldsymbol{x}_i,\boldsymbol{u}\rangle+\epsilon_i\right)x_{i,j}-\varphi(\boldsymbol{u})\frac{u_j}{\|\boldsymbol{u}\|}$ is $\mathcal{O}\left(\frac{1}{m}\right)$-sub-Gaussian random variable, since it is the sample average of $\operatorname{sign}\left(\langle\boldsymbol{x}_i,\boldsymbol{u}\rangle+\epsilon_i\right)x_{i,j}-\varphi(\boldsymbol{u})\frac{u_j}{\|\boldsymbol{u}\|}$.

Hence, via maximal inequality, we have that for $\forall t>0$,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\boldsymbol{u}\in\mathcal{A}}\left\|\frac{1}{m}\sum_{i=1}^{m}\operatorname{sign}\left(\langle\boldsymbol{x}_i,\boldsymbol{u}\rangle+\epsilon_i\right)\boldsymbol{x}_i-\varphi(\boldsymbol{u})\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|}\right\|_\infty\right. \\
&\qquad\left.\ge\sup_{y\in\mathcal{D}}\mathbb{E}\left[\sup_{\boldsymbol{u}\in\mathcal{A}}\frac{1}{m}\sum_{i=1}^{m}\operatorname{sign}\left(\langle\boldsymbol{x}_i,\boldsymbol{u}\rangle+\epsilon_i\right)\langle\boldsymbol{x}_i,\boldsymbol{y}\rangle-\varphi(\boldsymbol{u})\frac{\langle\boldsymbol{u},\boldsymbol{y}\rangle}{\|\boldsymbol{u}\|}\right]+t\right) \\
&\le 2de^{-cmt^2}.
\end{aligned}\tag{3.102}$$

Hence, it suffices to study $\mathbb{E}\left[\sup_{\boldsymbol{u}\in\mathcal{A}}\frac{1}{m}\sum_{i=1}^{m}\operatorname{sign}\left(\langle\boldsymbol{x}_i,\boldsymbol{u}\rangle+\epsilon_i\right)x_{i,1}-\varphi(\boldsymbol{u})\frac{u_1}{\|\boldsymbol{u}\|}\right]$. To this goal, we decompose it into two terms via triangle inequality.

$$\mathbb{E}\left[\sup_{\boldsymbol{u}\in\mathcal{A}}\frac{1}{m}\sum_{i=1}^{m}\operatorname{sign}\left(\langle\boldsymbol{x}_i,\boldsymbol{u}\rangle+\epsilon_i\right)x_{i,1}-\varphi(\boldsymbol{u})\frac{u_1}{\|\boldsymbol{u}\|}\right]\le(\mathrm{A})+(\mathrm{B}),\tag{3.103}$$

where

$$(\mathrm{A}):=\mathbb{E}\left[\sup_{\boldsymbol{u}\in\mathcal{B}_\zeta}\frac{1}{m}\sum_{i=1}^{m}\operatorname{sign}\left(\langle\boldsymbol{x}_i,\boldsymbol{u}\rangle+\epsilon_i\right)x_{i,1}-\varphi(\boldsymbol{u})\frac{u_1}{\|\boldsymbol{u}\|}\right],\tag{3.104}$$

and

(B)

$$:= \mathbb{E}\left[\sup_{(\boldsymbol{u},\boldsymbol{u}')\in\mathcal{C}} \frac{1}{m}\sum_{i=1}^{m}\left(\operatorname{sign}\left(\langle \boldsymbol{x}_i,\boldsymbol{u}\rangle+\epsilon_i\right)-\operatorname{sign}\left(\langle \boldsymbol{x}_i,\boldsymbol{u}'\rangle+\epsilon_i\right)\right)x_{i,1}-\varphi(\boldsymbol{u})\frac{u_1}{\|\boldsymbol{u}\|}+\varphi(\boldsymbol{u}')\frac{u_1'}{\|\boldsymbol{u}'\|}\right]. \tag{3.105}$$

We first control (A). To this goal, we apply the union bound. Note that $\frac{1}{m}\sum_{i=1}^{m}\operatorname{sign}\left(\langle \boldsymbol{x}_i,\boldsymbol{u}\rangle+\epsilon_i\right)x_{i,1}-\varphi(\boldsymbol{u})\frac{u_1}{\|\boldsymbol{u}\|}$ is $\mathcal{O}(\frac{1}{m})$-sub-Gaussian and $|\mathcal{B}_\zeta|\le\left(\frac{R}{\zeta}\right)^{Ck\log(d)}$. We then have

$$(\mathrm{A})\lesssim\sqrt{\frac{k\log(d)\log\left(\frac{R}{\zeta}\right)}{m}}. \tag{3.106}$$

Now we control (B). Via triangle inequality, we first obtain

$$\begin{aligned}(\mathrm{B})\le&\underbrace{\mathbb{E}\left[\sup_{(\boldsymbol{u},\boldsymbol{u}')\in\mathcal{C}}\frac{1}{m}\sum_{i=1}^{m}\left(\operatorname{sign}\left(\langle \boldsymbol{x}_i,\boldsymbol{u}\rangle+\epsilon_i\right)-\operatorname{sign}\left(\langle \boldsymbol{x}_i,\boldsymbol{u}'\rangle+\epsilon_i\right)\right)x_{i,1}\right]}_{(\mathrm{B}_1)}\\&+\underbrace{\sup_{(\boldsymbol{u},\boldsymbol{u}')\in\mathcal{C}}\left\{-\varphi(\boldsymbol{u})\frac{u_1}{\|\boldsymbol{u}\|}+\varphi(\boldsymbol{u}')\frac{u_1'}{\|\boldsymbol{u}'\|}\right\}}_{(\mathrm{B}_2)}.\end{aligned} \tag{3.107}$$

For the first part, applying Hölder's inequality leads to

$$\begin{aligned}(\mathrm{B}_1)&\le\mathbb{E}\left[\sup_{(\boldsymbol{u},\boldsymbol{u}')\in\mathcal{C}}\left(\frac{1}{m}\sum_{i=1}^{m}\left|\operatorname{sign}\left(\langle \boldsymbol{x}_i,\boldsymbol{u}\rangle+\epsilon_i\right)-\operatorname{sign}\left(\langle \boldsymbol{x}_i,\boldsymbol{u}'\rangle+\epsilon_i\right)\right|\right)\max_{1\le i\le m}|x_{i,1}|\right]\\&\le\mathbb{E}\left[\sup_{(\boldsymbol{u},\boldsymbol{u}')\in\mathcal{C}}\left(\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}\left(\left|\langle \boldsymbol{x}_i,u-\boldsymbol{u}'\rangle\right|\ge\left|\langle \boldsymbol{x}_i,\boldsymbol{u}\rangle+\epsilon_i\right|\right)\right)\max_{1\le i\le m}|x_{i,1}|\right]\\&\le\underbrace{\mathbb{E}\left[\sup_{\|\boldsymbol{\Delta u}\|\le\zeta}\left(\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}\left(\left|\langle \boldsymbol{x}_i,\boldsymbol{\Delta u}\rangle\right|\ge t\right)\right)\max_{1\le i\le m}|x_{i,1}|\right]}_{(\mathrm{B}_3)}\\&+\underbrace{\mathbb{E}\left[\sup_{\boldsymbol{u}\in\mathcal{B}_\zeta}\left(\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}\left(\left|\langle \boldsymbol{x}_i,\boldsymbol{u}\rangle+\epsilon_i\right|\le t\right)\right)\max_{1\le i\le m}|x_{i,1}|\right]}_{(\mathrm{B}_4)},\end{aligned} \tag{3.108}$$

where $t>0$ is a constant to be determined later. Here, we used the fact that $\mathbb{1}\left(\left|\langle \boldsymbol{x}_i,u-\boldsymbol{u}'\rangle\right|\ge\left|\langle \boldsymbol{x}_i,\boldsymbol{u}\rangle+\epsilon_i\right|\right)\le\mathbb{1}\left(\left|\langle \boldsymbol{x}_i,\boldsymbol{\Delta u}\rangle\right|\ge t\right)+\mathbb{1}\left(\left|\langle \boldsymbol{x}_i,\boldsymbol{u}\rangle+\epsilon_i\right|\le t\right)$ in the last in-

equality. We first bound (B_3)

$$\begin{aligned}(\mathrm{B}_3) &\leq \mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}\left(\zeta\left\|\boldsymbol{x}_i\right\|\geq t\right)\right)\max_{1\leq i\leq m}\left|x_{i,1}\right|\right]\\ &\leq \mathbb{E}\left[\mathbb{1}\left(\zeta\left\|\boldsymbol{x}_i\right\|\geq t\right)\right]\mathbb{E}\left[\max_{j\neq i}\left|x_{j,1}\right|\right]+\mathbb{E}\left[\mathbb{1}\left(\zeta\left\|\boldsymbol{x}_i\right\|\geq t\right)\left|x_{i,1}\right|\right]\\ &\lesssim e^{-C\frac{t^2}{\zeta^2}}\sqrt{\log\left(m\right)}+\mathbb{E}\left[\mathbb{1}\left(\zeta\left\|\boldsymbol{x}_i\right\|\geq t\right)\left|x_{i,1}\right|\right],\end{aligned} \tag{3.109}$$

provided that $\frac{t}{\zeta}\gtrsim\sqrt{d}$. Applying Cauchy-Schwarz inequality, we have

$$\mathbb{E}\left[\mathbb{1}\left(\zeta\left\|\boldsymbol{x}_i\right\|\geq t\right)\left|x_{i,1}\right|\right]\leq\sqrt{\mathbb{P}\left(\zeta\left\|\boldsymbol{x}_i\right\|\geq t\right)}\sqrt{\mathbb{E}\left[x_{i,1}^2\right]}\leq e^{-C\frac{t^2}{\zeta^2}}. \tag{3.110}$$

Hence, we conclude that $(\mathrm{B}_3)\lesssim e^{-C\frac{t^2}{\zeta^2}}\sqrt{\log\left(m\right)}$. Next we control (B_4). Note that $\max_i\left|x_{i,1}\right|$ is $\mathcal{O}\left(\log(m)\right)$-sub-Gaussian. Via union bound, we have

$$(\mathrm{B}_4)\leq\sup_{\boldsymbol{u}\in\mathcal{B}}\mathbb{E}\left[\mathbb{1}\left(\left|\left\langle\boldsymbol{x}_i,\boldsymbol{u}\right\rangle+\epsilon_i\right|\leq t\right)\max_{1\leq i\leq m}\left|x_{i,1}\right|\right]+C\sqrt{\frac{k\log(m)\log(d)\log(\frac{R}{\zeta})}{m}}. \tag{3.111}$$

For the first part, applying the similar decomposition method, we have

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}\left(\left|\left\langle\boldsymbol{x}_i,\boldsymbol{u}\right\rangle+\epsilon_i\right|\leq t\right)\max_{1\leq i\leq m}\left|x_{i,1}\right|\right]&\leq\mathbb{E}\left[\mathbb{1}\left(\left|\left\langle\boldsymbol{x}_i,\boldsymbol{u}\right\rangle+\epsilon_i\right|\leq t\right)\max_{j\neq i}\left|x_{i,1}\right|\right]\\ &\quad+\mathbb{E}\left[\mathbb{1}\left(\left|\left\langle\boldsymbol{x}_i,\boldsymbol{u}\right\rangle+\epsilon_i\right|\leq t\right)\left|x_{i,1}\right|\right]\\ &\lesssim\sqrt{\log(m)}\frac{t}{r}.\end{aligned} \tag{3.112}$$

Hence, we conclude that $(\mathrm{B}_4)\lesssim\sqrt{\log(m)}\frac{t}{r}+\sqrt{\frac{k\log(m)\log(d)\log(\frac{R}{\zeta})}{m}}$. For (B_2), we first have

$$\begin{aligned}\left|-\varphi(\boldsymbol{u})\frac{u_1}{\left\|\boldsymbol{u}\right\|}+\varphi(\boldsymbol{u}')\frac{u_1'}{\left\|\boldsymbol{u}'\right\|}\right|&=\left|\varphi(\boldsymbol{u}')-\varphi(\boldsymbol{u})\right|\frac{\left|u_1\right|}{\left\|\boldsymbol{u}\right\|}+\varphi(\boldsymbol{u}')\left|\frac{u_1'}{\left\|\boldsymbol{u}'\right\|}-\frac{u_1}{\left\|\boldsymbol{u}\right\|}\right|\\ &\lesssim\left|\varphi(\boldsymbol{u}')-\varphi(\boldsymbol{u})\right|+\zeta.\end{aligned} \tag{3.113}$$

For the first part, we use Mean Value Theorem to write

$$\left|\varphi(\boldsymbol{u}')-\varphi(\boldsymbol{u})\right|\leq\left\|\nabla\varphi(\boldsymbol{v})\right\|\left\|\boldsymbol{u}'-\boldsymbol{u}\right\|\leq\left\|\nabla\varphi(\boldsymbol{v})\right\|\zeta, \tag{3.114}$$

where v is a point between u and u'. Note that $\nabla\varphi(\boldsymbol{v})=\sqrt{\frac{2}{\pi}}p\mathbb{E}\left[\frac{\epsilon^2 v}{\left\|\boldsymbol{v}\right\|^4}e^{-\frac{\epsilon^2}{2\left\|\boldsymbol{v}\right\|^2}}\right]$. Hence, we

have

$$\sup_{\|\boldsymbol{v}\|\geq r}\|\nabla\varphi(\boldsymbol{v})\|\lesssim\sup_{\|\boldsymbol{v}\|\geq r}\mathbb{E}\left[\frac{\epsilon^2}{\|\boldsymbol{v}\|^3}e^{-\frac{\epsilon^2}{2\|\boldsymbol{v}\|^2}}\right]\leq\frac{1}{r}\sup_{\|\boldsymbol{v}\|\geq r}\mathbb{E}\left[\frac{\epsilon^2}{\|\boldsymbol{v}\|^2}e^{-\frac{\epsilon^2}{2\|\boldsymbol{v}\|^2}}\right]\lesssim\frac{1}{r}. \tag{3.115}$$

Overall, we have $(\mathrm{B}_2)\lesssim\frac{\zeta}{r}$, which results in

$$\begin{aligned}&\mathbb{E}\left[\sup_{\boldsymbol{u}\in\mathcal{A}}\frac{1}{m}\sum_{i=1}^{m}\operatorname{sign}\left(\langle\boldsymbol{x}_i,\boldsymbol{u}\rangle+\epsilon_i\right)x_{i,1}-\varphi(\boldsymbol{u})\frac{u_1}{\|\boldsymbol{u}\|}\right]\\&\lesssim\frac{\zeta}{r}+e^{-C\frac{t^2}{\zeta^2}}\sqrt{\log\left(m\right)}+\sqrt{\log(m)}\frac{t}{r}+\sqrt{\frac{k\log(m)\log(d)\log(\frac{R}{\zeta})}{m}}.\end{aligned} \tag{3.116}$$

Hence, once we set $\zeta\asymp\vartheta$, and $t\asymp\sqrt{d}\vartheta\log(m)$, together with the assumption that $r\gtrsim\sqrt{\frac{dm}{k}}\vartheta\log\left(\frac{1}{\vartheta}\right)$, we conclude that

$$\mathbb{E}\left[\sup_{\boldsymbol{u}\in\mathcal{A}}\frac{1}{m}\sum_{i=1}^{m}\operatorname{sign}\left(\langle\boldsymbol{x}_i,\boldsymbol{u}\rangle+\epsilon_i\right)x_{i,1}-\varphi(\boldsymbol{u})\frac{u_1}{\|\boldsymbol{u}\|}\right]\lesssim\sqrt{\frac{k\log^2(m)\log(d)\log(\frac{R}{\vartheta})}{m}}. \tag{3.117}$$

This leads to

$$\begin{aligned}&\mathbb{P}\left(\sup_{\boldsymbol{u}\in\mathcal{A}}\left\|\frac{1}{m}\sum_{i=1}^{m}\operatorname{sign}\left(\langle\boldsymbol{x}_i,\boldsymbol{u}\rangle+\epsilon_i\right)\boldsymbol{x}_i-\varphi(\boldsymbol{u})\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|}\right\|_\infty\geq C\sqrt{\frac{k\log^2(m)\log(d)\log(\frac{R}{\vartheta})}{m}}+\delta\right)\\&\leq 2de^{-cm\delta^2}.\end{aligned} \tag{3.118}$$

Therefore, the following inequality holds, provided that $m\gtrsim\frac{k\log^2(m)\log(d)\log(\frac{R}{\vartheta})}{(1-p)^2\delta^2}$

$$\mathbb{P}\left(\sup_{\boldsymbol{u}\in\mathcal{A}}\left\|\frac{1}{\varphi(\boldsymbol{u})}\frac{1}{m}\sum_{i=1}^{m}\operatorname{sign}\left(\langle\boldsymbol{x}_i,\boldsymbol{u}\rangle+\epsilon_i\right)\boldsymbol{x}_i-\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|}\right\|_\infty\geq\delta\right)\leq e^{-cm\delta^2}. \tag{3.119}$$

Now we turn to the case $m\gtrsim\frac{d}{(1-p)^2}$. Following the same technique, it suffices to bound $\mathbb{E}\left[\sup_{\boldsymbol{u}\in\mathbb{R}^d}\frac{1}{m}\sum_{i=1}^{m}\operatorname{sign}\left(\langle\boldsymbol{x}_i,\boldsymbol{u}\rangle+\epsilon_i\right)x_{i,1}-\varphi(\boldsymbol{u})\frac{u_1}{\|\boldsymbol{u}\|}\right]$. To this goal, we first notice that

$$\begin{aligned}&\mathbb{E}\left[\sup_{\boldsymbol{u}\in\mathbb{R}^d}\frac{1}{m}\sum_{i=1}^{m}\operatorname{sign}\left(\langle\boldsymbol{x}_i,\boldsymbol{u}\rangle+\epsilon_i\right)x_{i,1}-\varphi(\boldsymbol{u})\frac{u_1}{\|\boldsymbol{u}\|}\right]\\&=\underbrace{\mathbb{E}\left[\sup_{\|\boldsymbol{u}\|=1,\lambda\in\mathbb{R}}\frac{1}{m}\sum_{i=1}^{m}\operatorname{sign}\left(\langle\boldsymbol{x}_i,\boldsymbol{u}\rangle+\lambda\epsilon_i\right)x_{i,1}-\varphi(\boldsymbol{u})u_1\right]}_{(A)}.\end{aligned} \tag{3.120}$$

Similarly, applying one-step discretization, we have

$$(A) \leq \underbrace{\mathbb{E}\left[\sup_{\boldsymbol{u}\in\mathbb{S}_\epsilon,\lambda\in\mathbb{R}} \frac{1}{m}\sum_{i=1}^{m} \operatorname{sign}\left(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle + \lambda\epsilon_i\right) x_{i,1} - \phi(\lambda)u_1\right]}_{(B)}$$
$$+ \underbrace{\mathbb{E}\left[\sup_{\|\boldsymbol{u}-\boldsymbol{u}'\|\leq\epsilon,\lambda\in\mathbb{R}} \frac{1}{m}\sum_{i=1}^{m} \left(\operatorname{sign}\left(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle + \lambda\epsilon_i\right) - \operatorname{sign}\left(\langle \boldsymbol{x}_i, \boldsymbol{u}'\rangle + \lambda\epsilon_i\right)\right) x_{i,1} + \phi(\lambda)\left(u_1' - u_1\right)\right]}_{(C)}. \tag{3.121}$$

Here $\phi(\lambda) = \sqrt{\frac{2}{\pi}}(1-p) + \sqrt{\frac{2}{\pi}}p\mathbb{E}\left[e^{-\lambda^2\epsilon^2/2}\right]$ is the same as before. We first control (B). To this goal, we show that $\sup_{\lambda\in\mathbb{R}} \frac{1}{m}\sum_{i=1}^{m} \operatorname{sign}\left(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle + \lambda\epsilon_i\right) x_{i,1} - \phi(\lambda)u_1$ is $\mathcal{O}(1/m)$-sub-Gaussian. We prove it via checking the sub-Gaussian norm

$$\left\|\sup_{\lambda\in\mathbb{R}} \operatorname{sign}\left(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle + \lambda\epsilon_i\right) x_{i,1} - \phi(\lambda)u_1\right\|_{\psi_2} \leq \left\||x_{i,1}|\right\|_{\psi_2} + \sqrt{\frac{2}{\pi}} = \mathcal{O}(1). \tag{3.122}$$

Hence, via maximum inequality, we have

$$(B) \leq \underbrace{\mathbb{E}\left[\sup_{\lambda\in\mathbb{R}} \frac{1}{m}\sum_{i=1}^{m} \operatorname{sign}\left(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle + \lambda\epsilon_i\right) x_{i,1} - \phi(\lambda)u_1\right]}_{(D)} + \mathcal{O}\left(\sqrt{\frac{d\log\left(\frac{1}{\epsilon}\right)}{m}}\right). \tag{3.123}$$

To control (D), we further decompose it into two parts,

$$(D) \leq \underbrace{\mathbb{E}\left[\sup_{\nu\in[0,1]} \frac{1}{m}\sum_{i=1}^{m} \operatorname{sign}\left(\nu\left\langle \boldsymbol{x}_i, \boldsymbol{u}\right\rangle + \epsilon_i\right) x_{i,1} - \phi\left(\frac{1}{\nu}\right)u_1\right]}_{(D_1)}$$
$$+ \underbrace{\mathbb{E}\left[\sup_{\lambda\in[0,1]} \frac{1}{m}\sum_{i=1}^{m} \operatorname{sign}\left(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle + \lambda\epsilon_i\right) x_{i,1} - \phi(\lambda)u_1\right]}_{(D_2)}. \tag{3.124}$$

To control (D_1) and (D_2) we use arguments based on bracketing maximal inequality. We defer a preliminary discussion on bracketing maximal inequality to Appendix 3.H. We first control (D_1). Let $\mathbb{T}_\xi$ be defined as the ξ-net of the interval $[0,1]$. We show that for any $\nu, \nu' \in [0,1]$ such that $|\nu - \nu'| \leq \xi$, we can control $\left\|\left(\operatorname{sign}(\nu\left\langle \boldsymbol{x}_i, \boldsymbol{u}\right\rangle + \epsilon_i) - \operatorname{sign}(\nu'\left\langle \boldsymbol{x}_i, \boldsymbol{u}\right\rangle + \epsilon_i)\right) x_{i,1}\right\|_{L_2(\mathbb{P})}$.

To this goal, we first have

$$\begin{aligned}
&\mathbb{E}\left[\left(\operatorname{sign}(\nu\left\langle \boldsymbol{x}_i, \boldsymbol{u}\right\rangle + \epsilon_i) - \operatorname{sign}(\nu'\left\langle \boldsymbol{x}_i, \boldsymbol{u}\right\rangle + \epsilon_i)\right)^2 x_{i,1}^2\right] \\
&\lesssim \mathbb{E}\left[\left|\operatorname{sign}(\nu\left\langle \boldsymbol{x}_i, \boldsymbol{u}\right\rangle + \epsilon_i) - \operatorname{sign}(\nu'\left\langle \boldsymbol{x}_i, \boldsymbol{u}\right\rangle + \epsilon_i)\right|\right] \\
&\leq \mathbb{E}\left[\mathbb{1}\left(\left|\left(\nu - \nu'\right)\left\langle \boldsymbol{x}_i, \boldsymbol{u}\right\rangle\right| \geq t\right) + \mathbb{1}\left(\left|\nu\left\langle \boldsymbol{x}_i, \boldsymbol{u}\right\rangle + \epsilon_i\right| \leq t\right)\right] \\
&\lesssim e^{-C\frac{t^2}{\xi^2}} + t.
\end{aligned} \tag{3.125}$$

Upon picking $t \asymp \xi \log\left(\frac{1}{\xi}\right)$, we have

$$\left\|\left(\operatorname{sign}(\nu\left\langle \boldsymbol{x}_i, \boldsymbol{u}\right\rangle + \epsilon_i) - \operatorname{sign}(\nu'\left\langle \boldsymbol{x}_i, \boldsymbol{u}\right\rangle + \epsilon_i)\right) x_{i,1}\right\|_{L_2(\mathbb{P})} \lesssim \sqrt{\xi \log\left(\frac{1}{\xi}\right)}. \tag{3.126}$$

Therefore, the bracketing number is bounded by $N_{[]}(\varepsilon\|F\|, \mathcal{F}, \|\cdot\|) \lesssim C\frac{1}{\sqrt{\epsilon}}$, which in turn leads to an upper bound on the bracketing entropy $J_{[]}(1, \mathcal{F}, L_2(\mathbb{P})) \lesssim 1$. Applying Theorem 19 leads to

$$(D_1) \lesssim \sqrt{\frac{1}{m}}. \tag{3.127}$$

Similarly, we can show that $(D_2) \lesssim \sqrt{\frac{1}{m}}$. Therefore, we conclude that

$$(B) \lesssim \sqrt{\frac{d\log\left(\frac{1}{\epsilon}\right)}{m}}. \tag{3.128}$$

For (C), we can use the similar technique in the overparameterized setting ($m \ll d$), which leads to

$$(C) \lesssim \sqrt{\log(m)}\epsilon + \sqrt{\frac{d\log(m)}{m}}. \tag{3.129}$$

Therefore, once we set $\epsilon \asymp \sqrt{\frac{d}{m}}$, we immediately obtain

$$(A) \lesssim \sqrt{\frac{d\log(m)}{m}}. \tag{3.130}$$

Combining the derived bounds results in

$$\mathbb{P}\left(\sup_{\boldsymbol{u}\in\mathbb{R}^d}\left\|\frac{1}{m}\sum_{i=1}^{m}\operatorname{sign}\left(\left\langle \boldsymbol{x}_i, \boldsymbol{u}\right\rangle + \epsilon_i\right)\boldsymbol{x}_i - \varphi(\boldsymbol{u})\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|}\right\|_\infty \geq C\sqrt{\frac{d\log(m)}{m}} + \delta\right) \leq e^{c_1\log(d) - c_2 m\delta^2}. \tag{3.131}$$

Assuming $m \gtrsim \frac{d\log(m)}{(1-p)^2}$, the above bound reduces to

$$\mathbb{P}\left(\sup_{\boldsymbol{u}\in\mathbb{R}^d}\left\|\frac{1}{\varphi(\boldsymbol{u})}\frac{1}{m}\sum_{i=1}^{m}\operatorname{sign}\left(\langle\boldsymbol{x}_i,\boldsymbol{u}\rangle+\epsilon_i\right)\boldsymbol{x}_i-\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|}\right\|_\infty\geq\delta\right)\leq e^{-cm\delta^2}. \tag{3.132}$$

3.F Auxiliary Lemmas

Lemma 59. *Suppose $x_1,\cdots,x_m$ are i.i.d. standard Gaussian vectors with dimension d. Then, for arbitrary $\delta>0$ we have*

$$\mathbb{P}\left(\sup_{\|\boldsymbol{u}\|=1}\left|\frac{1}{m}\sum_{i=1}^{m}|\left\langle\boldsymbol{x}_i,\boldsymbol{u}\right\rangle|-\sqrt{\frac{2}{\pi}}\right|\geq C\sqrt{\frac{d}{m}}+\delta\right)\leq e^{-cm\delta^2}. \tag{3.133}$$

Here C,c are universal constants.

Proof. This lemma directly follows from the standard expectation and high probability bounds for sub-Gaussian process. See e.g., [Ma and Fattahi, 2023a, Lemma 4] for a simple proof. □

Lemma 60. *For two arbitrary vectors $\boldsymbol{a},\boldsymbol{b}\in\mathbb{R}^n$, we have*

$$\|\boldsymbol{a}\odot\boldsymbol{b}\|\leq\|\boldsymbol{a}\|_\infty\|\boldsymbol{b}\|. \tag{3.134}$$

3.G Deferred Proofs

3.G.1 Proof of Lemma 58

Proof. To prove this lemma, it suffices to show that, for any $\boldsymbol{u},\boldsymbol{v}\in\mathbb{R}^d$, we have

$$\mathbb{E}\left[\operatorname{sign}\left(\epsilon+\langle\boldsymbol{x},\boldsymbol{u}\rangle\right)\langle\boldsymbol{x},\boldsymbol{v}\rangle\right]=\sqrt{\frac{2}{\pi}}\mathbb{E}\left[e^{-\epsilon^2/2\|\boldsymbol{u}\|^2}\right]\left\langle\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|},\boldsymbol{v}\right\rangle. \tag{3.135}$$

Without loss of generality, we assume that $\|\boldsymbol{u}\| = \|\boldsymbol{v}\| = 1$. Let us denote $w := \langle \boldsymbol{x}, \boldsymbol{u} \rangle, z := \langle \boldsymbol{x}, \boldsymbol{v} \rangle, \rho := \mathrm{Cov}(w, z) = \langle \boldsymbol{u}, \boldsymbol{v} \rangle$. Then

$$\begin{aligned}
\mathbb{E}\left[\mathrm{sign}\left(\epsilon + \langle \boldsymbol{x}, \boldsymbol{u} \rangle\right) \langle \boldsymbol{x}, \boldsymbol{v} \rangle\right] &= \mathbb{E}\left[\mathrm{sign}\left(\epsilon + w\right) z\right] \\
&\overset{(a)}{=} \rho \mathbb{E}\left[\mathrm{sign}(w + \epsilon) w\right] \\
&= \rho \mathbb{E}_\epsilon \left[\int_{-\epsilon}^{\infty} t \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt - \int_{-\infty}^{-\epsilon} t \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt\right] \\
&= \rho \mathbb{E}_\epsilon \left[\int_{-\epsilon}^{\infty} t \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt + \int_{\epsilon}^{\infty} t \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt\right] \\
&= 2\rho \mathbb{E}_\epsilon \left[\int_{|\epsilon|}^{\infty} t \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt\right] \\
&= \sqrt{\frac{2}{\pi}} \langle \boldsymbol{u}, \boldsymbol{v} \rangle \mathbb{E}_\epsilon \left[\int_{|\epsilon|}^{\infty} d\left(-e^{-t^2/2}\right)\right] \\
&= \sqrt{\frac{2}{\pi}} \langle \boldsymbol{u}, \boldsymbol{v} \rangle \mathbb{E}_\epsilon \left[e^{-\epsilon^2/2}\right].
\end{aligned} \tag{3.136}$$

Here in (a) we use the fact that $z|w, \epsilon \sim \mathcal{N}(\rho w, 1 - \rho^2)$ since ϵ is independent of w, z. Hence, we have

$$\mathbb{E}\left[\mathrm{sign}\left(\epsilon + \langle \boldsymbol{x}, \boldsymbol{u} \rangle\right) \langle \boldsymbol{x}, \boldsymbol{v} \rangle\right] = \sqrt{\frac{2}{\pi}} \mathbb{E}\left[e^{-\epsilon^2/2\|\boldsymbol{u}\|^2}\right] \left\langle \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|}, \boldsymbol{v} \right\rangle \tag{3.137}$$

for any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$. On the other hand, it is easy to verify that $\mathbb{E}\left[\mathrm{sign}\left(\langle \boldsymbol{x}, \boldsymbol{u} \rangle\right) \langle \boldsymbol{x}, \boldsymbol{v} \rangle\right] = \sqrt{\frac{2}{\pi}} \left\langle \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|}, \boldsymbol{v} \right\rangle$. The proof is completed by noting that the corruption probability is p. □

3.H Preliminaries on the Uniform Concentration Bounds

In this section, we provide the preliminary probability tools for proving Proposition 11.

Definition 16 (Sub-Gaussian random variable)**.** *We say a random variable $X \in \mathbb{R}$ with expectation $\mathbb{E}[X] = \mu$ is σ^2-sub-Gaussian if for all $\lambda \in \mathbb{R}$, we have $\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\frac{\lambda^2\sigma^2}{2}}$. Moreover, the sub-Gaussian norm of X is defined as $\|X\|_{\psi_2} := \sup_{p \geq 1} \left\{p^{-1/2} (\mathbb{E}[|X|^p])^{1/p}\right\}$.*

According to [Wainwright, 2019], the following statements are equivalent:

- X is σ^2-sub-Gaussian.
- (Tail bound) For any $t > 0$, we have $\mathbb{P}(|X - \mu| \geq t) \leq 2e^{-\frac{t^2}{2\sigma^2}}$.
- (Moment bound) We have $\|X\|_{\psi_2} \lesssim \sigma$.

Next, we provide the definitions of the sub-Gaussian process, ϵ-net, and covering number.

Definition 17 (Sub-Gaussian process). *A zero mean stochastic process $\{\mathcal{X}_\theta, \theta \in \mathbb{T}\}$ is a σ^2-sub-Gaussian process with respect to a metric d on a set $\mathbb{T}$, if for every $\theta, \theta' \in \mathbb{T}$, the random variable $\mathcal{X}_\theta - \mathcal{X}_{\theta'}$ is $(\sigma d(\theta, \theta'))^2$-sub-Gaussian.*

Definition 18 (ϵ-net and covering number). *A set $\mathcal{N}$ is called an ϵ-net for $(\mathbb{T}, d)$ if for every $t \in \mathbb{T}$, there exists $\pi(t) \in \mathcal{N}$ such that $d(t, \pi(t)) \leq \epsilon$. The covering number $N(\mathbb{T}, d, \epsilon)$ is defined as the smallest cardinality of an ϵ-net for $(\mathbb{T}, d)$:*

$$N(\mathbb{T}, d, \epsilon) := \inf\{|\mathcal{N}| : \mathcal{N} \text{ is an } \epsilon\text{-net for } (\mathbb{T}, d)\}.$$

Definition 19 (Bracketing number, Definition 2.1.6 in [Van Der Vaart, van der Vaart, van der Vaart, and Wellner, 1996]). *Given two functions l and u, the bracket $[l, u]$ is the set of all functions f with $l \leq f \leq u$. An ε-bracket is a bracket $[l, u]$ with $\|u - l\| < \varepsilon$. The bracketing number $N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)$ is the minimum number of ε-brackets needed to cover $\mathcal{F}$. The bracketing entropy is the logarithm of the bracketing number. In the definition of the bracketing number, the upper and lower bounds u and l of the brackets need not belong to $\mathcal{F}$ themselves but are assumed to have finite norms.*

Bracketing number can be regarded as an analog of covering number, describing the geometric complexity of the underlining function space. Although bracketing number of a general function class is difficult to characterize, for some specific function classes, we can easily derive upper bounds for their bracketing number. In particular, we have the following result for Lipschitz functions.

Theorem 18 (Adapted from Theorem 2.7.11 in [Van Der Vaart et al., 1996]). *Let $\mathcal{F} = \{f_t : t \in T\}$ be a class of functions. Suppose that for arbitrary $s, t \in T$, we have*

$$|f_s(\boldsymbol{x}) - f_t(\boldsymbol{x})| \leq d(s, t)F(\boldsymbol{x}), \tag{3.138}$$

for some metric d on the index set, function F on the sample space, and every x. Then, for any norm $\|\cdot\|$,

$$N_{[]}(2\varepsilon\|F\|, \mathcal{F}, \|\cdot\|) \leq N(\varepsilon, T, d). \tag{3.139}$$

Theorem 19 (Adapted from Theorem 2.14.2 in [Van Der Vaart et al., 1996]). *For a given norm $\|\cdot\|$, define a bracketing integral of a class of functions $\mathcal{F}$ as*

$$J_{[]}(\delta, \mathcal{F}, \|\cdot\|) = \int_0^\delta \sqrt{1 + \log N_{[]}(\varepsilon\|F\|, \mathcal{F}, \|\cdot\|)}d\varepsilon. \tag{3.140}$$

Let $\mathcal{F}$ be a class of measurable functions with measurable envelope function F, we have

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}f(X_i)-\mathbb{E}\left[f(X)\right]\right]\lesssim J_{[]}(1,\mathcal{F},L_2(\mathbb{P}))\frac{\|F\|_{L_2(\mathbb{P})}}{\sqrt{n}},\tag{3.141}$$

where $\mathbb{P}$ is the distribution of X, and the $L_2(\mathbb{P})$-norm is defined as $\|f\|_{L_2(\mathbb{P})} := \left(\int_\Omega f^2(\omega)d\mathbb{P}(\omega)\right)^{1/2}$.

CHAPTER 4

Unregularized Matrix Completion

This chapter is based on the following paper:

- Jianhao Ma and Salar Fattahi. Convergence of gradient descent with small initialization for unregularized matrix completion. In *The Thirty Seventh Annual Conference on Learning Theory*, 2024.

4.1 Introduction

Matrix completion is a fundamental problem in the field of machine learning, where the objective is to reconstruct a positive semidefinite (PSD) matrix of rank-r, denoted as $\boldsymbol{X}^\star \in \mathbb{R}^{d\times d}$, from only a subset of its observed entries. The most natural approach to solve this problem involves minimizing the following mean squared error:

$$\min_{\boldsymbol{U}\in\mathbb{R}^{d\times r'}} f(\boldsymbol{U}) = \frac{1}{4p}\left\|\mathcal{P}_\Omega(\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{X}^\star)\right\|_{\mathrm{F}}^2. \tag{MC}$$

Here, p represents the probability of observing each entry in $\boldsymbol{X}^\star$, Ω denotes the set of observed entries, and $\mathcal{P}_\Omega$ shows the projection operation onto the set of matrices supported by Ω. When the true rank r is unknown, it is often over-estimated by the search rank $r' \geq r$, leading to what is referred to as *over-parameterized* matrix completion.

A prominent application of matrix completion is in collaborative filtering [Gleich and Lim, 2011]. Additionally, it has applications in other areas, including image reconstruction [Hu, Liu, and Jacob, 2018], fast kernel matrix approximation [Graepel, 2002, Paisley and Carin, 2010], and more recently, in teaching arithmetic to transformers [Lee, Sreenivasan, Lee, Lee, and Papailiopoulos, 2023].

Perhaps one of the most natural approaches for solving the above optimization problem is (vanilla) gradient descent (GD): given an initial point $\boldsymbol{U}_0$ and a fixed step-size $\eta > 0$, generate a sequence of iterates $\{\boldsymbol{U}_t\}_{t=1}^T$ according to $\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta\nabla f(\boldsymbol{U}_t)$. Despite its

simplicity and desirable practical performance (see Figure 4.1 and the experiments in [Zheng and Lafferty, 2016]), the conditions under which the GD converges globally to the ground truth $\boldsymbol{X}^\star$ have remained a long-standing mystery.

A line of research has been devoted to studying gradient-based algorithms with *explicit regularization* [Sun and Luo, 2016, Zheng and Lafferty, 2016, Jain, Netrapalli, and Sanghavi, 2013, Ge et al., 2016]. These methods typically incorporate either an $\ell_{2,\infty}$-norm regularizer or a projection step to constrain the iterates within a set with $\ell_{2,\infty}$-norm bounds to promote *incoherence* (see Definition 2). However, the use of $\ell_{2,\infty}$-norm regularization or projection techniques often introduce more tuning parameters, and has been found to be unnecessary in most cases [Zheng and Lafferty, 2016, Ma, Wang, Chi, and Chen, 2018][1].

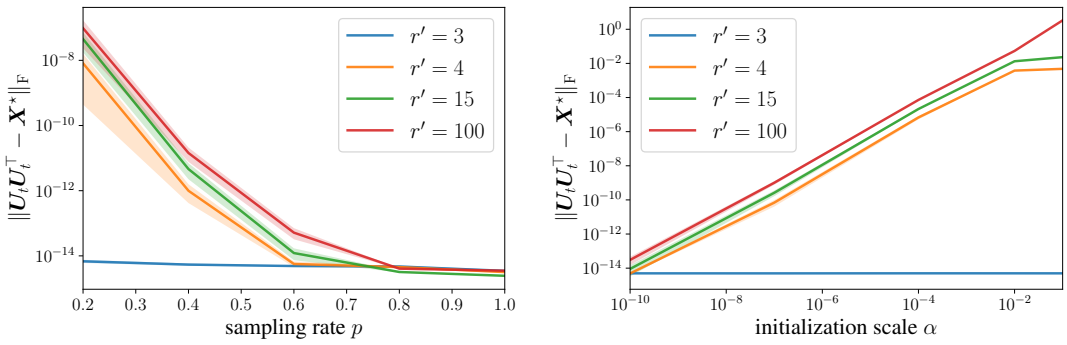
On the other hand, the convergence of GD without explicit regularization was initially tackled by Ma et al. [2018] in the context of symmetric matrix completion, and subsequently extended by Chen, Liu, and Li [2020] to asymmetric settings. However, these studies consider a very special case of matrix completion where GD is initialized sufficiently close to the ground truth, and the rank of the ground truth r is known. In practice, however, GD converges even if it is initialized far from the ground truth and the rank is over-parameterized $r' \gg r$ (see Figure 4.1).

Therefore, the following question still remains open:

> *Why does GD with a small initialization efficiently converge to the ground truth of MC in the absence of explicit regularization, even in the general rank-r case where $r' \geq r \geq 1$?*

Recently, Kim and Chung [2022] answered the above question for the special case of rank-1 symmetric matrix completion with $r' = r = 1$, but their proposed approach does not extend to the general rank-r case. It is also worth noting that the above question has been addressed for another class of matrix factorization problems satisfying a norm-preserving property called *restricted isometry property* (RIP). Problems that satisfy this property include matrix sensing [Li et al., 2018, Stöger and Soltanolkotabi, 2021, Ma and Fattahi, 2023a] and sparse recovery [Vaskevicius et al., 2019]. However, a significant challenge arises with matrix completion, as it *does not* satisfy the restricted isometry property. Consequently, the existing methodologies built upon RIP are not directly applicable to matrix completion.

[1] It is worth noting that explicit regularization can still offer significant advantages in scenarios with extremely limited sample sizes [Sun, 2015].



(a) (b)

Figure 4.1: The performance of vanilla GD with small initialization and constant step-size on MC without any explicit regularization. In all experiments, we set the dimension $d = 100$ and the true rank $r = 3$. (a): Higher sampling rates improve the final accuracy of GD for the over-parameterized MC, but have no impact on the exactly-parameterized MC. (b): Increasing the initialization scale hampers the final accuracy of GD for the over-parameterized MC, but has no impact on the exactly-parameterized MC.

4.1.1 Summary of Contributions

In this work, we provide a complete answer to the aforementioned question. A comparison of our results with other studies on unregularized matrix completion can be found in Table 4.1. The key contributions of our work are as follows:

- **Convergence of GD with small initialization in over-parameterized regime:** When the rank of the ground truth r is unknown and over-parameterized by $r' \geq r \geq 1$, we prove that GD with small initialization converges to the ground truth at a near-linear rate. Surprisingly, neither the convergence rate nor the final accuracy depends on the over-parameterized search rank r', and they are only governed by the true rank r. In particular, given an initial point that satisfies $\|\boldsymbol{U}_0\| \leq \mathcal{O}(\epsilon)$ for some $\epsilon > 0$ and a sampling rate of $p = \widetilde{\Omega}(r^9 \log^6(1/\epsilon)/d)$, GD converges to ϵ-neighborhood of $\boldsymbol{X}^\star$ in $\mathcal{O}(\log^4(1/\epsilon))$ iterations. Therefore, a smaller initial point or a larger sampling rate can improve the final error of GD. The empirical observation presented in Figure 4.1 provides further support for this result.

- **Improved results in the exactly-parameterized regime:** We show that GD enjoys an improved convergence when the rank of the ground truth r is known and $r' = r \geq 1$. In particular, given an initial point that satisfies $\|\boldsymbol{U}_0\| \leq \mathcal{O}(1/d)$ and a sampling rate of $p = \widetilde{\Omega}(r^9/d)$, GD converges to ϵ-neighborhood of $\boldsymbol{X}^\star$ in $\mathcal{O}(\log(1/\epsilon))$ iterations for any arbitrarily small $\epsilon > 0$. A key distinction from the over-parameterized setting is that the final error of GD remains unaffected by the initialization scale or the sample

Algorithm	Sample complexity	Computational complexity	Global	Exact	Over-param.
[Ma et al., 2018]	$dr^3\log^3(d)$	$\kappa^2\log\left(\frac{1}{\epsilon}\right)$	✘	✔	✘
[Chen et al., 2020]	$dr^2\log(d)$	$\kappa^2\log\left(\frac{1}{\epsilon}\right)$	✘	✔	✘
[Kim and Chung, 2022]	$d\log^{22}(d)$	$\log\left(\frac{1}{\epsilon}\right)$	✔	✱	✘
Ours (Theorem 21)	$dr^9\log^8(d)$	$\kappa^4\log\left(\frac{1}{\epsilon}\right)$	✔	✔	✘
Ours (Theorem 20)	$dr^9\log^2(d)\log^6\left(\frac{1}{\epsilon}\right)$	$\kappa^4\log^4\left(\frac{1}{\epsilon}\right)$	✔	✔	✔

Table 4.1: Comparisons between different algorithms for matrix completion without explicit regularization. ✱ The result only holds for $r' = r = 1$.

size, provided that they meet certain thresholds. This is also evident in Figure 4.1. When $r = \mathcal{O}(1)$, the resulting sample complexity is information-theoretically optimal (modulo logarithmic factors).

- **Weakly-coupled leave-one-out analysis:** In order to establish the implicit regularization of GD for MC, a pivotal technique is a decoupling mechanism known as leave-one-out analysis, a trick rooted in probability and random matrix theory. However, the current theory based on this technique is only effective when the iterates are sufficiently close to the ground truth. At the crux of our technical analysis lies an extension of the classical leave-one-out analysis to the global setting, which we term *weakly-coupled leave-one-out analysis*. In essence, our proposed method relaxes the requirement for the initial iterates to be sufficiently close to the ground truth, making it particularly suitable for the global convergence analysis of GD.

Notations. We use bold uppercase letters $\boldsymbol{X}, \boldsymbol{Y}$ to denote matrices and bold lowercase letters $\boldsymbol{x}, \boldsymbol{y}$ to denote vectors. For vectors, we use $\|\cdot\|$ to denote ℓ_2-norm, and for matrices we use $\|\cdot\|$ and $\|\cdot\|_{\mathrm{F}}$ to denote operator norm and Frobenius norm, respectively. For matrix $\boldsymbol{X} \in \mathbb{R}^{d_1\times d_2}$, we denote by $X_{i,j}$ the (i,j)-th element of $\boldsymbol{X}$, $\boldsymbol{X}_{i,\cdot}$ the i-th row, and $\boldsymbol{X}_{\cdot,j}$ the j-th column. The $\ell_{2,\infty}$-norm of $\boldsymbol{X}$, denoted as $\|\boldsymbol{X}\|_{2,\infty}$, is defined as $\max_i \|\boldsymbol{X}_{i,\cdot}\|$. Additionally, we define the singular values of $\boldsymbol{X} \in \mathbb{R}^{d_1\times d_2}$ as $\sigma_1(\boldsymbol{X}) \geq \sigma_2(\boldsymbol{X}) \geq \cdots \geq \sigma_{\min\{d_1,d_2\}}(\boldsymbol{X}) \geq 0$. The set of all orthogonal matrices is denoted by $\mathcal{O}_{d_1\times d_2} := \{\boldsymbol{O} \in \mathbb{R}^{d_1\times d_2} : \boldsymbol{O}^\top\boldsymbol{O} = \boldsymbol{I}\}$. For two matrices $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{d_1\times d_2}$, we define their Procrustes distance as $\mathrm{dist}(\boldsymbol{X}, \boldsymbol{Y}) = \min_{\boldsymbol{O}\in\mathcal{O}_{d_2\times d_2}} \|\boldsymbol{X} - \boldsymbol{Y}\boldsymbol{O}\|_{\mathrm{F}}$. The projection matrix onto the column space of an orthogonal matrix $\boldsymbol{V} \in \mathcal{O}_{d_1\times d_2}$ is defined as $\mathcal{P}_{\boldsymbol{V}} := \boldsymbol{V}\boldsymbol{V}^\top$.

We use the notation $f(n) \lesssim g(n)$ or $f(n) = \mathcal{O}(g(n))$ when a constant $C > 0$ exists such that $f(n) \leq Cg(n)$ for sufficiently large n. Conversely, $f(n) \gtrsim g(n)$ or $f(n) = \Omega(g(n))$ implies the existence of a constant $C > 0$ such that $f(n) \geq Cg(n)$ for sufficiently large n. Moreover,

we use the notations $\widetilde{\mathcal{O}}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ to hide logarithmic dependencies on the dimension or other parameters of the problem. Additionally, we use $f(n) \asymp g(n)$ or $f(n) = \Theta(g(n))$ when $f(n) \lesssim g(n)$ and $f(n) \gtrsim g(n)$.

4.2 Problem Setup and Main Results

Suppose that the singular value decomposition (SVD) of $\boldsymbol{X}^\star$ is given by $\boldsymbol{X}^\star = \boldsymbol{V}^\star \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top}$, where $\boldsymbol{V}^\star \in \mathcal{O}_{d\times r}$ and $\boldsymbol{\Sigma}^\star$ is an $r \times r$ diagonal matrix with diagonal elements in descending order $\sigma_1^\star \geq \cdots \geq \sigma_r^\star > 0$. We denote the condition number of $\boldsymbol{X}^\star$ as $\kappa = \sigma_1^\star/\sigma_r^\star$. Upon defining the symmetrized operator $\mathcal{R}_\Omega = \frac{1}{2p}(\mathcal{P}_\Omega + \mathcal{P}_\Omega^\top)$, the update rule for GD can be written as

$$\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta \nabla f(\boldsymbol{U}_t) = \boldsymbol{U}_t - \eta \mathcal{R}_\Omega(\boldsymbol{U}_t \boldsymbol{U}_t^\top - \boldsymbol{X}^\star)\boldsymbol{U}_t, \qquad \boldsymbol{U}_0 = \alpha \boldsymbol{Z} \tag{GD}$$

where $\alpha > 0$ is the initialization scale and $\boldsymbol{Z}$ is the initialization matrix satisfying $\|\boldsymbol{Z}\| = 1$. We assume that $\boldsymbol{Z}$ satisfies the following *alignment condition*.

Condition 1 (Alignment)**.** *We say the matrix $\boldsymbol{Z} \in \mathbb{R}^{d\times r'}$ with $\|\boldsymbol{Z}\| = 1$ satisfies the alignment condition if there exists a universal constant $c_0 > 0$ such that*

$$\sigma_r(\mathcal{P}_{\boldsymbol{V}^\star} \boldsymbol{Z}) \geq c_0. \tag{alignment condition}$$

Intuitively, this condition necessitates that the initialization matrix should have a non-negligible alignment with the column space of the ground truth. The following lemma reveals that this alignment condition is satisfied for common initialization strategies with overwhelming probability.

Lemma 61 (Sufficient condition for alignment)**.** *The following statements are satisfied:*

- ***Gaussian initialization.*** *Given $0.5d \leq r' \leq d$ and $\boldsymbol{Z} = \boldsymbol{G}/\|\boldsymbol{G}\|$, where $\boldsymbol{G}$ is a standard Gaussian matrix, alignment condition holds with $c_0 = 0.1$ with probability at least $1 - \exp\{-\Omega(d)\}$.*
- ***Orthogonal initialization.*** *Given $r' = d$ and $\boldsymbol{Z} = \boldsymbol{O}$ for any $\boldsymbol{O} \in \mathcal{O}_{d\times d}$, alignment condition is satisfied with $c_0 = 1$.*
- ***Spectral initialization.*** *Let $\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ be the eigendecomposition of the best rank-r' approximation of $\mathcal{R}_\Omega(\boldsymbol{X}^\star)$ measured in Frobenius norm. Given $r \leq r' \leq d$ and $\boldsymbol{Z} = \boldsymbol{U}/\|\boldsymbol{U}\|$, where $\boldsymbol{U} = \boldsymbol{V}\boldsymbol{\Sigma}^{1/2}$, alignment condition holds with $c_0 = \frac{1}{2\kappa}$ with probability at least $1 - \frac{1}{d^3}$.*

Before proceeding to the main theorem, we introduce two crucial conditions on the ground truth $\boldsymbol{X}^\star$ and the random observation set Ω.

Condition 2 (Incoherence)**.** *The rank-r PSD matrix $\boldsymbol{X}^\star \in \mathbb{R}^{d\times d}$ with SVD $\boldsymbol{X}^\star = \boldsymbol{V}^\star \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top}$ is μ-incoherent for some $\mu \geq 1$ if $\|\boldsymbol{V}^\star\|_{2,\infty} = \sqrt{\frac{\mu}{d}}\|\boldsymbol{V}^\star\|_{\mathrm{F}} = \sqrt{\frac{\mu r}{d}}$.*

Condition 3 (Random sampling model)**.** *Each entry of $\boldsymbol{X}^\star$ is observed independently with probability p. In other words, $\mathbb{P}((i,j)\in\Omega) = p$ independently for all $1 \leq i, j \leq d$.*

Chen [2015] shows that the incoherence condition is necessary for the recovery of the ground truth. Intuitively, the incoherence condition with $\mu = \mathcal{O}(1)$ entails that none of the columns of $\boldsymbol{V}^\star$ have significant alignment with the standard basis vectors. Such a condition ensures that the ground truth is far from being sparse. We note that successful recovery of a sparse $\boldsymbol{X}^\star$ is only achievable in a near-ideal scenario where $p \approx 1$ and nearly the entirety of $\boldsymbol{X}^\star$ is observed. On the other hand, when $\mu = \mathcal{O}(1)$, the recovery is possible even when p scales with $\widetilde{\Omega}(\mathrm{poly}(r)/d)$ [Chen, 2015].

With the aforementioned conditions in place, we can now present our main result, which establishes the global convergence of GD on MC for the over-parameterized setting.

Theorem 20 (Convergence of GD for over-parameterized MC)**.** *Let $\boldsymbol{X}^\star$ be rank-r, and Conditions 2 and 3 are satisfied with a sampling rate of $p \gtrsim \frac{\kappa^6\mu^4 r^9 \log^6(\frac{1}{\alpha})\log^2(d)}{d}$. Consider MC with search rank $r \leq r' \leq d$. Consider the iterates of GD with the step-size $\eta \asymp \frac{\mu r}{\sqrt{pd}\sigma_1^\star}$ and the initial point $\boldsymbol{U}_0 = \alpha\boldsymbol{Z}$, where $0 < \alpha \leq \sqrt{\frac{\sigma_1^\star}{d}}$ and $\boldsymbol{Z}$ satisfies Condition 1. With probability at least $1 - \frac{2}{d^3}$, after $T \lesssim \frac{1}{\eta\sigma_r^\star}\log\left(\frac{1}{\alpha}\right)$ iterations, we have*

$$\left\|\boldsymbol{U}_T\boldsymbol{U}_T^\top - \boldsymbol{X}^\star\right\|_{\mathrm{F}} \lesssim \sqrt{\frac{\sigma_1^\star\kappa^2\mu r^2}{p}}\alpha.$$

A few observations are in order based on the above theorem.

Computational complexity. The initialization scale α governs the final accuracy of GD. Therefore, to ensure that $\left\|\boldsymbol{U}_T\boldsymbol{U}_T^\top - \boldsymbol{X}^\star\right\|_{\mathrm{F}} \leq \epsilon$, it suffices to set the initialization scale to $\alpha \lesssim \sqrt{\frac{p}{\sigma_1^\star\kappa^2\mu r^2}}\epsilon$. Moreover, assuming that $\max\{\kappa, \mu, r\} = \mathcal{O}(1)$ and $\epsilon \leq 1/d$, this accuracy is achieved within $\widetilde{O}\left(\log^4\left(\frac{1}{\epsilon}\right)\right)$ iterations, which scales only poly-logarithmically with $1/\epsilon$.

Effect of over-parameterization. The level of over-parameterization in the search rank r' does not have any impact on either the sample complexity or the convergence of GD. As a result, our results hold even if $d^2p \ll dr'$. In such cases, MC has many global minima, some

of which may not satisfy $\boldsymbol{U}\boldsymbol{U}^\top \approx \boldsymbol{X}^\star$. This sheds light on the implicit regularization of the vanilla GD with small initialization toward low-rank solutions when applied to MC.

Sample complexity. The required sample complexity is given by $d^2 p \gtrsim dr^9\kappa^6\mu^4 \log^6\left(\frac{1}{\alpha}\right)\log^2(d)$, which is optimal with respect to the dimension d up to a logarithmic factor. This contrasts with the direct extension of the approach by Kim and Chung [2022], which necessitates a sample size on the order of $d^{1+\Theta(\kappa-1)}$. Moreover, Theorem 20 highlights that the sampling rate p must scale with $\log^6\left(\frac{1}{\epsilon}\right)$ to attain an accuracy level of ϵ, which in turn leads to a mild dependency of the final error on the sampling rate. In other words, given a fixed sampling rate p, GD achieves an accuracy in the order of $\exp(-\Omega(pd))$. We suspect that the observed dependency might be an artifact of our proof technique and could potentially be relaxed with a more detailed analysis, similar to that presented in [Stöger and Soltanolkotabi, 2021]. In our next theorem, we demonstrate that in the exact-parameterization regime, it is possible to relax this mild dependency, achieving a sample complexity that does not depend on the desired accuracy level or the initialization scale.

Theorem 21 (Convergence of GD for exactly-parameterized MC)**.** *Let $\boldsymbol{X}^\star$ be rank-r, and Conditions 2 and 3 are satisfied with a sampling rate of $p \gtrsim \frac{\kappa^6\mu^4 r^9 \log^8(d)}{d}$. Consider MC with search rank $r' = r$. Consider the iterates of GD with the step-size $\eta \asymp \frac{\mu r}{\sqrt{pd}\sigma_1^\star}$ and the initial point $\boldsymbol{U}_0 = \alpha\boldsymbol{Z}$, where $\alpha \asymp \frac{\sigma_r^\star}{\kappa^{1.5}d}$ and $\boldsymbol{Z}$ satisfies Condition 1. Given any accuracy $\epsilon > 0$, with probability at least $1 - O\left(\frac{1}{d^3}\right)$ and after $T \lesssim \frac{1}{\eta\sigma_r^\star}\log\left(\frac{1}{\epsilon}\right)$ iterations, we have*

$$\left\|\boldsymbol{U}_T\boldsymbol{U}_T^\top - \boldsymbol{X}^\star\right\|_{\mathrm{F}} \leq \epsilon. \tag{4.1}$$

We next outline the key distinctions between the two aforementioned theorems. In the exactly-parameterized regime, neither the sampling rate p nor the initialization scale α affect the final error ϵ, as long as they meet certain thresholds. In contrast, a smaller initialization scale or a larger sampling rate improves the final error in the over-parameterized regime. Moreover, the convergence rate of GD improves from $O\left(\log^4\left(\frac{1}{\epsilon}\right)\right)$ to $O\left(\log\left(\frac{1}{\epsilon}\right)\right)$. This is because the required sampling rate is smaller in the exactly-parameterized regime, allowing the algorithm to adopt a more aggressive step-size.

4.3 Proof Outline

In this section, we present the key ideas underpinning our proof techniques. We begin in Section 4.3.1 with a dynamic signal-residual decomposition. Next, in Section 4.3.2, we

introduce the weakly-coupled leave-one-out analysis which, together with our dynamic signal-residual decomposition, completes the proof for Theorem 20. Section 4.3.3 explains how these techniques can be further refined to yield improved results for the exactly-parameterized regime. Throughout this section, we occasionally omit the consideration of higher-order terms involving the step-size η. We highlight that while this omission serves to streamline the presentation, our rigorous proofs in the appendix carefully account for these higher-order terms.

4.3.1 Dynamic Signal-residual Decomposition

We employ a dynamic projection scheme akin to that described by Li et al. [2018], which decomposes the iterates $\boldsymbol{U}_t$ into two distinct components: a low-rank signal part, $\boldsymbol{S}_t$, and a residual part, $\boldsymbol{E}_t$. This decomposition is represented as follows:

$$\boldsymbol{U}_t = \boldsymbol{S}_t + \boldsymbol{E}_t, \quad \text{where} \quad \boldsymbol{S}_t = \mathcal{P}_{\boldsymbol{V}_t}\boldsymbol{U}_t, \text{ and } \boldsymbol{E}_t = \mathcal{P}^{\perp}_{\boldsymbol{V}_t}\boldsymbol{U}_t.$$

Upon defining $\boldsymbol{M}_t = \mathcal{R}_\Omega(\boldsymbol{X}^\star - \boldsymbol{U}_t\boldsymbol{U}_t^\top)$, the dynamic orthonormal matrix $\boldsymbol{V}_t$ is recursively defined as:

$$\boldsymbol{V}_{t+1} = \boldsymbol{Z}_{t+1}\left(\boldsymbol{Z}_{t+1}^\top\boldsymbol{Z}_{t+1}\right)^{-1/2} \quad \text{where} \quad \boldsymbol{Z}_{t+1} = \left(\boldsymbol{I} + \eta\boldsymbol{M}_t\right)\boldsymbol{V}_t \text{ and } \boldsymbol{V}_0 = \boldsymbol{V}^\star.$$

Upon defining the error matrix as $\boldsymbol{\Delta}_t := \boldsymbol{X}^\star - \boldsymbol{U}_t\boldsymbol{U}_t^\top$, our goal is to show that $\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}$ decreases efficiently to $\mathcal{O}(\alpha)$. To show the benefit of the proposed dynamic signal-residual decomposition in achieving this goal, we start by stating the one-step dynamic of the error matrix:

$$\|\boldsymbol{\Delta}_{t+1}\|_{\mathrm{F}}^2 = \|\boldsymbol{\Delta}_{t+1}\|_{\mathrm{F}}^2 - 4\eta\left\langle\boldsymbol{\Delta}_t, \boldsymbol{M}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top\right\rangle + \mathcal{O}(\eta^2). \tag{4.2}$$

Therefore, to establish the convergence of $\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}$, it suffices to provide a reasonable lower-bound for $\left\langle\boldsymbol{\Delta}_t, \boldsymbol{M}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top\right\rangle$. This can be achieved via the following descent lemma:

Lemma 62 (Descent lemma, informal)**.** *Suppose that* $\frac{\sqrt{\sigma_r^\star}}{2} \le \sigma_r(\boldsymbol{S}_t)$, $\|\boldsymbol{S}_t\| \le 2\sqrt{\sigma_1^\star}$, *and* $\|\boldsymbol{V}_t - \boldsymbol{V}^\star\| \le 0.1$. *Then, we have*

$$\left\langle\boldsymbol{\Delta}_t, \boldsymbol{M}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top\right\rangle \ge \frac{\sigma_r^\star}{15}\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}^2 - O\left(\sqrt{\frac{\sigma_r^{\star 3}\mu r^2}{p}}\|\boldsymbol{E}_t\| + \sqrt{r}\sigma_1^\star\|(\mathrm{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\|\right)\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}.$$

By combining the descent lemma with Equation (4.2), we arrive at the following expres-

sion:

$$\|\boldsymbol{\Delta}_{t+1}\|_{\mathrm{F}} \leq \left(1 - \frac{\eta\sigma_r^\star}{10}\right)\|\boldsymbol{\Delta}_t\|_{\mathrm{F}} + \mathcal{O}(\eta)\left(\sqrt{\frac{\sigma_r^{\star 3}\mu r^2}{p}}\,\|\boldsymbol{E}_t\| + \sqrt{r}\sigma_1^\star\,\|(\mathrm{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\|\right) + \mathcal{O}(\eta^2). \tag{4.3}$$

The above inequality holds once the conditions $\frac{\sqrt{\sigma_r^\star}}{2} \leq \sigma_r(\boldsymbol{S}_t) \leq \|\boldsymbol{S}_t\| \leq 2\sqrt{\sigma_1^\star}$ and $\|\boldsymbol{V}_t - \boldsymbol{V}^\star\| \leq 0.1$ are met. These conditions entail that during the initial phase of the algorithm, $\sigma_r(\boldsymbol{S}_t)$ must undergo a fast growth, whereas $\boldsymbol{V}_t$ must remain close to $\boldsymbol{V}^\star$. Under these conditions, GD enters a fast linear convergence phase, provided that $\|(\mathrm{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\| \ll \|\boldsymbol{\Delta}_t\|_{\mathrm{F}}$. In fact, we can readily establish that $\|(\mathrm{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\| \leq c\,\|\boldsymbol{\Delta}_t\|$ for some $c > 0$. However, the challenge lies in ensuring that c remains sufficiently small so as not to negate the effect of a constant factor improvement $1 - 0.1\eta\sigma_r^\star$ in Equation (4.3). This phase continues until GD reaches an error level controlled by $\|\boldsymbol{E}_t\|$. Therefore, to prove the convergence of GD, we need to establish the following properties:

- **Fast growth of $\boldsymbol{S}_t$:** Recall that $\max\{\|\boldsymbol{S}_0\|, \|\boldsymbol{E}_0\|\} = \mathcal{O}(\alpha)$. We need to ensure efficient growth of $\sigma_r(\boldsymbol{S}_t)$ from $\mathcal{O}(\alpha)$ to $\frac{\sqrt{\sigma_r^\star}}{2}$, while keeping $\|\boldsymbol{S}_t\|$ below $2\sqrt{\sigma_1^\star}$.
- **Slow growth of $\boldsymbol{E}_t$:** We need to show that while the signal term grows rapidly, the residual term $\boldsymbol{E}_t$ grows at a much slower rate. Specifically, we will demonstrate that $T = O\left(\frac{1}{\eta\sigma_r^\star}\log(\frac{1}{\alpha})\right)$ suffices to ensure $\frac{\sqrt{\sigma_r^\star}}{2} \leq \sigma_r(\boldsymbol{S}_t) \leq \|\boldsymbol{S}_t\| \leq 2\sqrt{\sigma_1^\star}$ while keeping $\|\boldsymbol{E}_t\| = \mathcal{O}(\alpha)$.
- **Small values of $\|(\mathrm{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\|$ and $\|\boldsymbol{V}_t - \boldsymbol{V}^\star\|$:** Equally important is maintaining control over $\|(\mathrm{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\|$ and $\|\boldsymbol{V}_t - \boldsymbol{V}^\star\|$. While $\|\boldsymbol{V}_t - \boldsymbol{V}^\star\| \leq 0.1$ is needed as a crucial condition for Equation (4.3), the value of $\|(\mathrm{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\|$ directly controls the convergence rate of GD.

To establish the above properties, we provide the one-step dynamics of $\boldsymbol{S}_t$, $\boldsymbol{E}_t$, and $(\mathrm{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)$.

Lemma 63 (One-step dynamics, informal)**.** *Under the conditions of Theorem 20, the fol-*

lowing inequalities hold with an overwhelming probability:

$$\|\boldsymbol{S}_{t+1}\| \le \left(1+\eta\left(\sigma_1^\star - \|\boldsymbol{S}_t\|^2 + O\left(\sigma_1^\star \|\boldsymbol{V}_t-\boldsymbol{V}^\star\| + \|(\mathrm{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\|\right)\right)\right)\|\boldsymbol{S}_t\|, \quad (4.4)$$

$$\sigma_r(\boldsymbol{S}_{t+1}) \ge \left(1+\eta\left(\sigma_r^\star - \sigma_r^2(\boldsymbol{S}_t) - O\left(\sigma_1^\star \|\boldsymbol{V}_t-\boldsymbol{V}^\star\| + \|(\mathrm{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\|\right)\right)\right)\sigma_r(\boldsymbol{S}_t) + \mathcal{O}(\eta)\cdot\left(\sigma_1^\star\|\boldsymbol{V}_t-\boldsymbol{V}^\star\| + \|(\mathrm{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\|\right)\|\boldsymbol{E}_t\|, \quad (4.5)$$

$$\|\boldsymbol{E}_{t+1}\| \le \left(1+\mathcal{O}(\eta)\cdot\left(\sigma_1^\star\|\boldsymbol{V}_t-\boldsymbol{V}^\star\| + \|(\mathrm{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\|\right)\right)\|\boldsymbol{E}_t\|, \quad (4.6)$$

$$\|(\mathrm{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\| \le O\left(\sqrt{\frac{d}{p}}\|\boldsymbol{\Delta}_t\|\left(\|\boldsymbol{V}^\star\|_{2,\infty}^2 + \|\boldsymbol{V}_t\|_{2,\infty}^2\right) + \sqrt{\frac{\sigma_1^\star\mu r}{p}}\|\boldsymbol{E}_t\|\right). \quad (4.7)$$

Next, we provide a high-level overview of how Lemma 63 can be used to establish the aforementioned properties. To this goal, we only focus on the initial phase of the algorithm, where both $\boldsymbol{S}_t$ and $\boldsymbol{E}_t$ are small. A more formal analysis for the entire trajectory is provided in Appendix 4.B.

To use Lemma 63, it suffices to control two key quantities: $\|\boldsymbol{V}_t\|_{2,\infty}$ and $\|\boldsymbol{V}_t-\boldsymbol{V}^\star\|$. To illustrate this, let us assume that $\boldsymbol{V}_t$ inherits the incoherence of $\boldsymbol{V}^\star$, that is, $\|\boldsymbol{V}_t\|_{2,\infty} \le \mathcal{O}(\sqrt{\mu r/d}) \ll 1$ for all $1 \le t \le T$. Then, Equation (4.7) suggests that $\|(\mathrm{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\| \le c\|\boldsymbol{\Delta}_t\| + \mathcal{O}(\alpha)$, where $c = \mathcal{O}(\mu r/\sqrt{pd}) \ll 1$, thereby ensuring the necessary control over $\|(\mathrm{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\|$.

On the other hand, the small values of $\|(\mathrm{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\|$ and $\|\boldsymbol{V}_t-\boldsymbol{V}^\star\|$ play crucial roles in controlling the behavior of the signal and residual terms. To illustrate this, let us assume that at a certain point, $\|\boldsymbol{S}_t\| \ge 1.5\sqrt{\sigma_1^\star}$. Given that we have considered $\|(\mathrm{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\|$ and $\|\boldsymbol{V}_t-\boldsymbol{V}^\star\|$ to be small, Equation (4.4) simplifies to $\|\boldsymbol{S}_{t+1}\| \le (1-\eta\Omega(\sigma_1^\star))\|\boldsymbol{S}_t\| + \mathcal{O}(\eta\sigma_1^\star\|\boldsymbol{E}_t\|)$, effectively preventing $\|\boldsymbol{S}_{t+1}\|$ from further growth. With a similar reasoning, Equation (4.5) simplifies to $\sigma_r(\boldsymbol{S}_{t+1}) \ge (1+\eta\Omega(\sigma_r^\star))\sigma_r(\boldsymbol{S}_t)$. Here, we have leveraged the assumption that $\sigma_r^2(\boldsymbol{S}_t) \ll \sigma_r^\star$ during the initial phase. This implies that $\sigma_r(\boldsymbol{S}_t)$ grows at a rate of $1+\eta\Omega(\sigma_r^\star)$. In contrast, Equation (4.6) implies that $\|\boldsymbol{E}_t\|$ grows at a rate of $1+\eta\mathcal{O}(\|(\mathrm{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\| + \sigma_1^\star\|\boldsymbol{V}_t-\boldsymbol{V}^\star\|)$, which is significantly slower than the growth rate of $\sigma_r(\boldsymbol{S}_t)$ because $\max\{\|(\mathrm{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\|, \sigma_1^\star\|\boldsymbol{V}_t-\boldsymbol{V}^\star\|\} \ll \sigma_r^\star$. It is due to this discrepancy in the growth rates of $\|\boldsymbol{S}_t\|$ and $\|\boldsymbol{E}_t\|$ that GD enters the local linear convergence rate and achieves a final error of $\mathcal{O}(\alpha)$.

4.3.2 Refined Leave-one-out Analysis with Weak Coupling

Indeed, it is not immediately evident why both $\|\boldsymbol{V}_t\|_{2,\infty}$ and $\|\boldsymbol{V}_t-\boldsymbol{V}^\star\|$ would remain small. In fact, our initial intuition might suggest the opposite: recall that GD takes an aggressively large step size. Consequently, even a single GD update has the potential to disrupt the

incoherence of $\boldsymbol{V}_t$. Our key contribution is to establish that such disruption does not occur, even when the iterates are arbitrarily far from the ground truth. In essence, we show that despite the gradient update $\|\eta\mathcal{R}_\Omega(\boldsymbol{\Delta}_t)\boldsymbol{U}_t\|$ potentially having a magnitude of $\widetilde{\Omega}(1)$, its impact on $\boldsymbol{V}_t$ is distributed fairly evenly across its elements. As a result, it has minimal influence on $\|\boldsymbol{V}_t\|_{2,\infty}$ and $\|\boldsymbol{V}_t - \boldsymbol{V}^\star\|$.

We start by showing that a small $\|\boldsymbol{V}_t\|_{2,\infty}$ implies a small $\|\boldsymbol{V}_t - \boldsymbol{V}^\star\|$.

Lemma 64 (Small $\|\boldsymbol{V}_t\|_{2,\infty}$ implies small $\|\boldsymbol{V}_t - \boldsymbol{V}^\star\|_{\mathrm{F}}$, informal)**.** *Suppose that* $\|\boldsymbol{S}_t\| \le 2\sqrt{\sigma_1^\star}$ *and* $\|\boldsymbol{V}_t - \boldsymbol{V}^\star\| \le \frac{1}{2\kappa}$. *With an overwhelming probability, we have*

$$\|\boldsymbol{V}_{t+1} - \boldsymbol{V}^\star\|_{\mathrm{F}} \le \|\boldsymbol{V}_t - \boldsymbol{V}^\star\|_{\mathrm{F}} + \mathcal{O}(\eta)\cdot\sigma_1^\star\sqrt{\frac{dr}{p}}\left(\|\boldsymbol{V}^\star\|_{2,\infty}^2 + \|\boldsymbol{V}_t\|_{2,\infty}^2\right) + \mathcal{O}(\eta^2).$$

Recall that, due to the incoherence of $\boldsymbol{V}^\star$, we have $\|\boldsymbol{V}^\star\|_{2,\infty} = \sqrt{\mu r/d}$. Now, suppose we can further establish that $\boldsymbol{V}_t$ enjoys a similar incoherence property. In such a case, the aforementioned lemma leads to $\|\boldsymbol{V}_t - \boldsymbol{V}^\star\| \le \|\boldsymbol{V}_t - \boldsymbol{V}^\star\|_{\mathrm{F}} \le O\left(T\cdot\eta\sigma_1^\star\sqrt{\frac{\mu^2 r^3}{pd}}\right) + \mathcal{O}(T\cdot\eta^2)$ for every $1 \le t \le T$. Given the provided bounds on T and p, this automatically establishes that $\|\boldsymbol{V}_t - \boldsymbol{V}^\star\|$ remains small throughout the iterations. Therefore, it suffices to control the incoherence of $\|\boldsymbol{V}_t\|_{2,\infty}$.

Controlling $\|\boldsymbol{V}_t\|_{2,\infty}$, which necessitates estimating the ℓ_2-norm of each row, requires a more fine-grained analysis than what is needed for the Frobenius norm. The primary challenge lies in the intricate correlations between the orthogonal matrix $\boldsymbol{V}_t$ and the random observation set Ω, which preclude the straightforward application of classical concentration inequalities. To effectively decouple these correlations, we propose a technique called *weakly-coupled leave-one-out analysis*. Before introducing our proposed methodology, it is essential to grasp the core principles of the classical leave-one-out analysis.

Local leave-one-out analysis. When the search rank is exactly parameterized ($r = r'$) and the initial point is sufficiently close to the ground truth $\boldsymbol{U}_0\boldsymbol{U}_0^\top \approx \boldsymbol{X}^\star$, Ma et al. [2018] established the incoherence of the iterates via the following leave-one-out sequences $\left\{\boldsymbol{U}_t^{(l)}\right\}_{t=0}^T$ for each $1 \le l \le d$:

$$\mathrm{dist}\left(\boldsymbol{U}_0, \boldsymbol{U}_0^{(l)}\right) \approx 0, \quad \text{and} \quad \boldsymbol{U}_{t+1}^{(l)} = \left(\boldsymbol{I} - \eta\mathcal{R}_{\Omega^{(l)}}\left(\boldsymbol{U}_t^{(l)}\boldsymbol{U}_t^{(l)\top} - \boldsymbol{X}^\star\right)\right)\boldsymbol{U}_t^{(l)}, \tag{4.8}$$

where $\mathcal{R}_{\Omega^{(l)}}$ is the leave-one-out projection operator defined by

$$\mathcal{R}_{\Omega^{(l)}} = \frac{1}{2p}\left(\mathcal{P}_{\Omega^{(l)}} + \mathcal{P}_{\Omega^{(l)}}^{\top}\right), \quad \text{and} \quad [\mathcal{P}_{\Omega^{(l)}}(\boldsymbol{X})]_{i,j} = \begin{cases} pX_{i,j} & \text{if } i = l \text{ or } j = l, \\ X_{i,j} & \text{if } (i,j) \in \Omega,\ i \neq l, \text{ and } j \neq l, \\ 0 & \text{otherwise.} \end{cases}$$

The sole distinction between the projection operators $\mathcal{R}_{\Omega^{(l)}}$ and $\mathcal{R}_{\Omega}$ is in their l-th row and l-th column: in contrast to $\mathcal{R}_{\Omega}$, the l-th row and l-th column of $\mathcal{R}_{\Omega^{(l)}}(\boldsymbol{X})$ are *deterministically* set to match the corresponding values of $\boldsymbol{X}$. This seemingly minor adjustment yields two important consequences: first, it ensures that $\boldsymbol{U}_t \approx \boldsymbol{U}_t^{(l)}$, and second, it guarantees that the behavior of $\boldsymbol{U}_t^{(l)}$ remains independent of the random measurements in the l-th row and l-th column. This decoupling technique is the key to controlling the deviation of $\|\boldsymbol{U}_t\|_{2,\infty}$. To formalize this intuition, let us define $\boldsymbol{U}^\star = \boldsymbol{V}^\star \boldsymbol{\Sigma}^{\star 1/2}$. One can write

$$\begin{aligned} \|\boldsymbol{U}_t\|_{2,\infty} &= \max_{1\le l\le d}\left\{\left\|\left[\boldsymbol{U}_t\boldsymbol{H}_t^{(l)}\right]_{l,\cdot}\right\|\right\} \\ &\le \max_{1\le l\le d}\left\{\left\|\left[\boldsymbol{U}^\star - \boldsymbol{U}_t^{(l)}\boldsymbol{R}_t^{(l)}\right]_{l,\cdot}\right\| + \left\|\left[\boldsymbol{U}_t\boldsymbol{H}_t^{(l)} - \boldsymbol{U}_t^{(l)}\boldsymbol{R}_t^{(l)}\right]_{l,\cdot}\right\| + \left\|[\boldsymbol{U}^\star]_{l,\cdot}\right\|\right\} \\ &\le \underbrace{\max_{1\le l\le d}\left\{\left\|\left(\boldsymbol{U}^\star - \boldsymbol{U}_t^{(l)}\boldsymbol{R}_t^{(l)}\right)_{l,\cdot}\right\|\right\}}_{\text{leave-one-out error}} + \underbrace{\max_{1\le l\le d}\left\{\left\|\boldsymbol{U}_t\boldsymbol{H}_t^{(l)} - \boldsymbol{U}_t^{(l)}\boldsymbol{R}_t^{(l)}\right\|_{\mathrm{F}}\right\}}_{\text{proximal error}} + \sqrt{\frac{\sigma_1^\star \mu r}{d}}. \end{aligned} \tag{4.9}$$

Here $\boldsymbol{R}_t^{(l)}$ and $\boldsymbol{H}_t^{(l)}$ are orthogonal matrices defined as $\boldsymbol{R}_t^{(l)} = \arg\min_{\boldsymbol{O}\in\mathcal{O}_{r\times r}} \left\|\boldsymbol{U}_t^{(l)}\boldsymbol{O} - \boldsymbol{U}^\star\right\|_{\mathrm{F}}$ and $\boldsymbol{H}_t^{(l)} = \arg\min_{\boldsymbol{O}\in\mathcal{O}_{r\times r}} \left\|\boldsymbol{U}_t^{(l)}\boldsymbol{O} - \boldsymbol{U}_t^{(l)}\boldsymbol{R}_t^{(l)}\right\|_{\mathrm{F}}$. Although it may not be immediately obvious, it can be shown that the l-th row of the matrix $\boldsymbol{U}^\star - \boldsymbol{U}_t^{(l)}\boldsymbol{R}_t^{(l)}$ is purely *deterministic*. Consequently, it becomes possible to effectively control the leave-one-out error. To tackle the proximal error, recall that the initial point $\boldsymbol{U}_0\boldsymbol{U}_0^\top$ is assumed to be close to $\boldsymbol{X}^\star$. Within this region, the local landscape exhibits restricted strong convexity. This ensures that the true iterates $\boldsymbol{U}_t$ and the leave-one-out versions $\boldsymbol{U}_t^{(l)}$ become increasingly close, leading to a small proximal error. By combining these two arguments, we can guarantee the incoherence of the true iterates. Furthermore, the incoherence of $\boldsymbol{U}_t$ automatically implies the incoherence of $\boldsymbol{V}_t$, given that $\sqrt{\sigma_r^\star}\|\boldsymbol{V}_t\|_{2,\infty} \lesssim \|\boldsymbol{U}_t\|_{2,\infty}$ when $\boldsymbol{U}_t\boldsymbol{U}_t^\top \approx \boldsymbol{X}^\star$. For more details, we refer interested readers to the discussions in [Ma et al., 2018].

While the classical leave-one-out analysis provides precise *local* guarantees within the *exactly-parameterized* regimes, we shed light on its limitations when applied *globally* in the *over-parameterized* settings. A significant challenge arises from the discrepancy of the sin-

gular values of $\boldsymbol{U}_t$ and $\boldsymbol{U}_t^{(l)}$: although they may remain close to the singular values of $\boldsymbol{U}^\star$ in the local regime, they can undergo substantial changes when positioned far from the true solution. Consequently, the original measure of proximal error based on $\mathrm{dist}(\boldsymbol{U}_t, \boldsymbol{U}_t^{(l)})$ loses its effectiveness as a reliable metric.

Instead, recall that we only require controlling $\boldsymbol{V}_t$, which unlike $\boldsymbol{U}_t$, has *unit* singular values. This motivates us to switch to a more stable metric—the divergence between the left column spaces of $\boldsymbol{U}_t$ and $\boldsymbol{U}_t^{(l)}$. However, an additional complication is that these left column spaces may also not align perfectly due to over-parameterization. Fortunately, by resorting to our proposed dynamic signal-residual decomposition, we can show that the iterates $\boldsymbol{U}_t$ are well-approximated by the low-rank signal $\boldsymbol{U}_t \approx \boldsymbol{S}_t$. Therefore, it suffices to focus on controlling the discrepancy in the column spaces of $\boldsymbol{S}_t$ and $\boldsymbol{S}_t^{(l)}$, i.e., $\mathrm{dist}(\boldsymbol{V}_t, \boldsymbol{V}_t^{(l)})$. However, the new proximal error $\mathrm{dist}(\boldsymbol{V}_t, \boldsymbol{V}_t^{(l)})$ can still grow exponentially. To explain the root cause of this exponential growth, we employ matrix Taylor expansion to derive the first-order approximations for $\boldsymbol{V}_{t+1}$ and $\boldsymbol{V}_{t+1}^{(l)}$:

$$\boldsymbol{V}_{t+1} = \boldsymbol{V}_t + \eta \mathcal{P}_{\boldsymbol{V}_t}^{\perp} \boldsymbol{M}_t \boldsymbol{V}_t + \mathcal{O}(\eta^2) \quad \text{and} \quad \boldsymbol{V}_{t+1}^{(l)} = \boldsymbol{V}_t^{(l)} + \eta \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp} \boldsymbol{M}_t^{(l)} \boldsymbol{V}_t^{(l)} + \mathcal{O}(\eta^2), \tag{4.10}$$

where we define $\boldsymbol{M}_t^{(l)} = \mathcal{R}_{\Omega^{(l)}}\big(\boldsymbol{X}^\star - \boldsymbol{U}_t^{(l)} \boldsymbol{U}_t^{(l)\top}\big)$. To effectively control the proximal error, it is crucial to establish an upper bound for $\big\|\boldsymbol{M}_t - \boldsymbol{M}_t^{(l)}\big\|$. This distance tends to concentrate around $\big\|\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{U}_t^{(l)}\boldsymbol{U}_t^{(l)\top}\big\| \approx \big\|\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{S}_t^{(l)}\boldsymbol{S}_t^{(l)\top}\big\|$ when the sampling rate p is sufficiently large. However, as previously noted, the singular values of $\boldsymbol{S}_t$ and $\boldsymbol{S}_t^{(l)}$ may diverge. This misalignment can lead to $\big\|\boldsymbol{M}_t - \boldsymbol{M}_t^{(l)}\big\| = \Omega(\sigma_1^\star)$ in the worst case. Hence, the proximal error can grow exponentially.

Weakly-coupled leave-one-out analysis. To remedy the alignment challenges identified earlier, we propose the following refined leave-one-out sequences $\big\{\widetilde{\boldsymbol{V}}_t^{(l)}\big\}_{t=0}^T$:

$$\widetilde{\boldsymbol{V}}_{t+1}^{(l)} = \widetilde{\boldsymbol{Z}}_{t+1}^{(l)} \Big(\widetilde{\boldsymbol{Z}}_{t+1}^{(l)\top} \widetilde{\boldsymbol{Z}}_{t+1}^{(l)}\Big)^{-1/2} \quad \text{where} \quad \widetilde{\boldsymbol{Z}}_{t+1}^{(l)} = \Big(\boldsymbol{I} + \eta \widetilde{\boldsymbol{M}}_t^{(l)}\Big) \widetilde{\boldsymbol{V}}_t^{(l)} \text{ and } \widetilde{\boldsymbol{V}}_0^{(l)} = \boldsymbol{V}^\star. \tag{4.11}$$

In this context, $\widetilde{\boldsymbol{M}}_t^{(l)}$ is defined as:

$$\widetilde{\boldsymbol{M}}_t^{(l)} = \mathcal{R}_{\Omega^{(l)}}\Big(\boldsymbol{X}^\star - \widetilde{\boldsymbol{V}}_t^{(l)} \boldsymbol{\Sigma}_t \widetilde{\boldsymbol{V}}_t^{(l)\top}\Big) \quad \text{where } \boldsymbol{\Sigma}_t = \boldsymbol{V}_t^\top \boldsymbol{U}_t \boldsymbol{U}_t^\top \boldsymbol{V}_t \in \mathbb{R}^{r\times r}. \tag{4.12}$$

Compared to the original $\boldsymbol{M}_t^{(l)}$, we replace $\boldsymbol{\Sigma}_t^{(l)} = \boldsymbol{V}_t^{(l)\top}\boldsymbol{U}_t^{(l)}\boldsymbol{U}_t^{(l)\top}\boldsymbol{V}_t^{(l)}$ by $\boldsymbol{\Sigma}_t = \boldsymbol{V}_t^\top \boldsymbol{U}_t \boldsymbol{U}_t^\top \boldsymbol{V}_t$ in the definition of $\widetilde{\boldsymbol{M}}_t^{(l)}$. Our analysis indicates that this adjustment significantly improves the control over the distance $\big\|\boldsymbol{M}_t - \widetilde{\boldsymbol{M}}_t^{(l)}\big\| = o(1)$ when the sampling rate p is sufficiently

large. Hence, the proximal error grows at a much slower rate.

Despite their promise, these refined leave-one-out sequences do introduce a trade-off: the statistical independence inherent in the original leave-one-out sequences is compromised due to the inclusion of $\boldsymbol{\Sigma}_t$. In other words, the l-th leave-one-out sequence $\left\{\widetilde{\boldsymbol{V}}_t^{(l)}\right\}_{t=0}^{T}$ is no longer independent of the random measurements in the l-th row and l-th column. Nonetheless, we demonstrate that the resulting correlation is relatively weak, primarily because $\boldsymbol{\Sigma}_t$ is a comparatively small $r \times r$ matrix. To control this statistical coupling, we employ a novel *adaptive covering argument*, which can be of independent interest. This approach effectively mitigates the statistical coupling while incurring a mild increase in the required sample complexity, which remains only polynomial in r.

To formalize our arguments, we can decompose the refined leave-one-out sequences $\|\boldsymbol{V}_t\|_{2,\infty}$ as:

$$\begin{aligned}\|\boldsymbol{V}_t\|_{2,\infty} &\le \|\boldsymbol{V}^\star - \boldsymbol{V}_t\|_{2,\infty} + \|\boldsymbol{V}^\star\|_{2,\infty} \\ &= \max_{1\le l\le d}\left\{\left\|(\boldsymbol{V}^\star - \boldsymbol{V}_t)_{l,\cdot}\right\|\right\} + \|\boldsymbol{V}^\star\|_{2,\infty} \\ &\le \underbrace{\max_{1\le l\le d}\left\{\left\|\left(\boldsymbol{V}^\star - \widetilde{\boldsymbol{V}}_t^{(l)}\right)_{l,\cdot}\right\|\right\}}_{\text{refined leave-one-out error (Proposition 12)}} + \underbrace{\max_{1\le l\le d}\left\{\left\|\boldsymbol{V}_t - \widetilde{\boldsymbol{V}}_t^{(l)}\right\|_{\mathrm{F}}\right\}}_{\text{refined proximal error (Proposition 13)}} + \sqrt{\frac{\mu r}{d}}.\end{aligned} \tag{4.13}$$

Next, we characterize the dynamic of the refined leave-one-out error.

Proposition 12 (Refined leave-one-out error)**.** *Suppose that* $p \gtrsim \frac{\log(d)}{d}$ *and* $\|\boldsymbol{V}^\star - \boldsymbol{V}_t\| \le \frac{1}{2\kappa}$. *With probability at least* $1 - \frac{1}{d^3}$, *for any* $1 \le t \le T \lesssim \frac{1}{\eta\sigma_r^\star}\log\left(\frac{1}{\alpha}\right)$ *and* $1 \le l \le d$, *we have*

$$\left\|\left(\boldsymbol{V}^\star - \widetilde{\boldsymbol{V}}_{t+1}^{(l)}\right)_{l,\cdot}\right\| \le (1 - 0.5\eta\sigma_r^\star)\left\|\left(\boldsymbol{V}^\star - \widetilde{\boldsymbol{V}}_t^{(l)}\right)_{l,\cdot}\right\| + \mathcal{O}(\eta)\cdot\sigma_1^\star\frac{\kappa\mu^{1.5}r^2\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd^2}}.$$

A simple inductive argument based on Proposition 12 reveals that the following inequality holds with an overwhelming probability for all $1 \le t \le T$:

$$\left\|\left(\boldsymbol{V}^\star - \widetilde{\boldsymbol{V}}_t^{(l)}\right)_{l,\cdot}\right\| \lesssim \frac{\kappa^2\mu^{1.5}r^2\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd^2}} \le \sqrt{\frac{\mu r}{4d}}, \quad \text{assuming} \quad p \gtrsim \frac{\kappa^4\mu^2 r^3\log^2\left(\frac{1}{\alpha}\right)\log(d)}{d}.$$

Next, we characterize the dynamic of the refined proximal error.

Proposition 13 (Refined proximal error)**.** *Suppose that* $p \gtrsim \frac{\kappa^6\mu^4 r^9\log^6(\frac{1}{\alpha})\log^2(d)}{d}$, $\|\boldsymbol{V}^\star - \boldsymbol{V}_t\| \le \frac{1}{2\kappa}$, *and* $\left\|\boldsymbol{V}_t - \widetilde{\boldsymbol{V}}_t^{(l)}\right\|_{\mathrm{F}} \le \sqrt{\frac{\mu r}{4d}}$. *With probability at least* $1 - \frac{1}{d^3}$, *for any* $1 \le t \le T \lesssim \frac{1}{\eta\sigma_r^\star}\log\left(\frac{1}{\alpha}\right)$

and $1 \le l \le d$*, we have*

$$\left\|\boldsymbol{V}_{t+1} - \widetilde{\boldsymbol{V}}_{t+1}^{(l)}\right\|_{\mathrm{F}} \le \left\|\boldsymbol{V}_t - \widetilde{\boldsymbol{V}}_t^{(l)}\right\|_{\mathrm{F}} + \mathcal{O}(\eta)\cdot\sigma_1^\star\sqrt{\frac{\kappa\mu^3 r^{5.5}\log\left(\frac{1}{\alpha}\right)\log(d)}{\sqrt{pd}\cdot d}}.$$

The above proposition implies that

$$\left\|\boldsymbol{V}_t - \widetilde{\boldsymbol{V}}_t^{(l)}\right\|_{\mathrm{F}} \lesssim \eta\sigma_1^\star\sqrt{\frac{\kappa\mu^3 r^{5.5}\log\left(\frac{1}{\alpha}\right)\log(d)}{\sqrt{pd}\cdot d}}\cdot T \lesssim \sqrt{\frac{\kappa^3\mu^3 r^{5.5}\log^3\left(\frac{1}{\alpha}\right)\log(d)}{\sqrt{pd}\cdot d}} \le \sqrt{\frac{\mu r}{4d}}.$$

Combining the above inequalities with the proposed decomposition in Equation (4.13) leads to:

$$\|\boldsymbol{V}_t\|_{2,\infty} \le \sqrt{\frac{\mu r}{4d}} + \sqrt{\frac{\mu r}{4d}} + \sqrt{\frac{\mu r}{d}} \le \sqrt{\frac{4\mu r}{d}}, \quad \text{with probably at least } 1 - \frac{2}{d^3}.$$

This establishes the incoherence of $\boldsymbol{V}_t$ for all $1 \le t \le T$.

4.3.3 Improved Results for Exact Parameterization

Finally, we show that our analysis in the over-parameterized regime, combined with the following local convergence result for the exact parameterization regime by Ma et al. [2018], readily establishes the proof of Theorem 21.

Theorem 22 (Local convergence of GD [Ma et al., 2018, Theorem 2])**.** *Consider MC with search rank* $r' = r$*. Suppose that the sampling rate satisfies* $p \gtrsim \frac{\mu^3 r^3 \log^3(d)}{d}$*. Consider the iterates of GD with step-size* $\eta \le \frac{2}{25\kappa\sigma_1^\star}$*. Suppose that there exists* $t_0 \ge 0$ *such that* $\boldsymbol{U}_{t_0}$ *and the leave-one-out sequences* $\left\{\boldsymbol{U}_{t_0}^{(l)}\right\}_{l=0}^{d}$ *defined in Equation* (4.8) *satisfy:*

$$\operatorname{dist}\left(\boldsymbol{U}_{t_0}, \boldsymbol{U}^\star\right) \le O\left(\sqrt{\frac{\sigma_r^\star\mu^3 r^3\log(d)}{pd^2}}\right), \tag{4.14}$$

$$\max\left\{\operatorname{dist}\left(\boldsymbol{U}_{t_0}, \boldsymbol{U}_{t_0}^{(l)}\right), \operatorname{dist}\left(\boldsymbol{U}_{t_0}^{(l)}, \boldsymbol{U}^\star\right)\right\} \le O\left(\sqrt{\frac{\sigma_r^\star\mu^3 r^3\log(d)}{pd^2}}\right), \quad \textit{for all } 1 \le l \le d. \tag{4.15}$$

With probability at least $1 - O\left(\frac{1}{d^3}\right)$*, for all* $t_0 \le t \le t_0 + \mathcal{O}(d^5)$*, we have*

$$\left\|\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{X}^\star\right\|_{\mathrm{F}} \le \left\|\boldsymbol{U}_{t_0}\boldsymbol{U}_{t_0}^\top - \boldsymbol{X}^\star\right\|_{\mathrm{F}}\left(1 - 0.2\eta\sigma_r^\star\right)^{t-t_0}.$$

To prove Theorem 21, it suffices to show that the conditions of the above theorem are met at a certain iteration $0 \leq t_0 \leq T$. This can be achieved by leveraging our result for the over-parameterized regime. In particular, upon choosing $\alpha = c \cdot \frac{\sigma_r^\star}{\kappa^{1.5} d}$ for sufficiently small $c > 0$ in Theorem 20, one can show that both Conditions (4.14) and (4.15) are satisfied with an overwhelming probability after $t_0 = \widetilde{O}\big(\frac{1}{\eta \sigma_r^\star}\big)$ iterations. From this iteration onward, Theorem 22 shows that the iterations of GD enter a local linear convergence regime, which readily establishes the final result of Theorem 21.

4.4 Conclusion and Future Directions

In this paper, we prove the convergence of vanilla gradient descent (GD) with small initialization for symmetric matrix completion. Existing convergence results for this problem typically require explicit regularization or precise initializations. However, our work proves that neither condition is necessary for GD to converge. Moreover, our results also apply to the over-parameterized regime, where the rank of the true solution is unknown and over-estimated instead.

Although our required sample complexity $\widetilde{\mathcal{O}}(dr^9)$ is optimal with respect to the dimension d, it remains sub-optimal with respect to the rank r. Specifically, it exceeds the sample complexity of regularized GD, which stands at $\widetilde{\mathcal{O}}(dr^2)$ [Chen and Wainwright, 2015]. We expect our analysis can be sharpened to achieve a similar sample complexity.

We anticipate that our findings will pave the way for broader results extending beyond symmetric matrix completion. In particular, our proposed weakly-coupled leave-one-out analysis relaxes several stringent conditions of classical leave-one-out analysis, making it highly applicable for the global analysis of GD. We believe that this approach, along with potential variations, holds promise for explaining the favorable performance of GD or its variants in various statistical learning problems.

4.A Preliminaries

4.A.1 Otuline of the Appendix

The structure of the appendix is as follows. In the remainder of this section, we introduce additional notation. Following this, we present key intermediate lemmas (Lemmas 65 to 68) crucial for our main proofs. Section 4.B delves into a detailed proof of the signal and residual dynamics, starting with a refined version of Lemma 63 (Proposition 14) that takes into account the incoherence of $\boldsymbol{V}_t$. Additionally, the proof of Lemma 62 is provided in this

section. Moving on to Section 4.C, we present the proofs of our main theorems. Section 4.D presents the key novelty of our paper, focusing on establishing the incoherence of $\boldsymbol{V}_t$ via weakly-coupled leave-one-out analysis. The validation of different initialization schemes, as presented in Lemma 61, is addressed in Section 4.E. In Section 4.F, we compile several known results on matrix completion crucial to our arguments. Lastly, Section 4.G collects several basic lemmas, which we include for completeness.

4.A.2 Additional Notations

We introduce some additional notations that will be used throughout the appendix. The max-norm of $\boldsymbol{X}$, denoted as $\|\boldsymbol{X}\|_{\max}$, is defined as $\max_{i,j}|X_{i,j}|$. We define the operator and Frobenius norm ball as $\mathcal{B}_{\mathrm{op}}^{d_1\times d_2}(r) := \{\boldsymbol{X} \in \mathbb{R}^{d_1\times d_2} : \|\boldsymbol{X}\| \le r\}$ and $\mathcal{B}_{\mathrm{F}}^{d_1\times d_2}(r) := \{\boldsymbol{X} \in \mathbb{R}^{d_1\times d_2} : \|\boldsymbol{X}\|_{\mathrm{F}} \le r\}$, respectively. For any matrix $\boldsymbol{X}$, we denote its SVD as $\boldsymbol{X} = \boldsymbol{L}_{\boldsymbol{X}}\boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{R}_{\boldsymbol{X}}^\top$. We denote $\mathcal{S}_{d\times d}$ as the set of all the symmetric matrices $\boldsymbol{X} \in \mathbb{R}^{d\times d}$. In the appendix, $\Gamma, \Gamma_1, \Gamma_2, \ldots$ denote fixed universal constants, while $C, C_1, C_2, c_1, c_2, \ldots$ represent universal constants whose specific values may vary depending on the context.

Throughout the appendix, our arguments are conditioned on the following good event without further explanation. We define the random observation matrix $\boldsymbol{\Omega}$ as

$$\boldsymbol{\Omega}_{i,j} = \begin{cases} 1 & \text{if } (i,j) \in \Omega, \\ 0 & \text{otherwise.} \end{cases} \tag{4.16}$$

Then, the good event can be defined as

$$E_{\mathrm{good}} = \left\{ \left\| \frac{\boldsymbol{\Omega} + \boldsymbol{\Omega}^\top}{2p} - \boldsymbol{J} \right\| \le \Gamma \sqrt{\frac{d}{p}} \right\}. \tag{4.17}$$

Here $\boldsymbol{J}$ is the all-one matrix. According to Lemma 75, we have $\mathbb{P}(E_{\mathrm{good}}) \ge 1 - \frac{1}{d^3}$.

4.A.3 Important Intermediate Lemmas

Next, we collect some useful intermediate results that will be directly used throughout our proofs. We also note that some of these intermediate results rely on concentration inequalities for matrix completion, which are thoroughly discussed in Appendix 4.F. Before proceeding, we define the following notations

$$\boldsymbol{\Delta}_t = \boldsymbol{X}^\star - \boldsymbol{U}_t\boldsymbol{U}_t^\top, \quad \boldsymbol{M}_t = \mathcal{R}_\Omega\left(\boldsymbol{X}^\star - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right), \quad \text{and} \quad \boldsymbol{\Lambda}_t = \boldsymbol{S}_t\boldsymbol{E}_t^\top + \boldsymbol{E}_t\boldsymbol{S}_t^\top + \boldsymbol{E}_t\boldsymbol{E}_t^\top. \tag{4.18}$$

Moreover, we introduce the following term

$$\boldsymbol{A}_t = -\eta^2 \boldsymbol{M}_t \boldsymbol{V}_t \boldsymbol{V}_t^\top \boldsymbol{M}_t \boldsymbol{V}_t - 0.5\eta^2 \boldsymbol{V}_t \boldsymbol{V}_t^\top \boldsymbol{M}_t^2 \boldsymbol{V}_t - 0.5\eta^3 \boldsymbol{M}_t \boldsymbol{V}_t \boldsymbol{V}^{\star\top} \boldsymbol{M}_t^2 \boldsymbol{V}_t + (\boldsymbol{I} + \eta \boldsymbol{M}_t) \boldsymbol{V}_t \boldsymbol{R}(\boldsymbol{Y}_t) \tag{4.19}$$

where $\boldsymbol{Y}_t = \boldsymbol{V}_t^\top \left(2\eta \boldsymbol{M}_t + \eta^2 \boldsymbol{M}_t^2\right) \boldsymbol{V}_t$ and $\boldsymbol{R}(\boldsymbol{X}) = \sum_{k=2}^{\infty} \frac{(-1)^k (2k)!}{4^k (k!)^2} \boldsymbol{X}^k$. This notion of $\boldsymbol{A}_t$ will be used when controlling the higher-order terms with respect to the step-size η. We are now ready to statement our helper lemmas.

Lemma 65 (Helper lemma for $\boldsymbol{U}_t$)**.** *Suppose that* $\|\boldsymbol{S}_t\| \leq 2\sqrt{\sigma_1^\star}$ *and* $\|\boldsymbol{E}_t\| \leq \sqrt{\frac{\sigma_1^\star \mu r}{d}}$*. Then, we have*

$$\left\|\boldsymbol{U}_t \boldsymbol{U}_t^\top\right\|_{\mathrm{F}} \leq 8\sqrt{r}\sigma_1^\star. \tag{4.20}$$

Proof. Applying triangle inequality, we have

$$\begin{aligned}
\left\|\boldsymbol{U}_t \boldsymbol{U}_t^\top\right\|_{\mathrm{F}} &\leq \left\|\boldsymbol{S}_t \boldsymbol{S}_t^\top\right\|_{\mathrm{F}} + 2\left\|\boldsymbol{S}_t \boldsymbol{E}_t^\top\right\|_{\mathrm{F}} + \left\|\boldsymbol{E}_t \boldsymbol{E}_t^\top\right\|_{\mathrm{F}} \\
&\leq \sqrt{r}\left\|\boldsymbol{S}_t\right\|^2 + 2\sqrt{r}\left\|\boldsymbol{S}_t\right\|\left\|\boldsymbol{E}_t\right\| + \sqrt{d}\left\|\boldsymbol{E}_t\right\|^2 \\
&\leq 8\sqrt{r}\sigma_1^\star.
\end{aligned} \tag{4.21}$$

Here in the last inequality, we use the assumptions $\|\boldsymbol{S}_t\| \leq 2\sqrt{\sigma_1^\star}$ and $\|\boldsymbol{E}_t\| \leq \sqrt{\frac{\sigma_1^\star \mu r}{d}}$ and the fact that $d \gg \mu r$. □

Lemma 66 (Helper lemma for $\boldsymbol{\Lambda}_t$)**.** *Suppose that* $\|\boldsymbol{S}_t\| \leq 2\sqrt{\sigma_1^\star}$*,* $\|\boldsymbol{V}_t\|_{2,\infty} \leq 2\sqrt{\frac{\mu r}{d}}$*, and* $\|\boldsymbol{E}_t\| \leq \sqrt{\frac{\sigma_1^\star \mu r}{81 d}}$*. Then, conditioned on* E_{good}*, we have*

$$\begin{aligned}
\|\boldsymbol{\Lambda}_t\| &\leq 5\sqrt{\sigma_1^\star}\left\|\boldsymbol{E}_t\right\|, \\
\left\|\mathcal{R}_\Omega(\boldsymbol{\Lambda}_t)\right\| &\leq 10\Gamma\sqrt{\frac{\sigma_1^\star \mu r}{p}}\left\|\boldsymbol{E}_t\right\|, \\
\left\|\left(\mathrm{I} - \mathcal{R}_\Omega\right)(\boldsymbol{\Lambda}_t)\right\| &\leq 9\Gamma\sqrt{\frac{\sigma_1^\star \mu r}{p}}\left\|\boldsymbol{E}_t\right\|.
\end{aligned} \tag{4.22}$$

Proof. First, we can bound $\|\boldsymbol{\Lambda}_t\|$ as follows

$$\|\boldsymbol{\Lambda}_t\| \leq \left\|\boldsymbol{S}_t \boldsymbol{E}_t^\top\right\| + \left\|\boldsymbol{E}_t \boldsymbol{S}_t^\top\right\| + \left\|\boldsymbol{E}_t \boldsymbol{E}_t^\top\right\| \leq 5\sqrt{\sigma_1^\star}\left\|\boldsymbol{E}_t\right\| \tag{4.23}$$

where we use the assumptions $\|\boldsymbol{S}_t\| \leq 2\sqrt{\sigma_1^\star}$ and $\|\boldsymbol{E}_t\| \leq \sqrt{\frac{\sigma_1^\star \mu r}{d}} \leq \sqrt{\sigma_1^\star}$. Next, we control

$\|(\mathrm{I}-\mathcal{R}_\Omega)(\boldsymbol{\Lambda}_t)\|$. To this end, we first apply triangle inequality to obtain

$$\begin{aligned}\|(\mathrm{I}-\mathcal{R}_\Omega)(\boldsymbol{\Lambda}_t)\| &\le 2\left\|(\mathrm{I}-\mathcal{R}_\Omega)\left(\boldsymbol{S}_t\boldsymbol{E}_t^\top\right)\right\| + \left\|(\mathrm{I}-\mathcal{R}_\Omega)\left(\boldsymbol{E}_t\boldsymbol{E}_t^\top\right)\right\| \\ &\overset{(a)}{\le} \Gamma\sqrt{\frac{d}{p}}\|\boldsymbol{E}_t\|_{2,\infty}\left(\|\boldsymbol{E}_t\|_{2,\infty}+2\|\boldsymbol{S}_t\|_{2,\infty}\right) \\ &\overset{(b)}{\le} \Gamma\sqrt{\frac{d}{p}}\|\boldsymbol{E}_t\|\left(\|\boldsymbol{E}_t\|+8\sqrt{\frac{\sigma_1^\star\mu r}{d}}\right) \\ &\le 9\Gamma\sqrt{\frac{\sigma_1^\star\mu r}{p}}\|\boldsymbol{E}_t\|.\end{aligned} \tag{4.24}$$

Here in (a), we apply Corollary 7. In (b), we use the facts that $\|\boldsymbol{E}_t\|_{2,\infty} \le \|\boldsymbol{E}_t\| \le \sqrt{\frac{\sigma_1^\star\mu r}{d}}$ and

$$\|\boldsymbol{S}_t\|_{2,\infty} = \left\|\boldsymbol{V}_t\boldsymbol{V}_t^\top\boldsymbol{U}_t\right\|_{2,\infty} \overset{\text{Corollary 7}}{\le} \|\boldsymbol{V}_t\|_{2,\infty}\|\boldsymbol{S}_t\| \le 2\sqrt{\frac{\mu r}{d}}\cdot 2\sqrt{\sigma_1^\star} = 4\sqrt{\frac{\sigma_1^\star\mu r}{d}}. \tag{4.25}$$

In the last inequality, we use the fact that $\|\boldsymbol{S}_t\| \le 2\sqrt{\sigma_1^\star}$ and $\|\boldsymbol{V}_t\|_{2,\infty} \le 2\sqrt{\frac{\mu r}{d}}$. Lastly, applying triangle inequality leads to

$$\begin{aligned}\|\mathcal{R}_\Omega(\boldsymbol{\Lambda}_t)\| &\le \|\boldsymbol{\Lambda}_t\| + \|(\mathrm{I}-\mathcal{R}_\Omega)(\boldsymbol{\Lambda}_t)\| \\ &\le 5\sqrt{\sigma_1^\star}\|\boldsymbol{E}_t\| + 9\Gamma\sqrt{\frac{\sigma_1^\star\mu r}{p}}\|\boldsymbol{E}_t\| \\ &\le 10\Gamma\sqrt{\frac{\sigma_1^\star\mu r}{p}}\|\boldsymbol{E}_t\|,\end{aligned} \tag{4.26}$$

where the last inequality is due to $p \le \frac{1}{25}\Gamma^2\mu r$. □

Lemma 67 (Helper lemma for $\boldsymbol{\Delta}_t$)**.** *Under the same conditions as Lemma 66 with the*

additional assumption that $\|\boldsymbol{V}_t - \boldsymbol{V}^\star\| \le \Gamma_1 \frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}$*, we have*

$$
\begin{aligned}
\|\boldsymbol{\Delta}_t\| &\le 5\sigma_1^\star, \\
\|(\mathrm{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\| &\le 5\Gamma\sqrt{\frac{\mu^2 r^2}{pd}}\|\boldsymbol{\Delta}_t\| + 10\Gamma\sqrt{\frac{\sigma_1^\star \mu r}{p}}\|\boldsymbol{E}_t\|, \\
\|(\mathrm{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\| &\le 21\Gamma\sigma_1^\star\sqrt{\frac{\mu^2 r^2}{pd}}, \\
\|\boldsymbol{M}_t\| &\le \left(1 + 5\Gamma\sqrt{\frac{\mu^2 r^2}{pd}}\right)\|\boldsymbol{\Delta}_t\| + 10\Gamma\sqrt{\frac{\sigma_1^\star\mu r}{p}}\|\boldsymbol{E}_t\|, \\
\|\boldsymbol{M}_t\| &\le 6\sigma_1^\star, \\
\left\|\boldsymbol{M}_t\mathcal{P}_{\boldsymbol{V}_t}^\perp\right\| &\le 2\Gamma_1\sigma_1^\star\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}.
\end{aligned}
\tag{4.27}
$$

Proof. We first control $\|\boldsymbol{\Delta}_t\|$ as follows

$$
\begin{aligned}
\|\boldsymbol{\Delta}_t\| &\le \left\|\boldsymbol{X}^\star - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| + \|\boldsymbol{\Lambda}_t\| \\
&\le \max\{\|\boldsymbol{X}^\star\|, \|\boldsymbol{S}_t\|^2\} + 5\sqrt{\sigma_1^\star}\|\boldsymbol{E}_t\| \\
&\le 4\sigma_1^\star + 5\sqrt{\sigma_1^\star}\|\boldsymbol{E}_t\| \\
&\le 5\sigma_1^\star.
\end{aligned}
\tag{4.28}
$$

Next, we control $\|(\mathrm{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\|$. To this end, applying triangle inequality leads to

$$
\|(\mathrm{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\| \le \underbrace{\left\|(\mathrm{I}-\mathcal{R}_\Omega)\left(\boldsymbol{X}^\star - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right\|}_{:=(\mathrm{I})} + \underbrace{\left\|(\mathrm{I}-\mathcal{R}_\Omega)(\boldsymbol{\Lambda}_t)\right\|}_{:=(\mathrm{II})}. \tag{4.29}
$$

For (I), applying Lemma 78, we have

$$
\begin{aligned}
(\mathrm{I}) &\le \Gamma\sqrt{\frac{d}{p}}\left\|\boldsymbol{X}^\star - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|\left(\|\boldsymbol{V}^\star\|_{2,\infty}^2 + \|\boldsymbol{V}_t\|_{2,\infty}^2\right) \\
&\overset{(a)}{\le} 5\Gamma\sqrt{\frac{\mu^2 r^2}{pd}}\left\|\boldsymbol{X}^\star - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| \\
&\le 5\Gamma\sqrt{\frac{\mu^2 r^2}{pd}}\left(\|\boldsymbol{\Delta}_t\| + \|\boldsymbol{\Lambda}_t\|\right) \\
&\overset{(b)}{\le} 5\Gamma\sqrt{\frac{\mu^2 r^2}{pd}}\|\boldsymbol{\Delta}_t\| + 25\Gamma\sqrt{\frac{\sigma_1^\star\mu^2 r^2}{pd}}\|\boldsymbol{E}_t\|.
\end{aligned}
\tag{4.30}
$$

Here in (a), we use the assumption that $\|\boldsymbol{V}_t\|_{2,\infty} \leq 2\sqrt{\frac{\mu r}{d}}$. In (b), we use the result from Lemma 66 that $\|\boldsymbol{\Lambda}_t\| \leq 5\sqrt{\sigma_1^\star}\,\|\boldsymbol{E}_t\|$. On the other hand, we know that (II) $\leq 9\Gamma\sqrt{\frac{\sigma_1^\star \mu r}{p}}\,\|\boldsymbol{E}_t\|$ due to Lemma 66. Overall, we conclude that

$$\begin{aligned}\|(\mathrm{I}-\mathcal{R}_\Omega)\left(\boldsymbol{\Delta}_t\right)\| &\leq 5\Gamma\sqrt{\frac{\mu^2 r^2}{pd}}\,\|\boldsymbol{\Delta}_t\| + 25\Gamma\sqrt{\frac{\sigma_1^\star \mu^2 r^2}{pd}}\,\|\boldsymbol{E}_t\| + 9\Gamma\sqrt{\frac{\sigma_1^\star \mu r}{p}}\,\|\boldsymbol{E}_t\| \\ &\leq 5\Gamma\sqrt{\frac{\mu^2 r^2}{pd}}\,\|\boldsymbol{\Delta}_t\| + 10\Gamma\sqrt{\frac{\sigma_1^\star \mu r}{p}}\,\|\boldsymbol{E}_t\|\,.\end{aligned} \tag{4.31}$$

In the final inequality, we make the assumption $d \geq 9\mu r$ without loss of generality. This assumption simplifies the presentation of the proof but does not impact the final result of the paper.

Furthermore, upon noticing that $\left\|\boldsymbol{X}^\star - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| \leq \max\{\|\boldsymbol{X}^\star\|\,, \|\boldsymbol{S}_t\|^2\} \leq 4\sigma_1^\star$, we have

$$\begin{aligned}\|(\mathrm{I}-\mathcal{R}_\Omega)\left(\boldsymbol{\Delta}_t\right)\| &\leq 5\Gamma\sqrt{\frac{\mu^2 r^2}{pd}}\left\|\boldsymbol{X}^\star - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| + \|(\mathrm{I}-\mathcal{R}_\Omega)(\boldsymbol{\Lambda}_t)\| \\ &\leq 20\Gamma\sigma_1^\star\sqrt{\frac{\mu^2 r^2}{pd}} + 9\Gamma\sqrt{\frac{\sigma_1^\star \mu r}{p}}\,\|\boldsymbol{E}_t\| \\ &\leq 21\Gamma\sigma_1^\star\sqrt{\frac{\mu^2 r^2}{pd}}.\end{aligned} \tag{4.32}$$

Next, combining the above two inequalities, we can control $\|\boldsymbol{M}_t\|$ as follows:

$$\begin{aligned}\|\boldsymbol{M}_t\| &\leq \|\boldsymbol{\Delta}_t\| + \|(\mathrm{I}-\mathcal{R}_\Omega)\left(\boldsymbol{\Delta}_t\right)\| \\ &\leq \|\boldsymbol{\Delta}_t\| + 5\Gamma\sqrt{\frac{\mu^2 r^2}{pd}}\,\|\boldsymbol{\Delta}_t\| + 10\Gamma\sqrt{\frac{\sigma_1^\star \mu r}{p}}\,\|\boldsymbol{E}_t\| \\ &\leq \left(1 + 5\Gamma\sqrt{\frac{\mu^2 r^2}{pd}}\right)\|\boldsymbol{\Delta}_t\| + 10\Gamma\sqrt{\frac{\sigma_1^\star \mu r}{p}}\,\|\boldsymbol{E}_t\|\,.\end{aligned} \tag{4.33}$$

Furthermore, we can also bound $\|\boldsymbol{M}_t\|$ as

$$\begin{aligned}\|\boldsymbol{M}_t\| &\leq \|\boldsymbol{\Delta}_t\| + \|(\mathrm{I}-\mathcal{R}_\Omega)\left(\boldsymbol{\Delta}_t\right)\| \\ &\leq 5\sigma_1^\star + 21\Gamma\sigma_1^\star\sqrt{\frac{\mu^2 r^2}{pd}} \\ &\leq 6\sigma_1^\star.\end{aligned} \tag{4.34}$$

Lastly, for $\left\|\boldsymbol{M}_t\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\right\|$, we have the following decomposition

$$\begin{aligned}
\left\|\boldsymbol{M}_t\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\right\| &\leq \left\|\left(\mathrm{I}-\mathcal{R}_{\Omega}\right)\left(\boldsymbol{\Delta}_t\right)\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\right\| + \left\|\left(\boldsymbol{X}^{\star}-\boldsymbol{U}_t\boldsymbol{U}_t^{\top}\right)\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\right\| \\
&\leq \left\|\left(\mathrm{I}-\mathcal{R}_{\Omega}\right)\left(\boldsymbol{\Delta}_t\right)\right\| + \left\|\boldsymbol{V}^{\star}\boldsymbol{\Sigma}\left(\boldsymbol{V}^{\star}-\boldsymbol{V}_t\right)^{\top}\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\right\| + \left\|\boldsymbol{U}_t\boldsymbol{E}_t^{\top}\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\right\| \\
&\leq 21\Gamma\sigma_1^{\star}\sqrt{\frac{\mu^2r^2}{pd}} + \sigma_1^{\star}\left\|\boldsymbol{V}^{\star}-\boldsymbol{V}_t\right\| + \left\|\boldsymbol{E}_t\right\|\left(\left\|\boldsymbol{S}_t\right\| + \left\|\boldsymbol{E}_t\right\|\right) \\
&\leq 22\Gamma\frac{\sigma_1^{\star}\mu r}{\sqrt{pd}} + \sigma_1^{\star}\left\|\boldsymbol{V}^{\star}-\boldsymbol{V}_t\right\| \\
&\overset{(a)}{\leq} 22\Gamma\frac{\sigma_1^{\star}\mu r}{\sqrt{pd}} + \Gamma_1\sigma_1^{\star}\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}} \\
&\leq 2\Gamma_1\sigma_1^{\star}\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}.
\end{aligned} \tag{4.35}$$

Here in (a), we apply Lemma 64. □

Lemma 68 (Helper lemma for $\boldsymbol{A}_t$)**.** *Under the same conditions as Lemma 66, we have*

$$\left\|\boldsymbol{A}_t\right\| \leq 300\eta^2\sigma_1^{\star 2}. \tag{4.36}$$

Proof. We first use triangle inequality to bound $\left\|\boldsymbol{A}_t\right\|$ as follows

$$\left\|\boldsymbol{A}_t\right\| \leq 1.5\eta^2\left\|\boldsymbol{M}_t\right\|^2 + 0.5\eta^3\left\|\boldsymbol{M}_t\right\|^3 + \left(1+\eta\left\|\boldsymbol{M}_t\right\|\right)\left\|\boldsymbol{R}(\boldsymbol{Y}_t)\right\|. \tag{4.37}$$

Next, Lemma 67 tells us that $\left\|\boldsymbol{M}_t\right\| \leq 6\sigma_1^{\star}$ conditioned on E_{good}. For $\left\|\boldsymbol{R}(\boldsymbol{Y}_t)\right\|$, we first have

$$\left\|\boldsymbol{R}(\boldsymbol{Y}_t)\right\| \leq \sum_{k=2}^{\infty}\frac{(2k)!}{4^k(k!)^2}\left\|\boldsymbol{Y}_t\right\|^k = \frac{2+\sqrt{1-\left\|\boldsymbol{Y}_t\right\|}}{2\sqrt{1-\left\|\boldsymbol{Y}_t\right\|}\left(1+\sqrt{1-\left\|\boldsymbol{Y}_t\right\|}\right)^2}\left\|\boldsymbol{Y}_t\right\|^2. \tag{4.38}$$

Note that $\left\|\boldsymbol{Y}_t\right\| \leq \left\|2\eta\boldsymbol{M}_t+\eta^2\boldsymbol{M}_t^2\right\| \leq 20\eta\sigma_1^{\star}$ and the right-hand side is an increasing function of $\left\|\boldsymbol{Y}_t\right\|$. Therefore, we derive that

$$\left\|\boldsymbol{R}(\boldsymbol{Y}_t)\right\| \leq \frac{1}{2}\left\|\boldsymbol{Y}_t\right\|^2 \leq 200\eta^2\sigma_1^{\star 2}. \tag{4.39}$$

This implies that

$$\left\|\boldsymbol{A}_t\right\| \leq 300\eta^2\sigma_1^{\star 2}. \tag{4.40}$$

□

4.B Proofs for Dynamic Signal-residual Decomposition

We first present a more precise version of the one-step dynamics of the signal and residual terms.

Proposition 14. *Suppose that* $\|\boldsymbol{S}_t\| \le 2\sqrt{\sigma_1^\star}$, $\|\boldsymbol{E}_t\| \le \sqrt{\frac{\sigma_1^\star}{d}}$, $\|\boldsymbol{V}_t\|_{2,\infty} \le 2\sqrt{\frac{\mu r}{d}}$ *and* $\|\boldsymbol{V}^\star - \boldsymbol{V}_t\|_{\mathrm{F}} \le \Gamma_1 \frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}$. *Then, the following dynamics hold conditioned on* E_{good}*:*

$$\sigma_r(\boldsymbol{S}_{t+1}) \ge \left(1 + 0.8\eta\sigma_r^\star - \eta\sigma_r^2(\boldsymbol{S}_t)\right)\sigma_r(\boldsymbol{S}_t) - 6\Gamma_1\eta\frac{\sigma_1^\star\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\|\boldsymbol{E}_t\|, \tag{minimal signal dynamic}$$

$$\|\boldsymbol{S}_{t+1}\| \le 2\sqrt{\sigma_1^\star}, \tag{maximal signal dynamic}$$

$$\|\boldsymbol{E}_{t+1}\| \le \left(1 + 2\Gamma_1\eta\frac{\sigma_1^\star\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\right)\|\boldsymbol{E}_t\|. \tag{residual dynamic}$$

Additionally, if $\sigma_r(\boldsymbol{S}_t) \ge \frac{\sqrt{\sigma_r^\star}}{2}$, *then*

$$\left\|\boldsymbol{X}^\star - \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^\top\right\|_{\mathrm{F}} \le \left(1 - \frac{1}{10}\eta\sigma_r^\star\right)\left\|\boldsymbol{X}^\star - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right\|_{\mathrm{F}} + \Gamma_6\eta\sqrt{\frac{\sigma_1^{\star 3}\mu r^2}{p}}\|\boldsymbol{E}_t\|. \tag{error dynamic}$$

The key distinction between the above proposition and Lemma 63 lies in a finer control over the one-step dynamics, accompanied by additional assumptions on $\|\boldsymbol{E}_t\|$, $\|\boldsymbol{V}_t\|_{2,\infty}$, and $\|\boldsymbol{V}_t - \boldsymbol{V}^\star\|_{\mathrm{F}}$. We emphasize that the one-step dynamics in Lemma 63 are derived from the proof of this proposition. Moreover, the proof of the error dynamics (last inequality in Proposition 14) will incorporate the proof of the descent lemma (Lemma 62).

4.B.1 Proof of Signal Dynamic

We notice that

$$\boldsymbol{S}_{t+1} = \mathcal{P}_{\boldsymbol{V}_{t+1}}\left(\boldsymbol{I} + \eta\boldsymbol{M}_t\right)\left(\boldsymbol{S}_t + \boldsymbol{E}_t\right) = \left(\boldsymbol{I} + \eta\boldsymbol{M}_t\right)\boldsymbol{S}_t + \mathcal{P}_{\boldsymbol{V}_{t+1}}\left(\boldsymbol{I} + \eta\boldsymbol{M}_t\right)\boldsymbol{E}_t. \tag{4.41}$$

Here the second equality follows from the definition of $\mathcal{P}_{\boldsymbol{V}_{t+1}}$.

Maximal signal dynamic. We first provide an upper-bound for $\|\boldsymbol{S}_{t+1}\|$ by

$$\|\boldsymbol{S}_{t+1}\| \le \underbrace{\left\|\left(\boldsymbol{I}+\eta\boldsymbol{M}_t\right)\boldsymbol{S}_t\right\|}_{:=(\mathrm{I})} + \underbrace{\left\|\mathcal{P}_{\boldsymbol{V}_{t+1}}\left(\boldsymbol{I}+\eta\boldsymbol{M}_t\right)\boldsymbol{E}_t\right\|}_{:=(\mathrm{II})}. \tag{4.42}$$

Next, we further control (I) by

$$\begin{aligned}
(\mathrm{I}) &\le \left\|\left(\boldsymbol{I}+\eta\boldsymbol{\Delta}_t\right)\boldsymbol{S}_t\right\| + \eta\left\|\left(\mathrm{I}-\mathcal{R}_\Omega\right)\left(\boldsymbol{\Delta}_t\right)\right\|\|\boldsymbol{S}_t\| \\
&\overset{(a)}{\le} \left\|\left(\boldsymbol{I}+\eta\boldsymbol{\Delta}_t\right)\boldsymbol{S}_t\right\| + \eta\cdot 21\Gamma\sigma_1^\star\sqrt{\frac{\mu^2 r^2}{pd}}\cdot 2\sqrt{\sigma_1^\star} \\
&\le \left\|\left(\boldsymbol{I}+\eta\left(\boldsymbol{X}^\star-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right)\boldsymbol{S}_t\right\| + \eta\left\|\boldsymbol{\Lambda}_t\right\|\|\boldsymbol{S}_t\| + 42\Gamma\eta\frac{\sigma_1^{\star 1.5}\mu r}{\sqrt{pd}} \\
&\overset{(b)}{\le} \left\|\left(\boldsymbol{I}+\eta\left(\boldsymbol{X}^\star-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right)\boldsymbol{S}_t\right\| + 10\eta\sigma_1^\star\|\boldsymbol{E}_t\| + 42\Gamma\eta\frac{\sigma_1^{\star 1.5}\mu r}{\sqrt{pd}} \\
&\le \underbrace{\left\|\left(\boldsymbol{I}+\eta\left(\boldsymbol{X}^\star-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right)\boldsymbol{S}_t\right\|}_{:=(\mathrm{I}_1)} + 43\Gamma\eta\frac{\sigma_1^{\star 1.5}\mu r}{\sqrt{pd}}.
\end{aligned} \tag{4.43}$$

Here we apply Lemma 67 in (a) and (b). In the last inequality, we use the assumption that $\|\boldsymbol{E}_t\| \le \frac{\Gamma\sqrt{\sigma_1^\star}\mu r}{\sqrt{pd}}$. For (I_1), we further decompose it via triangle inequality as follows

$$\begin{aligned}
(\mathrm{I}_1) &\le \left\|\mathcal{P}_{\boldsymbol{V}_t}\left(\boldsymbol{I}+\eta\left(\boldsymbol{X}^\star-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right)\boldsymbol{S}_t\right\| + \left\|\mathcal{P}_{\boldsymbol{V}_t^\perp}\left(\boldsymbol{I}+\eta\left(\boldsymbol{X}^\star-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right)\boldsymbol{S}_t\right\| \\
&= \underbrace{\left\|\left(\boldsymbol{I}+\eta\left(\boldsymbol{V}_t^\top\boldsymbol{V}^\star\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}\boldsymbol{V}_t-\bar{\boldsymbol{S}}_t\bar{\boldsymbol{S}}_t^\top\right)\right)\bar{\boldsymbol{S}}_t\right\|}_{:=(\mathrm{I}_{1,1})} + \eta\underbrace{\left\|\mathcal{P}_{\boldsymbol{V}_t^\perp}\left(\boldsymbol{V}^\star-\boldsymbol{V}_t\right)\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}\boldsymbol{S}_t\right\|}_{:=(\mathrm{I}_{1,2})}.
\end{aligned} \tag{4.44}$$

Here we define $\bar{\boldsymbol{S}}_t = \boldsymbol{V}_t^\top\boldsymbol{U}_t \in \mathbb{R}^{r\times d}$. Note that $\left\|\bar{\boldsymbol{S}}_t\right\| = \|\boldsymbol{S}_t\|$. For $(\mathrm{I}_{1,1})$, we have

$$\begin{aligned}
(\mathrm{I}_{1,1}) &\le \left\|\left(\boldsymbol{I}-\eta\bar{\boldsymbol{S}}_t\bar{\boldsymbol{S}}_t^\top\right)\bar{\boldsymbol{S}}_t\right\| + \eta\left\|\boldsymbol{V}_t^\top\boldsymbol{V}^\star\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}\boldsymbol{V}_t\bar{\boldsymbol{S}}_t\right\| \\
&\overset{(a)}{=} \left\|\bar{\boldsymbol{S}}_t\right\|\left(1-\eta\left\|\bar{\boldsymbol{S}}_t\right\|^2\right) + \eta\sigma_1^\star\left\|\bar{\boldsymbol{S}}_t\right\| \\
&= \|\boldsymbol{S}_t\|\left(1+\eta\sigma_1^\star-\eta\|\boldsymbol{S}_t\|^2\right).
\end{aligned} \tag{4.45}$$

where (a) follows from the fact that $\bar{\boldsymbol{S}}_t$ and $\bar{\boldsymbol{S}}_t\bar{\boldsymbol{S}}_t^\top$ share the same eigenvectors, and the assumption $\eta \lesssim 1/\sigma_1$. Next, we control $(\mathrm{I}_{1,2})$:

$$(\mathrm{I}_{1,2}) \le \sigma_1^\star\|\boldsymbol{V}^\star-\boldsymbol{V}_t\|\|\boldsymbol{S}_t\| \le 2\Gamma_1\frac{\kappa\mu(\sigma_1^\star r)^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}. \tag{4.46}$$

Here we use Lemma 64. Therefore, we can bound (I_1) by

$$(\mathrm{I}_1) \leq \|\boldsymbol{S}_t\| \left(1 + \eta\sigma_1^\star - \eta \|\boldsymbol{S}_t\|^2\right) + 2\Gamma_1\eta \frac{\kappa\mu(\sigma_1^\star r)^{1.5} \log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}. \tag{4.47}$$

This leads to

$$\begin{aligned}(\mathrm{I}) &\leq \|\boldsymbol{S}_t\| \left(1 + \eta\sigma_1^\star - \eta \|\boldsymbol{S}_t\|^2\right) + 2\Gamma_1\eta \frac{\kappa\mu(\sigma_1^\star r)^{1.5} \log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}} + 43\Gamma\eta \frac{\sigma_1^{\star 1.5}\mu r}{\sqrt{pd}} \\ &\leq \|\boldsymbol{S}_t\| \left(1 + \eta\sigma_1^\star - \eta \|\boldsymbol{S}_t\|^2\right) + 3\Gamma_1\eta \frac{\kappa\mu(\sigma_1^\star r)^{1.5} \log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}.\end{aligned} \tag{4.48}$$

Next, we control (II). To this end, we first notice that

$$\mathcal{P}_{\boldsymbol{V}_{t+1}} \left(\boldsymbol{I} + \eta \boldsymbol{M}_t\right) \mathcal{P}_{\boldsymbol{V}_t}^\perp = \left(\mathcal{P}_{\boldsymbol{V}_{t+1}} - \mathcal{P}_{\boldsymbol{V}_t}\right) \mathcal{P}_{\boldsymbol{V}_t}^\perp + \eta \mathcal{P}_{\boldsymbol{V}_{t+1}} \boldsymbol{M}_t \mathcal{P}_{\boldsymbol{V}_t}^\perp. \tag{4.49}$$

Hence, we can bound (II) by

$$\begin{aligned}(\mathrm{II}) &\leq \left\|\mathcal{P}_{\boldsymbol{V}_{t+1}} - \mathcal{P}_{\boldsymbol{V}_t}\right\| \|\boldsymbol{E}_t\| + \eta \left\|\boldsymbol{M}_t \mathcal{P}_{\boldsymbol{V}_t}^\perp\right\| \|\boldsymbol{E}_t\| \\ &\overset{(a)}{\leq} \left(2 \|\boldsymbol{V}_{t+1} - \boldsymbol{V}_t\| + 2\Gamma_1\eta\sigma_1^\star \frac{\kappa\mu r^{1.5} \log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\right) \|\boldsymbol{E}_t\|.\end{aligned} \tag{4.50}$$

Here we use Lemma 90 in (a). It remains to control $\|\boldsymbol{V}_{t+1} - \boldsymbol{V}_t\|$. To this end, upon noticing that $\boldsymbol{V}_{t+1} - \boldsymbol{V}_t = \eta \mathcal{P}_{\boldsymbol{V}_t}^\perp \boldsymbol{M}_t \boldsymbol{V}_t + \boldsymbol{A}_t$, one has

$$\begin{aligned}\|\boldsymbol{V}_{t+1} - \boldsymbol{V}_t\| &\leq \eta \left\|\boldsymbol{M}_t \mathcal{P}_{\boldsymbol{V}_t}^\perp\right\| + \|\boldsymbol{A}_t\| \\ &\overset{(a)}{\leq} 2\Gamma_1\eta\sigma_1^\star \frac{\kappa\mu r^{1.5} \log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}} + 300\eta^2\sigma_1^{\star 2} \\ &\leq 3\Gamma_1\eta\sigma_1^\star \frac{\kappa\mu r^{1.5} \log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}.\end{aligned} \tag{4.51}$$

Here we apply Lemma 67 and Lemma 68 in (a). Hence, we have

$$(\mathrm{II}) \leq 5\Gamma_1\eta\sigma_1^\star \frac{\kappa\mu r^{1.5} \log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}} \|\boldsymbol{E}_t\|. \tag{4.52}$$

Putting everything together, we obtain that

$$\begin{aligned}\|\boldsymbol{S}_{t+1}\| &\leq \|\boldsymbol{S}_t\| \left(1+\eta\sigma_1^\star - \eta\|\boldsymbol{S}_t\|^2\right) + 3\Gamma_1\eta\sigma_1^\star \frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}} \left(\sqrt{\sigma_1^\star} + 2\|\boldsymbol{E}_t\|\right) \\ &\leq \|\boldsymbol{S}_t\| \left(1+\eta\sigma_1^\star - \eta\|\boldsymbol{S}_t\|^2\right) + 6\Gamma_1\eta\sigma_1^{\star 1.5} \frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}.\end{aligned} \tag{4.53}$$

Next, we consider two cases separately. First, if $\|\boldsymbol{S}_t\| \leq 1.5\sqrt{\sigma_1^\star}$, then we simply have

$$\begin{aligned}\|\boldsymbol{S}_{t+1}\| &\leq \|\boldsymbol{S}_t\| \left(1+\eta\sigma_1^\star - \eta\|\boldsymbol{S}_t\|^2\right) + 6\Gamma_1\eta\sigma_1^{\star 1.5} \frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}} \\ &\leq 1.5\sqrt{\sigma_1^\star}\cdot(1+\eta\sigma_1^\star) + 6\Gamma_1\eta\sigma_1^{\star 1.5} \frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}} \\ &\leq 2\sqrt{\sigma_1^\star}.\end{aligned} \tag{4.54}$$

On the other hand, if $1.5\sqrt{\sigma_1^\star} \leq \|\boldsymbol{S}_t\| \leq 2\sqrt{\sigma_1^\star}$, then we have

$$\begin{aligned}\|\boldsymbol{S}_{t+1}\| &\leq \|\boldsymbol{S}_t\| \left(1+\eta\sigma_1^\star - \eta\|\boldsymbol{S}_t\|^2\right) + 6\Gamma_1\eta\sigma_1^{\star 1.5} \frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}} \\ &\leq 2\sqrt{\sigma_1^\star}\left(1-1.25\eta\sigma_1^\star\right) + 6\Gamma_1\eta\sigma_1^{\star 1.5} \frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}} \\ &\leq 2\sqrt{\sigma_1^\star}.\end{aligned} \tag{4.55}$$

This completes the proof for the maximal signal dynamic.

Minimal signal dynamic. We first provide a lower-bound for $\sigma_r(\boldsymbol{S}_{t+1})$ as follows

$$\sigma_r(\boldsymbol{S}_{t+1}) \geq \underbrace{\sigma_r((\boldsymbol{I}+\eta\boldsymbol{M}_t)\boldsymbol{S}_t)}_{\text{(I)}} - \underbrace{\left\|\mathcal{P}_{\boldsymbol{V}_{t+1}}(\boldsymbol{I}+\eta\boldsymbol{M}_t)\boldsymbol{E}_t\right\|}_{\text{(II)}}. \tag{4.56}$$

For (I), applying Lemma 91, we first obtain that (I) $\geq \sigma_r\big(\boldsymbol{V}_t^\top(\boldsymbol{I}+\eta\boldsymbol{M}_t)\boldsymbol{S}_t\big)$. Next, we decompose $\boldsymbol{V}_t^\top(\boldsymbol{I}+\eta\boldsymbol{M}_t)\boldsymbol{S}_t$ as follows

$$\begin{aligned}&\boldsymbol{V}_t^\top(\boldsymbol{I}+\eta\boldsymbol{M}_t)\boldsymbol{S}_t \\ &= \underbrace{\left(\boldsymbol{I}+\eta\boldsymbol{V}_t^\top\left(\boldsymbol{M}_t+\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\boldsymbol{V}_t\left(\boldsymbol{I}-\eta\boldsymbol{V}_t^\top\boldsymbol{S}_t\boldsymbol{S}_t^\top\boldsymbol{V}_t\right)^{-1}\right)}_{:=\boldsymbol{B}_t} \underbrace{\boldsymbol{V}_t^\top\boldsymbol{S}_t\left(\boldsymbol{I}-\eta\boldsymbol{S}_t^\top\boldsymbol{S}_t\right)}_{:=\boldsymbol{C}_t}.\end{aligned} \tag{4.57}$$

According to Lemma 92, we have

$$\sigma_r((\boldsymbol{I}+\eta\boldsymbol{M}_t)\,\boldsymbol{S}_t) \geq \sigma_r(\boldsymbol{B}_t)\sigma_r(\boldsymbol{C}_t) = \sigma_r(\boldsymbol{B}_t)\left(1-\eta\sigma_r^2(\boldsymbol{S}_t)\right)\sigma_r(\boldsymbol{S}_t), \tag{4.58}$$

where in the last equality, we use the fact that $\boldsymbol{S}_t$ and $\boldsymbol{S}_t\boldsymbol{S}_t^\top\boldsymbol{S}_t$ share the same singular space and hence

$$\sigma_r(\boldsymbol{C}_t) = \sigma_r(\boldsymbol{S}_t(\boldsymbol{I}-\eta\boldsymbol{S}_t^\top\boldsymbol{S}_t)) = \left(1-\eta\sigma_r^2(\boldsymbol{S}_t)\right)\sigma_r(\boldsymbol{S}_t). \tag{4.59}$$

Now, it suffices to provide a lower-bound for $\sigma_r(\boldsymbol{B}_t)$. To this end, we first notice that

$$\begin{aligned}\sigma_r(\boldsymbol{B}_t) &\geq 1+\eta\sigma_r\left(\boldsymbol{V}_t^\top\left(\boldsymbol{M}_t+\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\boldsymbol{V}_t\right)\sigma_r\Big(\left(\boldsymbol{I}-\eta\boldsymbol{V}_t^\top\boldsymbol{S}_t\boldsymbol{S}_t^\top\boldsymbol{V}_t\right)^{-1}\Big)\\ &\geq 1+\eta\sigma_r\left(\boldsymbol{V}_t^\top\left(\boldsymbol{M}_t+\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\boldsymbol{V}_t\right).\end{aligned} \tag{4.60}$$

Here the second inequality is due to $\boldsymbol{I}-\eta\boldsymbol{V}_t^\top\boldsymbol{S}_t\boldsymbol{S}_t^\top\boldsymbol{V}_t \preceq \boldsymbol{I}$. To proceed, notice that

$$\boldsymbol{M}_t+\boldsymbol{S}_t\boldsymbol{S}_t^\top = \boldsymbol{X}^\star + (\mathcal{R}_\Omega-\mathrm{I})\left(\boldsymbol{\Delta}_t\right) - \boldsymbol{\Lambda}_t. \tag{4.61}$$

Therefore, we have

$$\begin{aligned}\sigma_r\left(\boldsymbol{V}_t^\top\left(\boldsymbol{M}_t+\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\boldsymbol{V}_t\right) &\geq \sigma_r\left(\boldsymbol{V}_t^\top\boldsymbol{X}^\star\boldsymbol{V}_t\right) - \left\|(\mathcal{R}_\Omega-\mathrm{I})\left(\boldsymbol{\Delta}_t\right)\right\| - \left\|\boldsymbol{\Lambda}_t\right\|\\ &\geq \sigma_r\left(\boldsymbol{V}_t^\top\boldsymbol{X}^\star\boldsymbol{V}_t\right) - 21\Gamma\sigma_1^\star\sqrt{\frac{\mu^2r^2}{pd}} - 5\sqrt{\sigma_1^\star}\left\|\boldsymbol{E}_t\right\|\\ &\geq \sigma_r\left(\boldsymbol{V}_t^\top\boldsymbol{X}^\star\boldsymbol{V}_t\right) - 22\Gamma\sigma_1^\star\sqrt{\frac{\mu^2r^2}{pd}}.\end{aligned} \tag{4.62}$$

For the first term in the above inequality, we have

$$\begin{aligned}\sigma_r\left(\boldsymbol{V}_t^\top\boldsymbol{X}^\star\boldsymbol{V}_t\right) &\geq \sigma_r\left(\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}\boldsymbol{V}_t\right) - \sigma_1^\star\left\|\boldsymbol{V}^\star-\boldsymbol{V}_t\right\|\\ &\geq \sigma_r^\star - 2\sigma_1^\star\left\|\boldsymbol{V}^\star-\boldsymbol{V}_t\right\|\\ &\geq \sigma_r^\star - 2\Gamma_1\sigma_1^\star\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\\ &\geq 0.95\sigma_r^\star.\end{aligned} \tag{4.63}$$

Therefore, we obtain

$$\sigma_r\left(\boldsymbol{V}_t^\top\left(\boldsymbol{M}_t+\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\boldsymbol{V}_t\right) \geq 0.95\sigma_r^\star - 22\Gamma\sigma_1^\star\sqrt{\frac{\mu^2r^2}{pd}} \geq 0.9\sigma_r^\star. \tag{4.64}$$

Combining the above arguments, we have

$$\text{(I)} \geq (1 + 0.9\eta\sigma_r^\star)\left(1 - \eta\sigma_r^2(\boldsymbol{S}_t)\right)\sigma_r(\boldsymbol{S}_t) \geq \left(1 + 0.8\eta\sigma_r^\star - \eta\sigma_r^2(\boldsymbol{S}_t)\right)\sigma_r(\boldsymbol{S}_t). \tag{4.65}$$

On the other hand, we have already derived an upper bound for (II) in the maximal signal dynamic, which is

$$\text{(II)} \leq 5\Gamma_1\eta\sigma_1^\star\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\left\|\boldsymbol{E}_t\right\|. \tag{4.66}$$

Putting everything together, we have

$$\sigma_r(\boldsymbol{S}_{t+1}) \geq \left(1 + 0.8\eta\sigma_r^\star - \eta\sigma_r^2(\boldsymbol{S}_t)\right)\sigma_r(\boldsymbol{S}_t) - 5\Gamma_1\eta\frac{\sigma_1^\star\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\left\|\boldsymbol{E}_t\right\|. \tag{4.67}$$

4.B.2 Proof of Residual Dynamic

First, we can expand $\boldsymbol{E}_{t+1}$ as below

$$\boldsymbol{E}_{t+1} = \mathcal{P}_{\boldsymbol{V}_{t+1}}^{\perp}\left(\boldsymbol{I} + \eta\boldsymbol{M}_t\right)\left(\boldsymbol{S}_t + \boldsymbol{E}_t\right) = \mathcal{P}_{\boldsymbol{V}_{t+1}}^{\perp}\left(\boldsymbol{I} + \eta\boldsymbol{M}_t\right)\boldsymbol{E}_t, \tag{4.68}$$

where in the second equality, we use the fact that $\mathcal{P}_{\boldsymbol{V}_{t+1}}^{\perp}\left(\boldsymbol{I} + \eta\boldsymbol{M}_t\right)\boldsymbol{S}_t = 0$. Then, by triangle inequality, we obtain

$$\left\|\boldsymbol{E}_{t+1}\right\| \leq \left(1 + \eta\left\|\boldsymbol{M}_t\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\right\|\right)\left\|\boldsymbol{E}_t\right\| \leq \left(1 + 2\Gamma_1\eta\frac{\sigma_1^\star\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\right)\left\|\boldsymbol{E}_t\right\|. \tag{4.69}$$

Here in the last inequality, we use Lemma 67.

4.B.3 Proof of Error Dynamic

The core proof idea is adapted from the proof of Proposition 4.3 appeared in [Li et al., 2018]. We first expand $\left\|\boldsymbol{\Delta}_{t+1}\right\|_{\mathrm{F}}^2$ as

$$\begin{aligned}\left\|\boldsymbol{\Delta}_{t+1}\right\|_{\mathrm{F}}^2 &= \left\|\boldsymbol{X}^\star - \left(\boldsymbol{I} + \eta\boldsymbol{M}_t\right)\boldsymbol{U}_t\boldsymbol{U}_t^\top\left(\boldsymbol{I} + \eta\boldsymbol{M}_t\right)\right\|_{\mathrm{F}}^2 \\ &= \left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}^2 - 4\eta\underbrace{\left\langle\boldsymbol{X}^\star - \boldsymbol{U}_t\boldsymbol{U}_t^\top, \boldsymbol{M}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top\right\rangle}_{\text{(I)}} + \text{(II)},\end{aligned} \tag{4.70}$$

where

$$\begin{aligned}\text{(II)} = {} & 2\left\langle-\boldsymbol{\Delta}_t + \eta\boldsymbol{M}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top, \eta^2\boldsymbol{M}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{M}_t\right\rangle \\ & + \eta^2\left\|\boldsymbol{M}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top + \boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{M}_t\right\|_{\mathrm{F}}^2 + \eta^4\left\|\boldsymbol{M}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{M}_t\right\|_{\mathrm{F}}^2\end{aligned} \tag{4.71}$$

contains all the higher-order terms. We then provide a lower-bound for (I). To this goal, we notice that

$$\begin{aligned}
\text{(I)} &= \|\boldsymbol{\Delta}_t \boldsymbol{U}_t\|_{\mathrm{F}}^2 - \left\langle \boldsymbol{\Delta}_t, \left(\boldsymbol{I} - \mathcal{R}_\Omega\right)\left(\boldsymbol{\Delta}_t\right) \boldsymbol{U}_t \boldsymbol{U}_t^\top \right\rangle \\
&\geq \|\boldsymbol{\Delta}_t \boldsymbol{U}_t\|_{\mathrm{F}}^2 - \|\boldsymbol{\Delta}_t\|_{\mathrm{F}} \left\|\boldsymbol{U}_t \boldsymbol{U}_t^\top\right\|_{\mathrm{F}} \|(\mathrm{I} - \mathcal{R}_\Omega)\left(\boldsymbol{\Delta}_t\right)\| \\
&\geq \|\boldsymbol{\Delta}_t \boldsymbol{U}_t\|_{\mathrm{F}}^2 - \|\boldsymbol{\Delta}_t\|_{\mathrm{F}} \cdot 8\sqrt{r}\sigma_1^\star \cdot \left(5\Gamma \sqrt{\frac{\mu^2 r^2}{pd}} \|\boldsymbol{\Delta}_t\| + 10\Gamma \sqrt{\frac{\sigma_1^\star \mu r}{p}} \|\boldsymbol{E}_t\|\right) \\
&= \|\boldsymbol{\Delta}_t \boldsymbol{U}_t\|_{\mathrm{F}}^2 - 40\Gamma\sigma_1^\star \sqrt{\frac{\mu^2 r^3}{pd}} \|\boldsymbol{\Delta}_t\|_{\mathrm{F}}^2 - 80\Gamma \sqrt{\frac{\sigma_1^{\star 3} \mu r^2}{p}} \|\boldsymbol{E}_t\| \|\boldsymbol{\Delta}_t\|_{\mathrm{F}}.
\end{aligned} \tag{4.72}$$

Here in the last inequality, we apply Lemma 67. Next, we provide a lower-bound for $\|\boldsymbol{\Delta}_t \boldsymbol{U}_t\|_{\mathrm{F}}^2$. To this goal, we first notice that

$$\begin{aligned}
\|\boldsymbol{\Delta}_t \boldsymbol{U}_t\|_{\mathrm{F}} &= \left\|\left(\boldsymbol{X}^\star - \boldsymbol{S}_t \boldsymbol{S}_t^\top\right) \boldsymbol{S}_t - \boldsymbol{\Lambda}_t \boldsymbol{S}_t + \boldsymbol{\Delta}_t \boldsymbol{E}_t\right\|_{\mathrm{F}} \\
&\geq \left\|\left(\boldsymbol{X}^\star - \boldsymbol{S}_t \boldsymbol{S}_t^\top\right) \boldsymbol{S}_t\right\|_{\mathrm{F}} - \|\boldsymbol{\Lambda}_t\| \|\boldsymbol{S}_t\|_{\mathrm{F}} - \|\boldsymbol{\Delta}_t\|_{\mathrm{F}} \|\boldsymbol{E}_t\| \\
&\geq \left\|\left(\boldsymbol{X}^\star - \boldsymbol{S}_t \boldsymbol{S}_t^\top\right) \boldsymbol{S}_t\right\|_{\mathrm{F}} - 10\sqrt{r}\sigma_1^\star \|\boldsymbol{E}_t\| - \|\boldsymbol{\Delta}_t\|_{\mathrm{F}} \|\boldsymbol{E}_t\|.
\end{aligned} \tag{4.73}$$

In the last inequality we use the fact that $\|\boldsymbol{\Lambda}_t\| \leq 5\sqrt{\sigma_1^\star} \|\boldsymbol{E}_t\|$ and $\|\boldsymbol{S}_t\|_{\mathrm{F}} \leq \sqrt{r} \|\boldsymbol{S}_t\| \leq 2\sqrt{r\sigma_1^\star}$ from Lemma 66. Applying Lemma 82, we can further provide a lower-bound for $\left\|\left(\boldsymbol{X}^\star - \boldsymbol{S}_t \boldsymbol{S}_t^\top\right) \boldsymbol{S}_t\right\|_{\mathrm{F}}$ as

$$\left\|\left(\boldsymbol{X}^\star - \boldsymbol{S}_t \boldsymbol{S}_t^\top\right) \boldsymbol{S}_t\right\|_{\mathrm{F}} \geq \sigma_r(\boldsymbol{S}_t) \left\|\left(\boldsymbol{X}^\star - \boldsymbol{S}_t \boldsymbol{S}_t^\top\right) \boldsymbol{V}_t\right\|_{\mathrm{F}} \geq \frac{\sqrt{\sigma_r^\star}}{2} \left\|\left(\boldsymbol{X}^\star - \boldsymbol{S}_t \boldsymbol{S}_t^\top\right) \boldsymbol{V}_t\right\|_{\mathrm{F}}. \tag{4.74}$$

Next, we present the following intermediate lemma to control $\left\|\left(\boldsymbol{X}^\star - \boldsymbol{S}_t \boldsymbol{S}_t^\top\right) \boldsymbol{V}_t\right\|_{\mathrm{F}}$.

Lemma 69. *Suppose that $\|\boldsymbol{V}^\star - \boldsymbol{V}_t\| \leq 0.1$. Then, we have*

$$\left\|\left(\boldsymbol{X}^\star - \boldsymbol{S}_t \boldsymbol{S}_t^\top\right) \boldsymbol{V}_t\right\|_{\mathrm{F}}^2 \geq \frac{2}{5} \left\|\boldsymbol{X}^\star - \boldsymbol{S}_t \boldsymbol{S}_t^\top\right\|_{\mathrm{F}}^2. \tag{4.75}$$

We first use this lemma to finish the proof of the loss dynamic and defer the proof to the end of this section. Applying this lemma to Equation (4.74) yields

$$\begin{aligned}
\left\|\left(\boldsymbol{X}^\star - \boldsymbol{S}_t \boldsymbol{S}_t^\top\right) \boldsymbol{S}_t\right\|_{\mathrm{F}} &\geq \sqrt{\frac{\sigma_r^\star}{10}} \left\|\boldsymbol{X}^\star - \boldsymbol{S}_t \boldsymbol{S}_t^\top\right\|_{\mathrm{F}} \\
&\geq \sqrt{\frac{\sigma_r^\star}{10}} \|\boldsymbol{\Delta}_t\|_{\mathrm{F}} - \sqrt{\frac{\sigma_r^\star}{10}} \|\boldsymbol{\Lambda}_t\|_{\mathrm{F}} \\
&\geq \sqrt{\frac{\sigma_r^\star}{10}} \|\boldsymbol{\Delta}_t\|_{\mathrm{F}} - \sigma_1^\star \sqrt{\frac{5}{2\kappa}} \|\boldsymbol{E}_t\|.
\end{aligned} \tag{4.76}$$

Combining Equation (4.76) and Equation (4.73) leads to

$$\begin{aligned}\left\|\boldsymbol{\Delta}_t\boldsymbol{U}_t\right\|_{\mathrm{F}} &\geq \sqrt{\frac{\sigma_r^\star}{10}}\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}} - \sigma_1^\star\sqrt{\frac{5}{2\kappa}}\left\|\boldsymbol{E}_t\right\| - 10\sqrt{r}\sigma_1^\star\left\|\boldsymbol{E}_t\right\| - \left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}\left\|\boldsymbol{E}_t\right\| \\ &\geq \sqrt{\frac{\sigma_r^\star}{10}}\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}} - 20\sqrt{r}\sigma_1^\star\left\|\boldsymbol{E}_t\right\|.\end{aligned} \tag{4.77}$$

This implies that

$$\left\|\boldsymbol{\Delta}_t\boldsymbol{U}_t\right\|_{\mathrm{F}}^2 \geq \frac{\sigma_r^\star}{10}\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}^2 - 13\sqrt{\frac{r}{\kappa}}\sigma_1^{\star 1.5}\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}\left\|\boldsymbol{E}_t\right\|. \tag{4.78}$$

Overall, we obtain

$$\begin{aligned}\text{(I)} &\geq \frac{\sigma_r^\star}{10}\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}^2 - 13\sqrt{\frac{r}{\kappa}}\sigma_1^{\star 1.5}\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}\left\|\boldsymbol{E}_t\right\| - 40\Gamma\sigma_1^\star\sqrt{\frac{\mu^2 r^3}{pd}}\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}^2 - 80C\sqrt{\frac{\sigma_1^{\star 3}\mu r^2}{p}}\left\|\boldsymbol{E}_t\right\|\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}} \\ &\geq \frac{\sigma_r^\star}{15}\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}^2 - 81\Gamma\sqrt{\frac{\sigma_1^{\star 3}\mu r^2}{p}}\left\|\boldsymbol{E}_t\right\|\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}.\end{aligned} \tag{4.79}$$

Next, we control (II). To this end, we first notice that

$$\begin{aligned}-2\left\langle\boldsymbol{\Delta}_t, \boldsymbol{M}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{M}_t\right\rangle &\leq 2\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}\left\|\boldsymbol{M}_t\right\|^2\left\|\boldsymbol{U}_t\boldsymbol{U}_t^\top\right\|_{\mathrm{F}} \\ &\leq 16\sqrt{r}\sigma_1^\star\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}\left(2\left\|\boldsymbol{\Delta}_t\right\| + 10\Gamma\sqrt{\frac{\sigma_1^\star\mu r}{p}}\left\|\boldsymbol{E}_t\right\|\right)^2 \\ &\leq 128\sqrt{r}\sigma_1^\star\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}\left(\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}^2 + 25\Gamma^2\frac{\sigma_1^\star\mu r}{p}\left\|\boldsymbol{E}_t\right\|^2\right).\end{aligned} \tag{4.80}$$

Similarly, we have

$$\begin{aligned}\left\langle\boldsymbol{M}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top, \boldsymbol{M}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{M}_t\right\rangle &\leq \left\|\boldsymbol{M}_t\right\|^3\left\|\boldsymbol{U}_t\boldsymbol{U}_t^\top\right\|_{\mathrm{F}}^2 \\ &\leq C_1 r\sigma_1^{\star 2}\left(\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}^3 + \frac{(\sigma_1^\star\mu r)^{1.5}}{p^{1.5}}\left\|\boldsymbol{E}_t\right\|^3\right),\end{aligned} \tag{4.81}$$

$$\begin{aligned}\left\|\boldsymbol{M}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top + \boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{M}_t\right\|_{\mathrm{F}}^2 &\leq 4\left\|\boldsymbol{M}_t\right\|^2\left\|\boldsymbol{U}_t\boldsymbol{U}_t^\top\right\|_{\mathrm{F}}^2 \\ &\leq C_2 r\sigma_1^{\star 2}\left(\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}^2 + \frac{\sigma_1^\star\mu r}{p}\left\|\boldsymbol{E}_t\right\|^2\right).\end{aligned} \tag{4.82}$$

$$\left\|\boldsymbol{M}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{M}_t\right\|_{\mathrm{F}}^2 \leq C_3 r\sigma_1^{\star 2}\left(\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}^4 + \frac{\sigma_1^{\star 2}\mu^2 r^2}{p^2}\left\|\boldsymbol{E}_t\right\|^4\right). \tag{4.83}$$

These inequalities lead to

$$\text{(II)} \leq C_4 \eta^2 r \sigma_1^{\star 2} \left(\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}^2 + \Gamma^2 \frac{\sigma_1^\star \mu r}{p} \|\boldsymbol{E}_t\|^2 \right). \tag{4.84}$$

Overall, we have

$$\begin{aligned} \|\boldsymbol{\Delta}_{t+1}\|_{\mathrm{F}}^2 &\leq \left(1 - \frac{1}{5}\eta\sigma_r^\star\right) \|\boldsymbol{\Delta}_t\|_{\mathrm{F}}^2 + 324\Gamma\eta \sqrt{\frac{\sigma_1^{\star 3}\mu r^2}{p}} \|\boldsymbol{E}_t\| \|\boldsymbol{\Delta}_t\|_{\mathrm{F}} + C_5 \eta^2 r \frac{\sigma_1^{\star 3}\mu r}{p} \|\boldsymbol{E}_t\|^2 \\ &\leq \left(\left(1 - \frac{1}{10}\eta\sigma_r^\star\right) \|\boldsymbol{\Delta}_t\|_{\mathrm{F}} + C_6 \eta \sqrt{\frac{\sigma_1^{\star 3}\mu r^2}{p}} \|\boldsymbol{E}_t\| \right)^2, \end{aligned} \tag{4.85}$$

which implies that

$$\|\boldsymbol{\Delta}_{t+1}\|_{\mathrm{F}} \leq \left(1 - \frac{1}{10}\eta\sigma_r^\star\right) \|\boldsymbol{\Delta}_t\|_{\mathrm{F}} + C_6 \eta \sqrt{\frac{\sigma_1^{\star 3}\mu r^2}{p}} \|\boldsymbol{E}_t\|. \tag{4.86}$$

Lastly, we provide the proof of Lemma 69.

Lemma 69. First, we define $\boldsymbol{P} = \boldsymbol{V}^{\star\top}\boldsymbol{V}_t$ and note that $\boldsymbol{S}_t\boldsymbol{S}_t^\top = \boldsymbol{V}_t\boldsymbol{\Sigma}_t\boldsymbol{V}_t^\top$. This allows us to write

$$\begin{aligned} \left\|\left(\boldsymbol{X}^\star - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\boldsymbol{V}_t\right\|_{\mathrm{F}}^2 &= \|\boldsymbol{\Sigma}_t\|_{\mathrm{F}}^2 + \|\boldsymbol{\Sigma}^\star\boldsymbol{P}\|_{\mathrm{F}}^2 - 2\left\langle \boldsymbol{\Sigma}_t, \boldsymbol{P}^\top\boldsymbol{\Sigma}^\star\boldsymbol{P}\right\rangle, \\ \left\|\boldsymbol{X}^\star - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|_{\mathrm{F}}^2 &= \|\boldsymbol{\Sigma}_t\|_{\mathrm{F}}^2 + \|\boldsymbol{\Sigma}^\star\|_{\mathrm{F}}^2 - 2\left\langle \boldsymbol{\Sigma}_t, \boldsymbol{P}^\top\boldsymbol{\Sigma}^\star\boldsymbol{P}\right\rangle. \end{aligned} \tag{4.87}$$

Substituting the above equivalent forms into Equation (4.75), we need to show that

$$3\|\boldsymbol{\Sigma}_t\|_{\mathrm{F}}^2 + 5\|\boldsymbol{\Sigma}^\star\boldsymbol{P}\|_{\mathrm{F}}^2 \geq 2\|\boldsymbol{\Sigma}^\star\|_{\mathrm{F}}^2 + 6\left\langle \boldsymbol{\Sigma}_t, \boldsymbol{P}^\top\boldsymbol{\Sigma}^\star\boldsymbol{P}\right\rangle. \tag{4.88}$$

To this end, we first apply the Cauchy-Schwartz inequality, which gives us $2\left\langle \boldsymbol{\Sigma}_t, \boldsymbol{P}^\top\boldsymbol{\Sigma}^\star\boldsymbol{P}\right\rangle \leq \|\boldsymbol{\Sigma}_t\|_{\mathrm{F}}^2 + \left\|\boldsymbol{P}^\top\boldsymbol{\Sigma}^\star\boldsymbol{P}\right\|_{\mathrm{F}}^2$. Therefore, it suffices to show that

$$5\|\boldsymbol{\Sigma}^\star\boldsymbol{P}\|_{\mathrm{F}}^2 - 2\|\boldsymbol{\Sigma}^\star\|_{\mathrm{F}}^2 - 3\left\|\boldsymbol{P}^\top\boldsymbol{\Sigma}^\star\boldsymbol{P}\right\|_{\mathrm{F}}^2 \geq 0. \tag{4.89}$$

This follows from

$$5\|\boldsymbol{\Sigma}^\star\boldsymbol{P}\|_{\mathrm{F}}^2 - 2\|\boldsymbol{\Sigma}^\star\|_{\mathrm{F}}^2 - 3\left\|\boldsymbol{P}^\top\boldsymbol{\Sigma}^\star\boldsymbol{P}\right\|_{\mathrm{F}}^2 = \mathrm{tr}\left(\left(\boldsymbol{\Sigma}^\star\left(\boldsymbol{I} - \boldsymbol{P}\boldsymbol{P}^\top\right)\boldsymbol{\Sigma}^\star\right)\cdot\left(3\boldsymbol{P}\boldsymbol{P}^\top - 2\boldsymbol{I}\right)\right) \geq 0. \tag{4.90}$$

Here we use the facts that $\boldsymbol{\Sigma}^\star\left(\boldsymbol{I} - \boldsymbol{P}\boldsymbol{P}^\top\right)\boldsymbol{\Sigma}^\star \succeq 0$ since $\|\boldsymbol{P}\| \leq \|\boldsymbol{V}^\star\|\,\|\boldsymbol{V}_t\| \leq 1$, and $3\boldsymbol{P}\boldsymbol{P}^\top - 2\boldsymbol{I} \succeq 0$ since $\sigma_r(\boldsymbol{P}) \geq 1 - \|\boldsymbol{V}^\star - \boldsymbol{V}_t\| \geq 0.9$. This completes the proof. □

4.C Proofs for Main Theorems

In this section, we use the one-step dynamics in Proposition 14 to prove our main theorems under the conditions that $\|\boldsymbol{V}_t\|_{2,\infty} \le 2\sqrt{\frac{\mu r}{d}}$ and $\|\boldsymbol{V}^\star - \boldsymbol{V}_t\|_{\mathrm{F}} \le \Gamma_1 \frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}$ for all $0 \le t \le T$. These two conditions will be established later in Appendix 4.D.

4.C.1 Proof of Theorem 20

The proof is divided into three distinct steps.

Step 1. In the first step, we show that $\|\boldsymbol{S}_t\| \le 2\sqrt{\sigma_1^\star}$ and $\|\boldsymbol{E}_t\| \le 2\alpha$ hold for all $0 \le t \le T$.

We prove this by induction. First, in the base case where $t = 0$, these two conditions are naturally met because $\|\boldsymbol{S}_0\| \le \|\boldsymbol{U}_0\| \le \alpha \le 2\sqrt{\sigma_1^\star}$ and $\|\boldsymbol{E}_0\| \le 2\|\boldsymbol{E}_0\| \le 2\alpha$. Next, for the induction step, we assume that $\|\boldsymbol{S}_t\| \le 2\sqrt{\sigma_1^\star}$ and $\|\boldsymbol{E}_t\| \le 2\alpha$ hold for all $0 \le s \le t$, with $t \le T-1$. Utilizing Proposition 14, we can directly derive that $\|\boldsymbol{S}_{t+1}\| \le 2\sqrt{\sigma_1^\star}$. Regarding $\|\boldsymbol{E}_{t+1}\|$, we have

$$\|\boldsymbol{E}_{t+1}\| \le \left(1 + 2\Gamma_1\eta\frac{\sigma_1^\star\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\right)^{t+1}\alpha \overset{(a)}{\le} \left(1 + 4\Gamma_1\eta\frac{\sigma_1^\star\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\cdot(t+1)\right)\alpha \le 2\alpha. \tag{4.91}$$

Here in (a), we apply Lemma 93. This is valid since $4\Gamma_1\eta\frac{\sigma_1^\star\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\cdot(t+1) \le 1$ for any $t \le T-1 \lesssim \frac{1}{\eta\sigma_r^\star}\log\left(\frac{1}{\alpha}\right)$ provided that the sampling rate satisfies $p \gtrsim \frac{\kappa^4\mu^2 r^3\log^4\left(\frac{1}{\alpha}\right)}{d}$. This completes the induction step.

Step 2. This step demonstrates that the minimal signal $\sigma_r(\boldsymbol{S}_t)$ grows linearly to $\frac{\sqrt{\sigma_r^\star}}{2}$.

Given that we have already established $\|\boldsymbol{S}_t\| \le 2\sqrt{\sigma_1^\star}$ and $\|\boldsymbol{E}_t\| \le 2\alpha$ for all $0 \le t \le T$, we can simplify the minimal signal dynamic in Proposition 14 as

$$\begin{aligned}\sigma_r(\boldsymbol{S}_{t+1}) &\ge \left(1 + 0.8\eta\sigma_r^\star - \eta\sigma_r^2(\boldsymbol{S}_t)\right)\sigma_r(\boldsymbol{S}_t) - 12\Gamma_1\eta\frac{\sigma_1^\star\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\alpha \\ &\ge \left(1 + 0.4\eta\sigma_r^\star\right)\sigma_r(\boldsymbol{S}_t) - 12\Gamma_1\eta\frac{\sigma_1^\star\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\alpha.\end{aligned} \tag{4.92}$$

This holds for any t that satisfies $\sigma_r(\boldsymbol{S}_t) \le \frac{3\sqrt{\sigma_r^\star}}{4}$. By applying Lemma 94, we obtain

$$\sigma_r(\boldsymbol{S}_t) \ge \left(1 + 0.4\eta\sigma_r^\star\right)^t\left(\sigma_r(\boldsymbol{S}_t) - 30\Gamma_1\frac{\kappa^2\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\alpha\right). \tag{4.93}$$

At initialization, it is observed that

$$\sigma_r(\boldsymbol{S}_0) - 30\Gamma_1\frac{\kappa^2\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\alpha \geq c_0\alpha - 30\Gamma_1\frac{\kappa^2\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\alpha \geq \frac{c_0}{2}\alpha, \tag{4.94}$$

provided that $p \gtrsim \frac{\kappa^4\mu^2 r^3\log^2\left(\frac{1}{\alpha}\right)}{d}$. Consequently, within $T_1 \lesssim \frac{1}{\eta\sigma_r^\star}\log\left(\frac{\sigma_r^\star}{\alpha}\right)$ iterations, $\sigma_r(\boldsymbol{S}_t)$ reaches $\frac{\sqrt{\sigma_r^\star}}{2}$. It is also easy to show that $\sigma_r(\boldsymbol{S}_t) \geq \frac{\sqrt{\sigma_r^\star}}{2}$ holds true for all $t \geq T_1$.

Step 3. This step is dedicated to demonstrating that the error $\left\|\boldsymbol{X}^\star - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right\|_{\mathrm{F}}$ converges linearly to $\mathcal{O}(\alpha)$ once $\sigma_r(\boldsymbol{S}_t) \geq \frac{\sqrt{\sigma_r^\star}}{2}$.

Based on our arguments in **Step 2**, where we established that $\sigma_r(\boldsymbol{S}_t) \geq \frac{\sqrt{\sigma_r^\star}}{2}$ for $t \geq T_1$, and leveraging Proposition 14 along with Lemma 94, we can derive that

$$\left\|\boldsymbol{X}^\star - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right\|_{\mathrm{F}} \leq \left(1 - \frac{1}{5}\eta\sigma_r^\star\right)^{t-T_1}\left\|\boldsymbol{X}^\star - \boldsymbol{U}_{T_1}\boldsymbol{U}_{T_1}^\top\right\|_{\mathrm{F}} + 3240\Gamma\sqrt{\frac{\sigma_1^\star\kappa^2\mu r^2}{p}}\alpha. \tag{4.95}$$

Note that $\left\|\boldsymbol{X}^\star - \boldsymbol{U}_{T_1}\boldsymbol{U}_{T_1}^\top\right\|_{\mathrm{F}} \leq \left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}} + \left\|\boldsymbol{U}_{T_1}\boldsymbol{U}_{T_1}^\top\right\|_{\mathrm{F}} \leq 9\sqrt{r}\sigma_1^\star$ according to Lemma 65. Hence, within an additional $T_2 = \frac{1}{\eta\sigma_r^\star}\log\left(\frac{r\sigma_1^\star}{\alpha}\right)$ iterations, the error converges to $O\Big(\sqrt{\frac{\sigma_1^\star\kappa^2\mu r^2}{p}}\alpha\Big)$, thus concluding the proof of Theorem 20.

4.C.2 Proof of Theorem 21

To prove this result, we first apply Theorem 20 to output a solution $\boldsymbol{U}_{t_0}$ and its leave-one-out versions $\boldsymbol{U}_{t_0}^{(l)}$ that meet the initialization conditions in Theorem 22. Then, we apply Theorem 22 to obtain the desired result.

Establishing Condition (4.14). By choosing the initialization scale $\alpha = c\cdot\frac{\sigma_r^\star}{\kappa^{1.5}d}$ for sufficiently small $c > 0$ and assuming the sampling rate of $p \gtrsim \frac{\kappa^2\mu^4 r^9\log^4(d)}{d}$, Theorem 20 guarantees that, with probability at least $1-\frac{1}{d^2}$, the iterations of GD with step-size $\eta \asymp \frac{\mu r}{\sqrt{pd}\sigma_1^\star}$ satisfy

$$\left\|\boldsymbol{U}_{t_0}\boldsymbol{U}_{t_0}^\top - \boldsymbol{X}^\star\right\|_{\mathrm{F}} \leq 0.9\Gamma_4\sqrt{\frac{\sigma_1^\star\kappa^2\mu r^2}{p}}\frac{\sigma_r^\star}{\kappa^{1.5}d}, \quad \text{for some} \quad t_0 \lesssim \frac{1}{\eta\sigma_r^\star}\log\left(\frac{\kappa^{1.5}d}{\sigma_r^\star}\right). \tag{4.96}$$

On the other hand, Lemma 87 in the appendix implies that

$$\mathrm{dist}\left(\boldsymbol{U}_{t_0}, \boldsymbol{U}^\star\right) \leq \frac{1.1}{\sigma_r^\star}\left\|\boldsymbol{U}_{t_0}\boldsymbol{U}_{t_0}^\top - \boldsymbol{X}^\star\right\|_{\mathrm{F}} \leq \Gamma_4\sqrt{\frac{\sigma_r^\star\mu^3 r^3\log(d)}{pd^2}},$$

which establishes Condition (4.14).

Establishing Condition (4.15). The proof of Condition (4.15) follows a similar logic, recognizing that the leave-one-out sequences exhibit a stronger concentration than the original iterations. Consequently, they fulfill Condition (4.15) within at most $t_0 \lesssim \frac{1}{\eta\sigma_r^\star}\log\left(\frac{\kappa^{1.5}d}{\sigma_r^\star}\right)$ iterations. Further details of this argument are omitted for brevity.

This shows that the initial conditions of Theorem 22 are satisfied after t_0 iterations. From this iteration onward, Theorem 22 shows that the iterations of GD enter a local linear convergence regime, which readily establishes the final result of Theorem 21.

4.D Proofs for Incoherence Dynamic

In this section, we present our proofs for establishing the incoherence of $\boldsymbol{V}_t$. To simplify the presentation, we omit the "$\sim$" from our notations. Therefore, $\boldsymbol{V}_t^{(l)}, \boldsymbol{Z}_t^{(l)}, \boldsymbol{M}_t^{(l)}, \ldots$ in this section refer to $\widetilde{\boldsymbol{V}}_t^{(l)}, \widetilde{\boldsymbol{Z}}_t^{(l)}, \widetilde{\boldsymbol{M}}_t^{(l)}, \ldots$ defined in Section 4.3.2.

4.D.1 Proof of Lemma 64

We first state a finer variant of this lemma here.

Proposition 15 (Controlling $\|\boldsymbol{V}_t - \boldsymbol{V}^\star\|_{\mathrm{F}}$)**.** *Suppose that the stepsize satisfies $\eta \asymp \frac{\mu r}{\sqrt{pd}\sigma_1^\star}$. Moreover, suppose that $\|\boldsymbol{S}_t\| \le 2\sqrt{\sigma_1^\star}$, $\|\boldsymbol{E}_t\| \le \sqrt{\frac{\sigma_1^\star}{d}}$, $\|\boldsymbol{V}_t\|_{2,\infty} \le 2\sqrt{\frac{\mu r}{d}}$ and $\|\boldsymbol{V}^\star - \boldsymbol{V}_t\|_{\mathrm{F}} \le \Gamma_1\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}$ for all $t \le T \lesssim \frac{1}{\eta\sigma_r^\star}\log\left(\frac{1}{\alpha}\right)$. Then, conditioned on E_{good}, we have*

$$\|\boldsymbol{V}_t - \boldsymbol{V}^\star\|_{\mathrm{F}} \le \Gamma_1\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}} \quad \forall t \le T \lesssim \frac{1}{\eta\sigma_r^\star}\log\left(\frac{1}{\alpha}\right). \tag{4.97}$$

Proof. First notice that $\boldsymbol{V}_{t+1}$ can be rewritten as

$$\begin{aligned}\boldsymbol{V}_{t+1} &= \boldsymbol{Z}_{t+1}\left(\boldsymbol{Z}_{t+1}^\top\boldsymbol{Z}_{t+1}\right)^{-1/2} \\ &= \left(\boldsymbol{I} + \eta\boldsymbol{M}_t\right)\boldsymbol{V}_t\left(\boldsymbol{V}_t^\top\left(\boldsymbol{I} + \eta\boldsymbol{M}_t\right)^2\boldsymbol{V}_t\right)^{-1/2} \\ &= \left(\boldsymbol{I} + \eta\boldsymbol{M}_t\right)\boldsymbol{V}_t\left(\boldsymbol{I} + \boldsymbol{Y}_t\right)^{-1/2}\end{aligned} \tag{4.98}$$

where we denote $\boldsymbol{Y}_t = \boldsymbol{V}_t^\top\left(2\eta\boldsymbol{M}_t + \eta^2\boldsymbol{M}_t^2\right)\boldsymbol{V}_t$. Next, we apply Taylor expansion for the matrix-valued function $f(\boldsymbol{X}) = (\boldsymbol{I}+\boldsymbol{X})^{-1/2}$, which states that for any $\boldsymbol{X}$ satisfying $\|\boldsymbol{X}\| < 1$,

$$f(\boldsymbol{X}) = (\boldsymbol{I}+\boldsymbol{X})^{-1/2} = \boldsymbol{I} - \frac{1}{2}\boldsymbol{X} + \boldsymbol{R}(\boldsymbol{X}) \text{ where } \boldsymbol{R}(\boldsymbol{X}) = \sum_{k=2}^{\infty}\frac{(-1)^k(2k)!}{4^k(k!)^2}\boldsymbol{X}^k. \tag{4.99}$$

Then, upon setting $\boldsymbol{X} = \boldsymbol{Y}_t$ in the above equation and plugging it into Equation (4.98) and rearranging the subterms, we have

$$\boldsymbol{V}_{t+1} = (\boldsymbol{I} + \eta \boldsymbol{M}_t)\, \boldsymbol{V}_t \left(\boldsymbol{I} - \frac{1}{2} \boldsymbol{V}_t^\top \left(2\eta \boldsymbol{M}_t + \eta^2 \boldsymbol{M}_t^2 \right) \boldsymbol{V}_t + \boldsymbol{R}(\boldsymbol{Y}_t) \right) = \left(\boldsymbol{I} + \eta \mathcal{P}_{\boldsymbol{V}_t}^{\perp} \boldsymbol{M}_t \right) \boldsymbol{V}_t + \boldsymbol{A}_t, \tag{4.100}$$

where

$$\boldsymbol{A}_t = -\eta^2 \boldsymbol{M}_t \boldsymbol{V}_t \boldsymbol{V}_t^\top \boldsymbol{M}_t \boldsymbol{V}_t - 0.5\eta^2 \boldsymbol{V}_t \boldsymbol{V}_t^\top \boldsymbol{M}_t^2 \boldsymbol{V}_t - 0.5\eta^3 \boldsymbol{M}_t \boldsymbol{V}_t \boldsymbol{V}^{\star\top} \boldsymbol{M}_t^2 \boldsymbol{V}_t + (\boldsymbol{I} + \eta \boldsymbol{M}_t)\, \boldsymbol{V}_t \boldsymbol{R}(\boldsymbol{Y}_t) \tag{4.101}$$

contains all the higher-order terms. Next, according to triangle inequality, we can provide an upper bound for $\|\boldsymbol{V}^\star - \boldsymbol{V}_{t+1}\|_{\mathrm{F}}$ as follows

$$\|\boldsymbol{V}^\star - \boldsymbol{V}_{t+1}\|_{\mathrm{F}} \leq \underbrace{\left\| \boldsymbol{V}^\star - \left(\boldsymbol{I} + \eta \mathcal{P}_{\boldsymbol{V}_t}^{\perp} \boldsymbol{M}_t \right) \boldsymbol{V}_t \right\|_{\mathrm{F}}}_{:=(\mathrm{I})} + \|\boldsymbol{A}_t\|_{\mathrm{F}}\,. \tag{4.102}$$

We first control the leading term (I). To this goal, we apply triangle inequality and obtain

$$\begin{aligned} (\mathrm{I}) &\leq \left\| \boldsymbol{V}^\star - \left(\boldsymbol{I} + \eta \mathcal{P}_{\boldsymbol{V}_t}^{\perp} \left(\boldsymbol{X}^\star - \boldsymbol{U}_t \boldsymbol{U}_t^\top \right) \right) \boldsymbol{V}_t \right\|_{\mathrm{F}} + \eta \left\| (\mathrm{I} - \mathcal{R}_\Omega) \left(\boldsymbol{X}^\star - \boldsymbol{U}_t \boldsymbol{U}_t^\top \right) \boldsymbol{V}_t \right\|_{\mathrm{F}} \\ &\leq \underbrace{\left\| \boldsymbol{V}^\star - \left(\boldsymbol{I} + \eta \mathcal{P}_{\boldsymbol{V}_t}^{\perp} \boldsymbol{\Delta}_t \right) \boldsymbol{V}_t \right\|_{\mathrm{F}}}_{:=(\mathrm{I}_1)} + \eta \sqrt{r}\, \underbrace{\left\| (\mathrm{I} - \mathcal{R}_\Omega) \left(\boldsymbol{\Delta}_t \right) \right\|}_{:=(\mathrm{I}_2)}. \end{aligned} \tag{4.103}$$

To control (I_1), we further decompose it as

$$\begin{aligned} (\mathrm{I}_1) &\leq \left\| \boldsymbol{V}^\star - \left(\boldsymbol{I} + \eta \mathcal{P}_{\boldsymbol{V}_t}^{\perp} \boldsymbol{V}^\star \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top} \right) \boldsymbol{V}_t \right\|_{\mathrm{F}} + \eta \left\| \mathcal{P}_{\boldsymbol{V}_t}^{\perp} \boldsymbol{U}_t \boldsymbol{U}_t^\top \boldsymbol{V}_t \right\|_{\mathrm{F}} \\ &\overset{(a)}{\leq} \left\| \boldsymbol{V}^\star - \boldsymbol{V}_t - \eta \mathcal{P}_{\boldsymbol{V}_t}^{\perp} \left(\boldsymbol{V}^\star - \boldsymbol{V}_t \right) \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top} \boldsymbol{V}_t \right\|_{\mathrm{F}} + \eta \sqrt{r} \left\| \boldsymbol{E}_t \boldsymbol{U}_t^\top \boldsymbol{V}_t \right\| \\ &\overset{(b)}{\leq} \underbrace{\left\| \boldsymbol{V}^\star - \boldsymbol{V}_t - \eta \mathcal{P}_{\boldsymbol{V}_t}^{\perp} \left(\boldsymbol{V}^\star - \boldsymbol{V}_t \right) \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top} \boldsymbol{V}_t \right\|_{\mathrm{F}}}_{:=(\mathrm{I}_{1,1})} + 2\eta \sqrt{r \sigma_1^\star}\, \|\boldsymbol{E}_t\|\,. \end{aligned} \tag{4.104}$$

Here in (a), we use the fact that $\mathcal{P}_{\boldsymbol{V}_t}^{\perp} \boldsymbol{V}_t = 0$ and the definition $\boldsymbol{E}_t = \mathcal{P}_{\boldsymbol{V}_t}^{\perp} \boldsymbol{U}_t$. In (b), we use the assumption that $\|\boldsymbol{S}_t\| \leq 2\sqrt{\sigma_1^\star}$. Next, according to the orthogonality of $\boldsymbol{V}_t$ and $\boldsymbol{V}_t^\perp$, we can upper-bound $(\mathrm{I}_{1,1})$ as follows

$$\begin{aligned} (\mathrm{I}_{1,1})^2 &= \|\mathcal{P}_{\boldsymbol{V}_t}(\boldsymbol{V}^\star - \boldsymbol{V}_t)\|_{\mathrm{F}}^2 + \left\| \mathcal{P}_{\boldsymbol{V}_t}^{\perp} \left(\boldsymbol{V}^\star - \boldsymbol{V}_t \right) \left(\boldsymbol{I} - \eta \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top} \boldsymbol{V}_t \right) \right\|_{\mathrm{F}}^2 \\ &\leq \|\mathcal{P}_{\boldsymbol{V}_t}(\boldsymbol{V}^\star - \boldsymbol{V}_t)\|_{\mathrm{F}}^2 + \left\| \mathcal{P}_{\boldsymbol{V}_t}^{\perp} \left(\boldsymbol{V}^\star - \boldsymbol{V}_t \right) \right\|_{\mathrm{F}}^2 \left\| \boldsymbol{I} - \eta \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top} \boldsymbol{V}_t \right\|^2 \\ &\overset{(a)}{\leq} \|\mathcal{P}_{\boldsymbol{V}_t}(\boldsymbol{V}^\star - \boldsymbol{V}_t)\|_{\mathrm{F}}^2 + \left\| \mathcal{P}_{\boldsymbol{V}_t}^{\perp} \left(\boldsymbol{V}^\star - \boldsymbol{V}_t \right) \right\|_{\mathrm{F}}^2 \\ &= \|\boldsymbol{V}^\star - \boldsymbol{V}_t\|_{\mathrm{F}}^2\,. \end{aligned} \tag{4.105}$$

Here (a) is due to the fact that $\left\|\boldsymbol{I}-\eta\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}\boldsymbol{V}_t\right\| \le \|\boldsymbol{I}-\eta\boldsymbol{\Sigma}^\star\|+\eta\sigma_1^\star\|\boldsymbol{V}^\star-\boldsymbol{V}_t\| \le 1-\eta(\sigma_r^\star-\sigma_1^\star\|\boldsymbol{V}^\star-\boldsymbol{V}_t\|) \le 1$ since we assume $\|\boldsymbol{V}^\star-\boldsymbol{V}_t\| \le \frac{1}{2\kappa}$. Therefore, we derive that

$$(\mathrm{I}_1) \le \|\boldsymbol{V}^\star-\boldsymbol{V}_t\|_{\mathrm{F}} + 2\eta\sqrt{r\sigma_1^\star}\,\|\boldsymbol{E}_t\|\,. \tag{4.106}$$

On the other hand, Lemma 67 tells us that, conditioned on E_{good}, we have $(\mathrm{I}_2) \le 21\Gamma\sigma_1^\star\sqrt{\frac{\mu^2r^2}{pd}}$. Therefore, we can conclude that

$$(\mathrm{I}) \le \|\boldsymbol{V}^\star-\boldsymbol{V}_t\|_{\mathrm{F}} + 2\eta\sqrt{r\sigma_1^\star}\,\|\boldsymbol{E}_t\| + \eta\sqrt{r}\cdot 21\Gamma\sigma_1^\star\sqrt{\frac{\mu^2r^2}{pd}} \le \|\boldsymbol{V}^\star-\boldsymbol{V}_t\|_{\mathrm{F}} + 22\Gamma\eta\frac{\sigma_1^\star\mu r^{1.5}}{\sqrt{pd}}. \tag{4.107}$$

Next, according to Lemma 68, we can control $\|\boldsymbol{A}_t\|_{\mathrm{F}}$ as

$$\|\boldsymbol{A}_t\|_{\mathrm{F}} \le \sqrt{r}\,\|\boldsymbol{A}_t\| \le 300\sqrt{r}\eta^2\sigma_1^{\star 2}. \tag{4.108}$$

Putting everything together, we have

$$\begin{aligned}\|\boldsymbol{V}^\star-\boldsymbol{V}_{t+1}\|_{\mathrm{F}} &\le \|\boldsymbol{V}^\star-\boldsymbol{V}_t\|_{\mathrm{F}} + 22\Gamma\eta\frac{\sigma_1^\star\mu r^{1.5}}{\sqrt{pd}} + 300\sqrt{r}\eta^2\sigma_1^{\star 2}\\ &\le \|\boldsymbol{V}^\star-\boldsymbol{V}_t\|_{\mathrm{F}} + 23\Gamma\eta\frac{\sigma_1^\star\mu r^{1.5}}{\sqrt{pd}},\end{aligned} \tag{4.109}$$

provided that $\eta \asymp \frac{\mu r}{\sqrt{pd}\sigma_1^\star}$. This completes the proof. □

4.D.2 Proof of Proposition 12

We restate the proposition here for clarity.

Proposition 16 (Dynamic of $\left\|\left(\boldsymbol{V}^\star-\boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\|$)**.** *Under the same conditions as Proposition 15, for all $t \le T \lesssim \frac{1}{\eta\sigma_r^\star}\log\left(\frac{1}{\alpha}\right)$, we have*

$$\left\|\left(\boldsymbol{V}^\star-\boldsymbol{V}_{t+1}^{(l)}\right)_{l,\cdot}\right\| \le (1-0.5\eta\sigma_r^\star)\left\|\left(\boldsymbol{V}^\star-\boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\| + \Gamma_2\eta\sigma_1^\star\frac{\kappa\mu^{1.5}r^2\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd^2}}. \tag{4.110}$$

Proof. Similar to Equation (4.100), we can express $\boldsymbol{V}_{t+1}^{(l)}$ as

$$\boldsymbol{V}_{t+1}^{(l)} = \left(\boldsymbol{I}+\eta\mathcal{P}^\perp_{\boldsymbol{V}_t^{(l)}}\boldsymbol{M}_t^{(l)}\right)\boldsymbol{V}_t^{(l)} + \boldsymbol{A}_t^{(l)}, \tag{4.111}$$

where $\boldsymbol{M}_t^{(l)} = \mathcal{R}_{\Omega^{(l)}}\left(\boldsymbol{X}^\star - \boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}_t\boldsymbol{V}_t^{(l)\top}\right)$ and $\boldsymbol{A}_t^{(l)}$ is defined as

$$\begin{aligned}\boldsymbol{A}_t^{(l)} = &-\eta^2\boldsymbol{M}_t^{(l)}\boldsymbol{V}_t^{(l)}\boldsymbol{V}_t^{(l)\top}\boldsymbol{M}_t^{(l)}\boldsymbol{V}_t^{(l)} - 0.5\eta^2\boldsymbol{V}_t^{(l)}\boldsymbol{V}_t^{(l)\top}\boldsymbol{M}_t^{(l)2}\boldsymbol{V}_t^{(l)}\\ &- 0.5\eta^3\boldsymbol{M}_t^{(l)}\boldsymbol{V}_t^{(l)}\boldsymbol{V}^{(l)\top}\boldsymbol{M}_t^{(l)2}\boldsymbol{V}_t^{(l)} + (\boldsymbol{I}+\eta\boldsymbol{M}_t)\,\boldsymbol{V}_t^{(l)}\boldsymbol{R}\left(\boldsymbol{Y}_t^{(l)}\right)\end{aligned} \tag{4.112}$$

containing all the higher-order terms. Applying triangle inequality yields

$$\begin{aligned}\left\|\left(\boldsymbol{V}^\star - \boldsymbol{V}_{t+1}^{(l)}\right)_{l,\cdot}\right\| &\le \left\|\left(\boldsymbol{V}^\star - \left(\boldsymbol{I}+\eta\mathcal{P}^\perp_{\boldsymbol{V}_t^{(l)}}\boldsymbol{M}_t^{(l)}\right)\boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\| + \left\|\left(\boldsymbol{A}_t^{(l)}\right)_{l,\cdot}\right\|\\ &= \underbrace{\left\|\left(\boldsymbol{V}^\star - \left(\boldsymbol{I}+\eta\mathcal{P}^\perp_{\boldsymbol{V}_t^{(l)}}\boldsymbol{\Xi}_t^{(l)}\right)\boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\|}_{:=(\mathrm{I})} + \left\|\left(\boldsymbol{A}_t^{(l)}\right)_{l,\cdot}\right\|.\end{aligned} \tag{4.113}$$

Here in the last equality, we use the fact that $\mathcal{P}^\perp_{\boldsymbol{V}_t^{(l)}}\boldsymbol{M}_t^{(l)} = \mathcal{P}^\perp_{\boldsymbol{V}_t^{(l)}}\boldsymbol{\Xi}_t^{(l)}$ where $\boldsymbol{\Xi}_t^{(l)} = \boldsymbol{M}_t^{(l)} - \boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top} + \boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}_t\boldsymbol{V}_t^{(l)\top}$. Upon noticing that $\mathcal{P}^\perp_{\boldsymbol{V}_t^{(l)}} = \boldsymbol{I} - \mathcal{P}_{\boldsymbol{V}_t^{(l)}}$, we further decompose (I) as follows

$$(\mathrm{I}) \le \underbrace{\left\|\left(\boldsymbol{V}^\star - \left(\boldsymbol{I}+\eta\boldsymbol{\Xi}_t^{(l)}\right)\boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\|}_{:=(\mathrm{I}_1)} + \eta\underbrace{\left\|\left(\mathcal{P}_{\boldsymbol{V}_t^{(l)}}\boldsymbol{\Xi}_t^{(l)}\boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\|}_{:=(\mathrm{I}_2)}. \tag{4.114}$$

To control (I_1), notice that the l-th row of $\boldsymbol{M}_t^{(l)}$ is equal to the l-th row of $\boldsymbol{X}^\star - \boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}_t\boldsymbol{V}_t^{(l)\top}$ due to our choice of $\mathcal{R}_{\Omega^{(l)}}$. Therefore, we have

$$\begin{aligned}(\mathrm{I}_1) &= \left\|\left(\boldsymbol{V}^\star - \left(\boldsymbol{I}+\eta\left(\boldsymbol{X}^\star - \boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}_t\boldsymbol{V}_t^{(l)\top} - \boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top} - \boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}_t\boldsymbol{V}_t^{(l)\top}\right)\right)\boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\|\\ &= \left\|\left(\boldsymbol{V}^\star - \left(\boldsymbol{I}+\eta\left(\left(\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right)\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}\right)\right)\boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\|\\ &= \left\|\left(\left(\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right)\left(\boldsymbol{I} - \eta\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}\boldsymbol{V}_t^{(l)}\right)\right)_{l,\cdot}\right\|\\ &\le \left\|\left(\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\|\left\|\boldsymbol{I} - \eta\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}\boldsymbol{V}_t^{(l)}\right\|\\ &\overset{(a)}{\le} (1 - 0.5\eta\sigma_r^\star)\left\|\left(\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\|.\end{aligned} \tag{4.115}$$

Here in (a), we use the fact that $\left\|\boldsymbol{I} - \eta\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}\boldsymbol{V}_t^{(l)}\right\| \le \|\boldsymbol{I} - \eta\boldsymbol{\Sigma}^\star\| + \eta\sigma_1^\star\left\|\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right\| \le 1 - \eta(\sigma_r^\star - \sigma_1^\star\left\|\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right\|) \le 1 - 0.5\eta\sigma_r^\star$ since we have $\left\|\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right\| \le \frac{1}{2\kappa}$ according to the following proposition.

Lemma 70 (Frobenius norm control)**.** *Under the same conditions as Proposition 15, for*

any $1 \le t \le T = \frac{100}{\eta\sigma_r^\star}\log\left(\frac{1}{\alpha}\right)$*, for all* $1 \le l \le d$*, we have*

$$\left\|\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}} \le \Gamma_1 \frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}. \tag{4.116}$$

The proof of the above lemma is the same as that of Proposition 15 and hence omitted here. Next, for (I_2), we first have

$$(\mathrm{I}_2) = \left\|\left(\boldsymbol{V}_t^{(l)}\boldsymbol{V}_t^{(l)\top}\boldsymbol{\Xi}_t^{(l)}\boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\| \le \left\|\left(\boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\|\left\|\boldsymbol{V}_t^{(l)\top}\boldsymbol{\Xi}_t^{(l)}\boldsymbol{V}_t^{(l)}\right\| \le \left\|\left(\boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\|\left\|\boldsymbol{\Xi}_t^{(l)}\right\|. \tag{4.117}$$

For the first part, we have

$$\left\|\left(\boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\| \le \left\|\boldsymbol{V}_{l,\cdot}^\star\right\| + \left\|\left(\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\| \le \sqrt{\frac{\mu r}{d}} + \left\|\left(\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\| \le 2\sqrt{\frac{\mu r}{d}}. \tag{4.118}$$

Here we use the assumption that $\left\|\left(\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\| \le \sqrt{\frac{\mu r}{d}}$. For the second part, we have

$$\begin{aligned}
\left\|\boldsymbol{\Xi}_t^{(l)}\right\| &= \left\|\left(\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right)\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top} - \left(\mathrm{I} - \mathcal{R}_{\Omega^{(l)}}\right)\left(\boldsymbol{X}^\star - \boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}_t\boldsymbol{V}_t^{(l)\top}\right)\right\| \\
&\le \left\|\left(\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right)\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}\right\| + \left\|\left(\mathrm{I} - \mathcal{R}_{\Omega^{(l)}}\right)\left(\boldsymbol{X}^\star - \boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}_t\boldsymbol{V}_t^{(l)\top}\right)\right\| \\
&\overset{(a)}{\le} \sigma_1^\star\left\|\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right\| + 21\Gamma\frac{\sigma_1^\star\mu r}{\sqrt{pd}} \\
&\overset{(b)}{\le} \sigma_1^\star\cdot\Gamma_1\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}} + 21\Gamma\frac{\sigma_1^\star\mu r}{\sqrt{pd}} \\
&\le 2\Gamma_1\sigma_1^\star\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}.
\end{aligned} \tag{4.119}$$

Here in (a), we apply Lemma 67. In (b), we apply Lemma 70. Invoking Equations (4.118) and (4.119) in Equation (4.117), we obtain that

$$(\mathrm{I}_2) \le 2\sqrt{\frac{\mu r}{d}}\cdot 2\Gamma_1\sigma_1^\star\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}} = 4\Gamma_1\sigma_1^\star\frac{\kappa\mu^{1.5}r^2\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd^2}}. \tag{4.120}$$

Next, we control the higher-order term $\left\|\left(\boldsymbol{A}_t^{(l)}\right)_{l,\cdot}\right\|$. To this end, we first notice that the l-th

row of $\boldsymbol{A}_t^{(l)}$ is equal to the l-th row of

$$\begin{aligned}&-\eta^2\boldsymbol{\Delta}_t^{(l)}\mathcal{P}_{\boldsymbol{V}_t^{(l)}}\boldsymbol{V}_t^{(l)\top}\boldsymbol{M}_t^{(l)}\boldsymbol{V}_t^{(l)}-0.5\eta^2\mathcal{P}_{\boldsymbol{V}_t^{(l)}}\boldsymbol{M}_t^{(l)2}\boldsymbol{V}_t^{(l)}-0.5\eta^3\boldsymbol{\Delta}_t^{(l)}\mathcal{P}_{\boldsymbol{V}_t^{(l)}}\boldsymbol{M}_t^{(l)2}\boldsymbol{V}_t^{(l)}\\&\quad+\left(\boldsymbol{I}+\eta\boldsymbol{\Delta}_t^{(l)}\right)\boldsymbol{V}_t^{(l)}\boldsymbol{R}\left(\boldsymbol{Y}_t^{(l)}\right).\end{aligned}\tag{4.121}$$

Therefore, we can upper-bound its operator norm by

$$\begin{aligned}\left\|\left(\boldsymbol{A}_t^{(l)}\right)_{l,\cdot}\right\|\leq&\left\|\left(\boldsymbol{\Delta}_t^{(l)}\right)_{l,\cdot}\right\|\left(\eta^2\left\|\boldsymbol{M}_t^{(l)}\right\|+0.5\eta^3\left\|\boldsymbol{M}_t^{(l)}\right\|^2+\eta\left\|\boldsymbol{R}\left(\boldsymbol{Y}_t^{(l)}\right)\right\|\right)\\&+\left\|\left(\boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\|\left(0.5\eta^2\left\|\boldsymbol{M}_t^{(l)}\right\|^2+\left\|\boldsymbol{R}\left(\boldsymbol{Y}_t^{(l)}\right)\right\|\right).\end{aligned}\tag{4.122}$$

Next, we notice that $\left\|\left(\boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\|\leq 2\sqrt{\frac{\mu r}{d}}$ and

$$\begin{aligned}\left\|\left(\boldsymbol{\Delta}_t^{(l)}\right)_{l,\cdot}\right\|&\leq\left\|\left(\boldsymbol{X}^\star-\boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}_t\boldsymbol{V}_t^{(l)\top}\right)_{l,\cdot}\right\|\\&\leq\sigma_1^\star\sqrt{\frac{\mu r}{d}}+\left\|\left(\boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\|\|\boldsymbol{\Sigma}_t\|\\&\leq 5\sigma_1^\star\sqrt{\frac{\mu r}{d}}.\end{aligned}\tag{4.123}$$

On the other hand, according to Lemma 68, we have $\left\|\boldsymbol{M}_t^{(l)}\right\|\leq 6\sqrt{\sigma_1^\star}$ and $\left\|\boldsymbol{R}\left(\boldsymbol{Y}_t^{(l)}\right)\right\|\leq 200\eta^2\sigma_1^{\star 2}$. Therefore, we derive that

$$\left\|\left(\boldsymbol{A}_t^{(l)}\right)_{l,\cdot}\right\|\leq 470\eta^2\sigma_1^{\star 2}\sqrt{\frac{\mu r}{d}}.\tag{4.124}$$

Overall, we obtain

$$\begin{aligned}\left\|\left(\boldsymbol{V}^\star-\boldsymbol{V}_{t+1}^{(l)}\right)_{l,\cdot}\right\|&\leq(1-0.5\eta\sigma_r^\star)\left\|\left(\boldsymbol{V}^\star-\boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\|+4\Gamma_1\eta\sigma_1^\star\frac{\kappa\mu^{1.5}r^2\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd^2}}+470\eta^2\sigma_1^{\star 2}\sqrt{\frac{\mu r}{d}}\\&\leq(1-0.5\eta\sigma_r^\star)\left\|\left(\boldsymbol{V}^\star-\boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\|+8\Gamma_1\eta\sigma_1^\star\frac{\kappa\mu^{1.5}r^2\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd^2}},\end{aligned}\tag{4.125}$$

provided that $\eta\asymp\frac{\mu r}{\sqrt{pd}\sigma_1^\star}$. This completes the proof of Proposition 16. □

4.D.3 Proof of Proposition 13

We restate the proposition here for clarity.

Proposition 17 (One-step dynamic of $\left\|\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}$)**.** *Suppose that the sampling rate satisfies* $p \gtrsim \frac{\kappa^6\mu^4 r^9 \log^4(\frac{1}{\alpha})\log^2(d)}{d}$. *Suppose that* $\|\boldsymbol{V}^\star-\boldsymbol{V}_t\| \le \frac{1}{2\kappa}$ *and* $\left\|\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}} \le \sqrt{\frac{\mu r}{4d}}$. *With probability at least* $1-\frac{1}{d^3}$*, for any* $1 \le t \le T \lesssim \frac{1}{\eta\sigma_r^\star}\log\left(\frac{1}{\alpha}\right)$*, we have*

$$\left\|\boldsymbol{V}_{t+1}-\boldsymbol{V}_{t+1}^{(l)}\right\|_{\mathrm{F}} \le \left\|\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}} + \Gamma\eta\sigma_1^\star\sqrt{\frac{\kappa\mu^3 r^{5.5}\log\left(\frac{1}{\alpha}\right)\log(d)}{\sqrt{pd}\cdot d}}. \tag{4.126}$$

Proof. Note that

$$\boldsymbol{V}_{t+1} = \left(\boldsymbol{I}+\eta\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\boldsymbol{M}_t\right)\boldsymbol{V}_t + \boldsymbol{A}_t \quad \text{and} \quad \boldsymbol{V}_{t+1}^{(l)} = \left(\boldsymbol{I}+\eta\mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\boldsymbol{M}_t^{(l)}\right)\boldsymbol{V}_t^{(l)} + \boldsymbol{A}_t^{(l)}. \tag{4.127}$$

Hence, we can expand $\left\|\boldsymbol{V}_{t+1}-\boldsymbol{V}_{t+1}^{(l)}\right\|_{\mathrm{F}}^2$ as

$$\left\|\boldsymbol{V}_{t+1}-\boldsymbol{V}_{t+1}^{(l)}\right\|_{\mathrm{F}}^2 = \left\|\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}^2 + 2\eta\underbrace{\left\langle \boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}, \mathcal{P}_{\boldsymbol{V}_t}^{\perp}\boldsymbol{M}_t\boldsymbol{V}_t - \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\boldsymbol{M}_t^{(l)}\boldsymbol{V}_t^{(l)}\right\rangle}_{:=(\mathrm{I})} + (\mathrm{II}), \tag{4.128}$$

where

$$\begin{aligned}(\mathrm{II}) = {}& \eta^2\left\|\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\boldsymbol{M}_t\boldsymbol{V}_t - \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\boldsymbol{M}_t^{(l)}\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}^2 + \left\|\boldsymbol{A}_t-\boldsymbol{A}_t^{(l)}\right\|_{\mathrm{F}}^2 \\ & + 2\left\langle\left(\boldsymbol{I}+\eta\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\boldsymbol{M}_t\right)\boldsymbol{V}_t - \left(\boldsymbol{I}+\eta\mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\boldsymbol{M}_t^{(l)}\right)\boldsymbol{V}_t^{(l)}, \boldsymbol{A}_t-\boldsymbol{A}_t^{(l)}\right\rangle\end{aligned} \tag{4.129}$$

contains all the higher-order terms.

We first control (I). Notice that $\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\boldsymbol{M}_t = \mathcal{P}_{\boldsymbol{V}_t}^{\perp}\boldsymbol{\Xi}_t$ and $\mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\boldsymbol{M}_t^{(l)} = \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\boldsymbol{\Xi}_t^{(l)}$. Here we define $\boldsymbol{\Xi}_t = \boldsymbol{M}_t - \boldsymbol{V}_t\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top} + \boldsymbol{V}_t\boldsymbol{\Sigma}_t\boldsymbol{V}_t^\top$ and $\boldsymbol{\Xi}_t^{(l)} = \boldsymbol{M}_t^{(l)} - \boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top} + \boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}_t\boldsymbol{V}_t^{(l)\top}$, respectively. Therefore, we can decompose (I) as follows

$$\begin{aligned}(\mathrm{I}) &= \left\langle \boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}, \mathcal{P}_{\boldsymbol{V}_t}^{\perp}\boldsymbol{\Xi}_t\boldsymbol{V}_t - \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\boldsymbol{\Xi}_t^{(l)}\boldsymbol{V}_t^{(l)}\right\rangle \\ &= \underbrace{\left\langle \boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}, \left(\mathcal{P}_{\boldsymbol{V}_t}^{\perp}-\mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\right)\boldsymbol{\Xi}_t^{(l)}\boldsymbol{V}_t^{(l)}\right\rangle}_{:=(\mathrm{I}_1)} + \underbrace{\left\langle \boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}, \mathcal{P}_{\boldsymbol{V}_t}^{\perp}\boldsymbol{\Xi}_t\left(\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right)\right\rangle}_{:=(\mathrm{I}_2)} \\ &\quad + \underbrace{\left\langle \boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}, \mathcal{P}_{\boldsymbol{V}_t}^{\perp}\left(\boldsymbol{\Xi}_t-\boldsymbol{\Xi}_t^{(l)}\right)\boldsymbol{V}_t^{(l)}\right\rangle}_{:=(\mathrm{I}_3)}.\end{aligned} \tag{4.130}$$

Next, we provide upper bounds for these terms separately. For (I_1), applying Cauchy-Schwarz

inequality leads to

$$\begin{aligned}
(\mathrm{I}_1) &\leq \left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}} \left\|\left(\mathcal{P}_{\boldsymbol{V}_t}^{\perp} - \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\right)\boldsymbol{\Xi}_t^{(l)}\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}} \\
&\leq \left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}} \left\|\mathcal{P}_{\boldsymbol{V}_t}^{\perp} - \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\right\|_{\mathrm{F}} \left\|\boldsymbol{\Xi}_t^{(l)}\right\| \\
&\overset{(a)}{\leq} 2\left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}^2 \left\|\boldsymbol{\Xi}_t^{(l)}\right\| \\
&\overset{(b)}{\leq} 4\left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}^2 \cdot \Gamma_1\sigma_1^{\star}\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}.
\end{aligned} \tag{4.131}$$

Here in (a), we apply Lemma 90. In (b), we use the result from Equation (4.119). Similarly, (I_2), we have

$$(\mathrm{I}_2) \leq \left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}^2 \|\boldsymbol{\Xi}_t\| \leq 2\left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}^2 \cdot \Gamma_1\sigma_1^{\star}\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}. \tag{4.132}$$

Next, for (I_3), we further decompose it as

$$\begin{aligned}
(\mathrm{I}_3) &= \left\langle \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}, \mathcal{P}_{\boldsymbol{V}_t}^{\perp}\left(\boldsymbol{M}_t - \boldsymbol{M}_t^{(l)} + \boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}^{\star}\boldsymbol{V}^{\star\top} - \boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}_t\boldsymbol{V}_t^{(l)\top}\right)\boldsymbol{V}_t^{(l)}\right\rangle \\
&= \underbrace{\left\langle \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}, \mathcal{P}_{\boldsymbol{V}_t}^{\perp}\left(\mathcal{R}_{\Omega} - \mathcal{R}_{\Omega^{(l)}}\right)\left(\boldsymbol{X}^{\star} - \boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}_t\boldsymbol{V}_t^{(l)\top}\right)\boldsymbol{V}_t^{(l)}\right\rangle}_{:=(\mathrm{I}_{3,1})} \\
&\quad + \underbrace{\left\langle \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}, \mathcal{P}_{\boldsymbol{V}_t}^{\perp}\left(\mathrm{I} - \mathcal{R}_{\Omega}\right)\left(\boldsymbol{V}_t\boldsymbol{\Sigma}_t\left(\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right)^{\top}\right)\boldsymbol{V}_t^{(l)}\right\rangle}_{:=(\mathrm{I}_{3,2})} \\
&\quad + \underbrace{\left\langle \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}, \mathcal{P}_{\boldsymbol{V}_t}^{\perp}\left(\mathrm{I} - \mathcal{R}_{\Omega}\right)\left(\left(\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right)\boldsymbol{\Sigma}_t\boldsymbol{V}_t^{(l)\top}\right)\boldsymbol{V}_t^{(l)}\right\rangle}_{:=(\mathrm{I}_{3,3})} \\
&\quad + \underbrace{\left\langle \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}, \mathcal{P}_{\boldsymbol{V}_t}^{\perp}\boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}^{\star}\boldsymbol{V}^{\star\top}\boldsymbol{V}_t^{(l)}\right\rangle}_{:=(\mathrm{I}_{3,4})}.
\end{aligned} \tag{4.133}$$

We control these terms separately. First, we present the following key proposition, the proof of which is deferred to the end of this section.

Proposition 18. *For all $0 \leq t \leq T$, we have*

$$\left\|\left(\mathcal{R}_{\Omega} - \mathcal{R}_{\Omega^{(l)}}\right)\left(\boldsymbol{X}^{\star} - \boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}_t\boldsymbol{V}_t^{(l)\top}\right)\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}} \leq \Gamma_3\sigma_1^{\star}\sqrt{\frac{\kappa\mu^3 r^{5.5}\log\left(\frac{1}{\alpha}\right)\log\left(d\right)}{\sqrt{pd}\cdot d}}. \tag{4.134}$$

For $(\mathrm{I}_{3,1})$, we have

$$\begin{aligned}(\mathrm{I}_{3,1}) &\leq \left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}} \left\|(\mathcal{R}_\Omega - \mathcal{R}_{\Omega^{(l)}})\left(\boldsymbol{X}^\star - \boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}_t\boldsymbol{V}_t^{(l)\top}\right)\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}} \\ &\leq \Gamma_3\sigma_1^\star\sqrt{\frac{\kappa\mu^3 r^{5.5}\log\left(\frac{1}{\alpha}\right)\log(d)}{\sqrt{pd}\cdot d}}\left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}.\end{aligned} \tag{4.135}$$

Here in the last inequality, we use Proposition 18. Next, we apply Corollary 6 to control $(\mathrm{I}_{3,2})$. Specifically, upon setting $\boldsymbol{A} = \boldsymbol{V}_t$, $\boldsymbol{B} = \mathcal{P}_{\boldsymbol{V}_t}^{\perp}\left(\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right)$, $\boldsymbol{C} = \left(\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right)\boldsymbol{\Sigma}_t$, and $\boldsymbol{D} = \boldsymbol{V}_t^{(l)}$ in Corollary 6, with probability at least $1 - d^{-3}$, one has

$$\begin{aligned}(\mathrm{I}_{3,2}) &= \left\langle(\mathrm{I} - \mathcal{R}_\Omega)\left(\boldsymbol{A}\boldsymbol{C}^\top\right), \boldsymbol{B}\boldsymbol{D}^\top\right\rangle \\ &\leq \Gamma\sqrt{\frac{d}{p}}\|\boldsymbol{A}\|_{2,\infty}\|\boldsymbol{B}\|_{\mathrm{F}}\cdot\|\boldsymbol{C}\|_{\mathrm{F}}\|\boldsymbol{D}\|_{2,\infty} \\ &= \Gamma\sqrt{\frac{d}{p}}\|\boldsymbol{V}_t\|_{2,\infty}\left\|\boldsymbol{V}_t^{(l)}\right\|_{2,\infty}\left\|\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\left(\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right)\right\|_{\mathrm{F}}\left\|\left(\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right)\boldsymbol{\Sigma}_t\right\|_{\mathrm{F}} \\ &\leq 16\Gamma\sigma_1^\star\sqrt{\frac{\mu^2 r^2}{pd}}\left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}^2.\end{aligned} \tag{4.136}$$

Similarly, we apply Corollary 6 to control $(\mathrm{I}_{3,3})$. Upon setting $\boldsymbol{A} = \left(\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right)\boldsymbol{\Sigma}_t$, $\boldsymbol{B} = \mathcal{P}_{\boldsymbol{V}_t}^{\perp}\left(\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right)$, $\boldsymbol{C} = \boldsymbol{V}_t$, and $\boldsymbol{D} = \boldsymbol{V}_t^{(l)}$ in Corollary 6, with probability at least $1 - d^{-3}$, one has

$$\begin{aligned}(\mathrm{I}_{3,3}) &\leq \Gamma\sqrt{\frac{d}{p}}\|\boldsymbol{V}_t\|_{\mathrm{F}}\left\|\boldsymbol{V}_t^{(l)}\right\|_{2,\infty}\left\|\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\left(\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right)\right\|_{\mathrm{F}}\left\|\left(\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right)\boldsymbol{\Sigma}_t\right\|_{2,\infty} \\ &\leq 8\Gamma\sigma_1^\star\sqrt{\frac{\mu r^2}{p}}\left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}\left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{2,\infty}.\end{aligned} \tag{4.137}$$

To control $\left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{2,\infty}$, we apply Lemma 86:

$$\begin{aligned}\left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{2,\infty} &\leq \left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}\left\|\boldsymbol{L}_{\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}}\right\|_{2,\infty} \\ &\leq \left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}\sqrt{\|\boldsymbol{L}_{\boldsymbol{V}_t}\|_{2,\infty}^2 + \left\|\boldsymbol{L}_{\boldsymbol{V}_t^{(l)}}\right\|_{2,\infty}^2} \\ &\leq 3\sqrt{\frac{\mu r}{d}}\left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}.\end{aligned} \tag{4.138}$$

Hence, we have

$$(\mathrm{I}_{3,3}) \leq 24\Gamma\sigma_1^\star\sqrt{\frac{\mu^2 r^3}{pd}}\left\|\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}^2. \tag{4.139}$$

Lastly, for $(\mathrm{I}_{3,4})$, we notice that

$$\begin{aligned}
(\mathrm{I}_{3,4}) &= -\left\langle \boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}, \mathcal{P}_{\boldsymbol{V}_t}^{\perp}\left(\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right)\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}\boldsymbol{V}_t^{(l)}\right\rangle \\
&\leq -\left\langle \boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}, \mathcal{P}_{\boldsymbol{V}_t}^{\perp}\left(\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right)\boldsymbol{\Sigma}^\star\right\rangle + \sigma_1^\star\left\|\boldsymbol{V}^\star-\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}\left\|\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}^2 \\
&\overset{(a)}{\leq} \sigma_1^\star\left\|\boldsymbol{V}^\star-\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}\left\|\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}^2 \\
&\overset{(b)}{\leq} \Gamma_1\sigma_1^\star\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\left\|\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}^2.
\end{aligned} \tag{4.140}$$

Here in (a), we use the fact that

$$\left\langle \boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}, \mathcal{P}_{\boldsymbol{V}_t}^{\perp}\left(\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right)\boldsymbol{\Sigma}^\star\right\rangle = \left\|\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\left(\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right)\boldsymbol{\Sigma}^{\star 1/2}\right\|_{\mathrm{F}}^2 \geq 0. \tag{4.141}$$

In (b), we apply Proposition 15. Therefore, we obtain

$$(\mathrm{I}_3) \leq C_1\sigma_1^\star\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\left\|\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}^2 + \Gamma_3\sigma_1^\star\sqrt{\frac{\kappa\mu^3 r^{5.5}\log\left(\frac{1}{\alpha}\right)\log(d)}{\sqrt{pd}\cdot d}}\left\|\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}, \tag{4.142}$$

which implies that

$$(\mathrm{I}) \leq C_2\sigma_1^\star\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\left\|\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}^2 + \Gamma_3\sigma_1^\star\sqrt{\frac{\kappa\mu^3 r^{5.5}\log\left(\frac{1}{\alpha}\right)\log(d)}{\sqrt{pd}\cdot d}}\left\|\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}. \tag{4.143}$$

Next, we move on to controlling (II). First, by the basic inequality $2ab \leq a^2+b^2$, we bound (II) as

$$(\mathrm{II}) \leq 2\eta^2\left\|\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\boldsymbol{M}_t\boldsymbol{V}_t - \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\boldsymbol{M}_t^{(l)}\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}^2 + 2\left\|\boldsymbol{A}_t-\boldsymbol{A}_t^{(l)}\right\|_{\mathrm{F}}^2 + 2\left\langle \boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}, \boldsymbol{A}_t-\boldsymbol{A}_t^{(l)}\right\rangle. \tag{4.144}$$

Next, we provide the control over $\left\|\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\boldsymbol{M}_t\boldsymbol{V}_t - \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\boldsymbol{M}_t^{(l)}\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}$. The remaining two terms can be controlled in a similar fashion, so we omit their analysis. We first apply triangle inequality to obtain

$$\begin{aligned}
\left\|\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\boldsymbol{M}_t\boldsymbol{V}_t - \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\boldsymbol{M}_t^{(l)}\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}} &\leq 2\left\|\boldsymbol{M}_t\right\|\left\|\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}} + \left\|\left(\boldsymbol{M}_t-\boldsymbol{M}_t^{(l)}\right)\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}} \\
&\leq 12\sigma_1^\star\left\|\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}} + \left\|\left(\boldsymbol{M}_t-\boldsymbol{M}_t^{(l)}\right)\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}.
\end{aligned} \tag{4.145}$$

For the second term, we further decompose it as

$$\begin{aligned}\left\|\left(\boldsymbol{M}_t-\boldsymbol{M}_t^{(l)}\right)\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}} \leq & \left\|(\mathcal{R}_\Omega-\mathcal{R}_{\Omega^{(l)}})\left(\boldsymbol{X}^\star-\boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}_t\boldsymbol{V}_t^{(l)\top}\right)\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}} \\ & +\left\|(\mathrm{I}-\mathcal{R}_\Omega)\left(\boldsymbol{V}_t\boldsymbol{\Sigma}_t\left(\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right)^\top\right)\boldsymbol{V}_t\right\|_{\mathrm{F}} \\ & +\left\|(\mathrm{I}-\mathcal{R}_\Omega)\left(\left(\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right)\boldsymbol{\Sigma}_t\boldsymbol{V}_t^\top\right)\boldsymbol{V}_t\right\|_{\mathrm{F}} \\ & +4\sigma_1^\star\left\|\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}.\end{aligned} \tag{4.146}$$

The first term can be controlled by Proposition 18. For the second term, we have

$$\begin{aligned}\left\|(\mathrm{I}-\mathcal{R}_\Omega)\left(\boldsymbol{V}_t\boldsymbol{\Sigma}_t\left(\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right)^\top\right)\boldsymbol{V}_t\right\|_{\mathrm{F}} &= \sup_{\|\boldsymbol{Z}\|_{\mathrm{F}}\leq 1}\left\langle(\mathrm{I}-\mathcal{R}_\Omega)\left(\boldsymbol{V}_t\boldsymbol{\Sigma}_t\left(\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right)^\top\right),\boldsymbol{Z}\boldsymbol{V}_t\right\rangle \\ &\overset{(a)}{\leq}\Gamma\sqrt{\frac{d}{p}}\left\|\boldsymbol{V}_t\right\|_{2,\infty}^2\left\|\left(\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right)\boldsymbol{\Sigma}_t\right\|_{\mathrm{F}} \\ &\leq 8\Gamma\sigma_1^\star\sqrt{\frac{\mu^2r^2}{pd}}\left\|\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}.\end{aligned} \tag{4.147}$$

Here we apply Corollary 6 in (a). Similarly, we can also control the third term as follows

$$\begin{aligned}\left\|(\mathrm{I}-\mathcal{R}_\Omega)\left(\left(\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right)\boldsymbol{\Sigma}_t\boldsymbol{V}_t^\top\right)\boldsymbol{V}_t\right\|_{\mathrm{F}} &= \sup_{\|\boldsymbol{Z}\|_{\mathrm{F}}\leq 1}\left\langle(\mathrm{I}-\mathcal{R}_\Omega)\left(\left(\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right)\boldsymbol{\Sigma}_t\boldsymbol{V}_t^\top\right),\boldsymbol{Z}\boldsymbol{V}_t\right\rangle \\ &\leq 8\Gamma\sqrt{\frac{\mu r^2}{p}}\left\|\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right\|_{2,\infty} \\ &\leq 24\Gamma\sigma_1^\star\sqrt{\frac{\mu^2r^3}{pd}}\left\|\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}.\end{aligned} \tag{4.148}$$

Combining the above inequalities leads to

$$\left\|\mathcal{P}_{\boldsymbol{V}_t}^\perp\boldsymbol{M}_t\boldsymbol{V}_t-\mathcal{P}_{\boldsymbol{V}_t^{(l)}}^\perp\boldsymbol{M}_t^{(l)}\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}\leq C\sigma_1^\star\left\|\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}. \tag{4.149}$$

Similarly, we can derive that $\left\|\boldsymbol{A}_t-\boldsymbol{A}_t^{(l)}\right\|_{\mathrm{F}}\leq C\sigma_1^\star\left\|\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}+C\sigma_1^\star\sqrt{\frac{\kappa\mu^3r^{5.5}\log\left(\frac{1}{\alpha}\right)\log(d)}{\sqrt{pd}\cdot d}}$. Therefore, we have

$$(\mathrm{II})\leq C_3\eta^2\left(\sigma_1^\star\left\|\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}^2+\sigma_1^{\star 2}\frac{\kappa\mu^3r^{5.5}\log\left(\frac{1}{\alpha}\right)\log(d)}{\sqrt{pd}\cdot d}\right). \tag{4.150}$$

Putting everything together, we obtain

$$\begin{aligned}\left\|\boldsymbol{V}_{t+1}-\boldsymbol{V}_{t+1}^{(l)}\right\|_{\mathrm{F}} &\leq \left(1+\Gamma_1\eta\sigma_1^\star\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\right)\left\|\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}+\Gamma_3\eta\sigma_1^\star\sqrt{\frac{\kappa\mu^3 r^{5.5}\log\left(\frac{1}{\alpha}\right)\log\left(d\right)}{\sqrt{pd}\cdot d}} \\ &\leq \left\|\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}+2\Gamma_3\eta\sigma_1^\star\sqrt{\frac{\kappa\mu^3 r^{5.5}\log\left(\frac{1}{\alpha}\right)\log\left(d\right)}{\sqrt{pd}\cdot d}}.\end{aligned} \tag{4.151}$$

□

4.D.4 Proof of Proposition 18

To prove this proposition, we propose a novel argument based on *adaptive ϵ-nets*. Upon fixing an ϵ-net $\mathcal{N}_\epsilon$ with respect to Frobenius norm for the operator norm ball $\mathcal{B}_{\mathrm{op}}^{r\times r}(4\sigma_1^\star)$ with $\epsilon=\frac{c}{d}$[2], we first construct a series of adaptive ϵ-nets $\left\{\mathcal{V}_{\epsilon,t}^{(l)}\right\}_{t=0}^{T}$ in the following recursive manner.

$$\begin{aligned}\mathcal{V}_{\epsilon,t+1}^{(l)}=\Bigg\{\boldsymbol{Y}_{\epsilon,t+1}^{(l)}:\left\|\boldsymbol{Y}_{\epsilon,t+1}^{(l)}\right\|_{2,\infty}\leq 2\sqrt{\frac{\mu r}{d}},\ \text{and}\ \boldsymbol{Y}_{\epsilon,t+1}^{(l)}=\boldsymbol{Z}_{\epsilon,t+1}^{(l)}\left(\boldsymbol{Z}_{\epsilon,t+1}^{(l)\top}\boldsymbol{Z}_{\epsilon,t+1}^{(l)}\right)^{-1/2}\ \text{where} \\ \boldsymbol{Z}_{\epsilon,t+1}^{(l)}=\left(\boldsymbol{I}+\eta\mathcal{R}_{\Omega^{(l)}}\left(\boldsymbol{X}^\star-\boldsymbol{V}_{\epsilon,t}^{(l)}\boldsymbol{\Sigma}_{\epsilon,t}\boldsymbol{V}_{\epsilon,t}^{(l)\top}\right)\right)\boldsymbol{V}_{\epsilon,t}^{(l)},\forall\boldsymbol{V}_{\epsilon,t}^{(l)}\in\mathcal{V}_{\epsilon,t}^{(l)},\boldsymbol{\Sigma}_{\epsilon,t}\in\mathcal{N}_\epsilon\Bigg\}.\end{aligned} \tag{4.152}$$

[2]We call a set $\mathcal{N}_\epsilon$ an ϵ-net with respect to a norm $\|\cdot\|$ for a given set S if for any element $s\in S$, there exists an element $s'\in\mathcal{N}_\epsilon$ such that $\|s-s'\|\leq\epsilon$.

Moreover, we set $\mathcal{V}_{\epsilon,0}^{(l)} = \{\boldsymbol{V}^\star\}$. Then, we denote the best approximation of $\boldsymbol{V}_t^{(l)}$ in $\mathcal{V}_{\epsilon,t}^{(l)}$ as $\boldsymbol{V}_{\epsilon,t}^{(l)} = \arg\min_{\boldsymbol{V}\in\mathcal{V}_{\epsilon,t}^{(l)}} \left\|\boldsymbol{V}_t^{(l)} - \boldsymbol{V}\right\|_{\mathrm{F}}$. Hence, we have the following decomposition

$$
\begin{aligned}
&\left\|(\mathcal{R}_\Omega - \mathcal{R}_{\Omega^{(l)}})\left(\boldsymbol{X}^\star - \boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}_t\boldsymbol{V}_t^{(l)\top}\right)\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}} \\
&\le \underbrace{\left\|(\mathcal{R}_\Omega - \mathcal{R}_{\Omega^{(l)}})\left(\boldsymbol{X}^\star - \boldsymbol{V}_{\epsilon,t}^{(l)}\boldsymbol{\Sigma}_{\epsilon,t}\boldsymbol{V}_{\epsilon,t}^{(l)\top}\right)\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}}_{\text{(I)}} \\
&\quad + \underbrace{\left\|(\mathcal{R}_\Omega - \mathcal{R}_{\Omega^{(l)}})\left(\boldsymbol{X}^\star - \boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}_t\boldsymbol{V}_t^{(l)\top}\right)\left(\boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}\right)\right\|_{\mathrm{F}}}_{\text{(II)}} \\
&\quad + \underbrace{\left\|(\mathcal{R}_\Omega - \mathcal{R}_{\Omega^{(l)}})\left(\boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}_t\left(\boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}\right)^\top\right)\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}}_{\text{(III)}} \\
&\quad + \underbrace{\left\|(\mathcal{R}_\Omega - \mathcal{R}_{\Omega^{(l)}})\left(\boldsymbol{V}_t^{(l)}\left(\boldsymbol{\Sigma}_t - \boldsymbol{\Sigma}_{\epsilon,t}\right)\boldsymbol{V}_{\epsilon,t}^{(l)\top}\right)\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}}_{\text{(IV)}} \\
&\quad + \underbrace{\left\|(\mathcal{R}_\Omega - \mathcal{R}_{\Omega^{(l)}})\left(\left(\boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}\right)\boldsymbol{\Sigma}_{\epsilon,t}\boldsymbol{V}_{\epsilon,t}^{(l)\top}\right)\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}}_{\text{(V)}}.
\end{aligned}
\tag{4.153}
$$

To proceed, we require the following key lemmas.

Lemma 71. *For any $1 \le i \le j \le d$, let $p_{i,j}$ be a Bernoulli random variable taking the value $p_{i,j} = 1$ if and only if $(i,j) \in \Omega$. Let $r_{i,j} = \frac{1}{2p}(p_{i,j} + p_{j,i})$. The following statements hold:*

- **Independent case:** *Suppose the sampling rate $p \gtrsim \frac{\mu r}{d}\log\left(\frac{4r}{\delta}\right)$. Suppose that $\boldsymbol{X} \in \mathcal{S}_{d\times d}$ and $\boldsymbol{V} \in \mathcal{O}_{d\times r}$ with $\|\boldsymbol{V}\|_{2,\infty} \le 2\sqrt{\frac{\mu r}{d}}$ are independent of $r_{l,1},\ldots,r_{l,d}$. Then, with probability at least $1-\delta$, we have*

$$
\left\|(\mathcal{R}_\Omega - \mathcal{R}_{\Omega^{(l)}})(\boldsymbol{X})\boldsymbol{V}\right\|_{\mathrm{F}}^2 \le \frac{32\mu r\log(4r/\delta)}{p}\|\boldsymbol{X}\|_{\max}^2. \tag{4.154}
$$

- **General case:** *Suppose the sampling rate $p \gtrsim \frac{\log(d)}{d}$. For arbitrary $\boldsymbol{X} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{V} \in \mathbb{R}^{d\times r}$, with probability at least $1 - \frac{1}{d^3}$, we have*

$$
\left\|(\mathcal{R}_\Omega - \mathcal{R}_{\Omega^{(l)}})(\boldsymbol{X})\boldsymbol{V}\right\|_{\mathrm{F}}^2 \le \frac{2d}{p}\|\boldsymbol{X}\|_{\max}^2\|\boldsymbol{V}\|_{\mathrm{F}}^2. \tag{4.155}
$$

Lemma 72. *Suppose the sampling rate satisfies $p \gtrsim \frac{\mu r}{d}\log\left(\frac{4r}{\delta}\right)$. For any $0 \le t \le T =$*

$\frac{100}{\eta\sigma_r^\star}\log\left(\frac{1}{\alpha}\right), 1 \le l \le d$, *with probability at least* $1-\frac{1}{d^3}$, *we have*

$$\left\|\left(\mathcal{R}_\Omega - \mathcal{R}_{\Omega^{(l)}}\right)\left(\boldsymbol{X}^\star - \boldsymbol{V}_{\epsilon,t}^{(l)}\boldsymbol{\Sigma}_{\epsilon,t}\boldsymbol{V}_{\epsilon,t}^{(l)\top}\right)\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}^2 \le \Gamma\frac{\kappa\mu^3 r^{5.5}\log\left(\frac{1}{\alpha}\right)\log\left(\frac{r\sigma_1^\star}{\epsilon}\right)}{\sqrt{pd^3}}. \tag{4.156}$$

Lemma 73. *Under the same conditions as Lemma 72, we have*

$$\left\|\boldsymbol{V}_{t+1}^{(l)} - \boldsymbol{V}_{\epsilon,t+1}^{(l)}\right\|_{\mathrm{F}} \le \left(1+\Gamma\eta\sigma_1^\star\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\right)\left\|\boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}} + C_3\eta\frac{\sigma_1^\star\mu r}{\sqrt{pd}}\epsilon. \tag{4.157}$$

The proof of these lemmas is deferred to the end of this section. Now we use Lemma 73 to control the dynamic of $\left\|\boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}$. Note that $\left\|\boldsymbol{V}_0^{(l)} - \boldsymbol{V}_{\epsilon,0}^{(l)}\right\|_{\mathrm{F}} = 0$. Hence, applying Lemma 94, we have

$$\begin{aligned}\left\|\boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}} &\le C\left(\left(1+\Gamma\eta\sigma_1^\star\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\right)^t - 1\right)\frac{\eta\frac{\sigma_1^\star\mu r}{\sqrt{pd}}\epsilon}{\eta\sigma_1^\star\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}} \\ &\le C\eta\frac{\sigma_1^\star\mu r}{\sqrt{pd}}\epsilon\cdot t.\end{aligned} \tag{4.158}$$

Now we are ready to control (I) to (V) separately. First, Lemma 72 directly implies that

$$\text{(I)} \le \Gamma\frac{\sqrt{\kappa\mu^3 r^{5.5}\log\left(\frac{1}{\alpha}\right)\log\left(\frac{r\sigma_1^\star}{\epsilon}\right)}}{\sqrt[4]{pd^3}}. \tag{4.159}$$

On the other hand, applying Lemma 71 to (II) leads to

$$\text{(II)} \le \sqrt{\frac{2d}{p}}\left\|\boldsymbol{X}^\star - \boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}_t\boldsymbol{V}_t^{(l)\top}\right\|_{\max}\left\|\boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}} \le \Gamma\sqrt{\frac{\mu^2 r^2}{pd}}\left\|\boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}. \tag{4.160}$$

Similarly, we have

$$\begin{aligned}\text{(III)} &\le \sqrt{\frac{2d}{p}}\left\|\boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}_t\left(\boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}\right)^\top\right\|_{\max}\left\|\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}} \le \Gamma\sqrt{\frac{\mu r^2}{p}}\left\|\boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}, \\ \text{(IV)} &\le \sqrt{\frac{2d}{p}}\left\|\boldsymbol{V}_t^{(l)}\left(\boldsymbol{\Sigma}_t - \boldsymbol{\Sigma}_{\epsilon,t}\right)\boldsymbol{V}_{\epsilon,t}^{(l)\top}\right\|_{\max}\left\|\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}} \le \Gamma\sqrt{\frac{\mu^2 r^3}{pd}}\epsilon, \\ \text{(V)} &\le \sqrt{\frac{2d}{p}}\left\|\left(\boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}\right)\boldsymbol{\Sigma}_{\epsilon,t}\boldsymbol{V}_{\epsilon,t}^{(l)\top}\right\|_{\max}\left\|\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}} \le \Gamma\sqrt{\frac{\mu r^2}{p}}\left\|\boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}.\end{aligned} \tag{4.161}$$

Overall, we derive that

$$\begin{aligned}
&\left\|(\mathcal{R}_{\Omega}-\mathcal{R}_{\Omega^{(l)}})\left(\boldsymbol{X}^{\star}-\boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}_t\boldsymbol{V}_t^{(l)\top}\right)\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}\\
&\leq \Gamma\frac{\sqrt{\kappa\mu^3 r^{5.5}\log\left(\frac{1}{\alpha}\right)\log\left(\frac{r\sigma_1^{\star}}{\epsilon}\right)}}{\sqrt[4]{pd^3}}+\Gamma\sqrt{\frac{\mu r^2}{p}}\left\|\boldsymbol{V}_t^{(l)}-\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}\\
&\leq \Gamma\frac{\sqrt{\kappa\mu^3 r^{5.5}\log\left(\frac{1}{\alpha}\right)\log\left(\frac{r\sigma_1^{\star}}{\epsilon}\right)}}{\sqrt[4]{pd^3}}+\Gamma\sqrt{\frac{\mu r^2}{p}}\cdot\eta\frac{\sigma_1^{\star}\mu r}{\sqrt{pd}}\epsilon\cdot t\\
&\leq 2\Gamma\sigma_1^{\star}\frac{\sqrt{\kappa\mu^3 r^{5.5}\log\left(\frac{1}{\alpha}\right)\log\left(d\right)}}{\sqrt[4]{pd^3}}.
\end{aligned}\tag{4.162}$$

The last inequality follows from the fact that $\epsilon=\frac{c}{d}$. This completes the proof of Proposition 18. □

Next, we proceed to present the proofs of Lemma 71, Lemma 72, and Lemma 73.

Lemma 71. We first expand $\|(\mathcal{R}_{\Omega}-\mathcal{R}_{\Omega^{(l)}})(\boldsymbol{X})\boldsymbol{V}\|_{\mathrm{F}}^2$ as follows

$$\|(\mathcal{R}_{\Omega}-\mathcal{R}_{\Omega^{(l)}})(\boldsymbol{X})\boldsymbol{V}\|_{\mathrm{F}}^2=\sum_{j=1}^{r}\left(\sum_{k=1}^{d}(r_{l,k}-1)X_{l,k}V_{k,j}\right)^2+\sum_{i\neq l}(r_{i,l}-1)^2X_{i,l}^2\|\boldsymbol{V}_{l,\cdot}\|^2.\tag{4.163}$$

Next, we prove these two cases separately.

Independent case. First, we control $\sum_{k=1}^{d}(r_{l,k}-1)X_{l,k}V_{k,j}$ for all $1\leq j\leq r$ via Bernstein's inequality (Lemma 79). To this goal, for a fixed $1\leq j\leq r$, upon defining $Z_k=(r_{l,k}-1)X_{l,k}V_{k,j}$ with $\mathbb{E}[Z_k]=0$, we have

$$\begin{aligned}
M&:=\max_k|Z_k|\leq\frac{1}{p}\|\boldsymbol{X}\|_{\max}\|\boldsymbol{V}\|_{2,\infty},\\
\nu^2&:=\sum_{k=1}^{d}\mathrm{Var}\left[Z_k^2\right]\leq\frac{1}{p}\sum_{k=1}^{d}X_{l,k}^2V_{k,j}^2\leq\frac{1}{p}\|\boldsymbol{X}\|_{\max}^2\sum_{k=1}^{d}V_{k,j}^2=\frac{1}{p}\|\boldsymbol{X}\|_{\max}^2.
\end{aligned}\tag{4.164}$$

Here in the last equality we use the fact that $\sum_{k=1}^{d} V_{k,j}^2 = 1$ since $\boldsymbol{V} \in \mathcal{O}_{d\times r}$. Therefore, due to Bernstein's inequality, with probability at least $1 - \frac{\delta}{2r}$, one has

$$\begin{aligned}\left|\sum_{k=1}^{d} Z_k\right| &\leq 2\nu\sqrt{\log\left(\frac{4r}{\delta}\right)} + \frac{4}{3}M\log\left(\frac{4r}{\delta}\right) \\ &\leq \frac{2\left\|\boldsymbol{X}\right\|_{\max}}{\sqrt{p}}\sqrt{\log\left(\frac{4r}{\delta}\right)} + \frac{4\left\|\boldsymbol{X}\right\|_{\max}\left\|\boldsymbol{V}\right\|_{2,\infty}}{3p}\log\left(\frac{4r}{\delta}\right) \\ &\leq \frac{4\left\|\boldsymbol{X}\right\|_{\max}}{\sqrt{p}}\sqrt{\log\left(\frac{4r}{\delta}\right)},\end{aligned} \tag{4.165}$$

where we use the assumption $p \gtrsim \frac{\mu r}{d}\log\left(\frac{4r}{\delta}\right)$. Hence, via a union bound, we know that with probability at least $1 - \frac{\delta}{2}$, we have

$$\sum_{j=1}^{r}\left(\sum_{k=1}^{d}\left(r_{l,k}-1\right)X_{l,k}V_{k,j}\right)^2 \leq \frac{16r\left\|\boldsymbol{X}\right\|_{\max}^2}{p}\log\left(\frac{4r}{\delta}\right). \tag{4.166}$$

Next, we can control $\sum_{i\neq l}\left(r_{i,l}-1\right)^2 X_{i,l}^2\left\|\boldsymbol{V}_{l,\cdot}\right\|^2$ as

$$\sum_{i\neq l}\left(r_{i,l}-1\right)^2 X_{i,l}^2\left\|\boldsymbol{V}_{l,\cdot}\right\|^2 \leq \sum_{i=1}^{d}\left(r_{i,l}-1\right)^2\left\|\boldsymbol{X}\right\|_{\max}^2\left\|\boldsymbol{V}\right\|_{2,\infty}^2. \tag{4.167}$$

Then, we apply Bernstein's inequality to control $\sum_{k=1}^{d}\left(r_{l,k}-1\right)^2$. To this end, notice that

$$\begin{aligned}&\mathbb{E}\left[\left(r_{l,k}-1\right)^2\right] = \frac{1-p}{2p} \leq \frac{1}{2p}, \\ &M := \max_k\left(r_{l,k}-1\right)^2 \leq \left(\frac{1}{p}-1\right)^2 \leq \frac{1}{p^2}, \\ &\nu^2 := \sum_{k=1}^{d}\mathrm{Var}\left[\left(r_{l,k}-1\right)^2\right] \leq \frac{d}{p^3}.\end{aligned} \tag{4.168}$$

Therefore, with probability at least $1 - \frac{\delta}{2}$, we have

$$\sum_{k=1}^{d}\left(r_{l,k}-1\right)^2 \leq \frac{d}{2p} + \frac{4}{3}\frac{1}{p^2}\log\left(\frac{2}{\delta}\right) + 2\sqrt{\frac{d}{p^3}}\sqrt{\log\left(\frac{2}{\delta}\right)} \leq \frac{d}{p} \tag{4.169}$$

since we set $p \gtrsim \frac{\mu r}{d} \log\left(\frac{4r}{\delta}\right)$. This implies that with probability at least $1 - \frac{\delta}{2}$, we have

$$\sum_{i \neq l} \left(r_{i,l} - 1\right)^2 X_{i,l}^2 \left\|\boldsymbol{V}_{l,\cdot}\right\|^2 \leq \frac{d}{p} \left\|\boldsymbol{X}\right\|_{\max}^2 \left\|\boldsymbol{V}\right\|_{2,\infty}^2. \tag{4.170}$$

Finally, taking a union bound, we conclude that with probability at least $1 - \delta$, we have

$$\begin{aligned} \left\|\left(\mathcal{R}_{\Omega} - \mathcal{R}_{\Omega^{(l)}}\right)(\boldsymbol{X})\, \boldsymbol{V}\right\|_{\mathrm{F}}^2 &\leq \frac{16r \left\|\boldsymbol{X}\right\|_{\max}^2 \log(4r/\delta)}{p} + \frac{d}{p} \left\|\boldsymbol{X}\right\|_{\max}^2 \left\|\boldsymbol{V}\right\|_{2,\infty}^2 \\ &\leq \frac{32\mu r \log(4r/\delta)}{p} \left\|\boldsymbol{X}\right\|_{\max}^2. \end{aligned} \tag{4.171}$$

General case. First, we apply Cauchy-Schwarz inequality to obtain

$$\begin{aligned} \sum_{j=1}^{r} \left(\sum_{k=1}^{d} \left(r_{l,k} - 1\right) X_{l,k} V_{k,j}\right)^2 &\leq \sum_{j=1}^{r} \left(\sum_{k=1}^{d} \left(r_{l,k} - 1\right)^2\right) \left(\sum_{k=1}^{d} X_{l,k}^2 V_{k,j}^2\right) \\ &\leq \left(\sum_{k=1}^{d} \left(r_{l,k} - 1\right)^2\right) \cdot \sum_{j=1}^{r} \left\|\boldsymbol{X}\right\|_{\max}^2 \sum_{k=1}^{d} V_{k,j}^2 \\ &= \left(\sum_{k=1}^{d} \left(r_{l,k} - 1\right)^2\right) \cdot \left\|\boldsymbol{X}\right\|_{\max}^2 \left\|\boldsymbol{V}\right\|_{\mathrm{F}}^2. \end{aligned} \tag{4.172}$$

Next, we can control $\sum_{i \neq l} \left(r_{i,l} - 1\right)^2 X_{i,l}^2 \left\|\boldsymbol{V}_{l,\cdot}\right\|^2$ as follows

$$\begin{aligned} \sum_{i \neq l} \left(r_{i,l} - 1\right)^2 X_{i,l}^2 \left\|\boldsymbol{V}_{l,\cdot}\right\|^2 &\leq \sum_{i=1}^{d} \left(r_{i,l} - 1\right)^2 \left\|\boldsymbol{X}\right\|_{\max}^2 \left\|\boldsymbol{V}\right\|_{\mathrm{F}}^2 \\ &= \left(\sum_{k=1}^{d} \left(r_{l,k} - 1\right)^2\right) \left\|\boldsymbol{X}\right\|_{\max}^2 \left\|\boldsymbol{V}\right\|_{\mathrm{F}}^2. \end{aligned} \tag{4.173}$$

Here we use the fact that $r_{i,j} = r_{j,i}, \forall i, j \in [d]$. Next, we apply Bernstein's inequality to control $\sum_{k=1}^{d} \left(r_{l,k} - 1\right)^2$. To this end, notice that

$$\begin{aligned} &\mathbb{E}\left[\left(r_{l,k} - 1\right)^2\right] = \frac{1-p}{2p} \leq \frac{1}{2p}, \\ &M := \max_{k} \left(r_{l,k} - 1\right)^2 \leq \left(\frac{1}{p} - 1\right)^2 \leq \frac{1}{p^2}, \\ &\nu^2 := \sum_{k=1}^{d} \mathrm{Var}\left[\left(r_{l,k} - 1\right)^2\right] \leq \frac{d}{p^3}. \end{aligned} \tag{4.174}$$

Therefore, with probability at least $1-\delta$, we have

$$\sum_{k=1}^{d}\left(r_{l,k}-1\right)^2 \leq \frac{d}{2p}+\frac{4}{3}\frac{1}{p^2}\log\left(\frac{2}{\delta}\right)+2\sqrt{\frac{d}{p^3}}\sqrt{\log\left(\frac{2}{\delta}\right)}. \tag{4.175}$$

Specifically, upon setting $\delta=\frac{1}{d^3}$, we obtain that with probability at least $1-\frac{1}{d^3}$, one has

$$\sum_{k=1}^{d}\left(r_{l,k}-1\right)^2 \leq \frac{d}{p}, \tag{4.176}$$

since we set $p \gtrsim \frac{\log(d)}{d}$. Overall, we have

$$\left\|\left(\mathcal{R}_{\Omega}-\mathcal{R}_{\Omega^{(l)}}\right)(\boldsymbol{X})\,\boldsymbol{V}\right\|_{\mathrm{F}}^2 \leq 2\left(\sum_{k=1}^{d}\left(r_{l,k}-1\right)^2\right)\|\boldsymbol{X}\|_{\max}^2\|\boldsymbol{V}\|_{\mathrm{F}}^2 \leq \frac{2d}{p}\|\boldsymbol{X}\|_{\max}^2\|\boldsymbol{V}\|_{\mathrm{F}}^2, \tag{4.177}$$

with probability at least $1-\frac{1}{d^3}$. This completes the proof. □

Lemma 72. First, for fixed $\boldsymbol{\Sigma}_{\epsilon,t}\in\mathcal{N}_{\epsilon}$ and $\boldsymbol{V}_{\epsilon,t}^{(l)}\in\mathcal{V}_{\epsilon,t}^{(l)}$, Lemma 71 implies that with probability at least $1-\delta$, we have

$$\left\|\left(\mathcal{R}_{\Omega}-\mathcal{R}_{\Omega^{(l)}}\right)\left(\boldsymbol{X}^{\star}-\boldsymbol{V}_{\epsilon,t}^{(l)}\boldsymbol{\Sigma}_{\epsilon,t}\boldsymbol{V}_{\epsilon,t}^{(l)\top}\right)\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}^2 \leq \frac{32\mu r\log(4r/\delta)}{p}\left\|\boldsymbol{X}^{\star}-\boldsymbol{V}_{\epsilon,t}^{(l)}\boldsymbol{\Sigma}_{\epsilon,t}\boldsymbol{V}_{\epsilon,t}^{(l)\top}\right\|_{\max}^2. \tag{4.178}$$

Note that

$$\begin{aligned}\left\|\boldsymbol{X}^{\star}-\boldsymbol{V}_{\epsilon,t}^{(l)}\boldsymbol{\Sigma}_{\epsilon,t}\boldsymbol{V}_{\epsilon,t}^{(l)\top}\right\|_{\max} &\leq \|\boldsymbol{X}^{\star}\|_{\max}+\left\|\boldsymbol{V}_{\epsilon,t}^{(l)}\boldsymbol{\Sigma}_{\epsilon,t}\boldsymbol{V}_{\epsilon,t}^{(l)\top}\right\|_{\max}\\ &\overset{(a)}{\leq} \|\boldsymbol{\Sigma}^{\star}\|\,\|\boldsymbol{V}^{\star}\|_{2,\infty}^2+\|\boldsymbol{\Sigma}_{\epsilon,t}\|\left\|\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{2,\infty}^2\\ &\leq 9\sigma_1^{\star}\frac{\mu r}{d}.\end{aligned} \tag{4.179}$$

Therefore, with probability at least $1-\delta$, we have

$$\left\|\left(\mathcal{R}_{\Omega}-\mathcal{R}_{\Omega^{(l)}}\right)\left(\boldsymbol{X}^{\star}-\boldsymbol{V}_{\epsilon,t}^{(l)}\boldsymbol{\Sigma}_{\epsilon,t}\boldsymbol{V}_{\epsilon,t}^{(l)\top}\right)\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}^2 \leq \frac{288\mu^3r^3\log(4r/\delta)}{pd^2}. \tag{4.180}$$

Lastly, we apply the union bound to finalize the desired result. To this end, note that $\mathcal{B}_{\mathrm{op}}^{r\times r}(4\sigma_1^{\star})\subset\mathcal{B}_{\mathrm{F}}^{r\times r}(4\sqrt{r}\sigma_1^{\star})$. Hence, according to Lemma 89, we know that $|\mathcal{N}_{\epsilon}|\leq\left(\frac{12\sqrt{r}\sigma_1^{\star}}{\epsilon}\right)^{r^2}$ which implies that $\left|\mathcal{V}_{\epsilon,t}^{(l)}\right|\leq\left(\frac{6\sqrt{r}\sigma_1^{\star}}{\epsilon}\right)^{r^2t}$. Therefore, the total cardinality of

$\bigcup_{l=1}^{d}\bigcup_{t=0}^{T}\mathcal{V}_{\epsilon,t}^{(l)}$ is upper-bounded by

$$\left|\bigcup_{l=1}^{d}\bigcup_{t=0}^{T}\mathcal{V}_{\epsilon,t}^{(l)}\right| \leq d\cdot\sum_{t=0}^{T}\left(\frac{12\sqrt{r}\sigma_1^\star}{\epsilon}\right)^{r^2t} \leq 2d\left(\frac{12\sqrt{r}\sigma_1^\star}{\epsilon}\right)^{r^2T}. \tag{4.181}$$

Hence, once we set $\delta=\frac{1}{2d^4}\left(\frac{6\sqrt{r}\sigma_1^\star}{\epsilon}\right)^{-r^2T}$, we obtain that with probability at least $1-d^{-3}$, for any $0\leq t\leq T, 1\leq l\leq d$ and $\boldsymbol{V}_{\epsilon,t}^{(l)}\in\mathcal{V}_{\epsilon,t}^{(l)}$,

$$\begin{aligned}\left\|\left(\mathcal{R}_\Omega-\mathcal{R}_{\Omega^{(l)}}\right)\left(\boldsymbol{X}^\star-\boldsymbol{V}_{\epsilon,t}^{(l)}\boldsymbol{\Sigma}_{\epsilon,t}\boldsymbol{V}_{\epsilon,t}^{(l)\top}\right)\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}^2 &\leq \Gamma\frac{\mu^3r^3}{pd^2}\left(\log(d)+r^2T\left(\frac{r\sigma_1^\star}{\epsilon}\right)\right)\\ &\leq \Gamma\frac{\kappa\mu^3r^{5.5}\log\left(\frac{1}{\alpha}\right)\log\left(\frac{r\sigma_1^\star}{\epsilon}\right)}{\sqrt{pd^3}}.\end{aligned} \tag{4.182}$$

Here we use the fact that $T\leq\frac{100}{\eta\sigma_r^\star}\log(1/\alpha)$ and $\eta=\Theta\left(\frac{\mu}{\sigma_1^\star}\sqrt{\frac{r}{pd}}\right)$. This completes the proof. □

Lemma 73. The derivation of $\left\|\boldsymbol{V}_{t+1}^{(l)}-\boldsymbol{V}_{\epsilon,t+1}^{(l)}\right\|_{\mathrm{F}}$ is nearly the same as that of $\left\|\boldsymbol{V}_{t+1}-\boldsymbol{V}_{t+1}^{(l)}\right\|_{\mathrm{F}}$. First, note that

$$\boldsymbol{V}_{\epsilon,t+1}^{(l)}=\left(\boldsymbol{I}+\eta\mathcal{P}_{\boldsymbol{V}_{\epsilon,t}^{(l)}}^{\perp}\boldsymbol{M}_{\epsilon,t}^{(l)}\right)\boldsymbol{V}_{\epsilon,t}^{(l)}+\boldsymbol{A}_{\epsilon,t}^{(l)}, \tag{4.183}$$

where $\boldsymbol{M}_{\epsilon,t}^{(l)}$ and $\boldsymbol{A}_{\epsilon,t}^{(l)}$ are defined similar to $\boldsymbol{M}_t^{(l)}$ and $\boldsymbol{A}_t^{(l)}$. Hence, we can expand $\left\|\boldsymbol{V}_{t+1}^{(l)}-\boldsymbol{V}_{\epsilon,t+1}^{(l)}\right\|_{\mathrm{F}}^2$ as

$$\left\|\boldsymbol{V}_{t+1}^{(l)}-\boldsymbol{V}_{\epsilon,t+1}^{(l)}\right\|_{\mathrm{F}}^2=\left\|\boldsymbol{V}_t^{(l)}-\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}^2+2\eta\underbrace{\left\langle\boldsymbol{V}_t^{(l)}-\boldsymbol{V}_{\epsilon,t}^{(l)},\mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\boldsymbol{M}_t^{(l)}\boldsymbol{V}_t^{(l)}-\mathcal{P}_{\boldsymbol{V}_{\epsilon,t}^{(l)}}^{\perp}\boldsymbol{M}_{\epsilon,t}^{(l)}\boldsymbol{V}_{\epsilon,t}^{(l)}\right\rangle}_{:=(\mathrm{I})}+(\mathrm{II}). \tag{4.184}$$

where

$$\begin{aligned}(\mathrm{II})=\eta^2\left\|\mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\boldsymbol{M}_t^{(l)}\boldsymbol{V}_t^{(l)}-\mathcal{P}_{\boldsymbol{V}_{\epsilon,t}^{(l)}}^{\perp}\boldsymbol{M}_{\epsilon,t}^{(l)}\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}^2+\left\|\boldsymbol{A}_t^{(l)}-\boldsymbol{A}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}^2\\ +2\left\langle\left(\boldsymbol{I}+\eta\mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\boldsymbol{M}_t^{(l)}\right)\boldsymbol{V}_t^{(l)}-\left(\boldsymbol{I}+\eta\mathcal{P}_{\boldsymbol{V}_{\epsilon,t}^{(l)}}^{\perp}\boldsymbol{M}_{\epsilon,t}^{(l)}\right)\boldsymbol{V}_{\epsilon,t}^{(l)},\boldsymbol{A}_t^{(l)}-\boldsymbol{A}_{\epsilon,t}^{(l)}\right\rangle\end{aligned} \tag{4.185}$$

contains all the higher-order terms. Next, we further decompose (I) as follows

$$\begin{aligned}
\text{(I)} &= \left\langle \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}, \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp} \boldsymbol{\Xi}_t^{(l)} \boldsymbol{V}_t^{(l)} - \mathcal{P}_{\boldsymbol{V}_{\epsilon,t}^{(l)}}^{\perp} \boldsymbol{\Xi}_{\epsilon,t}^{(l)} \boldsymbol{V}_{\epsilon,t}^{(l)} \right\rangle \\
&= \underbrace{\left\langle \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}, \left(\mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp} - \mathcal{P}_{\boldsymbol{V}_{\epsilon,t}^{(l)}}^{\perp} \right) \boldsymbol{\Xi}_{\epsilon,t}^{(l)} \boldsymbol{V}_{\epsilon,t}^{(l)} \right\rangle}_{:=(\mathrm{I}_1)} + \underbrace{\left\langle \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}, \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp} \boldsymbol{\Xi}_t^{(l)} \left(\boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right) \right\rangle}_{:=(\mathrm{I}_2)} \\
&\quad + \underbrace{\left\langle \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}, \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp} \left(\boldsymbol{\Xi}_t^{(l)} - \boldsymbol{\Xi}_{\epsilon,t}^{(l)} \right) \boldsymbol{V}_{\epsilon,t}^{(l)} \right\rangle}_{:=(\mathrm{I}_3)}.
\end{aligned} \tag{4.186}$$

Here we define $\boldsymbol{\Xi}_t^{(l)} = \boldsymbol{M}_t^{(l)} - \boldsymbol{V}_t^{(l)} \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top} + \boldsymbol{V}_t^{(l)} \boldsymbol{\Sigma}_t \boldsymbol{V}_t^{(l)\top}$ and $\boldsymbol{\Xi}_{\epsilon,t}^{(l)} = \boldsymbol{M}_{\epsilon,t}^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top} + \boldsymbol{V}_{\epsilon,t}^{(l)} \boldsymbol{\Sigma}_{\epsilon,t} \boldsymbol{V}_{\epsilon,t}^{(l)\top}$. We control (I_1) as follows

$$(\mathrm{I}_1) \le \left\| \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\|_{\mathrm{F}} \left\| \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp} - \mathcal{P}_{\boldsymbol{V}_{\epsilon,t}^{(l)}}^{\perp} \right\|_{\mathrm{F}} \left\| \boldsymbol{\Xi}_{\epsilon,t}^{(l)} \right\| \le 2 \left\| \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\|_{\mathrm{F}}^2 \left\| \boldsymbol{\Xi}_{\epsilon,t}^{(l)} \right\|. \tag{4.187}$$

Here we apply Lemma 90 in the second inequality. For $\left\| \boldsymbol{\Xi}_{\epsilon,t}^{(l)} \right\|$, following the same analysis as in Equation (4.119), we have

$$\begin{aligned}
\left\| \boldsymbol{\Xi}_{\epsilon,t}^{(l)} \right\| &\le \sigma_1^\star \left\| \boldsymbol{V}^\star - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\| + 10\Gamma \frac{\sigma_1^\star \mu r}{\sqrt{pd}} \\
&\le \sigma_1^\star \left(\left\| \boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)} \right\| + \left\| \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\| \right) + 10\Gamma \frac{\sigma_1^\star \mu r}{\sqrt{pd}} \\
&\le \sigma_1^\star \left\| \boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)} \right\| + 20\Gamma \frac{\sigma_1^\star \mu r}{\sqrt{pd}}.
\end{aligned} \tag{4.188}$$

Here in the last inequality, we use the fact that $\left\| \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\| \le \epsilon \le 10\Gamma \frac{\mu r}{\sqrt{pd}}$. Plugging this inequality into Equation (4.187), we obtain

$$(\mathrm{I}_1) \le 2\sigma_1^\star \left\| \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\|_{\mathrm{F}}^2 \left(\sigma_1^\star \left\| \boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)} \right\| + 20\Gamma \frac{\sigma_1^\star \mu r}{\sqrt{pd}} \right). \tag{4.189}$$

Similar to our derivation for (I_1), it follows that $(\mathrm{I}_2) \le \sigma_1^\star \left\| \boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)} \right\|_{\mathrm{F}}^2 \left(\sigma_1^\star \left\| \boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)} \right\| + \right.$

$10\Gamma\frac{\sigma_1^\star\mu r}{\sqrt{pd}}\Big)$. Lastly, to bound (I_3), we further decompose it as

$$
\begin{aligned}
(\mathrm{I}_3) &= \left\langle \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}, \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\left(\boldsymbol{M}_t^{(l)} - \boldsymbol{M}_{\epsilon,t}^{(l)} + \boldsymbol{V}_{\epsilon,t}^{(l)}\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top} - \boldsymbol{V}_{\epsilon,t}^{(l)}\boldsymbol{\Sigma}_t^{(l)}\boldsymbol{V}_{\epsilon,t}^{(l)\top}\right)\boldsymbol{V}_{\epsilon,t}^{(l)}\right\rangle \\
&= \underbrace{\left\langle \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}, \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\left(\mathrm{I} - \mathcal{R}_{\Omega^{(l)}}\right)\left(\boldsymbol{V}_{\epsilon,t}^{(l)}\boldsymbol{\Sigma}_{\epsilon,t}^{(l)}\left(\boldsymbol{V}_{\epsilon,t}^{(l)} - \boldsymbol{V}_t^{(l)}\right)^\top\right)\boldsymbol{V}_{\epsilon,t}^{(l)}\right\rangle}_{:=(\mathrm{I}_{3,1})} \\
&\quad + \underbrace{\left\langle \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}, \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\left(\mathrm{I} - \mathcal{R}_{\Omega^{(l)}}\right)\left(\left(\boldsymbol{V}_{\epsilon,t}^{(l)} - \boldsymbol{V}_t^{(l)}\right)\boldsymbol{\Sigma}_{\epsilon,t}^{(l)}\boldsymbol{V}_t^{(l)\top}\right)\boldsymbol{V}_{\epsilon,t}^{(l)}\right\rangle}_{:=(\mathrm{I}_{3,2})} \\
&\quad + \underbrace{\left\langle \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}, \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\left(\mathrm{I} - \mathcal{R}_{\Omega^{(l)}}\right)\left(\boldsymbol{V}_t^{(l)}\left(\boldsymbol{\Sigma}_{\epsilon,t} - \boldsymbol{\Sigma}_t\right)\boldsymbol{V}_t^{(l)\top}\right)\boldsymbol{V}_{\epsilon,t}^{(l)}\right\rangle}_{:=(\mathrm{I}_{3,3})} \\
&\quad + \underbrace{\left\langle \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}, \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\boldsymbol{V}_{\epsilon,t}^{(l)}\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}\boldsymbol{V}_{\epsilon,t}^{(l)}\right\rangle}_{:=(\mathrm{I}_{3,4})}.
\end{aligned}
\tag{4.190}
$$

Next, we control these terms separately. First, we apply Corollary 6 to control $(\mathrm{I}_{3,1})$. Specifically, upon setting $\boldsymbol{A} = \boldsymbol{V}_{\epsilon,t}^{(l)}, \boldsymbol{B} = \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\left(\boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}\right), \boldsymbol{C} = \left(\boldsymbol{V}_{\epsilon,t}^{(l)} - \boldsymbol{V}_t^{(l)}\right)\boldsymbol{\Sigma}_{\epsilon,t}, \boldsymbol{D} = \boldsymbol{V}_{\epsilon,t}^{(l)}$ in Corollary 6, with probability at least $1 - d^{-3}$, we have

$$
\begin{aligned}
(\mathrm{I}_{3,1}) &= \left\langle \left(\mathrm{I} - \mathcal{R}_\Omega\right)\left(\boldsymbol{A}\boldsymbol{C}^\top\right), \boldsymbol{B}\boldsymbol{D}^\top\right\rangle \\
&\le \Gamma\sqrt{\frac{d}{p}}\left\|\boldsymbol{A}\right\|_{2,\infty}\left\|\boldsymbol{B}\right\|_{\mathrm{F}} \cdot \left\|\boldsymbol{C}\right\|_{\mathrm{F}}\left\|\boldsymbol{D}\right\|_{2,\infty} \\
&= \Gamma\sqrt{\frac{d}{p}}\left\|\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{2,\infty}^2\left\|\mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\left(\boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}\right)\right\|_{\mathrm{F}}\left\|\left(\boldsymbol{V}_{\epsilon,t}^{(l)} - \boldsymbol{V}_t^{(l)}\right)\boldsymbol{\Sigma}_{\epsilon,t}\right\|_{\mathrm{F}} \\
&\le 16\Gamma\sigma_1^\star\sqrt{\frac{\mu^2 r^2}{pd}}\left\|\boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}^2.
\end{aligned}
\tag{4.191}
$$

Here in the last inequality, we use the fact that $\|\boldsymbol{\Sigma}_{\epsilon,t}\| \le 4\sigma_1^\star$ and $\left\|\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{2,\infty} \le 2\sqrt{\frac{\mu r}{d}}$. In a manner akin to our derivation for $(\mathrm{I}_{3,1})$, we can also show that

$$
\begin{aligned}
(\mathrm{I}_{3,2}) &\le 16\Gamma\sigma_1^\star\sqrt{\frac{\mu^2 r^2}{pd}}\left\|\boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}^2 \\
(\mathrm{I}_{3,3}) &\le 4\Gamma\sigma_1^\star\sqrt{\frac{\mu^2 r^2}{pd}}\epsilon\left\|\boldsymbol{V}_{\epsilon,t}^{(l)} - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}.
\end{aligned}
\tag{4.192}
$$

Lastly, for $(\mathrm{I}_{3,4})$, we observe that

$$\begin{aligned}
(\mathrm{I}_{3,4}) &= -\left\langle \boldsymbol{V}_{\epsilon,t}^{(l)}, \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp} \boldsymbol{V}_{\epsilon,t}^{(l)} \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top} \boldsymbol{V}_{\epsilon,t}^{(l)} \right\rangle \\
&\leq -\left\langle \boldsymbol{V}_{\epsilon,t}^{(l)}, \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp} \boldsymbol{V}_{\epsilon,t}^{(l)} \boldsymbol{\Sigma}^\star \right\rangle + \sigma_1^\star \left\| \boldsymbol{V}^\star - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\|_{\mathrm{F}} \left\| \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\|_{\mathrm{F}}^2 \\
&\overset{(a)}{\leq} \sigma_1^\star \left\| \boldsymbol{V}^\star - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\|_{\mathrm{F}} \left\| \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\|_{\mathrm{F}}^2 \\
&\leq \sigma_1^\star \left(\left\| \boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)} \right\|_{\mathrm{F}} + \left\| \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\|_{\mathrm{F}} \right) \left\| \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\|_{\mathrm{F}}^2 \\
&\overset{(b)}{\leq} \sigma_1^\star \left(\Gamma_1 \frac{\kappa \mu r^{1.5} \log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}} + \left\| \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\|_{\mathrm{F}} \right) \left\| \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\|_{\mathrm{F}}^2 .
\end{aligned} \tag{4.193}$$

Here in (a), we use the fact that $\left\langle \boldsymbol{V}_{\epsilon,t}^{(l)}, \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp} \boldsymbol{V}_{\epsilon,t}^{(l)} \boldsymbol{\Sigma}^\star \right\rangle = \left\| \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp} \boldsymbol{V}_{\epsilon,t}^{(l)} \boldsymbol{\Sigma}^{\star 1/2} \right\|_{\mathrm{F}}^2 \geq 0$. In (b), we apply Proposition 15.

Putting everything together, we obtain that

$$(\mathrm{I}_3) \leq \left\| \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\|_{\mathrm{F}} \left(32\Gamma\sigma_1^\star \sqrt{\frac{\mu^2 r^2}{pd}} \left\| \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\|_{\mathrm{F}} + 4\Gamma\sigma_1^\star \sqrt{\frac{\mu^2 r^2}{pd}} \epsilon \right), \tag{4.194}$$

which in turn leads to

$$(\mathrm{I}) \leq \Gamma\sigma_1^\star \frac{\kappa \mu r^{1.5} \log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}} \left\| \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\|_{\mathrm{F}}^2 + \left\| \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\|_{\mathrm{F}} \cdot 4\Gamma\sigma_1^\star \sqrt{\frac{\mu^2 r^2}{pd}} \epsilon. \tag{4.195}$$

Therefore, our final bound is established as

$$\left\| \boldsymbol{V}_{t+1}^{(l)} - \boldsymbol{V}_{\epsilon,t+1}^{(l)} \right\|_{\mathrm{F}} \leq \left(1 + \Gamma\eta\sigma_1^\star \frac{\kappa \mu r^{1.5} \log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}} \right) \left\| \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\|_{\mathrm{F}} + C_3 \eta \frac{\sigma_1^\star \mu r}{\sqrt{pd}} \epsilon. \tag{4.196}$$

This completes the proof. □

4.E Proofs for Different Initialization Schemes

Gaussian initialization. Suppose the search rank $0.5d \leq r' \leq d$. Let $\boldsymbol{Z} = \boldsymbol{G} / \|\boldsymbol{G}\|$ where $\boldsymbol{G} \in \mathbb{R}^{d \times r'}$ is a standard Gaussian matrix. Before proceeding, we first introduce the following lemma, which characterizes the concentration of the largest and smallest singular values of a standard Gaussian matrix.

Lemma 74 (Adapted from Theorem 6.1 in [Wainwright, 2019]). *Suppose that* $\boldsymbol{G} \in \mathbb{R}^{d_1 \times d_2}$

is a standard Gaussian matrix where $d_1 \geq d_2$. Then, for any $\delta > 0$, we have

$$\begin{aligned} \mathbb{P}\left(\|\boldsymbol{G}\| \geq (2+\delta)\sqrt{d_1}\right) &\leq \exp\left\{-\frac{d_1\delta^2}{2}\right\}, \\ \mathbb{P}\left(\sigma_{\min}(\boldsymbol{G}) \geq (1-\delta)\sqrt{d_1} - \sqrt{d_2}\right) &\leq \exp\left\{-\frac{d_1\delta^2}{2}\right\}. \end{aligned} \tag{4.197}$$

Notice that $\sigma_r(\mathcal{P}_{\boldsymbol{V}^\star}\boldsymbol{G}) = \sigma_{\min}(\boldsymbol{V}^{\star\top}\boldsymbol{G})$ where $\boldsymbol{V}^{\star\top}\boldsymbol{G} \in \mathbb{R}^{r\times r'}$ is another standard Gaussian matrix. Hence, upon setting $\delta = \frac{1}{2}$ and noting that $r' \geq \frac{d}{2}$, Lemma 74 implies that with probability at least $1 - \exp\{d/16\}$, we have

$$\sigma_r(\mathcal{P}_{\boldsymbol{V}^\star}\boldsymbol{G}) \geq \frac{1}{2}\sqrt{r'} - \sqrt{r} \geq \frac{1}{4}\sqrt{d}. \tag{4.198}$$

On the other hand, upon setting $\delta = \frac{1}{2}$, Lemma 74 implies that with probability at least $1 - \exp\{d/8\}$, we have

$$\boldsymbol{G} \leq \frac{5}{2}\sqrt{d}. \tag{4.199}$$

Via a union bound, we know that with probability at least $1 - 2\exp\{d/16\}$, we have

$$\sigma_r(\mathcal{P}_{\boldsymbol{V}^\star}\boldsymbol{Z}) = \frac{\sigma_r(\mathcal{P}_{\boldsymbol{V}^\star}\boldsymbol{G})}{\|\boldsymbol{G}\|} \geq \frac{\sqrt{d}/4}{5\sqrt{d}/2} = \frac{1}{10}. \tag{4.200}$$

Hence, with probability at least $1 - 2\exp\{d/16\}$, Condition 1 is satisfied with $c_0 = 0.1$.

Orthogonal initialization. Suppose the search rank satisfies $r' = d$. Upon choosing $\boldsymbol{Z} = \boldsymbol{O}$ for some $\boldsymbol{O} \in \mathcal{O}_{d\times d}$, we have

$$\sigma_r(\mathcal{P}_{\boldsymbol{V}^\star}\boldsymbol{O}) = \sigma_r(\boldsymbol{V}^\star\boldsymbol{V}^{\star\top}) = 1. \tag{4.201}$$

Hence, Condition 1 is satisfied with $c_0 = 1$.

Spectral initialization. Let $\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ be the eigendecomposition of the best rank-r' approximation of $\mathcal{R}_\Omega(\boldsymbol{X}^\star)$ measured in Frobenius norm. Suppose $r \leq r' \leq d$ and $\boldsymbol{Z} = \boldsymbol{U}/\|\boldsymbol{U}\|$, where $\boldsymbol{U} = \boldsymbol{V}\boldsymbol{\Sigma}^{1/2}$. Corollary 7 tells us that, with probability at least $1 - \frac{1}{d^3}$, we have

$$\|\mathcal{R}_\Omega(\boldsymbol{X}^\star) - \boldsymbol{X}^\star\| \leq \Gamma\sigma_1^\star\sqrt{\frac{\mu^2 r^2}{pd}}. \tag{4.202}$$

Conditioned on this event, we have

$$\|\boldsymbol{U}\|^2 = \|\mathcal{R}_\Omega(\boldsymbol{X}^\star)\| \le \|\boldsymbol{X}^\star\| + \|\mathcal{R}_\Omega(\boldsymbol{X}^\star) - \boldsymbol{X}^\star\| \le 2\sigma_1^\star. \tag{4.203}$$

On the other hand, by Weyl's inequality, we have

$$\begin{aligned}
\sigma_r^2(\mathcal{P}_{\boldsymbol{V}^\star}\boldsymbol{U}) &= \sigma_r(\boldsymbol{V}^{\star\top}\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{V}^\star) \\
&\ge \sigma_r(\boldsymbol{V}^{\star\top}\mathcal{R}_\Omega(\boldsymbol{X}^\star)\boldsymbol{V}^\star) - \sigma_{r+1}(\mathcal{R}_\Omega(\boldsymbol{X}^\star)) \\
&\ge \sigma_r(\boldsymbol{V}^{\star\top}\boldsymbol{X}^\star\boldsymbol{V}^\star) - \|\mathcal{R}_\Omega(\boldsymbol{X}^\star) - \boldsymbol{X}^\star\| - \sigma_{r+1}(\mathcal{R}_\Omega(\boldsymbol{X}^\star)) \\
&\ge \sigma_r^\star - 2\,\|\mathcal{R}_\Omega(\boldsymbol{X}^\star) - \boldsymbol{X}^\star\| \\
&\ge 0.5\sigma_r^\star.
\end{aligned} \tag{4.204}$$

Combining the above two inequalities, we conclude that, with probability at least $1 - \frac{1}{d^3}$, Condition 1 is satisfied with $c_0 = \frac{1}{2\kappa}$.

4.F Concentration Inequalities for Matrix Completion

Recall that the sampling matrix $\boldsymbol{\Omega} \in \mathbb{R}^{d\times d}$ is defined as

$$\Omega_{i,j} = \begin{cases} 1 & \text{if } (i,j) \in \Omega, \\ 0 & \text{otherwise.} \end{cases} \tag{4.205}$$

The following lemma characterizes the concentration behavior of $\boldsymbol{\Omega}$.

Lemma 75 (Adapted from [Vu, 2018, Lemma 8])**.** *Suppose the sampling rate satisfies $p \gtrsim \frac{\log(d)}{d}$. There is a universal constant $\Gamma > 0$ such that, with probability at least $1 - \frac{1}{d^3}$, we have*

$$\left\| \frac{\boldsymbol{\Omega} + \boldsymbol{\Omega}^\top}{2p} - \boldsymbol{J} \right\| \le \Gamma\sqrt{\frac{d}{p}}. \tag{4.206}$$

Here $\boldsymbol{J}$ is the all-one matrix.

The original result appeared in [Vu, 2018, Lemma 8] only holds for symmetric Bernoulli model, i.e., $\boldsymbol{\Omega} = \boldsymbol{\Omega}^\top$. However, we can easily extend it to the asymmetric case via the dilation trick [Tropp et al., 2015]. Hence, we omit the details here. Next, we have the following extension to the leave-one-out sequences.

Corollary 5. *Suppose the sampling rate satisfies $p \gtrsim \frac{\log(d)}{d}$. Then, with probability at least*

$1-\frac{1}{d^3}$, *for all* $1 \le l \le d$, *we have*

$$\left\| \frac{\boldsymbol{\Omega}^{(l)} + \boldsymbol{\Omega}^{(l)\top}}{2p} - \boldsymbol{J} \right\| \le \Gamma \sqrt{\frac{d}{p}}. \tag{4.207}$$

Corollary 5. Note that for any $1 \le l \le d$, the matrix $\frac{\boldsymbol{\Omega}^{(l)}+\boldsymbol{\Omega}^{(l)\top}}{2p} - \boldsymbol{J}$ can be derived from $\frac{\boldsymbol{\Omega}+\boldsymbol{\Omega}^\top}{2p} - \boldsymbol{J}$ by zeroing out the l-th row and column. Hence, the proof follows by invoking Lemma 81 in Lemma 75. □

Lemma 76 ([Chen et al., 2020, Lemma A.1] and [Chen and Li, 2019, Lemma 8])**.** *For all* $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, *and* $\boldsymbol{D} \in \mathbb{R}^{d\times r}$, *we have*

$$\left| \left\langle \left(\mathrm{I} - \mathcal{R}_\Omega\right)\left(\boldsymbol{A}\boldsymbol{C}^\top\right), \boldsymbol{B}\boldsymbol{D}^\top \right\rangle \right| \le \left\| \frac{\boldsymbol{\Omega}+\boldsymbol{\Omega}^\top}{2p} - \boldsymbol{J} \right\| \cdot \|\boldsymbol{A}\|_{2,\infty} \|\boldsymbol{B}\|_{\mathrm{F}} \cdot \|\boldsymbol{C}\|_{\mathrm{F}} \|\boldsymbol{D}\|_{2,\infty}. \tag{4.208}$$

Moreover, for all $1 \le l \le d$, *we have*

$$\left| \left\langle \left(\mathrm{I} - \mathcal{R}_{\Omega^{(l)}}\right)\left(\boldsymbol{A}\boldsymbol{C}^\top\right), \boldsymbol{B}\boldsymbol{D}^\top \right\rangle \right| \le \left\| \frac{\boldsymbol{\Omega}^{(l)}+\boldsymbol{\Omega}^{(l)\top}}{2p} - \boldsymbol{J} \right\| \cdot \|\boldsymbol{A}\|_{2,\infty} \|\boldsymbol{B}\|_{\mathrm{F}} \cdot \|\boldsymbol{C}\|_{\mathrm{F}} \|\boldsymbol{D}\|_{2,\infty}. \tag{4.209}$$

As a special case, we have

Lemma 77. *For any matrix* $\boldsymbol{X} \in \mathbb{R}^{d\times d}$ *with the form* $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^\top$, *we have*

$$\left\|\left(\mathrm{I} - \mathcal{R}_{\Omega^{(l)}}\right)(\boldsymbol{X})\right\| \le \left\|\left(\mathrm{I} - \mathcal{R}_\Omega\right)(\boldsymbol{X})\right\| \le \left\| \frac{\boldsymbol{\Omega}+\boldsymbol{\Omega}^\top}{2p} - \boldsymbol{J} \right\| \|\boldsymbol{U}\|_{2,\infty} \|\boldsymbol{V}\|_{2,\infty}, \quad \forall 1 \le l \le d. \tag{4.210}$$

Note that the above two results are deterministic. Combining them with Lemma 75 leads to the following high-probability results.

Corollary 6 ([Chen et al., 2020, Lemma 4.3 and Lemma A.1])**.** *Suppose that the sampling rate satisfies* $p \gtrsim \frac{\log(d)}{d}$. *There exists a universal constant* $C > 0$ *such that, for any* $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D} \in \mathbb{R}^{d\times r}$, *with probability at least* $1 - \frac{1}{d^3}$, *we have*

$$\left| \left\langle \left(\mathrm{I} - \mathcal{R}_\Omega\right)\left(\boldsymbol{A}\boldsymbol{C}^\top\right), \boldsymbol{B}\boldsymbol{D}^\top \right\rangle \right| \le \Gamma \sqrt{\frac{d}{p}} \cdot \|\boldsymbol{A}\|_{2,\infty} \|\boldsymbol{B}\|_{\mathrm{F}} \cdot \|\boldsymbol{C}\|_{\mathrm{F}} \|\boldsymbol{D}\|_{2,\infty}. \tag{4.211}$$

Moreover, with the same probability, for any $1 \le l \le d$*, we have*

$$\left|\left\langle \left(\mathrm{I} - \mathcal{R}_{\Omega^{(l)}}\right)\left(\boldsymbol{A}\boldsymbol{C}^\top\right), \boldsymbol{B}\boldsymbol{D}^\top \right\rangle\right| \le \Gamma\sqrt{\frac{d}{p}} \cdot \|\boldsymbol{A}\|_{2,\infty} \|\boldsymbol{B}\|_{\mathrm{F}} \cdot \|\boldsymbol{C}\|_{\mathrm{F}} \|\boldsymbol{D}\|_{2,\infty}. \tag{4.212}$$

Corollary 7 ([Chen and Li, 2019, Lemma 9])**.** *Consider an arbitrary matrix* $\boldsymbol{X} \in \mathbb{R}^{d\times d}$ *decomposed as* $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^\top$*. There exists a universal constant* C *such that, with probability at least* $1 - \frac{1}{d^3}$*, we have*

$$\left\| \left(\mathrm{I} - \mathcal{R}_{\Omega}^{(l)}\right)(\boldsymbol{X}) \right\| \le \left\| \left(\mathrm{I} - \mathcal{R}_{\Omega}\right)(\boldsymbol{X}) \right\| \le \Gamma\sqrt{\frac{d}{p}} \|\boldsymbol{U}\|_{2,\infty} \|\boldsymbol{V}\|_{2,\infty}. \tag{4.213}$$

Lastly, we provide a finer result for a matrix of the form $\boldsymbol{X} - \boldsymbol{Y}$.

Lemma 78. *For two arbitrary symmetric matrices* $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{d\times d}$*, with probability at least* $1 - \frac{1}{d^3}$*, we have*

$$\left\| \left(\mathrm{I} - \mathcal{R}_{\Omega}^{(l)}\right)(\boldsymbol{X} - \boldsymbol{Y}) \right\| \le \left\| \left(\mathrm{I} - \mathcal{R}_{\Omega}\right)(\boldsymbol{X} - \boldsymbol{Y}) \right\| \le \Gamma\sqrt{\frac{d}{p}} \|\boldsymbol{X} - \boldsymbol{Y}\| \left(\|\boldsymbol{U}_{\boldsymbol{X}}\|_{2,\infty}^2 + \|\boldsymbol{U}_{\boldsymbol{Y}}\|_{2,\infty}^2 \right). \tag{4.214}$$

Proof. We denote $\boldsymbol{Z} = \boldsymbol{X} - \boldsymbol{Y}$ with an SVD $\boldsymbol{Z} = \boldsymbol{L}_{\boldsymbol{Z}}\boldsymbol{\Sigma}_{\boldsymbol{Z}}\boldsymbol{L}_{\boldsymbol{Z}}^\top$ (recall that $\boldsymbol{Z}$ is symmetric). Hence, with probability at least $1 - \frac{1}{d^3}$, we have

$$\begin{aligned}
\left\| \left(\mathrm{I} - \mathcal{R}_{\Omega}\right)(\boldsymbol{Z}) \right\| &\overset{(a)}{\le} \Gamma\sqrt{\frac{d}{p}} \|\boldsymbol{L}_{\boldsymbol{Z}}\boldsymbol{\Sigma}_{\boldsymbol{Z}}\|_{2,\infty} \|\boldsymbol{L}_{\boldsymbol{Z}}\|_{2,\infty} \\
&\overset{(b)}{\le} \Gamma\sqrt{\frac{d}{p}} \|\boldsymbol{\Sigma}_{\boldsymbol{Z}}\| \|\boldsymbol{L}_{\boldsymbol{Z}}\|_{2,\infty}^2 \\
&= \Gamma\sqrt{\frac{d}{p}} \|\boldsymbol{Z}\| \|\boldsymbol{L}_{\boldsymbol{Z}}\|_{2,\infty}^2 \\
&\overset{(c)}{\le} \Gamma\sqrt{\frac{d}{p}} \|\boldsymbol{Z}\| \left(\|\boldsymbol{L}_{\boldsymbol{X}}\|_{2,\infty}^2 + \|\boldsymbol{L}_{\boldsymbol{Y}}\|_{2,\infty}^2 \right).
\end{aligned} \tag{4.215}$$

Here in (a), we apply Corollary 7 upon setting $\boldsymbol{U} = \boldsymbol{L}_{\boldsymbol{Z}}\boldsymbol{\Sigma}_{\boldsymbol{Z}}$ and $\boldsymbol{V} = \boldsymbol{L}_{\boldsymbol{Z}}$. In (b), we apply Lemma 80. Lastly, in (c), we apply Lemma 86. This completes the proof. □

4.G Auxiliary Lemmas

4.G.1 Concentration Inequalities

Lemma 79 (Bernstein's inequality)**.** *Let $X_1, \cdots, X_n$ be independent zero-mean random variables. Suppose that $|X_i| \leq M$ almost surely, for all i and set $\nu^2 = \sum_{i=1}^n \operatorname{Var}[X_i^2]$. Then, for all positive t,*

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\nu^2 + \frac{2}{3}Mt}\right). \tag{4.216}$$

Or equivalently, with probability at least $1-\delta$, one has

$$\left|\sum_{i=1}^n X_i\right| \leq 2\nu\sqrt{\log(2/\delta)} + \frac{4}{3}M\log(2/\delta). \tag{4.217}$$

4.G.2 Matrix Norm Inequalities

Lemma 80 ([Cape, Tang, and Priebe, 2019, Proposition 6.5])**.** *For $\boldsymbol{A} \in \mathbb{R}^{d_1\times d_2}$, and $\boldsymbol{B} \in \mathbb{R}^{d_2\times d_3}$, we have*

$$\|\boldsymbol{AB}\|_{2,\infty} \leq \|\boldsymbol{A}\|_{2,\infty}\|\boldsymbol{B}\| \quad \textit{and} \quad \|\boldsymbol{AB}\|_{2,\infty} \leq \|\boldsymbol{A}\|_{\infty}\|\boldsymbol{B}\|_{2,\infty}. \tag{4.218}$$

Lemma 81 (Adapted from [Sun and Luo, 2016, Proposition A.3])**.** *For any matrix $\boldsymbol{A} \in \mathbb{R}^{d_1\times d_2}$, denote $\boldsymbol{A}_{-i,\cdot}$ ($\boldsymbol{A}_{\cdot,-i}$) as the matrix obtained by replacing the i-th row (column) of $\boldsymbol{A}$ by zeros, respectively. Then, we have*

$$\|\boldsymbol{A}_{-i,\cdot}\| \leq \|\boldsymbol{A}\| \quad \textit{and} \quad \|\boldsymbol{A}_{\cdot,-j}\| \leq \|\boldsymbol{A}\|, \forall i \in [d_1], j \in [d_2]. \tag{4.219}$$

Lemma 82 ([Sun and Luo, 2016, Proposition A.4])**.** *For any two matrices $\boldsymbol{A} \in \mathbb{R}^{d_1\times d_2}, \boldsymbol{B} \in \mathbb{R}^{d_2\times d_3}$, we have*

$$\|\boldsymbol{AB}\| \leq \|\boldsymbol{A}\|\|\boldsymbol{B}\| \quad \textit{and} \quad \|\boldsymbol{AB}\|_{\mathrm{F}} \leq \|\boldsymbol{A}\|\|\boldsymbol{B}\|_{\mathrm{F}}. \tag{4.220}$$

Furthermore, if $d_1 \geq d_2$, then

$$\sigma_{\min}(\boldsymbol{A})\|\boldsymbol{B}\|_{\mathrm{F}} \leq \|\boldsymbol{AB}\|_{\mathrm{F}} \quad \textit{and} \quad \sigma_{\min}(\boldsymbol{A})\|\boldsymbol{B}\| \leq \|\boldsymbol{AB}\|. \tag{4.221}$$

Lemma 83. *For arbitrary matrices $\boldsymbol{U} \in \mathbb{R}^{d_1\times d_2}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{d_2\times d_3}$ and $\boldsymbol{V} \in \mathbb{R}^{d_4\times d_3}$, we have*

$$\left\|\boldsymbol{U\Sigma V}^\top\right\|_{\max} \leq \|\boldsymbol{\Sigma}\|\|\boldsymbol{U}\|_{2,\infty}\|\boldsymbol{V}\|_{2,\infty}. \tag{4.222}$$

Proof. By Cauchy-Schwartz inequality, we have

$$\begin{aligned}\left\|\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top\right\|_{\max} &= \max_{i,j}\left|\sum_k [\boldsymbol{U}\boldsymbol{\Sigma}]_{i,k}V_{j,k}\right| \\ &\le \max_{i,j}\|(\boldsymbol{U}\boldsymbol{\Sigma})_{i,\cdot}\|\,\|\boldsymbol{V}_{j,\cdot}\| \\ &= \|\boldsymbol{U}\boldsymbol{\Sigma}\|_{2,\infty}\|\boldsymbol{V}\|_{2,\infty} \\ &\le \|\boldsymbol{\Sigma}\|\,\|\boldsymbol{U}\|_{2,\infty}\|\boldsymbol{V}\|_{2,\infty}.\end{aligned} \tag{4.223}$$

Here we apply Lemma 80 to derive the last inequality. This completes the proof. □

Lemma 84. *For any matrix $\boldsymbol{A}\in\mathbb{R}^{d_1\times d_2}$, we have*

$$\|\boldsymbol{A}\|_{2,\infty}^2 = \left\|\boldsymbol{A}\boldsymbol{A}^\top\right\|_{\max}. \tag{4.224}$$

Proof. We write $\boldsymbol{a}_1,\cdots,\boldsymbol{a}_{d_1}$ as the row vectors of $\boldsymbol{A}$. Then, we have

$$\|\boldsymbol{A}\|_{2,\infty}^2 = \max_{1\le i\le d_1}\|\boldsymbol{a}_i\|^2. \tag{4.225}$$

On the other hand, we can write $\left\|\boldsymbol{A}\boldsymbol{A}^\top\right\|_{\max}$ as

$$\left\|\boldsymbol{A}\boldsymbol{A}^\top\right\|_{\max} = \max_{1\le i,j\le d_1}|\langle\boldsymbol{a}_i,\boldsymbol{a}_j\rangle|. \tag{4.226}$$

First, we have $\left\|\boldsymbol{A}\boldsymbol{A}^\top\right\|_{\max}\le\|\boldsymbol{A}\|_{2,\infty}^2$ since

$$\left\|\boldsymbol{A}\boldsymbol{A}^\top\right\|_{\max} = \max_{1\le i,j\le d_1}|\langle\boldsymbol{a}_i,\boldsymbol{a}_j\rangle| \le \max_{1\le i,j\le d_1}\|\boldsymbol{a}_i\|\,\|\boldsymbol{a}_j\| = \max_{1\le i\le d_1}\|\boldsymbol{a}_i\|^2 = \|\boldsymbol{A}\|_{2,\infty}^2. \tag{4.227}$$

Second, we have $\left\|\boldsymbol{A}\boldsymbol{A}^\top\right\|_{\max}\ge\|\boldsymbol{A}\|_{2,\infty}^2$ upon noting that

$$\left\|\boldsymbol{A}\boldsymbol{A}^\top\right\|_{\max} = \max_{1\le i,j\le d_1}|\langle\boldsymbol{a}_i,\boldsymbol{a}_j\rangle| \ge \max_{i=j}|\langle\boldsymbol{a}_i,\boldsymbol{a}_j\rangle| = \max_{1\le i\le d_1}\|\boldsymbol{a}_i\|^2 = \|\boldsymbol{A}\|_{2,\infty}^2. \tag{4.228}$$

Therefore, we derive that $\left\|\boldsymbol{A}\boldsymbol{A}^\top\right\|_{\max} = \|\boldsymbol{A}\|_{2,\infty}^2$, which completes the proof. □

Lemma 85. *For two PSD matrices $\boldsymbol{A},\boldsymbol{B}\in\mathbb{R}^{d\times d}$ with $\boldsymbol{A}\preceq\boldsymbol{B}$, we have*

$$\|\boldsymbol{A}\|_{\max}\le\|\boldsymbol{B}\|_{\max}. \tag{4.229}$$

Proof. The proof follows by the fact that for any PSD matrix $\boldsymbol{A}$, we have $\|\boldsymbol{A}\|_{\max} =$

$\max_i\{A_{i,i}\}$. According to this fact, we immediately have

$$\|\boldsymbol{A}\|_{\max} = \max_i\{A_{i,i}\} \le \max_i\{B_{i,i}\} = \|\boldsymbol{B}\|_{\max} \tag{4.230}$$

since $\boldsymbol{A} \preceq \boldsymbol{B}$. Now we turn to prove this fact. Note that we can write any PSD matrix $\boldsymbol{A}$ as $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{P}^\top$. Then, according to Lemma 84, we have

$$\|\boldsymbol{A}\|_{\max} = \|\boldsymbol{P}\|_{2,\infty}^2 = \max_i \|\boldsymbol{p}_i\|^2 = \max_i\{A_{i,i}\}. \tag{4.231}$$

Here we write $\{\boldsymbol{p}_i\}$ as the row vectors of $\boldsymbol{P}$. This completes the proof. □

Lemma 86. *For $\boldsymbol{Z} = \boldsymbol{X} - \boldsymbol{Y}$, we have*

$$\|\boldsymbol{L}_{\boldsymbol{Z}}\|_{2,\infty}^2 \le \|\boldsymbol{L}_{\boldsymbol{X}}\|_{2,\infty}^2 + \|\boldsymbol{L}_{\boldsymbol{Y}}\|_{2,\infty}^2 . \tag{4.232}$$

Proof. We bound $\|\boldsymbol{L}_{\boldsymbol{Z}}\|_{2,\infty}^2$ as follows

$$\begin{aligned} \|\boldsymbol{L}_{\boldsymbol{Z}}\|_{2,\infty}^2 &\overset{(a)}{=} \left\|\boldsymbol{L}_{\boldsymbol{Z}}\boldsymbol{L}_{\boldsymbol{Z}}^\top\right\|_{\max} \\ &\overset{(b)}{\le} \left\|\boldsymbol{L}_{\boldsymbol{X}}\boldsymbol{L}_{\boldsymbol{X}}^\top + \boldsymbol{L}_{\boldsymbol{Y}}\boldsymbol{L}_{\boldsymbol{Y}}^\top\right\|_{\max} \\ &\le \left\|\boldsymbol{L}_{\boldsymbol{X}}\boldsymbol{L}_{\boldsymbol{X}}^\top\right\|_{\max} + \left\|\boldsymbol{L}_{\boldsymbol{Y}}\boldsymbol{L}_{\boldsymbol{Y}}^\top\right\|_{\max} \\ &\overset{(c)}{=} \|\boldsymbol{L}_{\boldsymbol{X}}\|_{2,\infty}^2 + \|\boldsymbol{L}_{\boldsymbol{Y}}\|_{2,\infty}^2 . \end{aligned} \tag{4.233}$$

Here (a) and (c) follow from Lemma 84. In (b), we apply Lemma 85 since $\boldsymbol{L}_{\boldsymbol{Z}}\boldsymbol{L}_{\boldsymbol{Z}}^\top \preceq \boldsymbol{L}_{\boldsymbol{X}}\boldsymbol{L}_{\boldsymbol{X}}^\top + \boldsymbol{L}_{\boldsymbol{Y}}\boldsymbol{L}_{\boldsymbol{Y}}^\top$. This is due to the fact that $\mathrm{col}(\boldsymbol{Z}) \subseteq \mathrm{col}(\boldsymbol{X}) \oplus \mathrm{col}(\boldsymbol{Y})$ which leads to $\boldsymbol{L}_{\boldsymbol{Z}}\boldsymbol{L}_{\boldsymbol{Z}}^\top = \mathcal{P}_{\boldsymbol{Z}} \preceq \mathcal{P}_{\boldsymbol{X}} + \mathcal{P}_{\boldsymbol{Y}}$. This completes the proof. □

Lemma 87 ([Tu et al., 2016, Lemma 5.4])**.** *For any $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{d\times r}$ with $\sigma_r(\boldsymbol{X}) > 0$, we have*

$$\mathrm{dist}^2(\boldsymbol{X}, \boldsymbol{Y}) \le \frac{1}{2\left(\sqrt{2}-1\right)\sigma_r^2(\boldsymbol{X})} \left\|\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{Y}\boldsymbol{Y}^\top\right\|_{\mathrm{F}}^2 . \tag{4.234}$$

Lemma 88. *Consider a matrix $\boldsymbol{U} \in \mathbb{R}^{d_1\times d_2}$ and a diagonal matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d_2\times d_2}$. We have*

$$\|\boldsymbol{U}\boldsymbol{\Sigma}\|_{2,\infty} \ge \sigma_r(\boldsymbol{\Sigma}) \|\boldsymbol{U}\|_{2,\infty} . \tag{4.235}$$

Proof. We first write $\boldsymbol{\Sigma} = \mathrm{Diag}\{\sigma_1, \cdots, \sigma_{d_2}\}$. Next, note that

$$\|(\boldsymbol{U}\boldsymbol{\Sigma})_{l,\cdot}\|^2 = \sum_{j=1}^{d_2} U_{l,j}^2\sigma_j^2 \ge \sigma_r^2(\boldsymbol{\Sigma}) \sum_{j=1}^{d_2} U_{l,j}^2 = \sigma_r^2(\boldsymbol{\Sigma}) \|\boldsymbol{U}_{l,\cdot}\|^2 . \tag{4.236}$$

Hence, taking the maximum over index l on both sides, we immediately obtain

$$\|\boldsymbol{U}\boldsymbol{\Sigma}\|_{2,\infty} \geq \sigma_r(\boldsymbol{\Sigma}) \|\boldsymbol{U}\|_{2,\infty}. \tag{4.237}$$

□

4.G.3 Other Useful Inequalities

Lemma 89 (Adapted from [Vershynin, 2018, Corollary 4.2.13])**.** *The covering number $\mathcal{N}_\epsilon$ of $\mathcal{B}_{\mathrm{F}}^{r\times r}(R)$ satisfies the following inequality for any $0 < \epsilon \leq 1$:*

$$\mathcal{N}_\epsilon \leq \left(\frac{3R}{\epsilon}\right)^{r^2}. \tag{4.238}$$

Lemma 90. *For two orthogonal matrices $\boldsymbol{V}_1, \boldsymbol{V}_2 \in \mathcal{O}_{d\times r}$, we have*

$$\left\|\boldsymbol{V}_1\boldsymbol{V}_1^\top - \boldsymbol{V}_2\boldsymbol{V}_2^\top\right\| \leq 2\|\boldsymbol{V}_1 - \boldsymbol{V}_2\| \quad \textit{and} \quad \left\|\boldsymbol{V}_1\boldsymbol{V}_1^\top - \boldsymbol{V}_2\boldsymbol{V}_2^\top\right\|_{\mathrm{F}} \leq 2\|\boldsymbol{V}_1 - \boldsymbol{V}_2\|_{\mathrm{F}}. \tag{4.239}$$

Proof. Note that $\boldsymbol{V}_1\boldsymbol{V}_1^\top - \boldsymbol{V}_2\boldsymbol{V}_2^\top = \boldsymbol{V}_1(\boldsymbol{V}_1 - \boldsymbol{V}_2)^\top + (\boldsymbol{V}_1 - \boldsymbol{V}_2)\boldsymbol{V}_2^\top$. Hence, we have

$$\left\|\boldsymbol{V}_1\boldsymbol{V}_1^\top - \boldsymbol{V}_2\boldsymbol{V}_2^\top\right\| \leq \|\boldsymbol{V}_1 - \boldsymbol{V}_2\|(\|\boldsymbol{V}_1\| + \|\boldsymbol{V}_2\|) \leq 2\|\boldsymbol{V}_1 - \boldsymbol{V}_2\|. \tag{4.240}$$

Similarly, for the Frobenius norm, we also have

$$\left\|\boldsymbol{V}_1\boldsymbol{V}_1^\top - \boldsymbol{V}_2\boldsymbol{V}_2^\top\right\|_{\mathrm{F}} \leq \|\boldsymbol{V}_1 - \boldsymbol{V}_2\|_{\mathrm{F}}(\|\boldsymbol{V}_1\| + \|\boldsymbol{V}_2\|) \leq 2\|\boldsymbol{V}_1 - \boldsymbol{V}_2\|_{\mathrm{F}}. \tag{4.241}$$

This completes the proof. □

Lemma 91. *For arbitrary matrix $\boldsymbol{X} \in \mathbb{R}^{d_1\times d_2}$ with $\mathrm{rank}(\boldsymbol{X}) = r$ and $\boldsymbol{O} \in \mathcal{O}_{d_1\times r}$, we have*

$$\sigma_r(\boldsymbol{X}) \geq \sigma_r\big(\boldsymbol{O}^\top\boldsymbol{X}\big). \tag{4.242}$$

Moreover, if $\boldsymbol{O} = \boldsymbol{U}_{\boldsymbol{X}}$, we have $\sigma_i(\boldsymbol{X}) = \sigma_i\left(\boldsymbol{O}^\top\boldsymbol{X}\right)$ for all $1 \leq i \leq r$.

Proof. We first prove the special case. Suppose $\boldsymbol{O} = \boldsymbol{U}_{\boldsymbol{X}}$, then we have $\boldsymbol{O}^\top\boldsymbol{X} = \boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{V}_{\boldsymbol{X}}^\top$. Note that this is the SVD of $\boldsymbol{O}^\top\boldsymbol{X}$ with the singular value matrix $\boldsymbol{\Sigma}_{\boldsymbol{X}}$. Hence, $\boldsymbol{O}^\top\boldsymbol{X}$ has

the same singluar values as $\boldsymbol{X}$. For the general case, it follows by

$$\begin{aligned}\sigma_r(\boldsymbol{O}^\top \boldsymbol{X}) &= \inf_{\boldsymbol{Y}\in\mathcal{M}_{\leq r-1}} \left\|\boldsymbol{O}^\top \boldsymbol{X} - \boldsymbol{Y}\right\|_{\mathrm{F}} \\ &\leq \left\|\boldsymbol{O}^\top \boldsymbol{X} - \boldsymbol{O}^\top \boldsymbol{U}_{\boldsymbol{X}} \boldsymbol{\Sigma}_{\boldsymbol{X},-1} \boldsymbol{V}_{\boldsymbol{X}}^\top\right\|_{\mathrm{F}} \\ &\leq \left\|\boldsymbol{O}^\top \sigma_r(\boldsymbol{X}) \boldsymbol{u}_{\boldsymbol{X},-1} \boldsymbol{v}_{\boldsymbol{X},-1}^\top\right\|_{\mathrm{F}} \\ &\leq \sigma_r(\boldsymbol{X}).\end{aligned} \tag{4.243}$$

Here we define $\mathcal{M}_{\leq r-1}$ as the set of all matrices of rank at most $r-1$. We also denote $\boldsymbol{\Sigma}_{\boldsymbol{X},-1} = \mathrm{Diag}\{\sigma_1^\star(\boldsymbol{X}), \cdots, \sigma_{r-1}(\boldsymbol{X}), 0\}$. Finally, $\boldsymbol{u}_{\boldsymbol{X},-1}, \boldsymbol{v}_{\boldsymbol{X},-1}$ refer to the last columns of $\boldsymbol{U}_{\boldsymbol{X}}, \boldsymbol{V}_{\boldsymbol{X}}$, respectively.

□

Lemma 92. *Consider two matrices $\boldsymbol{A} \in \mathbb{R}^{r\times r}, \boldsymbol{B} \in \mathbb{R}^{r\times d}$ where $\boldsymbol{A}$ is invertible and $\boldsymbol{B} \neq \boldsymbol{0}$. We have*

$$\sigma_r(\boldsymbol{AB}) \geq \sigma_r(\boldsymbol{A})\sigma_r(\boldsymbol{B}). \tag{4.244}$$

Proof. It directly follows by

$$\begin{aligned}\sigma_r(\boldsymbol{AB}) &= \min_{\|\boldsymbol{x}\|=1, \boldsymbol{x}\in\mathrm{range}(\boldsymbol{B}^\top)} \|\boldsymbol{ABx}\| \\ &\geq \sigma_r(\boldsymbol{A}) \min_{\|\boldsymbol{x}\|=1, \boldsymbol{x}\in\mathrm{range}(\boldsymbol{B}^\top)} \|\boldsymbol{Bx}\| \\ &= \sigma_r(\boldsymbol{A})\sigma_r(\boldsymbol{B}).\end{aligned} \tag{4.245}$$

□

Lemma 93. *For any $x \in [0, \frac{1}{2(r-1)})$ and $r > 1$, we have*

$$(1+x)^r \leq 1 + 2rx. \tag{4.246}$$

Lemma 94. *For the series $\{x_t\}_{t=0}^\infty$, the following two statements hold:*

- *Suppose that $x_{t+1} \leq Ax_t + B, \forall t \geq 0$ where $A > 0, A \neq 1$ and $x_0 + \frac{B}{A-1} \geq 0$. Then, we have*

$$x_t \leq A^t \left(x_0 + \frac{B}{A-1}\right) - \frac{B}{A-1}. \tag{4.247}$$

- *Suppose that $x_{t+1} \geq Ax_t - B, \forall t \geq 0$ where $A > 0$ and $x_0 - \frac{B}{A-1} \geq 0$. Then, we have*

$$x_t \geq A^t \left(x_0 - \frac{B}{A-1}\right) + \frac{B}{A-1}. \tag{4.248}$$

Proof. We first prove the first statement. Note that we can rewrite $x_{t+1} \leq Ax_t + B$ as $y_{t+1} \leq Ay_t$ where $y_t = x_t + \frac{B}{A-1}$. Note that $y_0 \geq 0, A > 0$ by our assumption. Hence, we derive $y_t \leq A^t y_0 = A^t \left(x_0 + \frac{B}{A-1}\right)$, which implies that $x_t = y_t - \frac{B}{A-1} \leq A^t \left(x_0 + \frac{B}{A-1}\right) - \frac{B}{A-1}$.

For the second statement, we first rewrite the condition as $y_{t+1} \geq Ay_t$ where $y_t = x_t - \frac{B}{A-1}$. Then, we have $y_t \geq A^t y_0$, which implies that $x_t = y_t + \frac{B}{A-1} \geq A^t y_0 + \frac{B}{A-1} = A^t \left(x_0 - \frac{B}{A-1}\right) + \frac{B}{A-1}$. □

CHAPTER 5

Function Basis Decomposition

This chapter is based on the following paper:

- Jianhao Ma, Lingjun Guo, and Salar Fattahi. Behind the scenes of gradient descent: A trajectory analysis via basis function decomposition. In *The Eleventh International Conference on Learning Representations*, 2023.

5.1 Introduction

Learning highly nonlinear models amounts to solving a nonconvex optimization problem, which is typically done via different variants of gradient descent (GD). But how does GD learn nonlinear models? Classical optimization theory asserts that, in the face of nonconvexity, GD and its variants may lack any meaningful optimality guarantee; they produce solutions that—while being first- or second-order optimal [Nesterov, 1998, Jin, Ge, Netrapalli, Kakade, and Jordan, 2017]—may not be globally optimal. In the rare event where the GD can recover a globally optimal solution, the recovered solution may correspond to an overfitted model rather than one with desirable generalization.

Inspired by the large empirical success of gradient-based algorithms in learning complex models, recent work has postulated that typical training losses have *benign landscapes*: they are devoid of spurious local minima and their global solutions coincide with true solutions—i.e., solutions corresponding to the true model. For instance, different variants of low-rank matrix factorization [Ge et al., 2016, 2017] and deep linear NNs [Kawaguchi, 2016] have benign landscapes. However, when spurious solutions *do exist* [Safran and Shamir, 2018] or global and true solutions *do not coincide* [Ma and Fattahi, 2023a], such a holistic view of the optimization landscape cannot explain the success of gradient-based algorithms. To address this issue, another line of research has focused on analyzing the solution trajectory of different algorithms. Analyzing the solution trajectory has been shown extremely powerful in sparse recovery [Vaskevicius et al., 2019], low-rank matrix factorization [Li et al., 2018,

Stöger and Soltanolkotabi, 2021], and linear DNNs [Arora, Cohen, Golowich, and Hu, 2018, Ma and Fattahi, 2022a]. However, these analyses are tailored to specific models and thereby cannot be directly generalized.

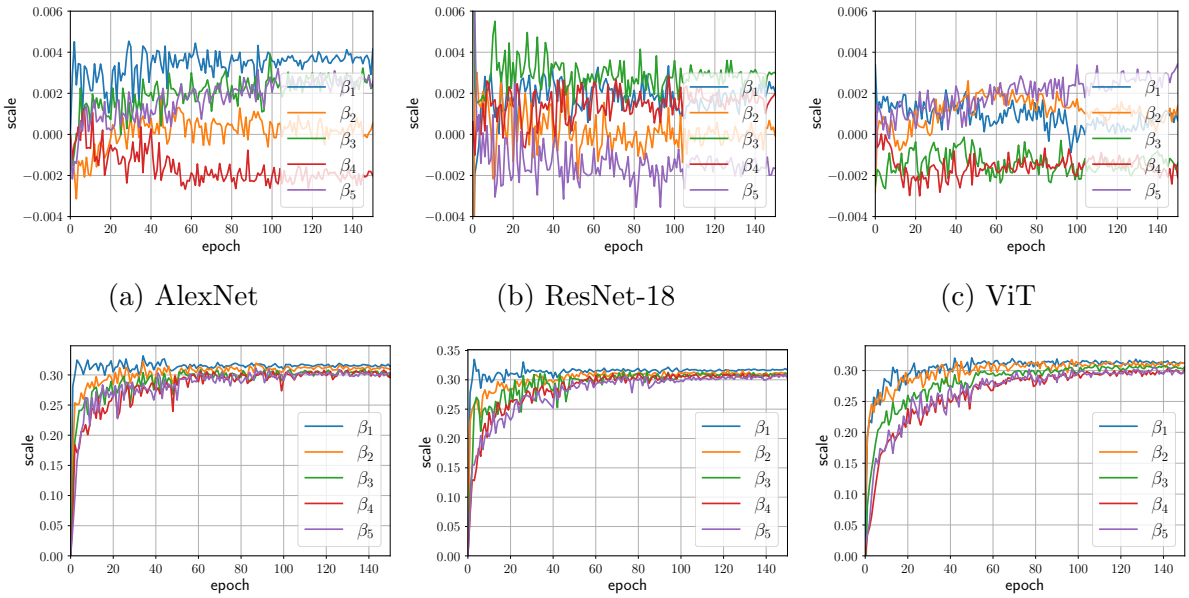
In this work, we propose a unifying framework for analyzing the optimization trajectory of GD based on a novel *basis function decomposition*. We show that, although the dynamics of GD may vary drastically on different models, they behave almost monotonically when projected onto an appropriate choice of orthonormal function basis.

Motivating example: Our first example illustrates this phenomenon on DNNs. We study the optimization trajectories of two adaptive gradient-based algorithms, namely AdamW and LARS, on three different DNN architectures, namely AlexNet [Krizhevsky, Sutskever, and Hinton, 2017], ResNet-18 [He et al., 2016], and Vision Transformer (ViT) [Dosovitskiy, Beyer, Kolesnikov, Weissenborn, Zhai, Unterthiner, Dehghani, Minderer, Heigold, Gelly, et al., 2020] with the CIFAR-10 dataset. The first row of the Figure 5.1 shows the top-5 coefficients of the solution trajectory when projected onto a randomly generated orthonormal basis. We see that the trajectories of the coefficients are highly non-monotonic and almost indistinguishable (they range between -0.04 to 0.06), implying that the energy of the obtained model is spread out on different orthogonal components. The second row of Figure 5.1 shows the same trajectory after projecting onto an orthogonal basis defined as the eigenvectors of the conjugate kernel after training [Long, 2021] (see Section 5.3.4 and Appendix 5.B for more details). Unlike the previous case, the top-5 coefficients carry more energy and behave monotonically (modulo the small fluctuations induced by the stochasticity in the algorithm) in all three architectures, until they plateau around their steady state. In other words, *the algorithm behaves more monotonically after projecting onto a correct choice of orthonormal basis.*

5.1.1 Main Contributions

The monotonicity of the projected solution trajectory motivates the use of an appropriate basis function decomposition to analyze the behavior of gradient-based algorithms. In this paper, we show how an appropriate basis function decomposition can be used to provide a much simpler convergence analysis for gradient-based algorithms on several representative learning problems, from simple kernel regression to complex DNNs. Our main contributions are summarized below:

- **Global convergence of GD via basis function decomposition:** We prove that GD learns the coefficients of an appropriate function basis that forms the true model. In



(a) AlexNet (b) ResNet-18 (c) ViT

(d) AlexNet (e) ResNet-18 (f) ViT

Figure 5.1: The solution trajectories of LARS on AlexNet and ResNet-18 and AdamW on ViT with ℓ_2-loss after projecting onto two different orthonormal bases. The first row shows the trajectories of the top-5 coefficients after projecting onto a randomly generated orthonormal basis. The second row shows the trajectories of the top-5 coefficients after projecting onto the eigenvectors of the conjugate kernel evaluated at the last epoch. More detail on our implementation can be found in Appendix 5.B.

particular, we show that GD learns the true model when applied to the expected ℓ_2-loss under certain gradient independence and gradient dominance conditions. Moreover, we characterize the convergence rate of GD, identifying conditions under which it enjoys linear or sublinear convergence rates. Our result does not require a benign landscape for the loss function and can be applied to both convex and nonconvex settings.

- **Application in learning problems:** We show that our general framework is well-suited for analyzing the solution trajectory of GD on different representative learning problems. Unlike the existing results, our proposed method leads to a much simpler trajectory analysis of GD for much broader classes of models. Using our technique, we improve the convergence of GD on the symmetric matrix factorization and provide an entirely new convergence result for GD on the orthogonal symmetric tensor decomposition. We also prove that GD enjoys an incremental learning phenomenon in both problems.

- **Empirical validation on DNNs:** We empirically show that our proposed framework applies to DNNs beyond GD. More specifically, we show that different gradient-based algorithms monotonically learn the coefficients of a particular function basis defined as the eigenvectors of the conjugate kernel after training (also known as "after kernel regime"). We show that this phenomenon happens across different architectures, datasets, solvers, and loss functions, strongly motivating the use of function basis decomposition to study deep learning.

5.2 General Framework: Function Basis Decomposition

We study the optimization trajectory of GD on the expected (population) ℓ_2-loss

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\theta}) := \frac{1}{2}\mathbb{E}_{\boldsymbol{x},y}\left[(f_{\boldsymbol{\theta}}(\boldsymbol{x}) - y)^2\right]. \qquad \text{(expected } \ell_2\text{-loss)}$$

Here the input $\boldsymbol{x} \in \mathbb{R}^d$ is drawn from an unknown distribution $\mathcal{D}$, and the output label y is generated as $y = f^\star(\boldsymbol{x}) + \epsilon$, where ϵ is an additive noise, independent of $\boldsymbol{x}$, with mean $\mathbb{E}[\epsilon] = 0$ and variance $\mathbb{E}[\epsilon^2] = \sigma_\epsilon^2 < \infty$. The model $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ is characterized by a parameter vector $\boldsymbol{\theta} \in \mathbb{R}^m$, which naturally induces a set of admissible models (model space for short) $\mathcal{F}_{\boldsymbol{\Theta}} := \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \mathbb{R}^m\}$. We do not require the true model $f^\star$ to lie within the model space; instead, we seek to obtain a model $f_{\boldsymbol{\theta}^\star} \in \mathcal{F}_{\boldsymbol{\Theta}}$ that is closest to $f^\star$ in $L^2(\mathcal{D})$-distance. In other

words, we consider $f^\star = f_{\boldsymbol{\theta}^\star}(\boldsymbol{x}) + f_\perp^\star(\boldsymbol{x})$, where $\boldsymbol{\theta}^\star = \arg\min_{\boldsymbol{\theta}} \|f_{\boldsymbol{\theta}} - f^\star\|_{L^2(\mathcal{D})}$. [1] To minimize the expected ℓ_2-loss, we use vanilla GD with constant step-size $\eta > 0$:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \nabla \mathcal{L}(\boldsymbol{\theta}_t). \tag{GD}$$

Definition 20 (Orthonormal function basis)**.** *A set of functions $\{\phi_i(\boldsymbol{x})\}_{i\in\mathcal{I}}$ forms an **orthonormal function basis** for the model space $\mathcal{F}_{\boldsymbol{\Theta}}$ with respect to the $L^2(\mathcal{D})$-metric if*

- *for any $i \in \mathcal{I}$, we have $\mathbb{E}_{x\sim\mathcal{D}}[\phi_i^2(\boldsymbol{x})] = 1$;*
- *for any $i, j \in \mathcal{I}$ such that $i \neq j$, we have $\mathbb{E}_{x\sim\mathcal{D}}[\phi_i(\boldsymbol{x})\phi_j(\boldsymbol{x})] = 0$;*
- *for any $f_{\boldsymbol{\theta}} \in \mathcal{F}_{\boldsymbol{\Theta}}$, there exists a unique sequence of basis coefficients $\{\beta_i(\boldsymbol{\theta})\}_{i\in\mathcal{I}}$ such that $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{i\in\mathcal{I}} \beta_i(\boldsymbol{\theta})\phi_i(\boldsymbol{x})$.*

Example 2 (Orthonormal basis for polynomials)**.** *Suppose that $\mathcal{F}_{\boldsymbol{\Theta}}$ is the class of all univariate real polynomials of degree at most n, that is, $\mathcal{F}_{\boldsymbol{\Theta}} = \{\sum_{i=1}^{n+1} \theta_i x^{i-1} : \boldsymbol{\theta} \in \mathbb{R}^{n+1}\}$. If $\mathcal{D}$ is a uniform distribution on $[-1, 1]$, then the so-called Legendre polynomials form an orthonormal basis for $\mathcal{F}_{\boldsymbol{\Theta}}$ with respect to the $L^2(\mathcal{D})$-metric [Olver, Lozier, Boisvert, and Clark, 2010, Chapter 14]. Moreover, if $\mathcal{D}$ is a normal distribution, then Hermite polynomials define an orthonormal basis for $\mathcal{F}_{\boldsymbol{\Theta}}$ with respect to the $L^2(\mathcal{D})$-metric [Olver et al., 2010, Chapter 18].*[2]

Example 3 (Orthonormal basis for symmetric matrix factorization)**.** *Suppose that the true model is defined as $f_{\boldsymbol{U}^\star}(\boldsymbol{X}) = \langle \boldsymbol{U}^\star \boldsymbol{U}^{\star\top}, \boldsymbol{X}\rangle$ with some rank-r matrix $\boldsymbol{U}^\star \in \mathbb{R}^{d\times r}$, and consider an "overparameterized" function class $\mathcal{F}_{\boldsymbol{\Theta}} = \{f_{\boldsymbol{U}}(\boldsymbol{X}) : \boldsymbol{U} \in \mathbb{R}^{d\times r'}\}$ where $r' \geq r$ is an overestimation of the rank. Moreover, suppose that the elements of $\boldsymbol{X} \sim \mathcal{D}$ are iid with zero mean and unit variance. Consider the eigenvalues of $\boldsymbol{U}^\star\boldsymbol{U}^{\star\top}$ as $\sigma_1 \geq \cdots \geq \sigma_d$ with $\sigma_{r+1} = \cdots = \sigma_d = 0$, and their corresponding eigenvectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_d$. It is easy to verify that the functions $\phi_{ij}(\boldsymbol{X}) = \langle \boldsymbol{z}_i\boldsymbol{z}_j^\top, \boldsymbol{X}\rangle$ for $1 \leq i, j \leq d$ define a valid orthogonal basis for $\mathcal{F}_{\boldsymbol{\Theta}}$ with respect to the $L^2(\mathcal{D})$-metric. Moreover, for any $f_{\boldsymbol{U}}(\boldsymbol{X})$, the basis coefficients can be obtained as $\beta_{ij}(\boldsymbol{U}) = \mathbb{E}\left[\left\langle \boldsymbol{U}\boldsymbol{U}^\top, \boldsymbol{X}\right\rangle \left\langle \boldsymbol{z}_i\boldsymbol{z}_j^\top, \boldsymbol{X}\right\rangle\right] = \left\langle \boldsymbol{z}_i\boldsymbol{z}_j^\top, \boldsymbol{U}\boldsymbol{U}^\top\right\rangle$. As will be shown in Section 5.3.2, this choice of orthonormal basis significantly simplifies the dynamics of GD for symmetric matrix factorization.*

Given the input distribution $\mathcal{D}$, we write $f_{\boldsymbol{\theta}^\star}(\boldsymbol{x}) = \sum_{i\in\mathcal{I}} \beta_i(\boldsymbol{\theta}^\star)\phi_i(\boldsymbol{x})$, where $\{\phi(\boldsymbol{x})\}_{i\in\mathcal{I}}$ is an orthonormal basis for $\mathcal{F}_{\boldsymbol{\Theta}}$ with respect to $L^2(\mathcal{D})$-metric, and $\{\beta_i(\boldsymbol{\theta}^\star)\}_{i\in\mathcal{I}}$ are the *true*

[1] Given a probability distribution $\mathcal{D}$, we define the $L_2(\mathcal{D})$-norm as $\|f\|_{L_2(\mathcal{D})}^2 = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[f^2(\mathbf{x})\right]$.

[2] Both Legendre and Hermite polynomials can be derived sequentially using Gram-Schmidt procedure. For instance, the first three Legendre polynomials are defined as $P_1(x) = 1/\sqrt{2}$, $P_2(x) = \sqrt{3/2}x$, and $P_3(x) = \sqrt{5/8}(3x^2 - 1)$.

basis coefficients. For short, we denote $\beta_i^\star = \beta_i(\boldsymbol{\theta}^\star)$. In light of this, the expected loss can be written as:

$$\mathcal{L}(\boldsymbol{\theta}) = \underbrace{\frac{1}{2}\sum_{i\in\mathcal{I}}(\beta_i(\boldsymbol{\theta}) - \beta_i^\star)^2}_{\text{optimization error}} + \underbrace{\frac{1}{2}\|f_\perp^\star\|_{L^2(\mathcal{D})}^2}_{\text{approximation error}} + \underbrace{\sigma_\epsilon^2/2}_{\text{noise}}. \tag{5.1}$$

Accordingly, GD takes the form

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \sum_{i\in\mathcal{I}} (\beta_i(\boldsymbol{\theta}_t) - \beta_i^\star)\, \nabla\beta_i(\boldsymbol{\theta}_t). \tag{GD dynamic}$$

Two important observations are in order based on GD dynamic: first, due to the decomposed nature of the expected loss, the solution trajectory becomes independent of the approximation error and noise. Second, in order to prove the global convergence of GD, it suffices to show the convergence of $\beta_i(\boldsymbol{\theta}_t)$ to $\beta_i^\star$. In fact, we will show that the coefficients $\beta_i(\boldsymbol{\theta}_t)$ enjoy simpler dynamics for particular choices of orthonormal basis that satisfy appropriate conditions.

Assumption 3 (Boundedness and smoothness)**.** *There exist constants $L_f, L_g, L_H > 0$ such that*

$$\|f_{\boldsymbol{\theta}}\|_{L^2(\mathcal{D})} \le L_f, \|\nabla f_{\boldsymbol{\theta}}\|_{L^2(\mathcal{D})} \le L_g, \left\|\nabla^2 f_{\boldsymbol{\theta}}\right\|_{L^2(\mathcal{D})} \le L_H. \tag{5.2}$$

The above assumptions are common in the optimization literature [Bubeck et al., 2015] and sometimes necessary to ensure the convergence of local-search algorithms [Patel and Berahas, 2022]. Moreover, although these assumptions may not hold globally, all of our subsequent results hold when Assumption 3 is satisfied within any bounded region for $\boldsymbol{\theta}$ that includes the solution trajectory. We will also relax these assumptions for several learning problems.

Proposition 19 (Dynamic of $\beta_i(\boldsymbol{\theta}_t)$)**.** *Under Assumption 3 and based on GD dynamic, we have*

$$\beta_i(\boldsymbol{\theta}_{t+1}) = \beta_i(\boldsymbol{\theta}_t) - \eta \sum_{j\in\mathcal{I}} \left(\beta_j(\boldsymbol{\theta}_t) - \beta_j^\star\right) \langle \nabla\beta_i(\boldsymbol{\theta}_t), \nabla\beta_j(\boldsymbol{\theta}_t)\rangle \pm \mathcal{O}\left(\eta^2 L_H L_f^2 L_g^2\right). \tag{5.3}$$

The above proposition holds for *any* valid choice of orthonormal basis $\{\phi_i(\boldsymbol{x})\}_{i\in\mathcal{I}}$. Indeed, there may exist multiple choices for the orthonormal basis, and not all of them would lead to equally simple dynamics for the coefficients. Examples of "good" and "bad" choices of orthonormal basis were presented for DNNs in our earlier motivating example. Indeed, an *ideal* choice of orthogonal basis should satisfy $\langle \nabla\beta_i(\boldsymbol{\theta}_t), \nabla\beta_j(\boldsymbol{\theta}_t)\rangle \approx 0$ for $i \neq j$, i.e., the gradients of the coefficients remain orthogonal along the solution trajectory. Under such

assumption, the dynamics of $\beta_i(\boldsymbol{\theta}_t)$ *almost* decompose over different indices:

$$\beta_i(\boldsymbol{\theta}_{t+1}) \approx \beta_i(\boldsymbol{\theta}_t) - \eta \left(\beta_i(\boldsymbol{\theta}_t) - \beta_i^\star\right) \|\nabla \beta_i(\boldsymbol{\theta}_t)\|^2 \pm \mathcal{O}(\eta^2), \tag{5.4}$$

where the last term accounts for the second-order interactions among the basis coefficients. If such an ideal orthonormal basis exists, then our next theorem shows that GD efficiently learns the true basis coefficients. To streamline the presentation, we assume that $\beta_1^\star \geq \cdots \geq \beta_k^\star > 0$ and $\beta_i^\star = 0, i > k$ for some $k < \infty$. We refer to the index set $\mathcal{S} = \{1, \ldots, k\}$ as *signal* and the index set $\mathcal{E} = \mathcal{I} \backslash \mathcal{S}$ as *residual*. When there is no ambiguity, we also refer to $\beta_i(\boldsymbol{\theta}_t)$ as a signal if $i \in \mathcal{S}$.

Theorem 23 (Convergence of GD with finite ideal basis)**.** *Suppose that the initial point $\boldsymbol{\theta}_0$ satisfies*

$$\begin{aligned} \beta_i(\boldsymbol{\theta}_0) &\geq C_1 \alpha, \qquad \text{for all } i \in \mathcal{S}, && \text{(lower bound on signals at } \boldsymbol{\theta}_0) \\ \|f_{\boldsymbol{\theta}_0}\|_{L^2(\mathcal{D})} &= \left(\sum_{i \in \mathcal{I}} \beta_i^2(\boldsymbol{\theta}_0)\right)^{1/2} \leq C_2 \alpha, && \text{(upper bound on energy at } \boldsymbol{\theta}_0) \end{aligned}$$

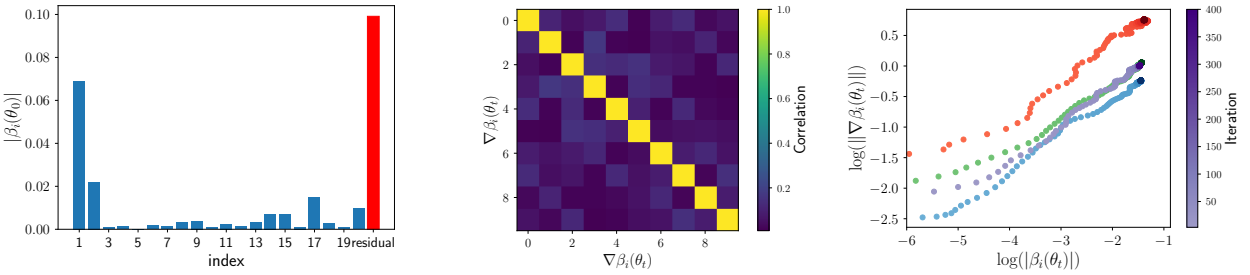
for $C_1, C_2 > 0$ and $\alpha \lesssim \beta_k^\star$. Moreover, suppose that the orthogonal function basis is finite, i.e., $|\mathcal{I}| = d$ for some finite d, and the gradients of the coefficients satisfy the following conditions for every $0 \leq t \leq T$:

$$\begin{aligned} \langle \nabla \beta_i(\boldsymbol{\theta}_t), \nabla \beta_j(\boldsymbol{\theta}_t) \rangle &= 0 && \text{for all } i \neq j, && \text{(gradient independence)} \\ \|\nabla \beta_i(\boldsymbol{\theta}_t)\| &\geq C \left|\beta_i(\boldsymbol{\theta}_t)\right|^\gamma && \text{for all } i \in \mathcal{S}, && \text{(gradient dominance)} \end{aligned}$$

for $C > 0$ and $1/2 \leq \gamma \leq 1$. Then, GD with step-size $\eta \lesssim \frac{\alpha^{2\gamma}}{\sqrt{d} C^2 L_H L_g^2 L_f^2} \beta_k^{\star 2\gamma} \log^{-1}\left(\frac{d\beta_k^\star}{C_1 \alpha}\right)$ satisfies:

- *If $\gamma = \frac{1}{2}$, then within $T = \mathcal{O}\left(\frac{1}{C^2 \eta \beta_k^\star} \log\left(\frac{\beta_k^\star}{C_1 \alpha}\right)\right)$ iterations, we have $\|f_{\boldsymbol{\theta}_T} - f_{\boldsymbol{\theta}^\star}\|_{L^2(\mathcal{D})} \lesssim \alpha$.*
- *If $\frac{1}{2} < \gamma \leq 1$, then within $T = \mathcal{O}\left(\frac{1}{C^2 \eta \beta_k^\star \alpha^{2\gamma - 1}}\right)$ iterations, we have $\|f_{\boldsymbol{\theta}_T} - f_{\boldsymbol{\theta}^\star}\|_{L^2(\mathcal{D})} \lesssim \alpha$.*

Theorem 23 shows that, under certain conditions on the basis coefficients and their gradients, GD with constant step-size converges to a model that is at most α-away from the true model. In particular, to achieve an ϵ-accurate solution for any $\epsilon > 0$, GD requires $\mathcal{O}((1/\epsilon)\log(1/\epsilon))$ iterations for $\gamma = 1/2$, and $\mathcal{O}(1/\epsilon^{2\gamma})$ iterations for $1/2 < \gamma \leq 1$ (ignoring the dependency on other problem-specific parameters). Due to its generality, our theorem inevitably relies on a small step-size and leads to a conservative convergence rate for GD.



(a) Small initialization (b) Gradient independence (c) Gradient dominance

Figure 5.2: The conditions of Theorem 23 are approximately satisfied for LARS on a 2-layer CNN with MNIST dataset. Here, the coefficients are obtained by projecting the solution trajectory onto the eigenvectors of the conjugate kernel after training. (a) The top-20 basis coefficients at a small random initial point along with residual energy on the remaining coefficients. (b) The maximum value of $|\cos(\nabla\beta_i(\boldsymbol{\theta}), \nabla\beta_j(\boldsymbol{\theta}))|$ for $1 \leq i < j \leq 10$ along the solution trajectory. (c) The scaling of $\|\nabla\beta_i(\boldsymbol{\theta}_t)\|$ with respect to $|\beta_i(\boldsymbol{\theta}_t)|$ for the top-4 coefficients.

Later, we will show how our proposed approach can be tailored to specific learning problems to achieve better convergence rates in each setting.

How realistic are the assumptions of Theorem 23? A natural question arises as to whether the conditions for Theorem 23 are realistic. We start with the conditions on the initial point. Intuitively, these assumptions entail that a non-negligible fraction of the energy is carried by the signal at the initial point. We note that these assumptions are mild and expected to hold in practice. For instance, We will show in Section 5.3 that, depending on the learning task, they are guaranteed to hold with fixed, random, or spectral initialization.[3] We have also empirically verified that these conditions are satisfied for DNNs with random or default initialization. For instance, Figure 5.2a illustrates the top-20 basis coefficients at $\boldsymbol{\theta}_0$ for LARS with random initialization on a realistic CNN. It can be seen that a non-negligible fraction of the energy at the initial point is carried by the first few coefficients.

The conditions on the coefficient gradients are indeed harder to satisfy; as will be shown later, the existence of an ideal orthonormal basis may not be guaranteed even for linear NNs. Nonetheless, we have empirically verified that, with an appropriate choice of the orthonormal basis, the gradients of the coefficients remain *approximately independent* throughout the solution trajectory. Figure 5.2b shows that, when the orthonormal basis is chosen as the eigenvectors of the conjugate kernel after training, the maximum value of $|\cos(\nabla\beta_i(\boldsymbol{\theta}_t), \nabla\beta_j(\boldsymbol{\theta}_t))|$ remains small throughout the solution trajectory. Finally, we turn to the gradient dominance condition. Intuitively, this condition entails that the gradient of each signal scales with its norm. We prove that this condition is guaranteed to hold for kernel regression, symmetric

[3] If $\boldsymbol{\theta}_0$ is selected from an isotropic Gaussian distribution, then C_1 and C_2 may scale with k and d. However, to streamline the presentation, we keep this dependency implicit.

matrix factorization, and symmetric tensor decomposition. Moreover, we have empirically verified that the gradient dominance holds across different DNN architectures. Figure 5.2c shows that this condition is indeed satisfied for the top-4 basis coefficients of the solution trajectory (other signal coefficients behave similarly). We also note that our theoretical result on GD may not naturally extend to LARS. Nonetheless, our extensive simulations suggest that our proposed analysis can be extended to other stochastic and adaptive variants of GD (see Section 5.3.4 and Appendix 5.B); a rigorous verification of this conjecture is left as future work.

5.3 Applications

In this section, we show how our proposed basis function decomposition can be used to study the performance of GD in different learning tasks, from simple kernel regression to complex DNNs. We start with the classical kernel regression, for which GD is known to converge linearly Karimi, Nutini, and Schmidt [2016]. Our purpose is to revisit GD through the lens of basis function decomposition, where there is a natural and simple choice for the basis functions. Next, we apply our approach to two important learning problems, namely symmetric matrix factorization and orthogonal symmetric tensor decomposition. In particular, we show how our proposed approach can be used to improve the convergence of GD for the symmetric matrix factorization and leads to a completely new convergence result for the orthogonal symmetric tensor decomposition. Finally, through extensive experiments, we showcase the promise of our proposed basis function decomposition on realistic DNNs.

5.3.1 Kernel Regression

In kernel regression (KR), the goal is to fit a regression model $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{i=1}^{d} \theta_i \phi_i(\boldsymbol{x})$ from the function class $\mathcal{F}_{\boldsymbol{\Theta}} = \{f_{\boldsymbol{\theta}}(\boldsymbol{x}) : \boldsymbol{\theta} \in \mathbb{R}^d\}$ to observation y, where $\{\phi_i(\boldsymbol{x})\}_{i=1}^{d}$ are some known kernel functions. Examples of KR are linear regression, polynomial regression (including those described in Example 2), and neural tangent kernel (NTK) [Jacot, Gabriel, and Hongler, 2018]. Without loss of generality, we may assume that the kernel functions $\{\phi_i(\boldsymbol{x})\}_{i=1}^{d}$ are orthonormal.[4] Under this assumption, the basis coefficients can be defined as $\beta_i(\boldsymbol{\theta}) = \theta_i$

[4]Suppose that $\{\phi_i(\boldsymbol{x})\}_{i=1}^{d}$ are not orthonormal. Let $\{\widetilde{\phi}_i(\boldsymbol{x})\}_{i\in\mathcal{I}}$ be any orthonormal basis for $\mathcal{F}_{\boldsymbol{\Theta}}$. Then, there exists a matrix A such that $\phi_i(\boldsymbol{x}) = \sum_j A_{ij}\widetilde{\phi}_j(\boldsymbol{x})$ for every $1 \le i \le d$. Therefore, upon defining $\widetilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^\top A$, one can write $f_{\widetilde{\boldsymbol{\theta}}^\star}(\boldsymbol{x}) = \sum_{i\in\mathcal{I}} \widetilde{\theta}_i^\star \widetilde{\phi}_i(\boldsymbol{x})$ which has the same form as the regression model.

and the expected loss can be written as

$$\mathcal{L}(\boldsymbol{\theta}_t) = \frac{1}{2}\mathbb{E}\left[(f_{\boldsymbol{\theta}_t}(\boldsymbol{x}) - f_{\boldsymbol{\theta}^\star}(\boldsymbol{x}))^2\right] = \frac{1}{2}\left\|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\right\|^2 = \frac{1}{2}\sum_{i=1}^{d}(\beta_i(\boldsymbol{\theta}_t) - \theta_i^\star)^2. \tag{5.5}$$

Moreover, the coefficients satisfy the gradient independence condition. Therefore, an adaptation of Proposition 19 reveals that the dynamics of the basis coefficients are independent of each other.

Proposition 20 (dynamics of $\beta_i(\boldsymbol{\theta}_t)$)**.** *Consider GD with a step-size that satisfies* $0 < \eta < 1$*. Then,*

- *for* $i \in \mathcal{S}$*, we have* $\beta_i(\boldsymbol{\theta}_t) = \beta_i^\star - (1-\eta)^t(\beta_i^\star - \beta_i(\boldsymbol{\theta}_0))$*,*
- *for* $i \notin \mathcal{S}$*, we have* $\beta_i(\boldsymbol{\theta}_t) = (1-\eta)^t\beta_i(\boldsymbol{\theta}_0)$*.*

Without loss of generality, we assume that $0 < \theta_k^\star \leq \cdots \leq \theta_1^\star \leq 1$ and $\left\|\boldsymbol{\theta}_0\right\|_\infty \leq \alpha$. Then, given Proposition 20, we have $|\beta_i(\boldsymbol{\theta}_t)| \leq 2 + \alpha$ for every $1 \leq i \leq k$. Therefore, the gradient dominance is satisfied with parameters $(C, \gamma) = (1/\sqrt{2+\alpha}, 1/2)$. Since both gradient independence and gradient dominance are satisfied, the convergence of GD can be established with an appropriate initial point.

Theorem 24. *Suppose that* $\boldsymbol{\theta}_0 = \alpha\mathbf{1}$*, where* $\alpha \lesssim k|\theta_k^\star|/d$*. Then, within* $T \lesssim (1/\eta)\log\left(k|\theta_1^\star|/\alpha\right)$ *iterations, GD with step-size* $0 < \eta < 1$ *satisfies* $\left\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^\star\right\| \lesssim \alpha$*.*

Theorem 24 reveals that GD with large step-size and small initial point converges linearly to an ϵ-accurate solution, provided that the initialization scale is chosen as $\alpha = \epsilon$. This is indeed better than our result on the convergence of GD for general models in Theorem 23.

5.3.2 Symmetric Matrix Factorization

In symmetric matrix factorization (SMF), the goal is to learn a model $f_{\boldsymbol{U}^\star}(\boldsymbol{X}) = \langle \boldsymbol{U}^\star\boldsymbol{U}^{\star\top}, \boldsymbol{X}\rangle$ with a low-rank matrix $\boldsymbol{U}^\star \in \mathbb{R}^{d\times r}$, where we assume that each element of $X \sim \mathcal{D}$ is iid with $\mathbb{E}[X_{ij}] = 0$ and $\mathbb{E}[X_{ij}^2] = 1$. Examples of SMF are matrix sensing Li et al. [2018] and completion Ge et al. [2016]. Given the eigenvectors $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_d\}$ of $\boldsymbol{U}^\star\boldsymbol{U}^{\star\top}$ and a function class $\mathcal{F}_{\boldsymbol{\Theta}} = \{f_{\boldsymbol{U}}(\boldsymbol{X}) : \boldsymbol{U} \in \mathbb{R}^{d\times r'}\}$ with $r' \geq r$, it was shown in Example 3 that the functions $\phi_{ij}(\boldsymbol{X}) = \langle \boldsymbol{z}_i\boldsymbol{z}_j^\top, \boldsymbol{X}\rangle$ define a valid orthogonal basis for $\mathcal{F}_{\boldsymbol{\Theta}}$ with coefficients $\beta_{ij}(\boldsymbol{U}) = \left\langle \boldsymbol{z}_i\boldsymbol{z}_j^\top, \boldsymbol{U}\boldsymbol{U}^\top\right\rangle$. Therefore, we have

$$\mathcal{L}(\boldsymbol{U}_t) = \frac{1}{4}\mathbb{E}\left[(f_{\boldsymbol{U}_t}(\boldsymbol{X}) - f_{\boldsymbol{U}^\star}(\boldsymbol{X}))^2\right] = \frac{1}{4}\mathbb{E}\left[\left(\left\langle \boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{U}^\star\boldsymbol{U}^{\star\top}, \boldsymbol{X}\right\rangle\right)^2\right] = \frac{1}{4}\sum_{i,j=1}^{r'}(\beta_{ij}(\boldsymbol{U}_t) - \beta_{ij}^\star)^2.$$

Here, the true basis coefficients are defined as $\beta^\star_{ii} = \sigma_i$ for $i \le r$, and $\beta^\star_{ij} = 0$ otherwise. Moreover, one can write $\|\nabla \beta_{ii}(\boldsymbol{U})\|_{\mathrm{F}} = 2\left\|\boldsymbol{z}_i \boldsymbol{z}_i^\top \boldsymbol{U}\right\|_{\mathrm{F}} = 2\sqrt{\beta_{ii}(\boldsymbol{U})}$. Therefore, gradient dominance holds with parameters $(C, \gamma) = (2, 1/2)$. However, gradient independence *does not* hold for this choice of function basis: given any pair (i, j) and (i, k) with $j \neq k$, we have $\langle \nabla \beta_{ij}(\boldsymbol{U}), \nabla \beta_{ik}(\boldsymbol{U})\rangle = \langle \boldsymbol{z}_j \boldsymbol{z}_k^\top, \boldsymbol{U}\boldsymbol{U}^\top\rangle$ which may not be zero. Despite the absence of gradient independence, our next proposition characterizes the dynamic of $\beta_{ij}(\boldsymbol{U}_t)$ via a finer control over the coefficient gradients.

Proposition 21. *Suppose that* $\gamma := \min_{1\le i\le r}\{\sigma_i - \sigma_{i+1}\} > 0$. *Let* $\boldsymbol{U}_0 = \alpha \boldsymbol{B}$, *where* $\alpha \lesssim \min\left\{(\eta\sigma_r^2)^{\Omega(\sigma_1/\gamma)}, (\sigma_r/d)^{\Omega(\sigma_1/\gamma)}, (\kappa \log^2(d))^{-\Omega(1/(\eta\sigma_r))}\right\}$ *and the entries of* $\boldsymbol{B}$ *are independently drawn from a standard normal distribution. Suppose that the step-size for GD satisfies* $\eta \lesssim 1/\sigma_1$. *Then, with probability of at least* $1 - \exp(-\Omega(r'))$*:*

- *For* $1 \le i \le r$, *we have* $0.99\sigma_i \le \beta_{ii}(\boldsymbol{U}_t) \le \sigma_i$ *within* $\mathcal{O}\left((1/(\eta\sigma_i)) \log(\sigma_i/\alpha)\right)$ *iterations.*
- *For* $t \ge 0$ *and* $i \neq j$ *or* $i, j > r$, *we have* $|\beta_{ij}(\boldsymbol{U}_t)| \lesssim \mathrm{poly}(\alpha)$.

Proposition 21 shows that GD with small random initialization learns larger eigenvalues before the smaller ones, which is commonly referred to as *incremental learning*. Incremental learning for SMF has been recently studied for gradient flow [Arora et al., 2019a, Li et al., 2020c], as well as GD with identical initialization for the special case $r' = d$ Chou, Gieshoff, Maly, and Rauhut [2020]. To the best of our knowledge, Proposition 21 is the first result that provides a full characterization of the incremental learning phenomenon for GD with random initialization on SMF.

Theorem 25. *Suppose that the conditions of Proposition 21 are satisfied. Then, with probability of at least* $1 - \exp(-\Omega(r'))$ *and within* $T \lesssim (1/(\eta\sigma_r)) \log(\sigma_r/\alpha)$ *iterations, GD satisfies*

$$\left\|\boldsymbol{U}_T \boldsymbol{U}_T^\top - \boldsymbol{M}^\star\right\|_{\mathrm{F}} \lesssim r' \log(d) d \alpha^2. \tag{5.6}$$

It has been shown in [Stöger and Soltanolkotabi, 2021, Thereom 3.3] that GD with small random initialization satisfies $\left\|\boldsymbol{U}_T \boldsymbol{U}_T^\top - \boldsymbol{M}^\star\right\|_{\mathrm{F}} \lesssim \left(d^2/r'^{15/16}\right) \alpha^{21/16}$ within the same number of iterations. Theorem 25 improves the dependency of the final error on the initialization scale α.

5.3.3 Orthogonal Symmetric Tensor Decomposition

We use our approach to provide a new convergence guarantee for GD on the orthogonal symmetric tensor decomposition (OSTD). In OSTD, the goal is to learn $f_{\boldsymbol{U}^\star}(\mathbf{X}) = \langle \mathbf{T}_{\boldsymbol{U}^\star}, \mathbf{X}\rangle$, where $\boldsymbol{U}^\star = [\boldsymbol{u}_1^\star, \ldots, \boldsymbol{u}_r^\star] \in \mathbb{R}^{d\times r}$ and $\mathbf{T}_{\boldsymbol{U}^\star} = \sum_{i=1}^r \boldsymbol{u}_i^{\star\otimes l} = \sum_{i=1}^d \sigma_i \boldsymbol{z}_i^{\otimes l}$ is a symmetric tensor

with order l and rank r. Here, $\sigma_1 \geq \cdots \geq \sigma_d$ are tensor eigenvalues with $\sigma_{r+1} = \cdots = \sigma_d = 0$, and $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_d$ are the corresponding tensor eigenvectors. The notation $\boldsymbol{u}^{\otimes l}$ refers to the l-time outer product of $\boldsymbol{u}$. We assume that $\mathbf{X} \sim \mathcal{D}$ is an l-order tensor whose elements are iid with zero mean and unit variance. Examples of OSTD are tensor regression [Tong, Ma, Prater-Bennette, Tripp, and Chi, 2022] and completion [Liu, Musialski, Wonka, and Ye, 2012].

When the rank of $\mathbf{T}_{\boldsymbol{U}^\star}$ is unknown, it must be overestimated. Even when the rank is known, its overestimation can improve the convergence of gradient-based algorithms [Wang, Wu, Lee, Ma, and Ge, 2020]. This leads to an *overparameterized* model $f_{\boldsymbol{U}}(\mathbf{X}) = \langle \mathbf{T}_{\boldsymbol{U}}, \mathbf{X} \rangle$, where $\mathbf{T}_{\boldsymbol{U}} = \sum_{i=1}^{r'} \boldsymbol{u}_i^{\otimes l}$ with an overestimated rank $r' \geq r$. Accordingly, the function class is defined as $\mathcal{F}_{\boldsymbol{\Theta}} = \left\{ f_{\boldsymbol{U}}(\boldsymbol{X}) : \boldsymbol{U} = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_{r'}] \in \mathbb{R}^{d \times r'} \right\}$. Upon defining a multi-index $\Lambda = (j_1, \cdots, j_l)$, the functions $\phi_\Lambda(\mathbf{X}) = \left\langle \otimes_{k=1}^{l} \boldsymbol{z}_{j_k}, \mathbf{X} \right\rangle$ for $1 \leq j_1, \ldots, j_l \leq d$ form an orthonormal basis for $\mathcal{F}_{\boldsymbol{\Theta}}$ with basis coefficients defined as

$$\beta_\Lambda(\boldsymbol{U}) = \mathbb{E}\left[\langle \mathbf{T}_{\boldsymbol{U}}, \mathbf{X} \rangle \left\langle \otimes_{k=1}^{l} \boldsymbol{z}_{j_k}, \mathbf{X} \right\rangle \right] = \sum_{i=1}^{r'} \left\langle \boldsymbol{u}_i^{\otimes l}, \otimes_{k=1}^{l} \boldsymbol{z}_{j_k} \right\rangle = \sum_{i=1}^{r'} \prod_{k=1}^{l} \langle \boldsymbol{u}_i, \boldsymbol{z}_{j_k} \rangle ,$$

and the expected loss can be written as

$$\mathcal{L}(\boldsymbol{U}) = \frac{1}{2}\mathbb{E}\left[(f_{\boldsymbol{U}}(\mathbf{X}) - f_{\boldsymbol{U}^\star}(\mathbf{X}))^2 \right] = \frac{1}{2} \left\| \sum_{i=1}^{r'} \boldsymbol{u}_i^{\otimes l} - \sum_{i=1}^{r} \sigma_i \boldsymbol{z}_i^{\otimes l} \right\|_{\mathrm{F}}^2 = \frac{1}{2} \sum_{\Lambda} \left(\beta_\Lambda(\boldsymbol{U}) - \beta_\Lambda^\star \right)^2 ,$$

where the true basis coefficients are $\beta_{\Lambda_i}^\star = \sigma_i$ for $\Lambda_i = (i, \ldots, i), 1 \leq i \leq r$, and $\beta_\Lambda^\star = 0$ otherwise. Unlike KR and SMF, neither gradient independence nor gradient dominance are satisfied for OSTD with a random or equal initialization. However, we show that these conditions are *approximately* satisfied throughout the solution trajectory, provided that the initial point is nearly aligned with the eigenvectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{r'}$; in other words, $\cos(\boldsymbol{u}_i(0), \boldsymbol{z}_i) \approx 1$ for every $1 \leq i \leq r'$.[5] Assuming that the initial point satisfies this alignment condition, we show that the *entire* solution trajectory remains aligned with these eigenvectors, i.e., $\cos(\boldsymbol{u}_i(t), \boldsymbol{z}_i) \approx 1$ for every $1 \leq i \leq r'$ and $1 \leq t \leq T$. Using this key result, we show that both gradient independence and gradient dominance are *approximately* satisfied throughout the solution trajectory. We briefly explain the intuition behind our approach for gradient dominance and defer our rigorous analysis for gradient independence to the appendix. Note that if $\cos(\boldsymbol{u}_i(t), \boldsymbol{z}_i) \approx 1$, then $\beta_{\Lambda_i}(\boldsymbol{U}_t) \approx \langle \boldsymbol{u}_i(t), \boldsymbol{z}_i \rangle^l$ and $\|\nabla \beta_{\Lambda_i}(\boldsymbol{U}_t)\|_{\mathrm{F}} \approx \|\nabla_{\boldsymbol{u}_i} \beta_{\Lambda_i}(\boldsymbol{U}_t)\| \approx l \langle \boldsymbol{u}_i(t), \boldsymbol{z} \rangle^{l-1}$. Therefore, gradient dominance holds with parameters $(C, \gamma) = (l, (l-1)/l)$. We will make this intuition rigorous in Appendix 5.G.

[5] We use the notations $\boldsymbol{u}_t$ or $\boldsymbol{u}(t)$ interchangeably to denote the solution at iteration t.

Proposition 22. *Suppose that the initial point $\boldsymbol{U}_0$ is chosen such that $\|\boldsymbol{u}_i(0)\| = \alpha^{1/l}$ and $\cos(\boldsymbol{u}_i(0), \boldsymbol{z}_i) \geq \sqrt{1-\gamma}$, for all $1 \leq i \leq r'$, where $\alpha \lesssim d^{-l^3}$ and $\gamma \lesssim (l\kappa)^{-l/(l-2)}$. Then, GD with step-size $\eta \lesssim 1/(l\sigma_1)$ satisfies:*

- *For $1 \leq i \leq r$, we have $0.99\sigma_i \leq \beta_{\Lambda_i}(\boldsymbol{U}_t) \leq 1.01\sigma_i$ within $\mathcal{O}\left((1/(\eta l \sigma_r))\alpha^{-\frac{l-2}{l}}\right)$ iterations.*
- *For $t \geq 0$ and $\Lambda \neq \Lambda_i$, we have $|\beta_\Lambda(\boldsymbol{U}_t)| = \mathrm{poly}(\alpha)$.*

Proposition 22 shows that, similar to SMF, GD learns the tensor eigenvalues incrementally. However, unlike SMF, we require a specific alignment for the initial point. We note that such initial point can be obtained in a pre-processing step via tensor power method within a number of iterations that is almost independent of d [Anandkumar, Ge, and Janzamin, 2017, Theorem 1]. We believe that Proposition 22 can be extended to random initialization; we leave the rigorous verification of this conjecture to future work. Equipped with this proposition, we next establish the convergence of GD on OSTD.

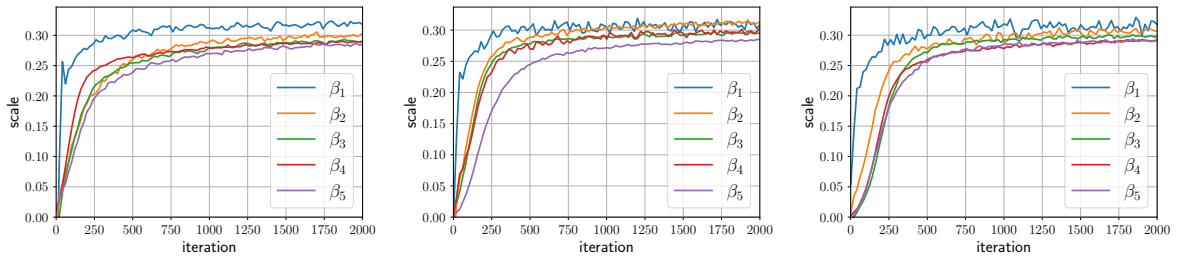
Theorem 26. *Suppose that the conditions of Proposition 22 are satisfied. Then, within $T \lesssim (1/(\eta l \sigma_r))\alpha^{-(l-2)/l}$ iterations, GD satisfies*

$$\|\mathbf{T}_{\boldsymbol{U}_T} - \mathbf{T}_{\boldsymbol{U}^\star}\|_{\mathrm{F}}^2 \lesssim r d^l \gamma \sigma_1^{\frac{l-1}{l}} \alpha^{\frac{1}{l}}. \tag{5.7}$$

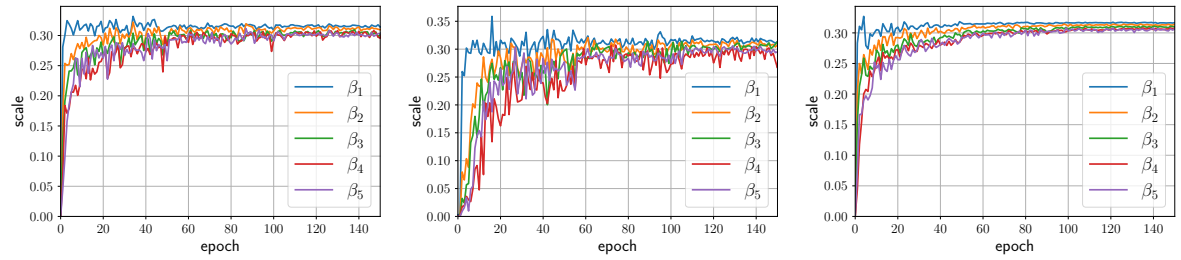
Theorem 26 shows that, with appropriate choices of η and α, GD converges to a solution that satisfies $\|\mathbf{T}_{\boldsymbol{U}_T} - \mathbf{T}_{\boldsymbol{U}^\star}\|_{\mathrm{F}}^2 \leq \epsilon$ within $\mathcal{O}(d^{l(l-2)}/\epsilon^{l-2})$ iterations. To the best of our knowledge, this is the first result establishing the convergence of GD with a large step-size on OSTD.

5.3.4 Empirical Verification on Neural Networks

In this section, we numerically show that the conjugate kernel after training (A-CK) can be used as a valid orthogonal basis for DNNs to capture the monotonicity of the solution trajectory of different optimizers on image classification tasks. To ensure consistency with our general framework, we use ℓ_2-loss, which is shown to have a comparable performance with the commonly-used cross-entropy loss Hui and Belkin [2020]. In Appendix 5.B, we extend our simulations to cross-entropy loss. The conjugate kernel (CK) is a method for analyzing the generalization performance of DNNs that uses the second to last layer (the layer before the last linear layer) at the *initial point* as the feature map [Daniely, Frostig, and Singer, 2016, Fan and Wang, 2020, Hu and Huang, 2021]. Recently, Long [2021] shows that A-CK, a variant of CK that is evaluated at the *last* epoch, better explains the generalization properties



(a) 2-block CNN (b) 3-block CNN (c) 4-block CNN



(d) LARS (e) SGD (f) AdamW

Figure 5.3: (First row) the projected trajectory of LARS on CNNs with MNIST dataset. The test accuracies for 2-Block, 3-Block, and 4-block CNN are $96.10\%, 96.48\%, 96.58\%$. (Second row) the projected trajectories of different optimizers on AlexNet with the CIFAR-10 dataset. The test accuracies for LARS, SGD, and AdamW are 90.54%, 91.03%, and 90.26%, respectively. We use the following settings for each optimizer: (d) LARS: learning rate of 2, Nesterov momentum of 0.9, and weight decay of 1×10^{-4}. (e) SGD: learning rate of 2 with "linear warm-up", Nesterov momentum of 0.9, weight decay of 1×10^{-4}. (f) AdamW: learning rate of 0.01.

of realistic DNNs. Surprisingly, we find that A-CK can be used not only to characterize the generalization performance but also to capture the underlining solution trajectory of different gradient-based algorithms.

To formalize the idea, note that any neural network whose last layer is linear can be characterized as $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{W}\psi(\boldsymbol{x})$, where $\boldsymbol{x} \in \mathbb{R}^d$ is the input drawn from the distribution $\mathcal{D}$, $\psi(\boldsymbol{x}) \in \mathbb{R}^m$ is the feature map with number of features m, and $\boldsymbol{W} \in \mathbb{R}^{k \times m}$ is the last linear layer with k referring to the number of classes. We denote the trained model, i.e., the model in the last epoch, by $f_{\boldsymbol{\theta}_\infty}(\boldsymbol{x}) = \boldsymbol{W}_\infty \psi_\infty(\boldsymbol{x})$. To form an orthogonal basis, we use SVD to obtain a series of basis functions $\phi_i(\boldsymbol{x}) = \boldsymbol{W}_{\infty,i}\psi_\infty(\boldsymbol{x})$ that satisfy $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\|\phi_i(\boldsymbol{x})\|^2] = 1$ and $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\langle \phi_i(\boldsymbol{x}), \phi_j(\boldsymbol{x})\rangle] = \delta_{ij}$ where δ_{ij} is the delta function. Hence, the coefficient $\beta_i(\boldsymbol{\theta}_t)$ at each epoch t can be derived as $\beta_i(\boldsymbol{\theta}_t) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\langle f_{\boldsymbol{\theta}_t}(\boldsymbol{x}), \phi_j(\boldsymbol{x})\rangle]$, where the expectation is estimated by its sample mean on the test set. More details on our implementation can be found in Appendix 5.B.

Performance on convolutional neural networks: We use LARS to train CNNs with varying depths on MNIST dataset. These networks are trained such that their test accuracies

are above 96%. Figures 5.3a-5.3c illustrate the evolution of the top-5 basis coefficients after projecting LARS onto the orthonormal basis obtained from A-CK. It can be observed that the basis coefficients are consistently monotonic across different depths, elucidating the generality of our proposed basis function decomposition. In the appendix, we discuss the connection between the convergence of the basis functions and the test accuracy for different architectures and loss functions.

Performance with different optimizers: The monotonic behavior of the projected solution trajectory is also observed across different optimizers. Figures 5.3d-5.3f show the solution trajectories of three optimizers, namely LARS, SGD, and AdamW, on AlexNet with the CIFAR-10 dataset. It can be seen that all three optimizers have a monotonic trend after projecting onto the orthonormal basis obtained from A-CK. Although our theoretical results only hold for GD, our simulations highlight the strength of the proposed basis function decomposition in capturing the behavior of other gradient-based algorithms on DNN.

5.A Discussion and Future Direction

Extension to empirical loss and implication on generalization. The most significant future direction is to extend our analysis of expected loss to the finite sample regime. It has been shown that incremental learning can drive generalization [Gissin et al., 2019]. Hence, our proposed framework is likely to explain the puzzling generalization ability of overparameterized machine learning models. For such an extension, the main technical difficulty is the relaxation of gradient independence. In the finite-sample regime, gradient independence does not hold exactly due to the randomness of samples. Therefore, an alternative approach would be to establish gradient independence approximately and with high probability. We have successfully applied our framework to matrix factorization and tensor decomposition, where gradient independence only holds approximately. Therefore, we believe that more general guarantees on the finite sample regime are not out of reach.

Extension to other optimization algorithms. In this paper, we mainly focused on GD as a representative of various optimization algorithms. Nonetheless, we believe that our analysis can be adapted to investigate other local-search optimization algorithms, such as GD with momentum [Nesterov, 1983], SGD [Robbins and Monro, 1951], and Adam [Kingma and Ba, 2014]. Overall, our approach can offer a unified framework for examining the implicit bias and incremental learning phenomena for different local-search algorithms.

5.B Additional Experiments

In this section, we provide more details on our simulation and further explore the empirical strength of the proposed basis function decomposition on different datasets, optimizers, loss functions, and batch sizes; see Table 5.1 for a summary of our simulations in this section.

Architectures	CNN	AlexNet	VGG11	ResNet-18	ResNet-34	ResNet-50	ViT
Datasets	MNIST	CIFAR-10	CIFAR-100				
Optimizers	SGD	AdamW	LARS				
Losses	ℓ_2-loss	CE loss					

Table 5.1: The summary of our experiments.

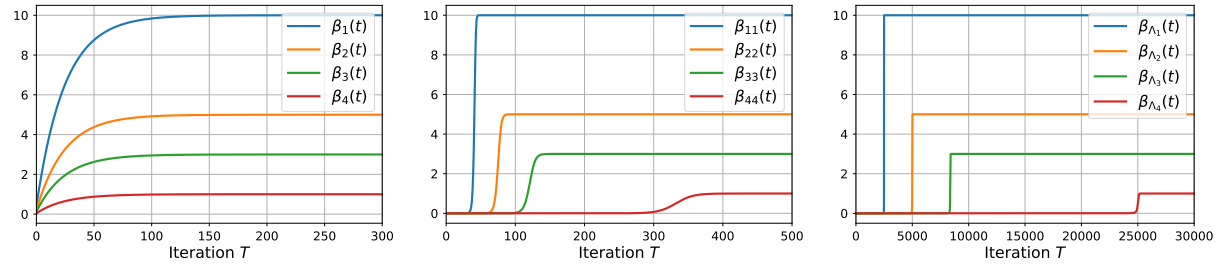
5.B.1 Numerical Verification of our Theoretical Results

In this section, we provide experimental evidence to support our theoretical results on kernel regression (KR), symmetric matrix factorization (SMF), and orthogonal symmetric tensor decomposition (OSTD). The results are presented in Figure 5.4.

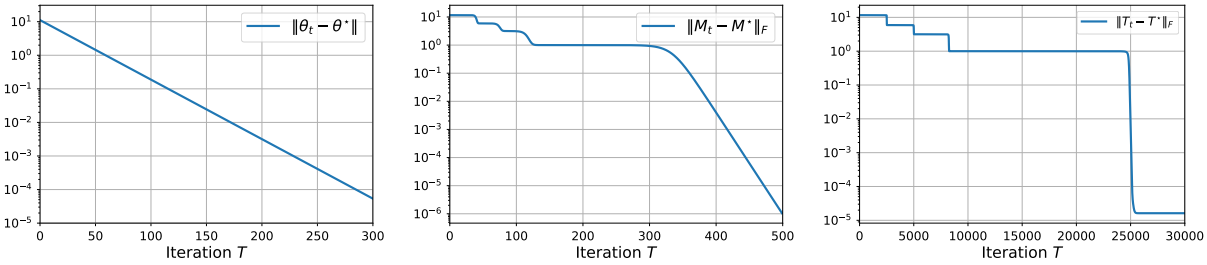
Kernel regression. We randomly generate 20 orthonormal kernel functions. The true model is comprised of 4 signal terms with basis coefficients $10, 5, 3, 1$. Figure 5.4a shows the trajectories of the top-4 basis coefficients of GD with initial point $\boldsymbol{\theta}_0 = 5 \times 10^{-7} \times \mathbf{1}$ and step-size $\eta = 0.4$. It can be seen that GD learns different coefficients at the same rate, which is in line with Proposition 20.

Symmetric matrix factorization. In this simulation, we aim to recover a rank-4 matrix $\boldsymbol{M}^\star = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^\top \in \mathbb{R}^{20\times 20}$. In particular, we assume that $\boldsymbol{V} \in \mathbb{R}^{20\times 4}$ is a randomly generated orthonormal matrix and $\boldsymbol{\Sigma} = \mathrm{Diag}\{10, 5, 3, 1\}$. We consider a fully over-parameterized model where $\boldsymbol{U} \in \mathbb{R}^{20\times 20}$ (i.e., $r' = 20$). Figure 5.4b illustrates the incremental learning phenomenon that was proved in Proposition 21 for GD with small Gaussian initialization $\boldsymbol{U}_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \alpha^2), \alpha = 5 \times 10^{-7}$ and step-size $\eta = 0.04$.

Orthogonal symmetric tensor decomposition. Finally, we present our simulations for OSTD. We aim to recover a rank-4 symmetric tensor of the form $\mathbf{T}^\star = \sum_{i=1}^4 \sigma_i \boldsymbol{z}_i^{\otimes 4}$ where σ_i are the nonzero eigenvalues with values $\{10, 5, 3, 1\}$ and $\boldsymbol{z}_i \in \mathbb{R}^{10}$ are the corresponding eigenvectors. We again consider a fully over-parameterized model with $r' = 10$. Figure 5.4c shows the incremental learning phenomenon for GD with an aligned initial point that satisfies $\cos(\boldsymbol{u}_i(0), \boldsymbol{z}_i) \geq 0.9983, 1 \leq i \leq r'$ and step-size $\eta = 0.001$.



(a) Solution trajectory for KR (b) Solution trajectory for SMF (c) Solution trajectory for OSTD



(d) Estimation error for KR (e) Estimation error for SMF (f) Estimation error for OSTD

Figure 5.4: Experimental verification to support Theorems 24, 25, and 26 in Section 5.3. The first row shows the projected trajectories of GD onto the specific basis functions we defined for each problem. The second row shows the estimation error.

5.B.2 Derivation of Basis Functions for DNNs

In this section, we provide more details on how we evaluate our proposed orthogonal basis induced by A-CK and calculate the corresponding coefficients $\beta_i(\boldsymbol{\theta}_t)$ for an arbitrary neural network. First, recall that any neural network whose last layer is linear can be characterized as $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{W}\psi(\boldsymbol{x})$, where $\boldsymbol{x} \in \mathbb{R}^d$ is the input drawn from the distribution $\mathcal{D}$, $\psi(\boldsymbol{x}) \in \mathbb{R}^m$ is the feature map with number of features m, and $\boldsymbol{W} \in \mathbb{R}^{k\times m}$ is the last linear layer with k referring to the number of classes. We denote the trained model, i.e., the model in the last epoch, by $f_{\boldsymbol{\theta}_\infty}(\boldsymbol{x}) = \boldsymbol{W}_\infty\psi_\infty(\boldsymbol{x})$. To form an orthogonal basis, we use SVD to obtain a series of basis functions $\phi_i(\boldsymbol{x}) = \boldsymbol{W}_{\infty,i}\psi_\infty(\boldsymbol{x})$ that satisfy $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\|\phi_i(\boldsymbol{x})\|^2] = 1$ and $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\langle\phi_i(\boldsymbol{x}), \phi_j(\boldsymbol{x})\rangle] = \delta_{ij}$ where δ_{ij} is the delta function. Hence, the coefficient $\beta_i(\boldsymbol{\theta}_t)$ at each epoch t can be derived as $\beta_i(\boldsymbol{\theta}_t) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\langle f_{\boldsymbol{\theta}_t}(\boldsymbol{x}), \phi_j(\boldsymbol{x})\rangle]$. In all of our implementation, we use the test dataset to approximate the population distribution.

Step 1: Obtaining the orthogonal basis $\phi_i(\boldsymbol{x})$. We denote $\boldsymbol{\Psi} = [\psi(\boldsymbol{x}_1), \cdots, \psi(\boldsymbol{x}_N)] \in \mathbb{R}^{m\times N}$ as the feature matrix where N is the number of the test data points. We write the SVD of $\boldsymbol{\Psi}$ as $\boldsymbol{\Psi} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$. The right singular vectors collected in $\boldsymbol{V}$ can be used to define the desired orthogonal basis of $\{\psi_i(\boldsymbol{x})\}_{i=1}^m$. To this goal, we write the prediction matrix as

$\boldsymbol{F} = \boldsymbol{W}\boldsymbol{\Psi} = \widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{\Psi}}$ where $\widetilde{\boldsymbol{W}} = \boldsymbol{W}\boldsymbol{U}\boldsymbol{\Sigma}$ and $\widetilde{\boldsymbol{\Psi}} = \boldsymbol{V}^\top$. Our goal is to define a set of matrices $\boldsymbol{A}_i$ such that $\phi_i(\boldsymbol{x}) = \boldsymbol{A}_i\widetilde{\psi}_i(\boldsymbol{x})$ form a valid orthonormal basis for $\boldsymbol{F} = \boldsymbol{W}\boldsymbol{\Psi} = \widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{\Psi}}$. Before designing such $\boldsymbol{A}_i$, first note that, due to the orthogonality of $\{\widetilde{\psi}_i(\boldsymbol{x})\}$, we have

$$\mathbb{E}\left[\|\phi_i(\boldsymbol{x})\|^2\right] = \|\boldsymbol{A}_i\|_{\mathrm{F}}^2\,, \quad \mathbb{E}\left[\langle\phi_i(\boldsymbol{x}), \phi_j(\boldsymbol{x})\rangle\right] = \langle\boldsymbol{A}_i, \boldsymbol{A}_j\rangle\,. \tag{5.8}$$

Therefore, it suffices to ensure that $\{\boldsymbol{A}_i\}$ are orthonormal. Consider the SVD of $\widetilde{\boldsymbol{W}}$ as $\widetilde{\boldsymbol{W}} = \sum_i \sigma_i \boldsymbol{u}_i\boldsymbol{v}_i^\top$. We define $\boldsymbol{A}_i = \boldsymbol{u}_i\boldsymbol{v}_i^\top$. Clearly, defined $\{\boldsymbol{A}_i\}$ are orthonormal. Moreover, it is easy to see that the basis coefficients (treated as the true basis coefficients) are exactly the singular values of $\widetilde{\boldsymbol{W}}$.

Step 2: Obtaining the basis coefficients $\beta_i(\boldsymbol{\theta}_t)$. After obtaining the desired orthonormal basis $\phi_i(\boldsymbol{x})$, we can calculate the coefficient $\beta_i(\boldsymbol{\theta}_t)$ for each epoch. Given the linear layer $\boldsymbol{W}_t$ and the feature matrix $\boldsymbol{\Phi}_t$ at epoch t, we can obtain the coefficients for the signal terms by projecting the prediction matrix $\boldsymbol{F}_t = \boldsymbol{W}_t\boldsymbol{\Psi}_t$ onto $\tilde{\boldsymbol{\Psi}} = \boldsymbol{V}^\top$. In particular, we write the prediction matrix as $\mathcal{P}_{\tilde{\psi}}\boldsymbol{F}_t = \mathcal{P}_{\tilde{\psi}}\boldsymbol{W}_t\boldsymbol{\Psi}_t = \boldsymbol{W}_t\boldsymbol{\Psi}_t\boldsymbol{V}\boldsymbol{V}^\top = \widetilde{\boldsymbol{W}}_t\widetilde{\boldsymbol{\Psi}}$ where $\widetilde{\boldsymbol{W}}_t = \boldsymbol{W}_t\boldsymbol{\Psi}_t\boldsymbol{V}$. Hence, the basis coefficients can be easily calculated as $\beta_i(\boldsymbol{\theta}_t) = \left\langle\widetilde{\boldsymbol{W}}_t, \boldsymbol{u}_i\boldsymbol{v}_i^\top\right\rangle$.

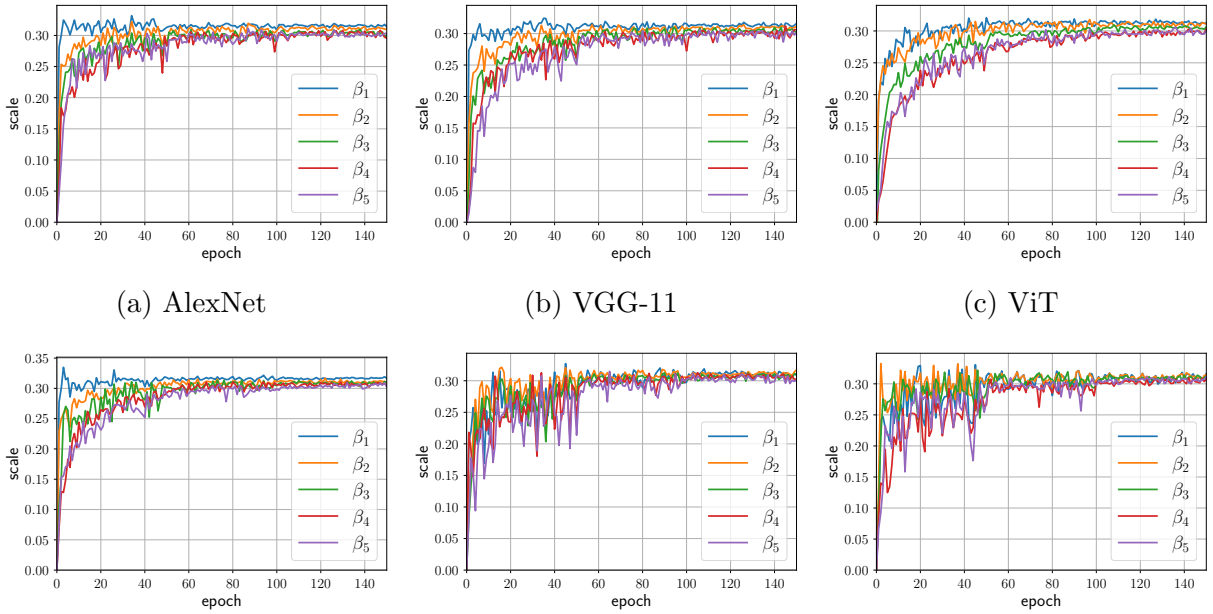
5.B.3 Further Details on the Experiments

In this section, we provide more details on our experiments presented in the main body of the paper and compare them with other DNN architectures.

All of our experiments are implemented in `Python 3.9, Pytorch 1.12.1` environment and run through a local server SLURM using NVIDIA Tesla with V100-PCIE-16GB GPUs. We use an additional NNGeometry package for calculating batch gradient, and our implemention of ViT is adapted from https://juliusruseckas.github.io/ml/cifar10-vit.html. To ensure consistency with our theoretical results, we drop the last softmax operator and use the ℓ_2-loss throughout this section. All of our training data are augmented by `RandomCrop` and `RandomHorizontalFlip`, and normalized by mean and standard deviation.

Experimental details for Figure 5.1. Here, we describe our implementation details for Figure 5.1, and present additional experiments on VGG-11, ResNet-34, and ResNet-50 with the CIFAR-10 dataset. The results can be seen in Figure 5.5. We use standard data augmentation for all architectures except for ViT. For ViT, we only use data normalization.

To obtain a stable A-CK, we trained the above models for 300 epochs. For ResNet-18, we used LARS with a learning rate of 0.5 and applied small initialization with $\alpha = 0.3$, i.e., we scale the default initial point by $\alpha = 0.3$. For ResNet-34 and ResNet-50, we choose the



(a) AlexNet (b) VGG-11 (c) ViT

(d) ResNet-18 (e) ResNet-34 (f) ResNet-50

Figure 5.5: Solution trajectories for different architectures trained on CIFAR-10. The test accuracies for Alexnet, VGG-11 and ViT are 90.54%, 91.11%, 77.99%, respectively. The test accuracies for ResNet-18, ResNet-34 and ResNet-50 are 88.31%, 93.93% and 94.06%, respectively.

default learning rate and apply small initialization with $\alpha = 0.3$. For ViT, we use AdamW with a learning rate of 0.01. The remaining parameters are set to their default values.

Experiments details for the first row of Figure 5.3. Here we conduct experiments on MNIST dataset with different CNN architectures. The CNNs are composed of l blocks of layers, followed by a single fully connected layer. A block of a CNN consists of a convolutional layer, an activation layer, and a pooling layer. In our experiments, we use ReLU activation and vary the depth of the network. For the first block, we used identity pooling. For the remaining blocks, we used max-pooling.

For 2-block CNN, we set the convolutional layer width to 256 and 64, respectively. For 3-block CNN, we set the convolutional layer width to $256, 128$, and 64, respectively. And for 4-block CNN, we set the convolutional layer width to $256, 128, 128$, and 64, respectively. To train these networks, we used LARS with the learning rate of 0.05. The remaining parameters are set to their default values. We run 20 epochs to calculate A-CK.

Experiments details for second row of Figure 5.3. We conduct experiments to compare the performance of different optimizers on the CIFAR-10 dataset. In particular, we use AlexNet to compare the performance of three optimizers, i.e., SGD, AdamW, LARS. For

SGD, we set the base learning rate to be 2 with Nesterov momentum of 0.9 and weight decay of 0.0001, together with the "linear warm-up" technique.[6] For AdamW, we set the learning rate to 0.01 and keep the remaining parameters unchanged. For LARS, we set the learning rate to 2 with Nesterov momentum of 0.9 and weight decay of 0.0001. The remaining parameters are set to the default setting.

5.B.4 Experiments for CIFAR-100

In this section, we conduct experiments using the CIFAR-100 dataset which is larger than both CIFAR-10 and MNIST. Our simulations are run on AlexNet, VGG-11, ViT, ResNet-18, ResNet-34, and ResNet-50. In particular, we use the "loss scaling trick" [Hui and Belkin, 2020] defined as follows: consider the datapoint $(\boldsymbol{x}, \boldsymbol{y})$ where $\boldsymbol{x} \in \mathbb{R}^d$ is the input and $\boldsymbol{y} \in \mathbb{R}^k$ is a one-hot vector with 1 at position i. Then, the scaled ℓ_2-loss is defined as

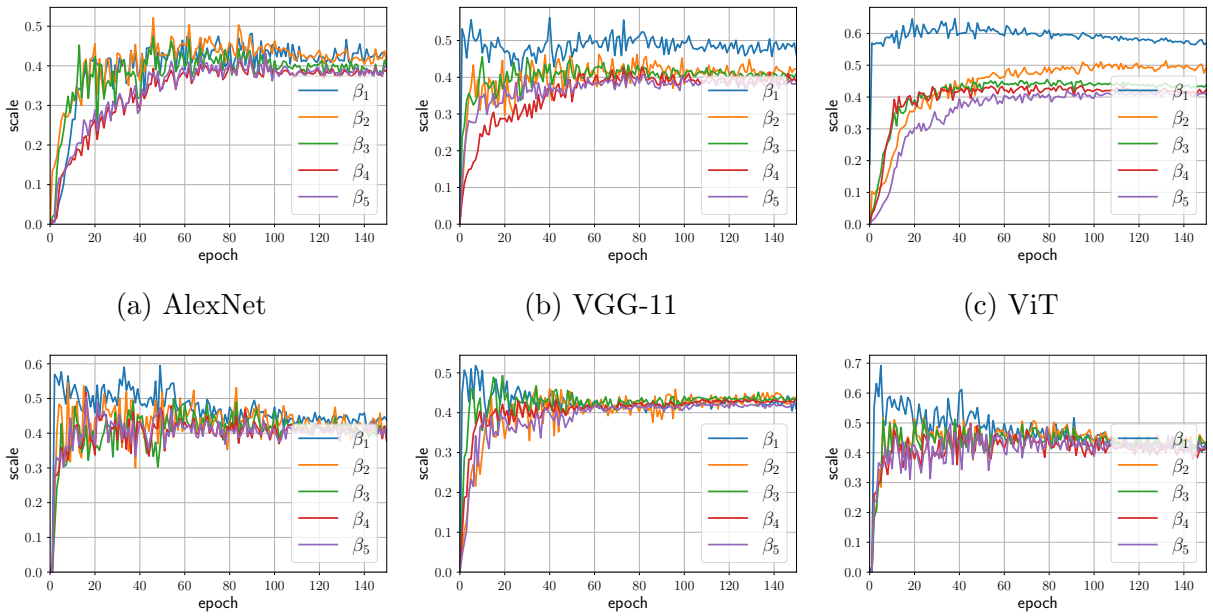
$$\ell_{2,\text{scaling}}(\boldsymbol{x}) = k \cdot (f_{\boldsymbol{\theta}}(\boldsymbol{x})[i] - M)^2 + \sum_{i' \neq i} (f_{\boldsymbol{\theta}}(\boldsymbol{x})[i'])^2, \tag{5.9}$$

for some constants $k, M > 0$. We set these parameters to $k = 1, M = 4$. For AlexNet and VGG-11, we use LARS with a base learning rate of $\eta = 1$. For ViT, we use AdamW with a base learning rate of 0.01 and batch size of 256. For ResNet architectures, we use SGD with a base learning rate of 0.3 and batch size of 64. We also add 5 warm-up epochs for ResNets. All the remaining parameters for the above architectures are set to their default values. The results can be seen in Figure 5.6. Our experiments highlight a trade-off between the monotonicity of the projected solution trajectories and the test accuracy: in order to obtain a higher test accuracy, one typically needs to pick a larger learning rate, which in turn results in more sporadic behavior of the solution trajectories. Nonetheless, even with large learning, the basis coefficients remain relatively monotonic after the first few epochs and converge to meaningful values.

5.B.5 Experiments for Different Losses

Next, we compare the projected solution trajectories on two loss functions, namely ℓ_2-loss and cross-entropy (CE) loss. We use LARS to train AlexNet on the CIFAR-10 dataset with both ℓ_2-loss and CE loss. In particular, we add the softmax operator before training the CE loss. For CE loss, we set the base learning rate of LARS to 1. For ℓ_2-loss, we use the base

[6] In "linear warm-up", we linearly increase the learning rate in the first 5 epochs. More precisely, we set the initial learning rate to 1×10^{-5} and linearly increase it to the selected learning rate in 5 epochs. After the first 5 epochs, the learning rate follows a regular decay scheme.



(a) AlexNet (b) VGG-11 (c) ViT

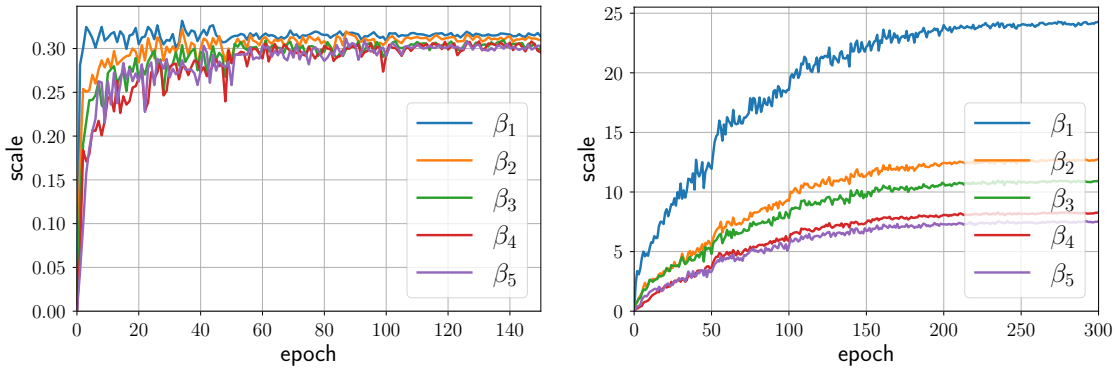
(d) ResNet-18 (e) ResNet-34 (f) ResNet-50

Figure 5.6: Solution trajectories for different architectures trained on CIFAR-100. The test accuracies for AlexNet, VGG-11, and ViT are 61.20%, 50.64%, and 50.41%, respectively. The test accuracies for ResNet-18, ResNet-34 and ResNet-50 are 78.11%, 74.26% and 80.05%, respectively.

learning rate of 2. The remaining parameters are set to their default values. The results can be seen in Figure 5.7. We observe that, similar to the ℓ_2-loss, the solution trajectory of the CE loss behaves monotonically after projecting onto the orthogonal basis induced by A-CK. Inspired by these observations, another venue for future research would be to extend our framework to general loss functions. Interestingly, the convergence of the basis coefficients is much slower than those of the ℓ_2-loss. This is despite the fact that CE loss can learn slightly faster than ℓ_2-loss in terms of test accuracy as shown by Hui and Belkin [2020].

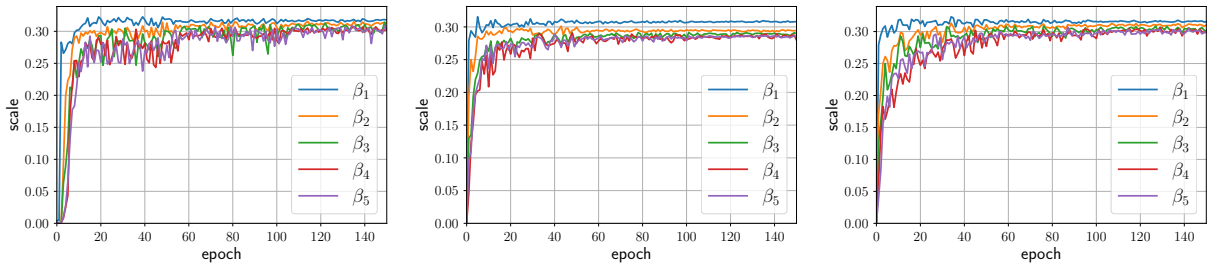
5.B.6 Experiments for Different Batch Size

Next, we study the effect of different batch sizes on the solution trajectory. We train AlexNet on the CIFAR-10 dataset. When testing for different batch sizes, we follow the "linear scaling" rule, i.e., the learning rate scales linearly with the batch size. For batch size of 32, we used SGD with a base learning rate $\eta = 0.1$ and 5 warm-up epochs. For batch size of 64, we used SGD with a base learning rate of $\eta = 0.2$ and 5 warm-up epochs. For batch size of 256, we used LARS with a base learning rate of 1. The remaining hyperparameters are set to their default values. The results are reported in Figure 5.8. We see that the projected solution trajectories share a similar monotonic behavior for different batch sizes.



(a) ℓ_2-loss (b) CE loss

Figure 5.7: Solution trajectories of LARS on the CIFAR-10 dataset with ℓ_2-loss and CE loss. The test accuracies are 90.54% for ℓ_2-loss and 91.09% for CE loss.



(a) Batch size 32 (b) Batch size 64 (c) Batch size 256

Figure 5.8: Solution trajectories for AlexNet on the CIFAR-10 dataset with different batch sizes. The test accuracies for batch size $32, 64$ and 256 are 91.50%, 91.79%, and 90.12%, respectively.

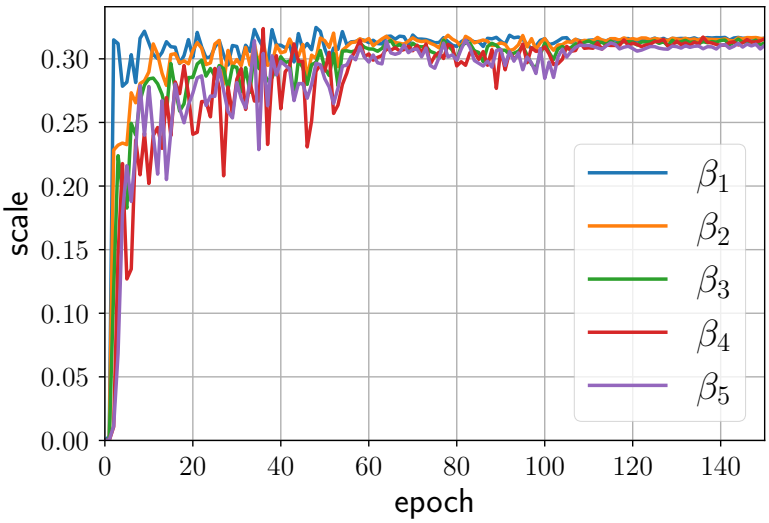


Figure 5.9: Solution trajectories for ResNet-18 on the CIFAR-10 dataset with SGD solver. The best test accuracy is 94.45%.

5.B.7 Experiments for ResNet-18 with SGD on CIFAR-10

Lastly, we train ResNet-18 with SGD on CIFAR-10. The results can be seen in Figure 5.9. To achieve good generalization, we use a large learning rate $\eta = 0.3$ (with 5 warm-up epochs starting at 1×10^{-5}), and decrease the learning rate by 0.33 for every 50 epochs. Moreover, We use a large batch size 512 to better imitate the trajectory of GD. Based on the simulation result, we can find an approximately monotonic (modulo the fluctuations caused by the randomness of the gradients) behavior of the dynamics of the top-5 components, which again validifies our theoretical results.

5.C Related Work

GD for general nonconvex optimization. Gradient descent and its stochastic or adaptive variants are considered as the "go-to" algorithms in large-scale (unconstrained) nonconvex optimization. Because of their first-order nature, they are known to converge to first-order stationary points [Nesterov, 1998]. Only recently it has been shown that GD [Lee, Panageas, Piliouras, Simchowitz, Jordan, and Recht, 2019, Panageas, Piliouras, and Wang, 2019] and its variants, such as perturbed GD [Jin et al., 2017] and SGD [Fang, Lin, and Zhang, 2019, Daneshmand, Kohler, Lucchi, and Hofmann, 2018], can avoid saddle points and converge to a second-order stationary point. However, these guarantees do not quantify the distance between the obtained solution and the globally optimal and/or true solutions.

To the best of our knowledge, the largest subclass of nonconvex optimization problems for which GD or its variants converge to meaningful solutions are those with benign landscapes. These problems include different variants of low-rank matrix optimization with exactly parameterized rank, namely matrix completion [Ge et al., 2016], matrix sensing [Ge et al., 2017, Zhang et al., 2021], dictionary learning [Sun et al., 2016], and robust PCA [Fattahi and Sojoudi, 2020], as well as deep linear neural networks [Kawaguchi, 2016]. However, benign landscape is too restrictive to hold in practice; for instance,Zhang [2021] shows that spurious local minima are ubiquitous in the low-rank matrix optimization, even under fairly mild conditions. Therefore, the notion of benign landscape cannot be used to explain the success of local search algorithms in more complex learning tasks.

GD for specific learning problems. Although there does not exist a unifying framework to study the global convergence of GD for general learning tasks, its convergence has been established in specific learning problems, such as kernel regression (which includes neural tangent kernel [Jacot et al., 2018]), sparse recovery [Vaskevicius et al., 2019], matrix factorization [Li et al., 2018], tensor decomposition [Wang et al., 2020, Ge, Ren, Wang, and Zhou, 2021], and linear neural network [Arora et al., 2018]. In what follows, we review specific learning tasks that are most related to our work.

The convergence of GD on kernel regression was studied far before the emergence of deep learning. Bauer, Pereverzev, and Rosasco [2007], Raskutti, Wainwright, and Yu [2014] establish the convergence of gradient descent on a special class of nonparametric kernel regression called reproducing kernel Hilbert space (RKHS). Recently, Jacot et al. [2018] discovered that under some conditions, neural networks can be approximated by a specific type of kernel models called neural tangent kernel (NTK). Later on, a series of papers studied the optimization [Allen-Zhu, Li, and Song, 2019b] and generalization [Allen-Zhu, Li, and Liang, 2019a, Arora, Du, Hu, Li, and Wang, 2019b] properties of NTK. As for the matrix factorization, Li et al. [2018], Stöger and Soltanolkotabi [2021] studied the global convergence of GD on the symmetric matrix sensing with noiseless measurements and overestimated rank. Later, these results were extended to noisy [Zhuo et al., 2021], asymmetric [Ye and Du, 2021], and robust [Ma and Fattahi, 2023a] variants. Wang et al. [2020], Ge et al. [2021] studied the dynamic of a modified GD for overcomplete nonconvex tensor decomposition. Moreover, Razin et al. [2021], Razin, Maman, and Cohen [2022] analyzed the implicit regularization and the incremental learning of gradient flow in hierarchical tensor decomposition and showed its connection to neural networks.

Conjugate kernel. Conjugate kernel (CK) at the initial point has been considered as one of the promising methods for studying the generalization properties of DNNs [Daniely et al., 2016, Hu and Huang, 2021, Fan and Wang, 2020]. However, similar to NTK, a major shortcoming of CK is that it cannot fully characterize the behavior of the practical neural networks [Vyas, Bansal, and Nakkiran, 2022]. Recent results have suggested that the conjugate kernel evaluated *after training* (for both NTK and CK) can better describe the generalization properties of DNNs [Fort, Dziugaite, Paul, Kharaghani, Roy, and Ganguli, 2020, Long, 2021]. In our work, we show that such "after kernel regime" can also be adapted to study the optimization trajectory of practical DNNs.

5.D Proofs for General Framework

5.D.1 Proof of Proposition 19

To prove this proposition, we first combine (5.1) and (GD):

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \nabla \mathcal{L}(\boldsymbol{\theta}_t) = \boldsymbol{\theta}_t - \eta \sum_{i \in \mathcal{I}} \left(\beta_i(\boldsymbol{\theta}_t) - \beta_i^\star\right) \nabla \beta_i(\boldsymbol{\theta}_t). \tag{5.10}$$

For notational simplicity, we denote $E(\boldsymbol{\theta}_t) = \frac{1}{2}\sum_{i \in \mathcal{E}} \beta_i^2(\boldsymbol{\theta}_t)$. Then, one can write

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \nabla \mathcal{L}(\boldsymbol{\theta}_t) = \boldsymbol{\theta}_t - \eta \sum_{i \in \mathcal{S}} \left(\beta_i(\boldsymbol{\theta}_t) - \beta_i^\star\right) \nabla \beta_i(\boldsymbol{\theta}_t) - \eta \nabla E(\boldsymbol{\theta}_t). \tag{5.11}$$

Due to the Mean-Value Theorem, there exists a $\xi \in \mathbb{R}^m$ such that

$$\begin{aligned}
\beta_i(\boldsymbol{\theta}_{t+1}) &= \beta_i \left(\boldsymbol{\theta}_t - \eta \sum_{j \in \mathcal{I}} \left(\beta_j(\boldsymbol{\theta}_t) - \beta_j^\star\right) \nabla \beta_j(\boldsymbol{\theta}_t) \right) \\
&= \beta_i(\boldsymbol{\theta}_t) - \eta \sum_{j \in \mathcal{I}} \left(\beta_j(\boldsymbol{\theta}_t) - \beta_j^\star\right) \langle \nabla \beta_i(\boldsymbol{\theta}_t), \nabla \beta_j(\boldsymbol{\theta}_t) \rangle + \frac{\eta^2}{2} \left\langle \nabla \mathcal{L}(\boldsymbol{\theta}_t), \nabla^2 \beta_i(\xi) \nabla \mathcal{L}(\boldsymbol{\theta}_t) \right\rangle.
\end{aligned} \tag{5.12}$$

On the other hand, one can write

$$\left| \left\langle \nabla \mathcal{L}(\boldsymbol{\theta}_t), \nabla^2 \beta_i(\xi) \nabla \mathcal{L}(\boldsymbol{\theta}_t) \right\rangle \right| \leq \sup_{\boldsymbol{\theta}} \left\| \nabla^2 \beta_i(\boldsymbol{\theta}) \right\| \left\| \nabla \mathcal{L}(\boldsymbol{\theta}_t) \right\|^2. \tag{5.13}$$

For $\sup_{\boldsymbol{\theta}} \|\nabla^2 \beta_i(\boldsymbol{\theta})\|$, we further have

$$\begin{aligned}
\sup_{\boldsymbol{\theta}} \left\|\nabla^2 \beta_i(\boldsymbol{\theta})\right\| &= \sup_{\boldsymbol{\theta}} \left\|\nabla^2 \mathbb{E}[f_{\boldsymbol{\theta}}(\boldsymbol{x})\phi(\boldsymbol{x})]\right\| \\
&= \sup_{\boldsymbol{\theta}} \left\|\mathbb{E}[\nabla^2 f_{\boldsymbol{\theta}}(\boldsymbol{x})\phi(\boldsymbol{x})]\right\| \\
&\leq \sup_{\boldsymbol{\theta}} \mathbb{E}\left[\left\|\nabla^2 f_{\boldsymbol{\theta}}(\boldsymbol{x})\right\| |\phi(\boldsymbol{x})|\right] \\
&\overset{(a)}{\leq} \sup_{\boldsymbol{\theta}} \left(\mathbb{E}\left[\left\|\nabla^2 f_{\boldsymbol{\theta}}(\boldsymbol{x})\right\|^2\right]\right)^{1/2} \left(\mathbb{E}\left[\phi^2(\boldsymbol{x})\right]\right)^{1/2} \\
&\overset{(b)}{\leq} L_H.
\end{aligned} \tag{5.14}$$

Here, we used Cauchy-Schwartz inequality for (a). Moreover, for (b), we used Assumption 3 and the definition of the orthonormal basis. Hence, we have

$$\beta_i(\boldsymbol{\theta}_{t+1}) = \beta_i(\boldsymbol{\theta}_t) - \eta \sum_{j\in\mathcal{I}} \left(\beta_j(\boldsymbol{\theta}_t) - \beta_j^\star\right) \langle \nabla\beta_i(\boldsymbol{\theta}_t), \nabla\beta_j(\boldsymbol{\theta}_t)\rangle \pm (1/2)\eta^2 L_H \left\|\nabla\mathcal{L}(\boldsymbol{\theta}_t)\right\|^2. \tag{5.15}$$

Now, it suffices to bound $\|\nabla\mathcal{L}(\boldsymbol{\theta}_t)\|^2$. Using Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\|\nabla\mathcal{L}(\boldsymbol{\theta}_t)\|^2 &= \left\|\nabla\mathbb{E}\left[\frac{1}{2}\left(f_{\boldsymbol{\theta}_t} - f_{\boldsymbol{\theta}^\star}\right)^2\right]\right\|^2 \\
&= \left\|\mathbb{E}\left[\left(f_{\boldsymbol{\theta}_t} - f_{\boldsymbol{\theta}^\star}\right)\nabla f_{\boldsymbol{\theta}_t}\right]\right\|^2 \\
&\leq \|f_{\boldsymbol{\theta}_t} - f_{\boldsymbol{\theta}^\star}\|_{L^2(\mathcal{D})}^2 \|\nabla f_{\boldsymbol{\theta}_t}\|_{L^2(\mathcal{D})}^2 \\
&\leq 4L_g^2 L_f^2.
\end{aligned} \tag{5.16}$$

Therefore, we conclude that

$$\beta_i(\boldsymbol{\theta}_{t+1}) = \beta_i(\boldsymbol{\theta}_t) - \eta \sum_{j\in\mathcal{I}} \left(\beta_j(\boldsymbol{\theta}_t) - \beta_j^\star\right) \langle \nabla\beta_i(\boldsymbol{\theta}_t), \nabla\beta_j(\boldsymbol{\theta}_t)\rangle \pm 2\eta^2 L_H L_g^2 L_f^2. \tag{5.17}$$

which completes the proof. □

5.D.2 Proof of Theorem 23

Proof. Invoking the gradient independence condition, Proposition 19 can be simplified as

$$\begin{aligned}
\beta_i(\boldsymbol{\theta}_{t+1}) &= \beta_i(\boldsymbol{\theta}_t) - \eta \sum_{j\in\mathcal{I}} \left(\beta_j(\boldsymbol{\theta}_t) - \beta_j^\star\right) \langle \nabla\beta_i(\boldsymbol{\theta}_t), \nabla\beta_j(\boldsymbol{\theta}_t)\rangle \pm 2\eta^2 L_H L_g^2 L_f^2 \\
&= \beta_i(\boldsymbol{\theta}_t) - \eta \left(\beta_i(\boldsymbol{\theta}_t) - \beta_i^\star\right) \|\nabla\beta_i(\boldsymbol{\theta}_t)\|^2 \pm 2\eta^2 L_H L_g^2 L_f^2.
\end{aligned} \tag{5.18}$$

We next provide upper and lower bounds for the residual and signal terms. Recall that $\mathcal{S} = \{i \in \mathcal{I} : \beta_i^\star \neq 0\}$, and $\mathcal{E} = \mathcal{I} \backslash \mathcal{S}$. We first consider the dynamic of the signal term $\beta_i(\boldsymbol{\theta}_t), i \in \mathcal{S}$. Without loss of generality, we assume $\beta_i^\star > 0$. Then, due to the gradient dominance condition, we have the following lower bound

$$\begin{aligned}\beta_i(\boldsymbol{\theta}_{t+1}) &\geq \beta_i(\boldsymbol{\theta}_t) - \eta \left(\beta_i(\boldsymbol{\theta}_t) - \beta_i^\star\right) \left\|\nabla \beta_i(\boldsymbol{\theta}_t)\right\|^2 - 2\eta^2 L_H L_g^2 L_f^2 \\ &\geq \left(1 + C^2 \eta \left(\beta_i^\star - \beta_i(\boldsymbol{\theta}_t)\right) \beta_i^{2\gamma-1}(\boldsymbol{\theta}_t)\right) \beta_i(\boldsymbol{\theta}_t) - 2\eta^2 L_H L_g^2 L_f^2, \quad i \in \mathcal{S}\end{aligned} \tag{5.19}$$

Next, for the dynamic of the residual term $\beta_i(\boldsymbol{\theta}_t), i \in \mathcal{E}$, we have

$$\beta_i(\boldsymbol{\theta}_{t+1}) = \left(1 - \eta \left\|\nabla \beta_i(\boldsymbol{\theta}_t)\right\|^2\right) \beta_i(\boldsymbol{\theta}_t) \pm 2\eta^2 L_H L_g^2 L_f^2. \tag{5.20}$$

Next, we show that $\|\nabla \beta_i(\boldsymbol{\theta}_t)\| \leq L_g$. One can write,

$$\begin{aligned}\|\nabla \beta_i(\boldsymbol{\theta}_t)\| &= \|\nabla \mathbb{E}\left[f_{\boldsymbol{\theta}_t}(\boldsymbol{x}) \phi_i(\boldsymbol{x})\right]\| \\ &\leq \mathbb{E}\left[\|\nabla f_{\boldsymbol{\theta}_t}(\boldsymbol{x})\| \left|\phi_i(\boldsymbol{x})\right|\right] \\ &\leq \left(\mathbb{E}\left[\|\nabla f_{\boldsymbol{\theta}_t}(\boldsymbol{x})\|^2\right]\right)^{1/2} \left(\mathbb{E}\left[\phi^2(\boldsymbol{x})\right]\right)^{1/2} \\ &\leq L_g.\end{aligned} \tag{5.21}$$

Due to our choice of the step-size, we have $\eta \lesssim \frac{1}{L_g^2}$, which in turn implies $\left|1 - \eta \left\|\nabla \beta_i(\boldsymbol{\theta}_t)\right\|^2\right| \leq 1$. Therefore, we have

$$\left|\beta_i(\boldsymbol{\theta}_{t+1})\right| \leq \left|\beta_i(\boldsymbol{\theta}_t)\right| + 2\eta^2 L_H L_g^2 L_f^2. \tag{5.22}$$

Now, we are ready to prove the theorem. We divide it into two cases.

Case 1: $\gamma = \frac{1}{2}$. In this case, since we set the step-size $\eta \lesssim \frac{\alpha}{\sqrt{d} C^2 L_H L_g^2 L_f^2} \beta_k^\star \log^{-1}\left(\frac{d\beta_k^\star}{C_1 \alpha}\right)$, we can simplify the dynamics of both signal and residual terms in Equation 5.19 and Equation 5.20 as follows

$$\begin{aligned}\beta_i(\boldsymbol{\theta}_{t+1}) &\geq \left(1 + 0.5 C^2 \eta \left(\beta_i^\star - \beta_i(\boldsymbol{\theta}_t)\right)\right) \beta_i(\boldsymbol{\theta}_t) \quad && \forall i \in \mathcal{S}, \\ \left|\beta_i(\boldsymbol{\theta}_{t+1})\right| &\leq \left|\beta_i(\boldsymbol{\theta}_t)\right| + 2\eta^2 L_H L_g^2 L_f^2 && \forall i \in \mathcal{E}.\end{aligned} \tag{5.23}$$

We first analyze the dynamic of signal $\beta_i(\boldsymbol{\theta}_t)$ for $i \in \mathcal{S}$. To this goal, we further divide this case into two phases. In the first phase, we assume $C_1 \alpha \leq \beta_i(\boldsymbol{\theta}_t) \leq \frac{1}{2}\beta_i^\star$. Under this assumption, we can simplify the dynamic of $\beta_i(\boldsymbol{\theta}_t)$ as

$$\beta_i(\boldsymbol{\theta}_{t+1}) \geq \left(1 + 0.25 C^2 \eta \beta_i^\star\right) \beta_i(\boldsymbol{\theta}_t). \tag{5.24}$$

Therefore, within $T_1 = \mathcal{O}\left(\frac{1}{C^2\eta\beta_i^\star}\log\left(\frac{\beta_i^\star}{C_1\alpha}\right)\right)$ iterations, $\beta_i(\boldsymbol{\theta}_t)$ becomes larger than $\frac{1}{2}\beta_i^\star$. In the second phase, we assume that $\beta_i(\boldsymbol{\theta}_t) \geq \beta_i^\star/2$ and define $y_t = \beta_i^\star - \beta_i(\boldsymbol{\theta}_t)$. One can write

$$\begin{aligned} y_{t+1} &\leq \left(1 - 0.5C^2\eta\beta_i(\boldsymbol{\theta}_t)\right) y_t \\ &\leq (1 - 0.25C^2\eta\beta_i^\star) y_t. \end{aligned} \tag{5.25}$$

Hence, with additional $T_2 = \mathcal{O}\left(\frac{1}{C^2\eta\beta_i^\star}\log\left(\frac{d\beta_i^\star}{\alpha}\right)\right)$, we have $y_t \leq \frac{\alpha}{\sqrt{d}}$ which implies $\beta_i(\boldsymbol{\theta}_t) \geq \beta_i^\star - \frac{\alpha}{\sqrt{d}}$. Next, we show that there exists a time $t^\star$ such that $\beta_i^\star - \frac{1}{\sqrt{d}}\alpha \leq \beta_i(\boldsymbol{\theta}_{t^\star}) \leq \beta_i^\star + \frac{1}{\sqrt{d}}\alpha$. Without loss of generality, we assume that $t^\star$ is the first time that $\beta_i(\boldsymbol{\theta}_t) \geq \beta_i^\star - \frac{1}{\sqrt{d}}\alpha$. Due to the dynamic of $\beta_i(\boldsymbol{\theta}_t)$, the distance between two adjacent iterations can be upper bounded as

$$\begin{aligned} |\beta_i(\boldsymbol{\theta}_{t+1}) - \beta_i(\boldsymbol{\theta}_t)| &\leq \eta\,|\beta_i^\star - \beta_i(\boldsymbol{\theta}_t)|\,\|\nabla\beta_i(\boldsymbol{\theta}_t)\|^2 + 2\eta^2 L_H L_g^2 L_f^2 \\ &\leq \eta L_g^2\,|\beta_i^\star - \beta_i(\boldsymbol{\theta}_t)| + 2\eta^2 L_H L_g^2 L_f^2. \end{aligned} \tag{5.26}$$

In particular, for $t = t^\star - 1$, we have $\beta_i(\boldsymbol{\theta}_{t^\star-1}) \leq \beta_i^\star - \frac{1}{\sqrt{d}}\alpha$, which in turn implies

$$\begin{aligned} \beta_i(\boldsymbol{\theta}_{t^\star}) &\leq \beta_i(\boldsymbol{\theta}_{t^\star-1}) + \eta L_g^2\,|\beta_i^\star - \beta_i(\boldsymbol{\theta}_{t^\star-1})| + 2\eta^2 L_H L_g^2 L_f^2 \\ &\leq \beta_i^\star + 2\eta^2 L_H L_g^2 L_f^2 \\ &\leq \beta_i^\star + \frac{1}{\sqrt{d}}\alpha. \end{aligned} \tag{5.27}$$

Therefore, for each $i \in \mathcal{S}$, we have $|\beta_i(\boldsymbol{\theta}_t) - \beta_i^\star| \leq \frac{1}{\sqrt{d}}\alpha$ within $T_i = \mathcal{O}\left(\frac{1}{C^2\eta\beta_i^\star}\log\left(\frac{d\beta_i^\star}{C_1\alpha}\right)\right)$ iterations. Meanwhile, we can show that the residual term $|\beta_i\boldsymbol{\theta}_t)|, \forall i \in \mathcal{E}$ remains small for $\max_{i\in\mathcal{S}} T_i = \mathcal{O}\left(\frac{1}{C^2\eta\beta_k^\star}\log\left(\frac{d\beta_k^\star}{C_1\alpha}\right)\right)$ iterations:

$$\begin{aligned} |\beta_i(\boldsymbol{\theta}_t)| &\leq |\beta_i(\boldsymbol{\theta}_0)| + \max_{i\in\mathcal{S}} T_i \cdot 2\eta^2 L_H L_g^2 L_f^2 \\ &= |\beta_i(\boldsymbol{\theta}_0)| + \mathcal{O}\left(\frac{1}{\beta_k^\star}\eta\log\left(\frac{\beta_k^\star}{C_1\alpha}\right) L_H L_g^2 L_f^2\right) \\ &= |\beta_i(\boldsymbol{\theta}_0)| + \mathcal{O}\left(\frac{1}{\sqrt{d}}\alpha\right). \end{aligned} \tag{5.28}$$

Therefore, we have that within $T = \mathcal{O}\left(\frac{1}{\eta\beta_k^\star}\log\left(\frac{d\beta_k^\star}{C_1\alpha}\right)\right)$ iterations:

$$\|f_{\boldsymbol{\theta}_T} - f_{\boldsymbol{\theta}^\star}\|_{L^2(\mathcal{D})}^2 = \sum_{i\in\mathcal{S}} \left(\beta_i^\star - \beta_i(\boldsymbol{\theta}_T)\right)^2 + \sum_{i\in\mathcal{E}} \beta_i^2(\boldsymbol{\theta}_T) \lesssim \alpha^2. \tag{5.29}$$

Case 2: $\frac{1}{2} < \gamma \leq 1$**.** In this case, we have the following bounds for the signal and residual terms

$$\begin{aligned} \beta_i(\boldsymbol{\theta}_{t+1}) &\geq \left(1 + C^2\eta\left(\beta_i^\star - \beta_i(\boldsymbol{\theta}_t)\right)\beta_i^{2\gamma-1}(\boldsymbol{\theta}_t)\right)\beta_i(\boldsymbol{\theta}_t) - 2\eta^2 L_H L_g^2 L_f^2 \quad \forall i \in \mathcal{S}, \\ |\beta_i(\boldsymbol{\theta}_{t+1})| &\leq |\beta_i(\boldsymbol{\theta}_t)| + 2\eta^2 L_H L_g^2 L_f^2 \qquad\qquad\qquad\qquad \forall i \in \mathcal{E}. \end{aligned} \tag{5.30}$$

We first analyze the dynamic of the signal term $\beta_i(\boldsymbol{\theta}_t)$ for $i \in \mathcal{S}$. We will show that $|\beta_i(\boldsymbol{\theta}_t) - \beta_i^\star| \leq \frac{\alpha}{\sqrt{k}}$ within $T = \mathcal{O}\left(\frac{1}{C^2\eta\beta_i^\star\alpha^{2\gamma-1}}\right)$ iterations. Due to $\eta \lesssim \frac{\alpha^{2\gamma}}{\sqrt{d}C^2 L_H L_g^2 L_f^2}\beta_k^{\star 2\gamma}$, we can further simplify the dynamic of $\beta_i(\boldsymbol{\theta}_t)$ as

$$\beta_i(\boldsymbol{\theta}_{t+1}) \geq \left(1 + 0.5C^2\eta\left(\beta_i^\star - \beta_i(\boldsymbol{\theta}_t)\right)\beta_i^{2\gamma-1}(\boldsymbol{\theta}_t)\right)\beta_i(\boldsymbol{\theta}_t). \tag{5.31}$$

Next, we divide our analysis into two phases. In the first phase, we have $\beta_i(\boldsymbol{\theta}_t) \leq \frac{1}{2}\beta_i^\star$. We denote the number of iterations for this phase as $T_{i,1}$. We further divide this period into $\lceil \log\left(\beta_i^\star/2\alpha\right)\rceil$ substages. In each Substage k, we have $C_1 2^{k-1}\alpha \leq \beta_i(\boldsymbol{\theta}_t) \leq C_1 2^k \alpha$. Let t_k be the number of iterations in Substage k. We first provide an upper bound for t_k. To this goal, note that at this substage

$$\begin{aligned} \beta_i(\boldsymbol{\theta}_{t+1}) &\geq \left(1 + 0.5C^2\eta\left(\beta_i^\star - \beta_i(\boldsymbol{\theta}_t)\right)\beta_i^{2\gamma-1}(\boldsymbol{\theta}_t)\right)\beta_i(\boldsymbol{\theta}_t) \\ &\geq \left(1 + 0.25C^2\eta\beta_i^\star\beta_i^{2\gamma-1}(\boldsymbol{\theta}_t)\right)\beta_i(\boldsymbol{\theta}_t). \end{aligned} \tag{5.32}$$

Hence, we have

$$t_k \leq \frac{\log(2)}{\log\left(1 + 0.25C^2\eta\beta_i^\star(C_1 2^{k-1}\alpha)^{2\gamma-1}\right)}. \tag{5.33}$$

Summing over t_k, we obtain an upper bound for $T_{i,1}$

$$T_{i,1} = \sum_{i=1}^{\lceil \log\left(\beta_1^\star/2\alpha\right)\rceil} t_k \lesssim \sum_{k=1}^{\infty} \frac{1}{C^2\eta\beta_i^\star(C_1 2^{k-1}\alpha)^{2\gamma-1}} \lesssim \frac{1}{C^2\eta\beta_i^\star(C_1\alpha)^{2\gamma-1}}. \tag{5.34}$$

Via a similar argument, we can show that in the second phase, we have $|\beta_i(\boldsymbol{\theta}_t) - \beta_i^\star| \lesssim \frac{1}{\sqrt{d}}\alpha$ within additional $T_{i,2} = \mathcal{O}\left(\frac{1}{C^2\eta\beta_i^\star(C_1\alpha)^{2\gamma-1}}\right)$ iterations. Therefore, for each $i \in \mathcal{S}$, we conclude that $|\beta_i(\boldsymbol{\theta}_t) - \beta_i^\star| \leq \frac{\alpha}{\sqrt{d}}$ within $T_i = T_{i,1} + T_{i,2} = \mathcal{O}\left(\frac{1}{C^2\eta\beta_i^\star(C_1\alpha)^{2\gamma-1}}\right)$ iterations. Meanwhile, for the residual term $\beta_i(\boldsymbol{\theta}_t), i \in \mathcal{E}$, we have

$$\begin{aligned} |\beta_i(\boldsymbol{\theta}_t)| &\leq |\beta_i(\boldsymbol{\theta}_0)| + \max_{i\in\mathcal{S}} T_i \cdot 2\eta^2 L_H L_g^2 L_f^2 \\ &= |\beta_i(\boldsymbol{\theta}_0)| + \mathcal{O}\left(\frac{1}{\sqrt{d}}\alpha\right). \end{aligned} \tag{5.35}$$

Therefore,

$$\|f_{\boldsymbol{\theta}_T} - f_{\boldsymbol{\theta}^\star}\|_{L^2(\mathcal{D})}^2 = \sum_{i\in\mathcal{S}} (\beta_i^\star - \beta_i(\boldsymbol{\theta}_T))^2 + \sum_{i\in\mathcal{E}} \beta_i^2(\boldsymbol{\theta}_T) \lesssim \alpha^2, \tag{5.36}$$

within $T = \mathcal{O}\left(\frac{1}{C^2\eta\beta_k^\star(C_1\alpha)^{2\gamma-1}}\right)$ iterations. This completes the proof. □

5.E Proofs for Kernel Regression

5.E.1 Proof of Proposition 20

Note that $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta(\boldsymbol{\theta}_t - \boldsymbol{\theta}^\star)$, and $\beta_i(\boldsymbol{\theta}) = \theta_i$. Hence, for every $1 \le i \le k$, we have

$$\beta_i(\boldsymbol{\theta}_{t+1}) = \beta_i(\boldsymbol{\theta}_t) + \eta(\theta_i^\star - \beta_i(\boldsymbol{\theta}_t)). \tag{5.37}$$

This in turn implies

$$\beta_i^\star - \beta_i(\boldsymbol{\theta}_{t+1}) = (1-\eta)\left(\beta_i^\star - \beta_i(\boldsymbol{\theta}_t)\right) \implies \beta_i^\star - \beta_i(\boldsymbol{\theta}_t) = (1-\eta)^t\left(\beta_i^\star - \beta_i(\boldsymbol{\theta}_0)\right). \tag{5.38}$$

For $i > k$, we have

$$\beta_i(\boldsymbol{\theta}_{t+1}) = (1-\eta)\beta_i(\boldsymbol{\theta}_t) \implies \beta_i(\boldsymbol{\theta}_t) = (1-\eta)^t\beta_i(\boldsymbol{\theta}_0), \tag{5.39}$$

which completes the proof. □

5.E.2 Proof of Theorem 24

Due to our choice of initial point $\boldsymbol{\theta}_0 = \alpha\mathbf{1}, \alpha \lesssim |\beta_k^\star|$, we have $|\beta_i^\star - \beta_i(\boldsymbol{\theta}_0)| \le 2|\beta_i^\star|$. Hence, by Proposition 20, we have

$$|\beta_i^\star - \beta_i(\boldsymbol{\theta}_t)| = (1-\eta)^t\,|\beta_i^\star - \beta_i(\boldsymbol{\theta}_0)| \le 2(1-\eta)^t\beta_i^\star, \quad i \in \mathcal{S} \tag{5.40}$$

$$|\beta_i(\boldsymbol{\theta}_t)| \le (1-\eta)^t\alpha, \quad i \in \mathcal{E}. \tag{5.41}$$

Therefore, to prove $|\beta_i^\star - \beta_i(\boldsymbol{\theta}_t)| \le \frac{\alpha}{\sqrt{k}}, i \in \mathcal{S}$, it suffices to have

$$2(1-\eta)^t\beta_i^\star \le \frac{\alpha}{\sqrt{k}} \quad \implies \quad t \gtrsim \frac{1}{\eta}\log\left(\frac{k\theta_i}{\alpha}\right). \tag{5.42}$$

On the other hand, to ensure $|\beta_i(\boldsymbol{\theta}_t)| \leq \frac{\alpha}{\sqrt{d}}, i \in \mathcal{E}$, it suffices to have

$$(1-\eta)^t \alpha \leq \frac{\alpha}{\sqrt{d}} \quad \Longrightarrow \quad t \gtrsim \frac{1}{\eta} \log(d). \tag{5.43}$$

Recall that $\alpha \lesssim \frac{k|\theta_k|}{d}$ and $|\theta_1| \geq |\theta_2| \geq \cdots \geq |\theta_k| > 0$. Therefore, within $T = \mathcal{O}\left(\frac{1}{\eta} \log\left(\frac{k|\theta_1|}{\alpha}\right)\right)$ iterations, we have

$$\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^\star\| \leq \sqrt{\sum_{i=1}^{k} \frac{\alpha^2}{k} + \sum_{i=k+1}^{d} \frac{\alpha^2}{d}} \leq \sqrt{2}\alpha, \tag{5.44}$$

which completes the proof. □

5.F Proofs for Symmetric Matrix Factorization

5.F.1 Initialization

We start by proving that both lower bound on signals at $\boldsymbol{\theta}_0$ and upper bound on energy at $\boldsymbol{\theta}_0$ are satisfied with high probability. Recall that each element of $\boldsymbol{U}_0$ is drawn from $\mathcal{N}(0, \alpha^2)$. The following proposition characterizes the upper and lower bounds for different coefficients $\beta_{ij}(\boldsymbol{U}_0)$.

Proposition 23 (Initialization)**.** *With probability at least* $1 - e^{-\Omega(r')}$*, we have*

$$\beta_{ii}(\boldsymbol{U}_0) = \left\langle \boldsymbol{z}_i \boldsymbol{z}_i^\top, \boldsymbol{U}_0 \boldsymbol{U}_0^\top \right\rangle \geq \frac{1}{4} r' \alpha^2, \quad \textit{for all } 1 \leq i \leq r, \tag{5.45}$$

and

$$|\beta_{ij}(\boldsymbol{U}_0)| = \left|\left\langle \boldsymbol{z}_i \boldsymbol{z}_j^\top, \boldsymbol{U}_0 \boldsymbol{U}_0^\top \right\rangle\right| \leq 4\log(d) r' \alpha^2, \quad \textit{for all } i \neq j \textit{ or } i, j > r. \tag{5.46}$$

Proof. First note that $\boldsymbol{U}_0^\top \boldsymbol{z}_i \sim \mathcal{N}(0, \alpha^2 I_{r' \times r'})$. Hence, a standard concentration bound on Gaussian random vectors implies

$$\mathbb{P}\left(\left|\left\|\boldsymbol{U}_0^\top \boldsymbol{z}_i\right\| - \alpha\sqrt{r'}\right| \geq \alpha\delta\right) \leq 2\exp\{-c\delta^2\}. \tag{5.47}$$

Via a union bound, we have that with probability of at least $1 - e^{-Cr'}$:

$$\left\|\boldsymbol{U}_0^\top \boldsymbol{z}_i\right\| \leq 2\alpha\sqrt{r'\log(d)}, \quad 1 \leq i \leq d, \qquad \left\|\boldsymbol{U}_0^\top \boldsymbol{z}_i\right\| \geq 0.5\alpha\sqrt{r'}, \quad \forall 1 \leq i \leq r. \tag{5.48}$$

Given these bounds, one can write

$$\beta_{ii}(\boldsymbol{U}_0) = \left\langle \boldsymbol{z}_i \boldsymbol{z}_i^\top, \boldsymbol{U}_0 \boldsymbol{U}_0^\top \right\rangle = \left\| \boldsymbol{U}_0^\top \boldsymbol{z}_i \right\|^2 \geq \frac{1}{4} r' \alpha^2 \quad \forall 1 \leq i \leq r, \tag{5.49}$$

and

$$|\beta_{ij}(\boldsymbol{U}_0)| = \left| \left\langle \boldsymbol{z}_i \boldsymbol{z}_j^\top, \boldsymbol{U}_0 \boldsymbol{U}_0^\top \right\rangle \right| \leq \left\| \boldsymbol{U}_0^\top \boldsymbol{z}_i \right\| \left\| \boldsymbol{U}_0^\top \boldsymbol{z}_j \right\| \leq 4 \log(d) r' \alpha^2 \quad \forall 1 \leq i, j \leq d, \tag{5.50}$$

which completes the proof. □

5.F.2 One-step Dynamics

In this section, we characterize the one-step dynamics of the basis coefficients. To this goal, we first provide a more precise statement of Proposition 21 along with its proof.

Proposition 24. *For the diagonal element $\beta_{ii}(\boldsymbol{U}_t)$, we have*

$$\begin{aligned} \beta_{ii}(\boldsymbol{U}_{t+1}) = {} & \left(1 + 2\eta \left(\sigma_i - \beta_{ii}(\boldsymbol{U}_t)\right)\right) \beta_{ii}(\boldsymbol{U}_t) - 2\eta \sum_{j \neq i} \beta_{ij}^2(\boldsymbol{U}_t) \\ & + \eta^2 \left(\sum_{j,k} \beta_{ij}(\boldsymbol{U}_t) \beta_{ik}(\boldsymbol{U}_t) \beta_{jk}(\boldsymbol{U}_t) - 2\sigma_i \sum_j \beta_{ij}^2(\boldsymbol{U}_t) + \sigma_i^2 \beta_{ii}(\boldsymbol{U}_t) \right). \end{aligned} \tag{5.51}$$

where $\sigma_i = 0$ for $r < i \leq d$. Moreover, for every $i \neq j$, we have

$$\begin{aligned} \beta_{ij}(\boldsymbol{U}_{t+1}) = {} & \left(1 + \eta \left(\sigma_i + \sigma_j - 2\beta_{ii}(\boldsymbol{U}_t) - 2\beta_{jj}(\boldsymbol{U}_t)\right)\right) \beta_{ij}(\boldsymbol{U}_t) + 2\eta \sum_{k \neq i,j} \beta_{ik}(\boldsymbol{U}_t) \beta_{jk}(\boldsymbol{U}_t) \\ & + \eta^2 \left(\sum_{k,l} \beta_{ik}(\boldsymbol{U}_t) \beta_{kl}(\boldsymbol{U}_t) \beta_{lj}(\boldsymbol{U}_t) - (\sigma_i + \sigma_j) \sum_k \beta_{ik}(\boldsymbol{U}_t) \beta_{kj}(\boldsymbol{U}_t) + \sigma_i \sigma_j \beta_{ij}(\boldsymbol{U}_t) \right). \end{aligned} \tag{5.52}$$

Proof. The iterations of GD on SMF take the form

$$\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta (\boldsymbol{U}_t \boldsymbol{U}_t^\top - \boldsymbol{M}^\star) \boldsymbol{U}_t. \tag{5.53}$$

This leads to

$$\begin{aligned} \boldsymbol{U}_{t+1} \boldsymbol{U}_{t+1}^\top = {} & \boldsymbol{U}_t \boldsymbol{U}_t^\top - \eta (\boldsymbol{U}_t \boldsymbol{U}_t^\top - \boldsymbol{M}^\star) \boldsymbol{U}_t \boldsymbol{U}_t^\top - \eta \boldsymbol{U}_t \boldsymbol{U}_t^\top (\boldsymbol{U}_t \boldsymbol{U}_t^\top - \boldsymbol{M}^\star) \\ & + \eta^2 (\boldsymbol{U}_t \boldsymbol{U}_t^\top - M^\star) \boldsymbol{U}_t \boldsymbol{U}_t^\top (\boldsymbol{U}_t \boldsymbol{U}_t^\top - \boldsymbol{M}^\star). \end{aligned} \tag{5.54}$$

Recall that $\beta_{ij}(\boldsymbol{U}_t) = \langle \boldsymbol{U}_t \boldsymbol{U}_t^\top, \boldsymbol{z}_i \boldsymbol{z}_j^\top \rangle$, $\boldsymbol{U}_t \boldsymbol{U}_t^\top = \sum_{i,j} \beta_{ij}(\boldsymbol{U}_t) \boldsymbol{z}_i \boldsymbol{z}_j^\top$, and $\boldsymbol{M}^\star = \sum_{i=1}^r \sigma_i \boldsymbol{z}_i \boldsymbol{z}_i^\top$.

Based on these definitions, one can write

$$\begin{aligned}\beta_{ij}(\boldsymbol{U}_{t+1}) &= \langle \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^\top, \boldsymbol{z}_i\boldsymbol{z}_j^\top\rangle \\ &= \langle \boldsymbol{U}_t\boldsymbol{U}_t^\top, \boldsymbol{z}_i\boldsymbol{z}_j^\top\rangle - \eta\langle(\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star)\boldsymbol{U}_t\boldsymbol{U}_t^\top, \boldsymbol{z}_i\boldsymbol{z}_j^\top\rangle - \eta\langle\boldsymbol{U}_t\boldsymbol{U}_t^\top(\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star), \boldsymbol{z}_i\boldsymbol{z}_j^\top\rangle \\ &\quad + \eta^2\langle(\boldsymbol{U}_t\boldsymbol{U}_t^\top - M^\star)\boldsymbol{U}_t\boldsymbol{U}_t^\top(\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star), \boldsymbol{z}_i\boldsymbol{z}_j^\top\rangle.\end{aligned} \tag{5.55}$$

In light of the above equality and the orthogonality of $\{\boldsymbol{z}_i\}_{i\in[d]}$, the RHS of Equation 5.55 can be written in terms of $\beta_{kl}(\boldsymbol{U}_t), 1 \le k, l \le d$. In particular

$$\begin{aligned}\left\langle(\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star)\boldsymbol{U}_t\boldsymbol{U}_t^\top, \boldsymbol{z}_i\boldsymbol{z}_j^\top\right\rangle &= \left\langle\left(\sum_{i,j}\left(\beta_{ij}(\boldsymbol{U}_t) - \beta_{ij}^\star\right)\boldsymbol{z}_i\boldsymbol{z}_j^\top\right)\sum_{i,j}\beta_{ij}(\boldsymbol{U}_t)\boldsymbol{z}_i\boldsymbol{z}_j^\top, \boldsymbol{z}_i\boldsymbol{z}_j^\top\right\rangle \\ &= \left\langle\sum_{i,j}\sum_k\left(\beta_{ik}(\boldsymbol{U}_t) - \beta_{ik}^\star\right)\beta_{kj}(\boldsymbol{U}_t)\boldsymbol{z}_i\boldsymbol{z}_j^\top, \boldsymbol{z}_i\boldsymbol{z}_j^\top\right\rangle \\ &= \sum_k\left(\beta_{ik}(\boldsymbol{U}_t) - \beta_{ik}^\star\right)\beta_{kj}(\boldsymbol{U}_t) \\ &= \left(\beta_{ii}(\boldsymbol{U}_t) - \sigma_i\right)\beta_{ij}(\boldsymbol{U}_t) + \sum_{k\neq i}\left(\beta_{ik}(\boldsymbol{U}_t) - \beta_{ik}^\star\right)\beta_{kj}(\boldsymbol{U}_t).\end{aligned} \tag{5.56}$$

Other terms in Equation 5.55 can be written in terms of $\beta_{ij}(\boldsymbol{U}_t)$ in an identical fashion. Substituting these derivations back in Equation 5.55, we obtain

$$\begin{aligned}\beta_{ij}(\boldsymbol{U}_{t+1}) &= \left(1 + \eta\left(\sigma_i + \sigma_j\right)\right)\beta_{ij}(\boldsymbol{U}_t) - 2\eta\sum_k\beta_{ik}(\boldsymbol{U}_t)\beta_{kj}(\boldsymbol{U}_t) \\ &\quad + \eta^2\left(\sum_{k,l}\beta_{ik}(\boldsymbol{U}_t)\beta_{kl}(\boldsymbol{U}_t)\beta_{lj}(\boldsymbol{U}_t) - \left(\sigma_i + \sigma_j\right)\sum_k\beta_{ik}(\boldsymbol{U}_t)\beta_{kj}(\boldsymbol{U}_t) + \sigma_i\sigma_j\beta_{ij}(\boldsymbol{U}_t)\right).\end{aligned} \tag{5.57}$$

Note that the above equality holds for any $1 \le i, j \le r'$. In particular, for $i = j$, we further have

$$\begin{aligned}\beta_{ii}(\boldsymbol{U}_{t+1}) &= \left(1 + 2\eta\left(\sigma_i - \beta_{ii}(\boldsymbol{U}_t)\right)\right)\beta_{ii}(\boldsymbol{U}_t) - 2\eta\sum_{j\neq i}\beta_{ij}^2(\boldsymbol{U}_t) \\ &\quad + \eta^2\left(\sum_{j,k}\beta_{ij}(\boldsymbol{U}_t)\beta_{ik}(\boldsymbol{U}_t)\beta_{jk}(\boldsymbol{U}_t) - 2\sigma_i\sum_j\beta_{ij}^2(\boldsymbol{U}_t) + \sigma_i^2\beta_{ii}(\boldsymbol{U}_t)\right),\end{aligned} \tag{5.58}$$

which completes the proof. □

5.F.3 Proofs of Proposition 21 and Theorem 25

To streamline the presentation, we prove Proposition 21 and Theorem 25 simultaneously. The main idea behind our proof technique is to divide the solution trajectory into r substages: in Substage i, the basis coefficient $\beta_{ii}(\boldsymbol{U}_t)$ converges linearly to σ_i while all the remaining coefficients remain almost unchanged. More precisely, suppose that Substage i lasts from iteration $t_{i,s}$ to $t_{i,e}$. We will show that $\beta_{ii}(\boldsymbol{U}_{t_{i,e}}) \approx \sigma_i$ and $\beta_{ij}(\boldsymbol{U}_{t_{i,e}}) \approx \beta_{ij}(\boldsymbol{U}_{t_{i,s}})$. Recall that $\gamma = \min_{1\leq i\leq r} \sigma_i - \sigma_{i+1}$ is the eigengap of the true model, which we assume is strictly positive.

Substage 1. In the first stage, we show that $\beta_{11}(\boldsymbol{U}_t)$ approaches σ_1 and $|\beta_{ij}(\boldsymbol{U}_t)|, i, j \geq 2$ remains in the order of poly(α) within $T_1 = \mathcal{O}\left(\frac{1}{\eta\sigma_1}\log\left(\frac{\sigma_1}{\alpha}\right)\right)$ iterations. To formalize this idea, we further divide this substage into two phases. In the first phase (which we refer to as the warm-up phase), we show that $\beta_{11}(\boldsymbol{U}_t)$ will quickly dominate the remaining terms $\beta_{ij}(\boldsymbol{U}_t), \forall (i,j) \neq (1,1)$ within $\mathcal{O}\left(\frac{1}{\eta\gamma}\log\log(d)\right)$ iterations. This is shown in the following lemma.

Lemma 95 (Warm-up phase)**.** *Suppose that the initial point satisfies Equation 5.45 and Equation 5.46. Then, within $\mathcal{O}\left(\frac{1}{\eta\gamma}\log\log(d)\right)$ iterations, we have*

$$|\beta_{ij}(\boldsymbol{U}_t)| \leq \beta_{11}(\boldsymbol{U}_t) \lesssim r'\alpha^2\log(d)^{1+\sigma_1/\gamma}. \tag{5.59}$$

Proof. To show this, we use an inductive argument. Due to our choice of the initial point, we have $|\beta_{ij}(\boldsymbol{U}_0)| \lesssim r'\alpha^2\log(d)^{1+\sigma_1/\gamma}, 1 \leq i, j \leq d$. Now, suppose that at time t, we have $|\beta_{ij}(\boldsymbol{U}_t)| \lesssim r'\alpha^2\log(d)^{1+\sigma_1/\gamma}, 1 \leq i, j \leq d$. Then, by Proposition 24, we have

$$\begin{aligned}
\beta_{11}(\boldsymbol{U}_{t+1}) &= \left(1 + 2\eta\left(\sigma_1 - \beta_{11}(\boldsymbol{U}_t)\right)\right)\beta_{11}(\boldsymbol{U}_t) - 2\eta\sum_{j\neq 1}\beta_{1j}^2(\boldsymbol{U}_t) \\
&\quad + \eta^2\left(\sum_{j,k}\beta_{1j}(\boldsymbol{U}_t)\beta_{1k}(\boldsymbol{U}_t)\beta_{jk}(\boldsymbol{U}_t) - 2\sigma_1\sum_{j}\beta_{1j}^2(\boldsymbol{U}_t) + \sigma_1^2\beta_{11}(\boldsymbol{U}_t)\right) \\
&\geq \left(1 + 2\eta\sigma_1 - \mathcal{O}\left(\eta d r'\alpha^2\log(d)^{1+\sigma_1/\gamma}\right)\right)\beta_{11}(\boldsymbol{U}_t).
\end{aligned} \tag{5.60}$$

Similarly, for the remaining coefficients $\beta_{ij}(\boldsymbol{U}_t), \forall (i,j) \neq (1,1)$, we have

$$|\beta_{ij}(\boldsymbol{U}_{t+1})| \leq \left(1 + \eta\left(\sigma_1 + \sigma_2 + \mathcal{O}\left(dr'\alpha^2\log(d)^{1+\sigma_1/\gamma}\right)\right)\right)|\beta_{ij}(\boldsymbol{U}_t)|. \tag{5.61}$$

Note that the stepsize satisfies $\eta \lesssim \frac{1}{\sigma_1}$, and $\sigma_1 - \sigma_2 \geq \gamma \gtrsim dr'\alpha^2 \log(d)^{1+\sigma_1/\gamma}$. Hence, we have

$$\begin{aligned}
\frac{\beta_{11}(\boldsymbol{U}_{t+1})}{|\beta_{ij}(\boldsymbol{U}_{t+1})|} &\geq \frac{1 + 2\eta\sigma_1 - \mathcal{O}\left(\eta dr'\alpha^2 \log(d)^{1+\sigma_1/\gamma}\right)}{1 + \eta\left(\sigma_1 + \sigma_2 + \mathcal{O}\left(dr'\alpha^2 \log(d)^{1+\sigma_1/\gamma}\right)\right)} \frac{\beta_{11}(\boldsymbol{U}_t)}{|\beta_{ij}(\boldsymbol{U}_t)|} \\
&= \left(1 + \frac{\eta(\sigma_1 - \sigma_2 - \mathcal{O}(dr'\alpha^2 \log(d)^{1+\sigma_1/\gamma}))}{1 + \eta(\sigma_1 + \sigma_2 + \mathcal{O}(dr'\alpha^2 \log(d)^{1+\sigma_1/\gamma}))}\right) \frac{\beta_{11}(\boldsymbol{U}_t)}{|\beta_{ij}(\boldsymbol{U}_t)|} \\
&\geq \left(1 + \frac{\eta\gamma}{1 + 0.5\eta(\sigma_1 + \sigma_2)}\right) \frac{\beta_{11}(\boldsymbol{U}_t)}{|\beta_{ij}(\boldsymbol{U}_t)|} \\
&\geq (1 + 0.5\eta\gamma) \frac{\beta_{11}(\boldsymbol{U}_t)}{|\beta_{ij}(\boldsymbol{U}_t)|}.
\end{aligned} \tag{5.62}$$

This further implies

$$\frac{\beta_{11}(\boldsymbol{U}_t)}{|\beta_{ij}(\boldsymbol{U}_t)|} \geq (1 + 0.5\eta\gamma)^t \frac{\beta_{11}(\boldsymbol{U}_0)}{|\beta_{ij}(\boldsymbol{U}_0)|}. \tag{5.63}$$

On the other hand, Equation 5.49 and Equation 5.50 imply that

$$\frac{\beta_{11}(\boldsymbol{U}_0)}{|\beta_{ij}(\boldsymbol{U}_0)|} \geq \frac{1}{16\log(d)}. \tag{5.64}$$

Hence, within $\mathcal{O}\left(\frac{1}{\eta\gamma}\log\log(d)\right)$ iterations, we have $\beta_{ij}(\boldsymbol{U}_t) \geq |\beta_{11}(\boldsymbol{U}_t)|, \forall (i,j) \neq (1,1)$. Moreover, we have that during this phase,

$$|\beta_{ij}(\boldsymbol{U}_t)| \leq \beta_{11}(\boldsymbol{U}_t) \leq r'\alpha^2 \log(d)\,(1 + 2\eta\sigma_1)^{\mathcal{O}\left(\frac{1}{\eta\gamma}\log\log(d)\right)} \lesssim r'\alpha^2 \log(d)^{1+\sigma_1/\gamma}, \tag{5.65}$$

which completes the proof. □

After the warm-up phase, we show that $\beta_{11}(\boldsymbol{U}_t)$ quickly approaches σ_1 while the remaining coefficients remain small.

Lemma 96 (Fast growth)**.** *After the warm-up phase followed by $\mathcal{O}\left(\frac{1}{\eta\sigma_1}\log\left(\frac{\sigma_1}{\alpha}\right)\right)$ iterations, we have*

$$0.99\sigma_1 \leq \beta_{11}(\boldsymbol{U}_t) \leq \sigma_1. \tag{5.66}$$

Moreover, for $|\beta_{ij}(\boldsymbol{U}_t)|, \forall (i,j) \neq (1,1)$, we have

$$|\beta_{ij}(\boldsymbol{U}_t)| \lesssim \sigma_1 r' \log(d) \alpha^{\frac{2\sigma_1 - \sigma_i - \sigma_j}{2\sigma_1}}. \tag{5.67}$$

Before providing the proof of Lemma 96 we analyze an intermediate logistic map which, as will be shown later, closely resembles the dynamic of $\beta_{ij}(\boldsymbol{U}_t)$:

$$x_{t+1} = (1 + \eta\sigma - \eta x_t)x_t, \quad x_0 = \alpha. \tag{logistic map}$$

The following two lemmas characterize the dynamic of a single logistic map, as well as the dynamic of the ratio between two different logistic maps.

Lemma 97 (Iteration complexity of logistic map)**.** *Suppose that $\alpha \le \epsilon \le 0.1\sigma$. Then, for the logistic map, we have $x_T \ge \sigma - \epsilon$ within $T = \frac{1}{\log(1+\eta\sigma)}\left(\log(4\sigma/\alpha) + \log(4\sigma/\epsilon)\right)$ iterations.*

Lemma 98 (Separation between two logistic maps)**.** *Let σ_1, σ_2 be such that $\sigma_1 - \sigma_2 > 0$, and*

$$x_{t+1} = (1 + \eta\sigma_1 - \eta x_t)x_t, \quad x_0 = \alpha,$$
$$y_{t+1} = (1 + \eta\sigma_2 - \eta y_t)y_t, \quad y_0 = \alpha.$$

Then, within $T = \frac{1}{\log(1+\eta\sigma_1)} \log\left(\frac{16\sigma_1^2}{\epsilon\alpha}\right)$ iterations, we have

$$\sigma_1 - \epsilon \le x_T \le \sigma_1, \quad y_T \le \frac{16\sigma_1^2}{\epsilon}\alpha^{\frac{\sigma_1-\sigma_2}{\sigma_1+\sigma_2}}.$$

The proofs of Lemmas 97 and 98 are deferred to Appendix 5.F.4. We are now ready to provide the proof of Lemma 96.

Proof of Lemma 96. Similar to the proof of Lemma 95, we use an inductive argument. Suppose that t_0 is when the second phase starts. According to Lemma 95, we have $|\beta_{ij}(\boldsymbol{U}_{t_0})| \le \beta_{11}(\boldsymbol{U}_{t_0}) \lesssim r'\alpha^2\log(d)^{1+\sigma_1/\gamma} \lesssim \sigma_1 r'\log(d)\alpha^{\frac{2\sigma_1-\sigma_i-\sigma_j}{2\sigma_1}}$. Therefore, the base case of our induction holds. Next, suppose that at some time t within the second phase, we have $|\beta_{ij}(\boldsymbol{U}_t)| \lesssim \sigma_1\alpha^{\frac{2\sigma_1-\sigma_i-\sigma_j}{2\sigma_1}}, \forall (i,j) \ne (1,1)$. Our goal is to show that $|\beta_{ij}(\boldsymbol{U}_{t+1})| \lesssim \sigma_1\alpha^{\frac{2\sigma_1-\sigma_i-\sigma_j}{2\sigma_1}}$. To this goal, we consider two cases.

Case I: $i \le r$ or $j \le r$ and $(i,j) \ne (1,1)$. We have

$$\begin{aligned}
|\beta_{ij}(\boldsymbol{U}_{t+1})| &\le \left(1 + \eta\left(\sigma_i + \sigma_j - 2\beta_{ii}(\boldsymbol{U}_t) - 2\beta_{jj}(\boldsymbol{U}_t)\right)\right)|\beta_{ij}(\boldsymbol{U}_t)| + 2\eta\sum_{k\ne i,j}|\beta_{ik}(\boldsymbol{U}_t)\beta_{jk}(\boldsymbol{U}_t)| \\
&\quad + \eta^2\left|\left(\sum_{k,l}\beta_{ik}(\boldsymbol{U}_t)\beta_{kl}(\boldsymbol{U}_t)\beta_{lj}(\boldsymbol{U}_t) - (\sigma_i+\sigma_j)\sum_k \beta_{ik}(\boldsymbol{U}_t)\beta_{kj}(\boldsymbol{U}_t) + \sigma_i\sigma_j\beta_{ij}(\boldsymbol{U}_t)\right)\right| \\
&\overset{(a)}{\le} \left(1 + \eta\left(\sigma_i + \sigma_j + \eta\sigma_i\sigma_j + \mathcal{O}\left(d\alpha^{\frac{\gamma}{\sigma_1}}\right)\right)\right)|\beta_{ij}(\boldsymbol{U}_t)|.
\end{aligned}$$

Here in (a) we used the assumption $|\beta_{ij}(\boldsymbol{U}_t)| \lesssim \sigma_1\alpha^{\frac{2\sigma_1-\sigma_i-\sigma_j}{2\sigma_1}} \lesssim \sigma_1\alpha^{\frac{\gamma}{2\sigma_1}}, \forall (i,j) \ne (1,1)$. Hence, by Lemma 98, we have

$$|\beta_{ij}(\boldsymbol{U}_{t+1})| \lesssim \sigma_1\alpha^{\frac{2\sigma_1-\sigma_i-\sigma_j}{2\sigma_1}}, \quad \text{where } 1 \le i \le r \text{ or } 1 \le j \le r. \tag{5.68}$$

Case II: $i, j \geq r+1$. For $\beta_{ij}(\boldsymbol{U}_t)$ such that $i, j \geq r+1$, its dynamic is characterized by

$$\begin{aligned}
|\beta_{ij}(\boldsymbol{U}_{t+1})| &\leq \left(1-\eta\left(2\beta_{ii}(\boldsymbol{U}_t)+2\beta_{jj}(\boldsymbol{U}_t)\right)\right)|\beta_{ij}(\boldsymbol{U}_t)| + 2\eta\sum_{k\neq i,j}|\beta_{ik}(\boldsymbol{U}_t)\beta_{jk}(\boldsymbol{U}_t)| \\
&\quad + \eta^2\sum_{k,l}|\beta_{ik}(\boldsymbol{U}_t)\beta_{kl}(\boldsymbol{U}_t)\beta_{lj}(\boldsymbol{U}_t)| \\
&\leq \left(1+\eta\mathcal{O}\left(d\alpha^{\frac{\gamma}{\sigma_1}}\right)\right)|\beta_{ij}(\boldsymbol{U}_t)|.
\end{aligned} \tag{5.69}$$

Hence, for $t \lesssim \frac{1}{\eta\sigma_1}\log\left(\frac{\sigma_1}{\alpha}\right)$, we have $|\beta_{ij}(\boldsymbol{U}_t)| \leq \left(1+\eta\mathcal{O}\left(d\alpha^{\frac{\gamma}{\sigma_1}}\right)\right)^{\mathcal{O}\left(\frac{1}{\eta\sigma_1}\log\left(\frac{\sigma_1}{\alpha}\right)\right)}|\beta_{ij}(\boldsymbol{U}_0)| \lesssim |\beta_{ij}(\boldsymbol{U}_0)|$ since we assume $\alpha \lesssim \left(\frac{\sigma_1}{d}\right)^{\sigma_1/\gamma}$. This completes our inductive proof for $|\beta_{ij}(\boldsymbol{U}_t)| \lesssim \sigma_1\alpha^{\frac{2\sigma_1-\sigma_i-\sigma_j}{2\sigma_1}}, \forall (i,j)\neq(1,1)$ in the second phase. Finally, we turn to $\beta_{11}(\boldsymbol{U}_t)$. One can write

$$\begin{aligned}
\beta_{11}(\boldsymbol{U}_{t+1}) &= \left(1+2\eta\left(\sigma_1-\beta_{11}(\boldsymbol{U}_t)\right)\right)\beta_{11}(\boldsymbol{U}_t) - 2\eta\sum_{j\neq 1}\beta_{1j}^2(\boldsymbol{U}_t) \\
&\quad + \eta^2\left(\sum_{j,k}\beta_{1j}(\boldsymbol{U}_t)\beta_{1k}(\boldsymbol{U}_t)\beta_{jk}(\boldsymbol{U}_t) - 2\sigma_1\sum_j\beta_{1j}^2(\boldsymbol{U}_t) + \sigma_1^2\beta_{11}(\boldsymbol{U}_t)\right) \\
&\overset{(a)}{\geq} \left(1+2\eta\left(\sigma_1+0.5\eta\sigma_1^2-\beta_{11}(\boldsymbol{U}_t)-\mathcal{O}\left(d\alpha^{\frac{\gamma}{2\sigma_1}}\right)\right)\right)\beta_{11}(\boldsymbol{U}_t) \\
&\overset{(b)}{\geq} \left(1+2\eta\left(0.9995\sigma_1+0.5\eta\sigma_1^2-\beta_{11}(\boldsymbol{U}_t)\right)\right)\beta_{11}(\boldsymbol{U}_t).
\end{aligned} \tag{5.70}$$

Here in (a) we used the fact that $|\beta_{ij}(\boldsymbol{U}_t)| \lesssim \sigma_1 r'\log(d)\alpha^{\frac{2\sigma_1-\sigma_i-\sigma_j}{2\sigma_1}} \lesssim \sigma_1\alpha^{\frac{\gamma}{2\sigma_1}}, \quad \forall (i,j)\neq(1,1)$. In (b), we used the assumption that $\alpha \lesssim \frac{\sigma_r}{d}2^{\sigma_1/\gamma}$. The above inequality together with Lemma 97 entails that within $\frac{1}{\log(1+\eta\sigma_1)}\log\left(\frac{1600\sigma_1}{\alpha}\right) = \mathcal{O}\left(\frac{1}{\eta\sigma_1}\log\left(\frac{\sigma_1}{\alpha}\right)\right)$ iterations, we have $\beta_{11}(\boldsymbol{U}_t) \geq 0.99\sigma_1$. This completes the proof of Lemma 96 and marks the end of Substage 1. □

Next, we move on to Substage 2.

Substage 2. In Substage 2, we show that the second component $\beta_{22}(\boldsymbol{U}_t)$ converges to σ_2 within $\mathcal{O}\left(\frac{1}{\eta\sigma_2}\log\left(\frac{\sigma_2}{\alpha}\right)\right)$ iterations while the other coefficients remain small. To this goal, we first study the one-step dynamic of $\beta_{22}(\boldsymbol{U}_t)$:

$$\begin{aligned}
\beta_{22}(\boldsymbol{U}_{t+1}) &= \left(1+2\eta\left(\sigma_2-\beta_{22}(\boldsymbol{U}_t)\right)\right)\beta_{22}(\boldsymbol{U}_t) - 2\eta\sum_{j\neq 2}\beta_{2j}^2(\boldsymbol{U}_t) \\
&\quad + \eta^2\left(\sum_{j,k}\beta_{2j}(\boldsymbol{U}_t)\beta_{2k}(\boldsymbol{U}_t)\beta_{jk}(\boldsymbol{U}_t) - 2\sigma_2\sum_j\beta_{2j}^2(\boldsymbol{U}_t) + \sigma_2^2\beta_{22}(\boldsymbol{U}_t)\right).
\end{aligned} \tag{5.71}$$

Different from the dynamic of $\beta_{11}(\boldsymbol{U}_t)$, not all the coefficients $\beta_{ij}(\boldsymbol{U}_t)$ with $i = 2$ or $j = 2$ are smaller than $\beta_{22}(\boldsymbol{U}_t)$ at the beginning of Substage 2. In particular, the basis coefficient $|\beta_{12}(\boldsymbol{U}_t)|$ may be much larger than $\beta_{22}(\boldsymbol{U}_t)$ at the beginning of Substage 2. To see this, note that, according to Equation 5.68, we have $\beta_{12}(\boldsymbol{U}_t) \lesssim \alpha^{\frac{\sigma_1-\sigma_2}{2\sigma_1}}$ and $\beta_{22}(\boldsymbol{U}_t) \lesssim \alpha^{\frac{2(\sigma_1-\sigma_2)}{2\sigma_1}}$. Hence, it may be possible to have $|\beta_{12}(\boldsymbol{U}_t)| \asymp \sqrt{\beta_{22}(\boldsymbol{U}_t)} \gg \beta_{22}(\boldsymbol{U}_t)$. Therefore, the term $2\eta\beta_{12}^2(\boldsymbol{U}_t)$ in Equation 5.76 must be handled with extra care. Note that if we can show $\sigma_2\beta_{22}(\boldsymbol{U}_t) \gg \beta_{12}^2(\boldsymbol{U}_t)$, then $2\eta\beta_{12}^2(\boldsymbol{U}_t)$ can be combined with the first term in the RHS of Equation 5.76 and the argument made in Substage 1 can be repeated to complete the proof of Substage 2. However, our provided bound in Equation 5.68 can only imply $\beta_{12}(\boldsymbol{U}_t)^2 \asymp \beta_{22}(\boldsymbol{U}_t)$. Therefore, we need to provide a tighter analysis to show that $\beta_{12}^2(\boldsymbol{U}_t) \ll \sigma_2\beta_{22}(\boldsymbol{U}_t)$ along the trajectory. Upon controlling $\beta_{12}^2(\boldsymbol{U}_t)$, we can then show the convergence of $\beta_{22}(\boldsymbol{U}_t)$ similar to our analysis for $\beta_{11}(\boldsymbol{U}_t)$ in Substage 1.

To control the behavior of $\beta_{12}^2(\boldsymbol{U}_t)$, we study the ratio $\omega(t) := \frac{\beta_{12}^2(\boldsymbol{U}_t)}{\beta_{22}(\boldsymbol{U}_t)}$. We will show that $\omega(t) \ll \sigma_2$ along the trajectory. To this goal, we will show that $\omega(t)$ can only increase for $\mathcal{O}(\frac{1}{\eta\sigma_1}\log(\frac{\sigma_1}{\alpha}))$ iterations. Therefore, its maximum along the solution trajectory happens at $T = \mathcal{O}(\frac{1}{\eta\sigma_1}\log(\frac{\sigma_1}{\alpha}))$. Therefore, by bounding the maximum, we can show that $\omega(t)$ remains small throughout the solution trajectory.

First, at the initial point, we have $\omega(0) \le 64\log^2(d)r'\alpha^2$, which satisfies our claim. We next provide an upper bound for $\omega(t+1)$ based on $\omega(t)$. Note that

$$\begin{aligned}
\omega(t+1) &\le \frac{(1+\eta(\sigma_1+\sigma_2-2\beta_{11}(\boldsymbol{U}_t)+\mathrm{poly}(\alpha)))^2}{(1+\eta(\sigma_2-\mathrm{poly}(\alpha)))^2}\cdot\frac{\beta_{12}^2(\boldsymbol{U}_t)}{\beta_{22}(\boldsymbol{U}_t)} \\
&= \left(1+\frac{\eta(\sigma_1-2\beta_{11}(\boldsymbol{U}_t)+\mathrm{poly}(\alpha))}{1+\eta(\sigma_2-\mathrm{poly}(\alpha))}\right)^2\omega(t) \\
&\le \left(1+\eta\left(\sigma_1-2\beta_{11}(\boldsymbol{U}_t)\right)-0.5\eta^2\sigma_1\sigma_2\right)^2\omega(t) \\
&\le \left((1+\eta\sigma_1)^2-2\eta\beta_{11}(\boldsymbol{U}_t)-0.4\eta^2\sigma_1\sigma_2\right)\omega(t).
\end{aligned} \tag{5.72}$$

Due to the first inequality, $\omega(t)$ can be increasing only until $\beta_{11}(\boldsymbol{U}_t) \ge \frac{\sigma_1}{2} \pm \mathrm{poly}(\alpha)$. On the other hand, due to the dynamic of β_{11} in Substage 1, we can show that $\beta_{11}(\boldsymbol{U}_t) \ge \frac{\sigma_1}{2} \pm \mathrm{poly}(\alpha)$ in at most $T = \mathcal{O}\left(\frac{1}{\eta\sigma_1}\log\left(\frac{\sigma_1}{\alpha}\right)\right)$ iterations. Therefore, $\omega(t)$ takes its maximum at $T = \mathcal{O}\left(\frac{1}{\eta\sigma_1}\log\left(\frac{\sigma_1}{\alpha}\right)\right)$. On the other hand, we know that $\beta_{11}(\boldsymbol{U}_t)$ satisfies

$$\beta_{11}(\boldsymbol{U}_{t+1}) = \left((1+\eta\sigma_1)^2-2\eta\beta_{11}(\boldsymbol{U}_t)\pm\eta\,\mathrm{poly}(\alpha)\right)\beta_{11}(\boldsymbol{U}_t). \tag{5.73}$$

Hence, we can bound $\omega(t)$ as

$$\begin{aligned}\omega(T) &\leq \prod_{t=0}^{T-1} \left((1+\eta\sigma_1)^2 - 2\eta\beta_{11}(\boldsymbol{U}_t) - 0.4\eta^2\sigma_1\sigma_2\right)\omega(0) \\ &\leq \prod_{t=0}^{T-1} \frac{(1+\eta\sigma_1)^2 - 2\eta\beta_{11}(\boldsymbol{U}_t) - 0.4\eta^2\sigma_1\sigma_2}{(1+\eta\sigma_1)^2 - 2\eta\beta_{11}(\boldsymbol{U}_t) \pm \eta\,\mathrm{poly}(\alpha)} \cdot \frac{\beta_{11}(\boldsymbol{U}_{t+1})}{\beta_{11}(\boldsymbol{U}_t)}\omega(0) \\ &\overset{(a)}{=} \prod_{t=0}^{T-1}(1-\Omega(\eta^2\sigma_1\sigma_2))\frac{\beta_{11}(\boldsymbol{U}_{t+1})}{\beta_{11}(\boldsymbol{U}_t)}\omega(0) \\ &\leq 256(1-\Omega(\eta^2\sigma_1\sigma_2))^T\beta_{11}(\boldsymbol{U}_T)\log^2(d).\end{aligned} \tag{5.74}$$

Here in (a) we used the fact that $\alpha \lesssim (\eta\sigma_r^2)^{\sigma_1/\gamma}$. Hence, we have

$$\omega(T) \lesssim (1-\Omega(\eta^2\sigma_1\sigma_2))^T\sigma_1\log^2(d) \lesssim \left(\frac{\alpha}{\sigma_1}\right)^{\Omega(\eta\sigma_2)}\sigma_1\log^2(d). \tag{5.75}$$

Due to our assumption $\alpha \lesssim \left(\kappa\log^2(d)\right)^{-\Omega\left(\frac{1}{\eta\sigma_r}\right)}$, we conclude that $\omega(t) \leq 0.01\sigma_2$. Therefore, equation 5.76 can be lower bounded as

$$\begin{aligned}\beta_{22}(\boldsymbol{U}_{t+1}) \geq{}& \left(1+1.99\eta\left(\sigma_2-\beta_{22}(\boldsymbol{U}_t)\right)\right)\beta_{22}(\boldsymbol{U}_t) - 2\eta\sum_{j>2}\beta_{2j}^2(\boldsymbol{U}_t) \\ &+\eta^2\left(\sum_{j,k}\beta_{2j}(\boldsymbol{U}_t)\beta_{2k}(\boldsymbol{U}_t)\beta_{jk}(\boldsymbol{U}_t) - 2\sigma_2\sum_j\beta_{2j}^2(\boldsymbol{U}_t) + \sigma_2^2\beta_{22}(\boldsymbol{U}_t)\right).\end{aligned} \tag{5.76}$$

The rest of the proof is a line by line reconstruction of Substage 1 and hence omitted for brevity.

Substage $3 \leq k \leq r$. Via an identical argument to Substage 2, we can show that for each Substage $3 \leq k \leq r$, we have

$$0.99\sigma_k \leq \beta_{kk}(\boldsymbol{U}_t) \leq \sigma_k. \tag{5.77}$$

within $T_k = \mathcal{O}\left(\frac{1}{\eta\sigma_k}\log\left(\frac{\sigma_1}{\alpha}\right)\right)$ iterations. This completes the proof of the first statement of Proposition 21.

To prove the second statement of Proposition 21 as well as Theorem 25, we next control the residual terms. First, we consider the residual term $\beta_{ij}(\boldsymbol{U}_t)$ where either $i \leq r$ or $j \leq r$. Note that $\beta_{ij}(\boldsymbol{U}_t) = \beta_{ji}(\boldsymbol{U}_t)$ and hence we can assume $i \leq r$ without loss of generality. We will show that $\beta_{ij}(\boldsymbol{U}_t)$ decreases linearly once the corresponding signal $\beta_{ii}(\boldsymbol{U}_t)$ converges to the vicinity of σ_i. To this goal, it suffices to control the largest component $\beta_{\max}(U_t) :=$

$\max_{i\neq j, i\leq r}|\beta_{ij}(U_t)|$. Without loss of generality, we assume that the index (i,j) attains the maximum at time t, i.e., $\beta_{\max}(U_t)=|\beta_{ij}(U_t)|$. One can write

$$\begin{aligned}
&|\beta_{ij}(\boldsymbol{U}_{t+1})| \\
&\leq \left(1+\eta\left(\sigma_i+\sigma_j-2\beta_{ii}(\boldsymbol{U}_t)-2\beta_{jj}(\boldsymbol{U}_t)\right)\right)|\beta_{ij}(\boldsymbol{U}_t)| + 2\eta\sum_{k\neq i,j}|\beta_{ik}(\boldsymbol{U}_t)\beta_{jk}(\boldsymbol{U}_t)| \\
&\quad + \eta^2\left|\left(\sum_{k,l}\beta_{ik}(\boldsymbol{U}_t)\beta_{kl}(\boldsymbol{U}_t)\beta_{lj}(\boldsymbol{U}_t) - (\sigma_i+\sigma_j)\sum_{k}\beta_{ik}(\boldsymbol{U}_t)\beta_{kj}(\boldsymbol{U}_t) + \sigma_i\sigma_j\beta_{ij}(\boldsymbol{U}_t)\right)\right| \qquad (5.78)\\
&\overset{(a)}{\leq} (1-\eta 0.9(\sigma_i+\sigma_j))|\beta_{ij}(\boldsymbol{U}_t)| \\
&\leq (1-\eta 0.9\sigma_r)|\beta_{ij}(\boldsymbol{U}_t)|.
\end{aligned}$$

Here in (a) we used the fact that $0.99\sigma_i \leq \beta_{ii}(\boldsymbol{U}_t)\leq \sigma_i, 1\leq i\leq r$ and the fact that $\beta_{\max}(U_t)=|\beta_{ij}(U_t)|$. Hence, we conclude that $\beta_{\max}(\boldsymbol{U}_{t+1})\leq(1-0.9\eta\sigma_r)\beta_{\max}(\boldsymbol{U}_t)$. Therefore, within additional $\mathcal{O}\left(\frac{1}{\eta\sigma_r}\log\left(\frac{1}{\alpha}\right)\right)$ iterations, we have $|\beta_{ij}(\boldsymbol{U}_t)|\leq r'\log(d)\alpha^2$ for all (i,j) such that $i\leq r, i\neq j$.

The remaining residual terms, i.e., those coefficients $\beta_{ij}(\boldsymbol{U}_t)$ for which $i,j>r$, can be bounded via the same approach in Case II of substage 1. In particular, we can show that $|\beta_{ij}(\boldsymbol{U}_t)|\lesssim|\beta_{ij}(\boldsymbol{U}_0)|\lesssim r'\log(d)\alpha^2, \forall i,j>r$. For brevity, we omit this step. This completes the proof of the second statement of Proposition 21.

Finally, to prove Theorem 25, we show that once $|\beta_{ij}(\boldsymbol{U}_t)|\leq r'\log(d)\alpha^2, \forall(i,j)\neq(k,k), 1\leq k\leq r$, the signals $\beta_{kk}(\boldsymbol{U}_t)$ will further converge to $\sigma_k\pm\mathcal{O}(\alpha^2)$ within $\mathcal{O}\left(\frac{1}{\eta\sigma_k}\log\left(\frac{\sigma_k}{\alpha}\right)\right)$ iterations. To see this, we simplify the dynamic of $\beta_{kk}(\boldsymbol{U}_t)$ as

$$\begin{aligned}
\beta_{kk}(\boldsymbol{U}_{t+1}) &= \left(1+2\eta\left(\sigma_k-\beta_{kk}(\boldsymbol{U}_t)\right)\right)\beta_{kk}(\boldsymbol{U}_t) - 2\eta\sum_{j\neq k}\beta_{kj}^2(\boldsymbol{U}_t) \\
&\quad + \eta^2\left(\sum_{j,l}\beta_{kj}(\boldsymbol{U}_t)\beta_{kl}(\boldsymbol{U}_t)\beta_{jl}(\boldsymbol{U}_t) - 2\sigma_k\sum_{j}\beta_{kj}^2(\boldsymbol{U}_t) + \sigma_k^2\beta_{kk}(\boldsymbol{U}_t)\right) \qquad (5.79)\\
&= \left(1+\eta\left(\beta_{kk}(\boldsymbol{U}_t)-\sigma_k\right)\right)^2\beta_{kk}(\boldsymbol{U}_t) - \mathcal{O}\left(\eta d r'^2\log^2(d)\alpha^4\right),
\end{aligned}$$

which leads to

$$\begin{aligned}
\sigma_k-\beta_{kk}(\boldsymbol{U}_{t+1}) &= \left(1-\eta\beta_{kk}(\boldsymbol{U}_t)\left(2+\eta(\sigma_k-\beta_{kk}(\boldsymbol{U}_t))\right)\right)(\sigma_k-\beta_{kk}(\boldsymbol{U}_t)) + \mathcal{O}\left(\eta d r'^2\log^2(d)\alpha^4\right) \\
&\leq (1-1.98\eta\sigma_k)(\sigma_k-\beta_{kk}(\boldsymbol{U}_t)) + \mathcal{O}\left(\eta d r'^2\log^2(d)\alpha^4\right).
\end{aligned} \qquad (5.80)$$

Hence, within additional $\mathcal{O}\left(\frac{1}{\eta\sigma_k}\log\left(\frac{\sigma_k}{\alpha}\right)\right)$ iterations, we have $|\sigma_k-\beta_{kk}(\boldsymbol{U}_{t+1})| = \mathcal{O}\left(\frac{1}{\sigma_k}dr'^2\log^2(d)\alpha^4\right)$ for every $1\leq k\leq r$.

In conclusion, we have

$$\left\|\boldsymbol{U}_T\boldsymbol{U}_T^\top - \boldsymbol{M}^\star\right\|_{\mathrm{F}}^2 = \sum_{i,j=1}^{d} \left(\beta_{ij}(\boldsymbol{U}_T) - \beta_{ij}^\star\right)^2 \leq r'^2 \log^2(d) d^2 \alpha^4. \tag{5.81}$$

within $\mathcal{O}\left(\frac{1}{\eta\sigma_r}\log\left(\frac{\sigma_r}{\alpha}\right)\right)$ iterations. This completes the proof of Theorem 25. □

5.F.4 Analysis of the Logistic Map

In this section, we provide the proofs of Lemmas 97 and 98.

Upper Bound of Iteration Complexity

Recall the logistic map

$$x_{t+1} = (1 + \eta\sigma - \eta x_t)x_t, \quad x_0 = \alpha. \tag{5.82}$$

Here the initial value satisfies $0 < \alpha \ll \sigma$. Vaskevicius et al. [2019] provide both upper and lower bounds for x_t that follows the above logistic map. However, their bounds are not directly applicable to our setting. Hence, we need to develop a new proof for Lemma 97.

Proof of Lemma 97. We divide the dynamic into two stages: (a) $x_t \leq \frac{1}{2}\sigma$, and (b) $x_t \geq \frac{1}{2}\sigma$.

Stage 1: $x_t \leq \frac{1}{2}\sigma$. We consider $K = \lceil \log\left(\frac{1}{2}\sigma/\alpha\right) \rceil$ substages, where in each Substage k, we have $\alpha e^k \leq x_t \leq \alpha e^{k+1}$. Suppose that t_k is the number of iterations in Substage k. One can write

$$\begin{aligned} x_{t+1} &= (1 + \eta\sigma - \eta x_t)x_t \\ &\geq (1 + \eta\sigma - \eta\alpha e^{k+1})x_t. \end{aligned} \tag{5.83}$$

Hence, it suffices to find the smallest $t = t_{\min}$ such that

$$\alpha e^k (1 + \eta\sigma - \eta\alpha e^{k+1})^t \geq \alpha e^{k+1}. \tag{5.84}$$

Solving this inequality leads to

$$t_{\min} = \frac{1}{\log(1 + \eta\sigma - \eta\alpha e^{k+1})}. \tag{5.85}$$

Based on the above equality, we provide an upper bound for $t_{\min}$:

$$\begin{aligned} t_{\min} &= \frac{1}{\log(1+\eta\sigma) + \log(1-\eta\alpha e^{k+1}/(1+\eta\sigma))} \\ &\overset{(a)}{\leq} \frac{1}{\log(1+\eta\sigma) - \frac{\eta\alpha e^{k+1}}{1+\eta\sigma-\eta\alpha e^{k+1}}} \\ &\leq \frac{1+\frac{\eta\alpha e^{k+1}}{1+\eta\sigma-\eta\alpha e^{k+1}}}{\log(1+\eta\sigma)} \\ &\overset{(b)}{\leq} \frac{1+\eta\alpha e^{k+1}}{\log(1+\eta\sigma)}. \end{aligned} \tag{5.86}$$

Here in (a) we used the fact that $\log(1+x) \leq \frac{x}{1+x}, \forall x > -1$ and in (b) we used the fact that $x_t \leq \frac{1}{2}\sigma$. Hence, we have $t_k \leq \frac{1+\eta\alpha e^{k+1}}{\log(1+\eta\sigma)}$. Therefore, the total iteration complexity of Stage 1 is upper bounded by

$$T_1 = \sum_{k=0}^{K-1} t_k \leq \sum_{k=0}^{K-1} \frac{1+\eta\alpha e^{k+1}}{\log(1+\eta\sigma)} \leq \frac{\lceil \log\left(\frac{1}{2}\sigma/\alpha\right) \rceil + \eta\sigma}{\log(1+\eta\sigma)} \leq \frac{\log(4\sigma/\alpha)}{\log(1+\eta\sigma)}. \tag{5.87}$$

Stage 2: $x_t \geq \frac{1}{2}\sigma$**.** In this stage, we rewrite the equation 5.82 as

$$\sigma - x_{t+1} = (1-\eta x_t)(\sigma - x_t). \tag{5.88}$$

Via a similar trick, we can show that within additional $T_2 = \frac{\log(4\sigma/\epsilon)}{\log(1+\eta\sigma)}$ iterations, we have $x_t \geq \sigma - \epsilon$, which can be achieved in the total number of iterations $T = T_1 + T_2 = \frac{1}{\log(1+\eta\sigma)} (\log(4\sigma/\alpha) + \log(4\sigma/\epsilon))$ iterations. This completes the proof of Lemma 97 □

Separation between Two Independent Signals

In this section, we show that there is a sharp separation between two logistic maps with signals σ_1, σ_2 provided that $\sigma_1 \neq \sigma_2$. In particular, suppose that $\sigma_1 - \sigma_2 \geq \gamma > 0$ and

$$\begin{aligned} x_{t+1} &= (1+\eta\sigma_1 - \eta x_t)x_t, \quad x_0 = \alpha, \\ y_{t+1} &= (1+\eta\sigma_2 - \eta y_t)y_t, \quad y_0 = \alpha. \end{aligned} \tag{5.89}$$

Proof of Lemma 98. By Lemma 97, we have $x_T \geq \sigma_1 - \epsilon$ within $T = \frac{1}{\log(1+\eta\sigma_1)} \log\left(\frac{16\sigma_1^2}{\epsilon\alpha}\right)$ iterations. Therefore, it suffices to show that y_t remains small for $t \leq T$. To this goal, note that

$$y_{t+1} = (1+\eta\sigma_2 - \eta y_t)y_t \leq (1+\eta\sigma_2)y_t \leq (1+\eta\sigma_2)^{t+1}\alpha. \tag{5.90}$$

Hence, we need to bound $\Gamma = (1+\eta\sigma_2)^T$. Taking logarithm of both sides, we have

$$\begin{aligned}\log(\Gamma) &= T\log(1+\eta\sigma_2) \\ &= \log\left(\frac{16\sigma_1^2}{\epsilon\alpha}\right)\frac{\log(1+\eta\sigma_2)}{\log(1+\eta\sigma_1)}.\end{aligned} \tag{5.91}$$

Now, we provide a lower bound for the ratio $\log(1+\eta\sigma_1)/\log(1+\eta\sigma_2)$:

$$\begin{aligned}\frac{\log(1+\eta\sigma_1)}{\log(1+\eta\sigma_2)} &= 1 + \frac{\log(1+\frac{\eta(\sigma_1-\sigma_2)}{1+\eta\sigma_2})}{\log(1+\eta\sigma_2)} \\ &\geq 1 + \frac{\frac{\eta(\sigma_1-\sigma_2)}{1+\eta\sigma_2}/\left(1+\frac{\eta(\sigma_1-\sigma_2)}{1+\eta\sigma_2}\right)}{\eta\sigma_2} \\ &\overset{(a)}{\geq} 1 + \frac{\sigma_1-\sigma_2}{2\sigma_2} \\ &= \frac{\sigma_1+\sigma_2}{2\sigma_2},\end{aligned} \tag{5.92}$$

where (a) follows from the assumption $\eta \leq \frac{1}{4\sigma_1}$. Therefore, we have

$$\Gamma = \exp\left\{\log\left(\frac{16\sigma_1^2}{\epsilon\alpha}\right)\frac{2\sigma_2}{\sigma_1+\sigma_2}\right\} = \left(\frac{16\sigma_1^2}{\epsilon\alpha}\right)^{2\sigma_2/(\sigma_1+\sigma_2)}. \tag{5.93}$$

which implies that

$$y_T \leq \alpha\left(\frac{16\sigma_1^2}{\epsilon\alpha}\right)^{2\sigma_2/(\sigma_1+\sigma_2)} \leq \frac{16\sigma_1^2}{\epsilon}\alpha^{\frac{\sigma_1-\sigma_2}{\sigma_1+\sigma_2}}. \tag{5.94}$$

This completes the proof of Lemma 98. □

5.G Proof for Tensor Decomposition

In this section, we prove our results for the orthonormal symmetric tensor decomposition (OSTD). Different from matrix factorization, we use a special initialization that aligns with the ground truth. In particular, for all $1 \leq i \leq r'$, we assume that $\sin(\boldsymbol{u}_i(0), \boldsymbol{z}_i) \leq \gamma$ for some small γ. We will show that $\boldsymbol{u}_i(t)$ aligns with $\boldsymbol{z}_i$ along the whole optimization trajectory. To this goal, we define $v_{ij}(t) = \langle \boldsymbol{u}_i(t), \boldsymbol{z}_j\rangle$ for every $1 \leq i \leq r'$ and $1 \leq j \leq d$. Recall that Λ is a multi-index with length l. We define $|\Lambda|_k$ as the number of times index k appears as one of the elements of Λ.[7] Evidently, we have $0 \leq |\Lambda|_k \leq l$. Based on these definitions, one can

[7]For instance, assume that $\Lambda = (1,1,2)$. Then, $|\Lambda|_1 = 2$ and $|\Lambda|_3 = 0$.

write

$$\beta_\Lambda(\boldsymbol{U}) = \sum_{i=1}^{r'} \prod_{k=1}^{d} \langle \boldsymbol{u}_i, \boldsymbol{z}_k \rangle^{|\Lambda|_k} = \sum_{i=1}^{r'} \prod_{k=1}^{d} v_{ik}^{|\Lambda|_k}. \tag{5.95}$$

Now, it suffices to study the dynamic of $v_{ij}(t)$. In particular, we will show that $v_{ij}(t)$ remains small except for the top-r diagonal elements $v_{jj}(t), 1 \le j \le r$, which will approach $\sigma_j^{1/l}$. To make this intuition more concrete, we divide the terms $\{v_{ij}(t)\}$ into three parts:

- **signal terms** defined as $v_{jj}(t), 1 \le j \le r$,
- **diagonal residual terms** defined as $v_{jj}(t), r+1 \le j \le d$, and
- **off-diagonal residual terms** defined as $v_{ij}(t), \forall i \ne j$.

Moreover, we define $V(t) = \max_{i \ne j} |v_{ij}(t)|$ as the maximum element of the off-diagonal residual terms at every iteration t. When there is no ambiguity, we will omit the dependence on iteration t. For example, we write $v_{ij} = v_{ij}(t)$ and $V = V(t)$. Similarly, when there is no ambiguity, we write $\beta_\Lambda(t)$ or β_Λ in lieu of $\beta_\Lambda(\boldsymbol{U}(t))$.

Our next lemma characterizes the relationship between β_Λ and v_{ij}.

Lemma 99. *Suppose that* $\max_{j \ge r+1} \left|v_{jj}^l\right| \lesssim \sigma_1^{\frac{l-1}{l}} V$, *and* $V \le \sigma_1^{1/l} d^{-\frac{1}{l-1}}$. *Then,*

- *For* β_{Λ_j} *with* $\Lambda_j = (j, \ldots, j)$, *we have*

$$\left|\beta_{\Lambda_j} - v_{jj}^l\right| \le r' V^l. \tag{5.96}$$

- *For* β_Λ *with at least two different indices in* Λ, *we have*

$$|\beta_\Lambda| \le 2r\sigma_1^{\frac{l-1}{l}} V. \tag{5.97}$$

The proof of this lemma is deferred to Appendix 5.G.1. Lemma 99 reveals that the magnitude of β_Λ can be upper bounded by $\max_{i \ne j} |v_{ij}|$. Next, we control v_{ij} by providing both lower and upper bounds on its dynamics.

Proposition 25 (One-step dynamics for $v_{ij}(t)$)**.** *Suppose that we have* $V(t) \lesssim \sigma_1^{1/l} d^{-\frac{1}{l-1}}$ *and* $v_{ii}(t) \le \sigma_1^{1/l}, \forall 1 \le i \le d$. *Moreover, suppose that the step-size satisfies* $\eta \lesssim \frac{1}{l\sigma_1}$. *Then,*

- *For the signal term* $v_{ii}(t), 1 \le i \le r$, *we have*

$$v_{ii}(t+1) \ge v_{ii}(t) + \eta l \left(\sigma_i - v_{ii}^l(t) - 2d^{l-1} v_{ii}^{l-2}(t) V^2(t)\right) v_{ii}^{l-1}(t) - l d^l \eta \sigma_1^{\frac{l-1}{l}} V^l(t). \tag{5.98}$$

- *For the diagonal residual term $v_{ii}(t), r+1 \le i \le d$, we have*

$$v_{ii}(t+1) \le v_{ii}(t) - \eta l v_{ii}^{2l-1}(t) + 2\eta l d^l \sigma_1^{\frac{l-1}{l}} V^l(t). \tag{5.99}$$

- *For the off-diagonal term $V(t)$, we have*

$$V(t+1) \le V(t) + 3\eta l \sigma_1 V(t)^{l-1}. \tag{5.100}$$

The proof of this proposition is deferred to Appendix 5.G.2. Equipped with the above one-step dynamics, we next provide a bound on the growth rate of v_{ij}.

Proposition 26. *Suppose that the initial point satisfies $\sin(\boldsymbol{u}_i(0), \boldsymbol{z}_i) \le \gamma$ and $\|\boldsymbol{u}_i(0)\| = \alpha^{1/l}$ with $\alpha \lesssim \frac{1}{d^{l^3}}$, $\gamma \lesssim \frac{1}{l\kappa^{\frac{l}{l-2}}}$. Moreover, suppose that the step-size satisfies $\eta \lesssim \frac{1}{l\sigma_1}$. Then, within $t^\star = \frac{8}{\eta l \sigma_r}\alpha^{-\frac{l-2}{l}}$ iterations,*

- *For the signal term $v_{ii}(t), 1 \le i \le r$, we have*

$$\left|v_{ii}^l(t^\star) - \sigma_i\right| \le 8 d^{l-1} \sigma_i^{\frac{l-2}{l}} \alpha^{2/l} \gamma^2. \tag{5.101}$$

- *For the diagonal residual term $v_{ii}(t), r+1 \le i \le d$, we have*

$$|v_{ii}(t^*)| \le 2\alpha^{1/l}. \tag{5.102}$$

- *For the off-diagonal term $V(t)$, we have*

$$V(t^*) \le 2^{1/l} V(0) \le (2\alpha)^{1/l}\gamma. \tag{5.103}$$

The proof of this proposition is deferred to Appendix 5.G.3. With the above proposition, we are ready to prove Theorem 26.

Proof of Thereom 26. We have the following decomposition

$$\|\mathbf{T} - \mathbf{T}^\star\|_{\mathrm{F}}^2 = \sum_{\Lambda} \left(\beta_\Lambda - \beta_\Lambda^\star\right)^2. \tag{5.104}$$

Hence, it suffices to bound each $|\beta_\Lambda - \beta_\Lambda^\star|$. Combining Lemma 99 and Proposition 26, we

have for every $\left|\beta_{\Lambda_j} - \beta^\star_{\Lambda_j}\right|, 1 \le j \le r$

$$\begin{aligned}\left|\beta_{\Lambda_j}(t^\star) - \beta^\star_{\Lambda_j}\right| &\overset{\text{Lemma 99}}{\le} \left|v^l_{jj}(t^\star) - \sigma_j\right| + r' V^l(t^\star) \\ &\overset{\text{Proposition 26}}{\le} 8d^{l-1}\sigma_j^{\frac{l-2}{l}}\alpha^{2/l}\gamma^2 + 2d\alpha\gamma^l \\ &\le 16d^{l-1}\sigma_j^{\frac{l-2}{l}}\alpha^{2/l}\gamma^2,\end{aligned} \tag{5.105}$$

where in the last inequality, we used $\sigma_j \ge 1$, $\alpha \le 1$, and $\gamma \le 1$. For the remaining diagonal elements $\beta_{\Lambda_j}, r+1 \le j \le d$, we have

$$\begin{aligned}\left|\beta_{\Lambda_j}(t^\star)\right| &\overset{\text{Lemma 99}}{\le} \left|v^l_{jj}(t^\star)\right| + r' V^l(t^\star) \\ &\overset{\text{Proposition 26}}{\le} 2^l\alpha + 2d\alpha\gamma^l.\end{aligned} \tag{5.106}$$

For the general β_Λ with at least two different indices in the multi-index Λ, we have

$$\begin{aligned}|\beta_\Lambda(t^\star)| &\overset{\text{Lemma 99}}{\le} 2r\sigma_1^{\frac{l-1}{l}} V(t^\star) \\ &\overset{\text{Proposition 26}}{\le} 4r\sigma_1^{\frac{l-1}{l}}\alpha^{1/l}\gamma.\end{aligned} \tag{5.107}$$

Hence, we conclude

$$\begin{aligned}\|\mathbf{T}(t^\star) - \mathbf{T}^\star\|_{\mathrm{F}}^2 &\le \sum_{i=1}^{r} 16d^{l-1}\sigma_i^{\frac{l-2}{l}}\alpha^{2/l}\gamma^2 + \sum_{i=r+1}^{d}\left(2^l\alpha + 2d\alpha\gamma^l\right) + d^l \cdot 4r\sigma_1^{\frac{l-1}{l}}\alpha^{1/l}\gamma \\ &\le 8rd^l\gamma\sigma_1^{\frac{l-1}{l}}\alpha^{1/l},\end{aligned} \tag{5.108}$$

which completes the proof of the theorem. □

5.G.1 Proof of Lemma 99

Proof. We first analyze β_{Λ_j}. Note that

$$\beta_{\Lambda_j} = \sum_{i=1}^{r'} \langle \boldsymbol{u}_i, \boldsymbol{z}_j\rangle^l = \sum_{i=1}^{r'} v^l_{ij} = v^l_{jj} + \sum_{i\neq j} v^l_{ij}. \tag{5.109}$$

Hence,

$$\left|\beta_{\Lambda_j} - v^l_{jj}\right| = \left|\sum_{i\neq j} v^l_{ij}\right| \le r' V^l, \tag{5.110}$$

where we used the definition of V. For general β_Λ where there are at least two different elements in the multi-index Λ, we have

$$\begin{aligned}
|\beta_\Lambda| &= \left|\sum_{j=1}^{r'}\prod_{k\in\Lambda} v_{jk}^{|\Lambda|_k}\right| \\
&\le \sum_{j=1}^{r'}\left|v_{jj}^{|\Lambda|_j}V^{l-|\Lambda|_j}\right| \\
&= \sum_{j=1}^{r}\left|v_{jj}^{|\Lambda|_j}V^{l-|\Lambda|_j}\right| + \sum_{j=r+1}^{r'}\left|v_{jj}^{|\Lambda|_j}V^{l-|\Lambda|_j}\right| \\
&\le r\sigma_1^{\frac{l-1}{l}}V + (r'-r)\left(\max_{j\ge r+1}\left|v_{jj}^l\right| + V^l\right) \\
&\le 2r\sigma_1^{\frac{l-1}{l}}V,
\end{aligned} \tag{5.111}$$

where in the last inequality, we used the assumption that $V \lesssim \sigma_1^{1/l}d^{-\frac{1}{l-1}}$ and $\max_{j\ge r+1}\left|v_{jj}^l\right| \lesssim \sigma_1^{\frac{l-1}{l}}V$. This completes the proof. □

5.G.2 Proof of Proposition 25

In this section, we provide the proof for Proposition 25. For simplicity and whenever there is no ambiguity, we omit the iteration t and show iteration $t+1$ with superscript '+'. For instance, we write $v_{ij} = v_{ij}(t)$ and $v_{ij}^+ = v_{ij}(t+1)$.

Recall that $v_{ij} = \langle \boldsymbol{u}_i, \boldsymbol{z}_j\rangle$. For simplicity, we denote $\mu = \sigma_1^{1/l}$. Hence, by our assumption, we have $v_{ii} \le \mu, \forall 1 \le i \le r$. We first provide the exact dynamic of v_{ij} in the following lemma.

Lemma 100. *The one-step dynamic of v_{ij} takes the following form*

$$\begin{aligned}
v_{ij}^+ = v_{ij} &+ \eta l(\sigma_j - v_{jj}^l)v_{ij}^{l-1} - \eta l\sum_{k\in[r'],k\ne j} v_{kj}^l v_{ij}^{l-1} \\
&- \eta\sum_{s\in[l-1]} s\sum_{\Lambda:|\Lambda|_j=s}\beta_\Lambda\left(\prod_{k\in\Lambda,k\ne j} v_{ik}^{|\Lambda|_k}\right)v_{ij}^{s-1}.
\end{aligned} \tag{5.112}$$

Proof. Recall that $\mathcal{L} = \frac{1}{2}\sum_\Lambda(\beta_\Lambda - \beta_\Lambda^\star)^2$ and $\beta_\Lambda = \sum_{i=1}^{r'}\prod_{k\in\Lambda}\langle \boldsymbol{u}_i, \boldsymbol{z}_k\rangle^{|\Lambda|_k}$. Moreover, we have

$\beta^\star_{\Lambda_j} = \sigma_j$ for $1 \le j \le r$, and $\beta^\star_\Lambda = 0$ otherwise. We first calculate

$$\begin{aligned}\nabla_{\boldsymbol{u}_i}\beta_\Lambda &= \sum_{s\in\Lambda} |\Lambda|_s \left(\prod_{k\in\Lambda,k\neq s} \langle \boldsymbol{u}_i, \boldsymbol{z}_k\rangle^{|\Lambda|_k}\right) \langle \boldsymbol{u}_i, \boldsymbol{z}_s\rangle^{|\Lambda|_s-1} \boldsymbol{z}_s \\ &= \sum_{s\in\Lambda} |\Lambda|_s \left(\prod_{k\in\Lambda,k\neq s} v_{ik}^{|\Lambda|_k}\right) v_{is}^{|\Lambda|_s-1}\boldsymbol{z}_s.\end{aligned} \tag{5.113}$$

Hence, the partial derivative of $\mathcal{L}(t)$ with respect to $\boldsymbol{u}_i$ is

$$\nabla_{\boldsymbol{u}_i}\mathcal{L} = \sum_{\forall\Lambda}(\beta_\Lambda - \beta^\star_\Lambda)\nabla_{\boldsymbol{u}_i}\beta_\Lambda = \sum_{\Lambda}(\beta_\Lambda - \beta^\star_\Lambda)\sum_{s\in\Lambda}|\Lambda|_s\left(\prod_{k\in\Lambda,k\neq s} v_{ik}^{|\Lambda|_k}\right) v_{is}^{|\Lambda|_s-1}\boldsymbol{z}_s.$$

Note that $\{\boldsymbol{z}_j\}_{j\in[d]}$ are unit orthogonal vectors. Hence, we have

$$\langle \nabla_{\boldsymbol{u}_i}\beta_\Lambda, \boldsymbol{z}_j\rangle = \begin{cases} |\Lambda|_j \left(\prod_{k\in\Lambda,k\neq j} v_{ik}^{|\Lambda|_k}\right) v_{ij}^{|\Lambda|_j-1} & \text{if} \quad j\in\Lambda \\ 0 & \text{if} \quad j\notin\Lambda. \end{cases} \tag{5.114}$$

By the definition of v_{ij}, its update rule can be written in the following way

$$\begin{aligned} v_{ij}^+ &= \left\langle \boldsymbol{u}_i^+, \boldsymbol{z}_j\right\rangle \\ &\overset{(a)}{=} v_{ij} - \eta\left\langle \nabla_{\boldsymbol{u}_i}\mathcal{L}, \boldsymbol{z}_j\right\rangle \\ &= v_{ij} + \eta\sum_{\Lambda}(\beta^\star_\Lambda - \beta_\Lambda)\left\langle \nabla_{\boldsymbol{u}_i}\beta_\Lambda, \boldsymbol{z}_j\right\rangle \\ &\overset{(b)}{=} v_{ij} + \eta\sum_{\Lambda:|\Lambda|_j\ge 1}(\beta^\star_\Lambda - \beta_\Lambda)|\Lambda|_j\left(\prod_{k\in\Lambda,k\neq j} v_{ik}^{|\Lambda|_k}\right) v_{ij}^{|\Lambda|_j-1} \\ &\overset{(c)}{=} v_{ij} + \eta\sum_{s\in[l]}\sum_{\Lambda:|\Lambda|_j=s}|\Lambda|_j(\beta^\star_\Lambda - \beta_\Lambda)\left(\prod_{k\in\Lambda,k\neq j} v_{ik}^{|\Lambda|_k}\right) v_{ij}^{s-1}. \end{aligned} \tag{5.115}$$

Here in (a), we used the update rule for $\boldsymbol{u}_i$. In (b), we applied equation 5.114 to exclude those Λ without j. In (c), we simply rearranged the above equation according to the cardinality

$|\Lambda|_j$. We further isolate the term that only has v_{ij}:

$$\begin{aligned} v_{ij}^+ &\overset{(a)}{=} v_{ij} + \eta l(\sigma_j - \beta_{\Lambda_j}) v_{ij}^{l-1} - \eta \sum_{s\in[l-1]} s \sum_{\Lambda:|\Lambda|_j=s} \beta_\Lambda \left(\prod_{k\in\Lambda, k\neq j} v_{ik}^{|\Lambda|_k} \right) v_{ij}^{s-1} \\ &\overset{(b)}{=} v_{ij} + \eta l(\sigma_j - v_{jj}^l) v_{ij}^{l-1} - \eta l \sum_{k\in\Lambda, k\neq j} v_{kj}^l v_{ij}^{l-1} \\ &\quad - \eta \sum_{s\in[l-1]} s \sum_{\Lambda:|\Lambda|_j=s} \beta_\Lambda \left(\prod_{k\in\Lambda, k\neq j} v_{ik}^{|\Lambda|_k} \right) v_{ij}^{s-1}. \end{aligned} \tag{5.116}$$

Here in (a), we rearranged terms and isolated the term with $|\Lambda|_j = l$. Note that the remaining terms must satisfy $1 \leq |\Lambda|_j \leq l-1$, which indicates that there must be at least 2 different indexes in Λ which in turn implies $\beta_\Lambda^\star = 0$. In (b), we used the definition of β_{Λ_j}. This completes the proof of Lemma 100. □

Equipped with Lemma 100, we are ready to prove Proposition 25.

Proof of Proposition 25. The proof is divided into three parts:

Signal Term: $v_{ii}(t), 1 \leq i \leq r$

We first consider the signal terms $v_{ii}(t), 1 \leq j \leq r$. First, upon setting $i = j$ in Lemma 100, we have

$$\begin{aligned} v_{ii}^+ &= v_{ii} + \eta l \left(\sigma_i - v_{ii}^l - \sum_{k\neq i} v_{ki}^l \right) v_{ii}^{l-1} - \eta \sum_{s=1}^{l-1} s \sum_{\Lambda:|\Lambda|_i=s} \beta_\Lambda \left(\prod_{k\in\Lambda, k\neq i} v_{ik}^{|\Lambda|_k} \right) v_{ii}^{s-1} \\ &\geq v_{ii} + \eta l \left(\sigma_i - v_{ii}^l - mV^l \right) v_{ii}^{l-1} - \eta \sum_{s=1}^{l-1} s \sum_{\Lambda:|\Lambda|_i=s} \beta_\Lambda \left(\prod_{k\in\Lambda, k\neq i} v_{ik}^{|\Lambda|_k} \right) v_{ii}^{s-1}. \end{aligned} \tag{5.117}$$

Now we aim to control $(A) = \beta_\Lambda \left(\prod_{k\in\Lambda, k\neq i} v_{ik}^{|\Lambda|_k}\right) v_{ii}^{s-1}$ for $|\Lambda|_i = s \in \{1, \ldots, l-1\}$. We have

$$\begin{aligned}
(A) &= \sum_{j=1}^{r'} \left(\prod_{h\in\Lambda} v_{jh}^{|\Lambda|_h}\right) \left(\prod_{k\in\Lambda, k\neq i} v_{ik}^{|\Lambda|_k}\right) v_{ii}^{|\Lambda|_i - 1} \\
&\overset{(a)}{\leq} \sum_{j=1}^{r'} v_{jj}^{|\Lambda|_j} V^{l-|\Lambda|_j} V^{l-s} v_{ii}^{s-1} \\
&\overset{(b)}{\leq} v_{ii}^{2s-1} V^{2l-2s} + r' \max_{j\neq i} \left| v_{jj}^{|\Lambda|_j} V^{2l-s-|\Lambda|_j} v_{ii}^{s-1} \right| \\
&\overset{(c)}{\leq} v_{ii}^{2l-3} V^2 + r' \max_{j\neq i} \mu^{|\Lambda|_j + s - 1} V^{2l-s-|\Lambda|_j} \\
&\leq v_{ii}^{2l-3} V^2 + r' \mu^{l-1} V^l.
\end{aligned} \tag{5.118}$$

In (a), we used the fact that $|v_{ij}| \leq V, \forall i \neq j$. In (b), we isolated the term with $j = i$ and bounded the remaining terms with their maximum value. In (c), we used the fact that $V \leq v_{ii} \leq \mu$.

After substituting equation 5.118 into equation 5.117, we have

$$\begin{aligned}
v_{ii}^+ &\geq v_{ii} + \eta l \left(\sigma_i - v_{ii}^l - r' V^l\right) v_{ii}^{l-1} - \eta \sum_{s=1}^{l-1} s \sum_{\Lambda:|\Lambda|_i = s} \left(v_{ii}^{2l-3} V^2 + r' \mu^{l-1} V^l\right) \\
&\geq v_{ii} + \eta l \left(\sigma_i - v_{ii}^l - r' V^l\right) v_{ii}^{l-1} - \eta \left(v_{ii}^{2l-3} V^2 + r' \mu^{l-1} V^l\right) \sum_{s=1}^{l-1} s C_l^s (d-1)^{l-s},
\end{aligned} \tag{5.119}$$

where $C_l^s = \binom{l}{s}$. Note that $\sum_{s=1}^{l-1} s C_l^s (d-1)^{l-s} = l d^{l-1} - l \leq l d^{l-1}$. Therefore, we have

$$\begin{aligned}
v_{ii}(t+1) &\geq v_{ii} + \eta l \left(\sigma_i - v_{ii}^l - r' V^l\right) v_{ii}^{l-1} - l d^{l-1} \eta \left(v_{ii}^{2l-3} V^2 + r' \mu^{l-1} V^l\right) \\
&\geq v_{ii} + \eta l \left(\sigma_i - v_{ii}^l - r' V^l - d^{l-1} v_{ii}^{l-2} V^2\right) v_{ii}^{l-1} - l d^{l-1} \eta r' \mu^{l-1} V^l \\
&\geq v_{ii} + \eta l \left(\sigma_i - v_{ii}^l - 2 d^{l-1} v_{ii}^{l-2} V^2\right) v_{ii}^{l-1} - l d^{l-1} \eta r' \mu^{l-1} V^l \\
&\geq v_{ii} + \eta l \left(\sigma_i - v_{ii}^l - 2 d^{l-1} v_{ii}^{l-2} V^2\right) v_{ii}^{l-1} - l d^l \eta \sigma_1^{\frac{l-1}{l}} V^l,
\end{aligned} \tag{5.120}$$

where in the last inequality, we used the fact that $r' \leq d$.

Diagonal Residual Term: $v_{ii}, r+1 \leq i \leq d$

In this case we consider the terms v_{ii} with $r+1 \leq i \leq d$, which is similar to the case $1 \leq i \leq r$. Without loss of generality, we assume that $v_{ii} \geq 0$. The case $v_{ii} \leq 0$ can be

argued in an identical fashion. By equation 5.117, we have

$$\begin{aligned} v_{ii}^{+} &= v_{ii} - \eta l \left(v_{ii}^{l} - \sum_{k \neq i} v_{ki}^{l} \right) v_{ii}^{l-1} - \eta \sum_{s=1}^{l-1} s \sum_{\Lambda : |\Lambda|_i = s} \beta_\Lambda \left(\prod_{k \in \Lambda, k \neq i} v_{ik}^{|\Lambda|_k} \right) v_{ii}^{s-1} \\ &\leq v_{ii} - \eta l \left(v_{ii}^{l} - r' V^{l} \right) v_{ii}^{l-1} - \eta \sum_{s=1}^{l-1} s \sum_{\Lambda : |\Lambda|_i = s} \beta_\Lambda \left(\prod_{k \in \Lambda, k \neq i} v_{ik}^{|\Lambda|_k} \right) v_{ii}^{s-1}. \end{aligned} \tag{5.121}$$

For $(A) = \beta_\Lambda \left(\prod_{k \in \Lambda, k \neq i} v_{ik}^{|\Lambda|_k} \right) v_{ii}^{s-1}$, we further have

$$\begin{aligned} (A) &= \sum_{j=1}^{r'} \left(\prod_{h \in \Lambda} v_{jh}^{|\Lambda|_h} \right) \left(\prod_{k \in \Lambda, k \neq i} v_{ik}^{|\Lambda|_k} \right) v_{ii}^{s-1} \\ &\geq \left(\prod_{h \in \Lambda} v_{ih}^{|\Lambda|_h} \right) \left(\prod_{k \in \Lambda, k \neq i} v_{ik}^{|\Lambda|_k} \right) v_{ii}^{s-1} - \sum_{j \neq i} \left(\prod_{h \in \Lambda} v_{jh}^{|\Lambda|_h} \right) \left(\prod_{k \in \Lambda, k \neq i} v_{ik}^{|\Lambda|_k} \right) v_{ii}^{s-1} \\ &\geq \underbrace{\left(\prod_{k \in \Lambda, k \neq i} v_{ik}^{2|\Lambda|_k} \right) v_{ii}^{2s-1}}_{\geq 0} - r' \max_{j \neq i} \left| v_{jj}^{|\Lambda|_j} V^{2l-s-|\Lambda|_j} v_{ii}^{s-1} \right| \\ &\geq -r' \mu^{l-1} V^{l}. \end{aligned} \tag{5.122}$$

Therefore, we obtain

$$\begin{aligned} v_{ii}^{+} &\leq v_{ii} - \eta l \left(v_{ii}^{l} - r' V^{l} \right) v_{ii}^{l-1} + \eta l d^{l-1} r' \mu^{l-1} V^{l} \\ &\leq v_{ii} - \eta l v_{ii}^{2l-1} + 2 \eta l d^{l-1} r' \mu^{l-1} V^{l}. \end{aligned} \tag{5.123}$$

Off-diagonal Residual Term: $V(t)$

Finally, we characterize the dynamic of $V(t) = \max_{i \neq j} |v_{ij}(t)|$. To this goal, we first consider the dynamic of each v_{ij} such that $i \neq j$. Without loss of generality, we assume that $v_{ij} \geq 0$. One can write

$$\begin{aligned} v_{ij}^{+} &= v_{ij} + \eta l \left(\sigma_j - \sum_{i \in [r']} v_{ij}^{l} \right) v_{ij}^{l-1} - \eta \sum_{s \in [l-1]} s \sum_{\Lambda : |\Lambda|_j = s} \beta_\Lambda \left(\prod_{k \in \Lambda, k \neq j} v_{ik}^{|\Lambda|_k} \right) v_{ij}^{s-1} \\ &\leq V + \eta l \sigma_1 V^{l-1} + \eta l r' V^{2l-1} - \eta \sum_{s \in [l-1]} s \sum_{\Lambda : |\Lambda|_j = s} \beta_\Lambda \left(\prod_{k \in \Lambda, k \neq j} v_{ik}^{|\Lambda|_k} \right) v_{ij}^{s-1} \\ &\leq V + 2 \eta l \sigma_1 V^{l-1} - \eta \sum_{s \in [l-1]} s \sum_{\Lambda : |\Lambda|_j = s} \beta_\Lambda \left(\prod_{k \in \Lambda, k \neq j} v_{ik}^{|\Lambda|_k} \right) v_{ij}^{s-1}, \end{aligned} \tag{5.124}$$

where in the last inequality we use the assumption $V \lesssim \sigma_1^{1/l} d^{-\frac{1}{l-1}}$. Similar to the previous case, we next bound $(A) = \beta_\Lambda \left(\prod_{k\in\Lambda, k\neq j} v_{ik}^{|\Lambda|_k}\right) v_{ij}^{s-1}$. First note that $|\Lambda|_j = s \in [l-1]$. Hence, we have that

$$\begin{aligned}
(A) &= \sum_{s=1}^{r'} \prod_{k\in\Lambda} v_{sk}^{|\Lambda|_k} \left(\prod_{k\in\Lambda, k\neq j} v_{ik}^{|\Lambda|_k}\right) v_{ij}^{s-1} \\
&= \underbrace{\left(\prod_{k\in\Lambda, k\neq j} v_{ik}^{2|\Lambda|_k}\right) v_{ij}^{2s-1}}_{\geq 0} - r' \max_{h\neq i} \left| \prod_{k\in\Lambda} v_{hk}^{|\Lambda|_k} \left(\prod_{k\in\Lambda, k\neq j} v_{ik}^{|\Lambda|_k}\right) v_{ij}^{h-1} \right| \\
&\geq -r' \max_{h\neq i} \left| \prod_{k\in\Lambda} v_{hk}^{|\Lambda|_k} \left(\prod_{k\in\Lambda, k\neq j} v_{ik}^{|\Lambda|_k}\right) v_{ij}^{h-1} \right| \\
&\geq -r' \mu^{|\Lambda|_s + |\Lambda|_i} V^{2l - |\Lambda|_s - |\Lambda|_i - 1}.
\end{aligned} \tag{5.125}$$

We further note that $|\Lambda|_h + |\Lambda|_i \leq l - s$. Hence, it can be lower bounded by

$$(A) \geq -r' \mu^{l-s} V^{l+s-1} \geq -r' \mu^{l-1} V^l. \tag{5.126}$$

Pluggin our estimation for (A) into equation 5.124, we finally have

$$v_{ij}^+ \leq V + 2\eta l \sigma_1 V^{l-1} + \eta l d^{l-1} r' \mu^{l-1} V^l \leq V + 3\eta l \sigma_1 V^{l-1}. \tag{5.127}$$

The last inequality comes from the assumption that $V \leq \frac{1}{d^l}$. This completes the proof of Proposition 25.

5.G.3 Proof of Proposition 26

In this section, we provide the proof of Proposition 26. We will show that the signal terms $v_{ii}(t), 1 \leq i \leq r$ quickly converge to $\sigma_i^{1/l}$, and the residual terms remain small. To this goal, we first study the dynamic of $V(t)$.

Iteration complexity of the off-diagonal residual term $V(t)$.

To start with the proof, we first study the time required for the off-diagonal term $V(t)$ to go from $V(0)$ to $2^{1/l}V(0)$, i.e., $T_1 = \min_{t\geq 0}\{V(t) \geq 2^{1/l} V(0)\}$. By Proposition 25, we know

$$V(t+1) \leq V(t) + 3\eta\sigma_1 l V^{l-1}(t) = \left(1 + 3\eta l \sigma_1 V^{l-2}(t)\right) V(t). \tag{5.128}$$

Hence, for $0 \leq t \leq T_1$, we have

$$V(t) \leq \prod_{s=0}^{t-1} \left(1 + 3\eta l \sigma_1 V^{l-2}(s)\right) V(0) \leq \left(1 + 6\eta l \sigma_1 V^{l-2}(0)\right)^t V(0).$$

Note that $V(T_1) \geq 2^{1/l} V(0)$. Therefore,

$$\left(1 + 6\eta l \sigma_1 V^{l-2}(0)\right)^{T_1} V(0) \geq 2^{1/l} V(0). \tag{5.129}$$

Solving the above inequality for T_1, we obtain

$$T_1 \geq \frac{\log(2^{1/l})}{\log(1 + 6\eta l \sigma_1 V^{l-2}(0))} \geq \frac{\log(2)}{6\eta l^2 \sigma_1 V^{l-2}(0)}. \tag{5.130}$$

On the other hand, our initial point satisfies $\sin(\boldsymbol{u}_i(0), \boldsymbol{z}_i) \leq \gamma$ and $\|\boldsymbol{u}_i(0)\| = \alpha^{1/l}$. Hence, we have $V(0) \leq \alpha^{1/l}\gamma$. Substituting this into the above equation, we conclude that

$$T_1 \geq \frac{\log(2)}{6\eta l^2 \sigma_1 \gamma^{l-2}} \alpha^{-\frac{l-2}{l}}. \tag{5.131}$$

Note that $T_1 \geq \frac{\log(2)}{6\eta l^2 \sigma_1 \gamma^{l-2}} \alpha^{-\frac{l-2}{l}} \gg t^\star$. Hence, we have $V(t^\star) \leq 2^{1/l} V(0) \leq (2\alpha)^{1/l}\gamma$.

Iteration complexity of the diagonal residual term $v_{ii}(t), r+1 \leq i \leq d$.

Next, we show that $v_{ii}(t), r+1 \leq i \leq d$ will remain small during $0 \leq t \leq t^\star = \frac{8}{\eta l \sigma_r} \alpha^{-\frac{l-2}{l}}$. First by equation 5.123 in the proof of Lemma 100, we have for every $0 \leq t \leq t^\star$

$$v_{ii}(t+1) \leq v_{ii}(t) + 2\eta l d^l \sigma_1^{l-1} V^l(t). \tag{5.132}$$

Note that $V(t) \leq 2^{1/l} V(0), \forall t \leq t^\star$, which leads to

$$\begin{aligned} v_{ii}(t) &\leq v_{ii}(0) + 4\eta l d^l \sigma_1^{l-1} \gamma^l \alpha t \\ &\leq v_{ii}(0) + 4\eta l d^l \sigma_1^{l-1} \gamma^l \alpha t^\star \\ &\leq \alpha^{1/l} + \mathcal{O}\left(d^l \kappa \alpha^{2/l} \gamma^l\right) \\ &\leq 2\alpha \end{aligned} \tag{5.133}$$

for $0 \leq t \leq t^\star$. Here we used the assumption that $\alpha \lesssim \frac{1}{d^{l^3}}$ and $\gamma \lesssim \frac{1}{\kappa l}^{\frac{1}{l-2}}$.

Iteration complexity of the signal term $v_{ii}(t), 1 \le i \le r$.

As the last piece of the proof, we show that by iteration $t^\star = \frac{8}{\eta l \sigma_r} \alpha^{-\frac{l-2}{l}}$, the signal $v_{ii}, 1 \le i \le r$ will converge to the eigenvalue $\sigma_i^{1/l}$. First, recall that

$$\begin{aligned} v_{ii}(t+1) &\ge v_{ii}(t) + \eta l \left(\sigma_i - v_{ii}^l(t) - 2d^{l-1} v_{ii}^{l-2}(t) V^2(t)\right) v_{ii}^{l-1}(t) - 2ld^l \eta \sigma_1^{l-1} V^l(t) \\ &\ge v_{ii}(t) + \eta l \left(\sigma_i - v_{ii}^l(t) - 4d^{l-1} v_{ii}^{l-2}(t) \alpha^{2/l} \gamma^2\right) v_{ii}^{l-1}(t) - 4ld^l \eta \sigma_1^{l-1} \gamma^l \alpha, \end{aligned} \tag{5.134}$$

where in the last inequality we used the fact that $V(t) \le 2^{1/l} V(0)$ for $0 \le t \le t^\star$. In light of the above inequality, we characterize the convergence of v_{ii} using a similar method as in [Ma and Fattahi, 2022a]. In particular, we divide our analysis into two phases.

Phase 1. In the first phase, we have $v_{ii} \le (0.5\sigma_i)^{1/l}$. First, since $v_{ii}(0) \ge \alpha^{1/l}\sqrt{1-\gamma^2}$, we can easily conclude that $v_{ii}(t+1) \ge v_{ii}(t)$ by induction. Hence, we can simplify the dynamic as

$$v_{ii}(t+1) \ge \left(1 + \eta l (0.99\sigma_i - v_{ii}^l) v_{ii}^{l-2}(t)\right) v_{ii}(t) \ge \left(1 + 0.49\eta l \sigma_i v_{ii}^{l-2}(t)\right) v_{ii}(t). \tag{5.135}$$

Next, we further split the interval $\mathcal{I} = \left[0, 0.5\sigma_i^{1/l}\right]$ into $N = \mathcal{O}\left(\log\left(0.5\sigma_i^{1/l}/\alpha\right)\right)$ sub-intervals $\{\mathcal{I}_0, \cdots, \mathcal{I}_{N-1}\}$, where $\mathcal{I}_k = [2^k v_{ii}(0), 2^{k+1} v_{ii}(0))$. Let $\mathcal{T}_k$ collect the iterations that v_{ii} spends in $\mathcal{I}_k$. Accordingly, let $|\mathcal{T}_k| = t_k$ be the number of iterations that v_{ii} spends within $\mathcal{I}_k$. First note that $v_{ii}(t) \ge 2^k v_{ii}(0)$ for every $t \in \mathcal{T}_k$. Hence, we have

$$\left(1 + 0.49\eta l \sigma_i 2^{(l-2)k} v_{ii}^{l-2}(0)\right)^{t_k} \ge 2. \tag{5.136}$$

which implies

$$t_k \le \frac{\log(2)}{0.49\eta l \sigma_i v_{ii}^{l-2}(0)} 2^{-(l-2)k}. \tag{5.137}$$

By summing over $k = 0, \cdots, N-1$, we can upper bound the required number of iterations T_3

$$T_3 \le \sum_{k=0}^{\infty} t_k \le \sum_{k=0}^{\infty} \frac{\log(2)}{0.49\eta l \sigma_i v_{ii}^{l-2}(0)} 2^{-(l-2)k} \le \frac{4}{\eta l \sigma_i \alpha^{l-2}} \le \frac{4}{\eta l \sigma_r \alpha^{l-2}} \ll T_1, \tag{5.138}$$

where the last inequality is due to our assumption $\gamma \lesssim \frac{1}{\kappa l}^{\frac{1}{l-2}}$.

Phase 2. In the second phase, we have $v_{ii} \geq 0.5\sigma_i^{1/l}$. We further simplify equation 5.134 as

$$\begin{aligned} v_{ii}(t+1) &\geq v_{ii}(t) + \eta l \left(\sigma_i - 8d^{l-1}\sigma_i^{\frac{l-2}{l}}\alpha^2\gamma^2 - v_{ii}^l(t)\right) v_{ii}^{l-1}(t) \\ &\geq v_{ii}(t) + \eta l \left(\tilde{\sigma}_i - v_{ii}^l(t)\right) v_{ii}^{l-1}(t), \end{aligned} \tag{5.139}$$

where we denote $\tilde{\sigma}_i = \sigma_i - 8d^{l-1}\sigma_i^{\frac{l-2}{l}}\alpha^{2/l}\gamma^2$. Then, via a similar trick, within additional $T_4 \leq \frac{4}{\eta l \sigma_r}\alpha^{-\frac{l-2}{l}}$ iterations, we have $v_{ii}^l(t) \geq \sigma_i - 8d^{l-1}\sigma_i^{\frac{l-2}{l}}\alpha^{2/l}\gamma^2$. A similar argument on the upper bound shows $v_{ii}^l(t) \leq \sigma_i + 8d^{l-1}\sigma_i^{\frac{l-2}{l}}\alpha^{2/l}\gamma^2$, which completes the proof. □

5.H Auxiliary Lemmas

Lemma 101 (Bernoulli inequality)**.** *For $0 \leq x < \frac{1}{r-1}$, and $r > 1$, we have*

$$(1+x)^r \leq 1 + \frac{rx}{1-(r-1)x}. \tag{5.140}$$

CHAPTER 6

Conclusion

This dissertation has explored the implicit regularization properties of gradient descent in realistic machine learning settings, where data may be noisy, inputs are often non-isotropic, and models are typically overparameterized. Our findings challenge conventional assumptions and extend the theoretical understanding of gradient-based optimization in three key areas: robust learning under outlier label noise, learning with non-isotropic input distributions, and a general framework for explaining implicit regularization through the lens of function basis decomposition.

First, we investigated robust learning in the presence of outlier label noise. Our analysis of robust matrix sensing showed that the subgradient method with ℓ_1-loss implicitly preserves low-rank structures and achieves nearly linear convergence under near-optimal sample complexity. We also examined robust training in deep linear networks, where we demonstrated that gradient descent can recover ground-truth parameters even under adversarial label corruptions. Together, these results offer new insights into the role of optimization geometry—such as strict saddle points—and the implicit regularization effects that enable resilience to outliers in both shallow and deep models.

Second, we studied the challenges posed by non-isotropic input distributions. Using unregularized matrix completion as a representative example, we developed a novel statistical decoupling tool to handle highly dependent and structured data. This allowed us to establish near-optimal sample complexity bounds for gradient descent, highlighting the robustness of gradient-based methods in practical, non-idealized data regimes. Our analysis underscores the significance of initialization and the implicit trajectory-dependent regularization that emerges during optimization.

Finally, we proposed a general framework to explain implicit regularization through the lens of function basis decomposition. This perspective reveals that optimization trajectories often follow a nearly monotonic progression when projected onto a suitable function basis. The framework unifies various phenomena observed in gradient-based learning and

strengthens theoretical guarantees across diverse applications, including neural networks and structured matrix factorization.

Our work bridges the gap between theory and practice by rigorously characterizing the behavior of gradient descent in complex learning environments. Several open directions remain for future exploration. First, extending our results to broader classes of nonconvex problems—including deep nonlinear networks—is a natural next step. Second, improving the sample complexity bounds for matrix completion, especially in overparameterized regimes, may lead to more efficient algorithms. Lastly, leveraging the function basis decomposition framework to study generalization in deep learning remains an exciting and largely open challenge.

By advancing the theoretical understanding of implicit regularization, this dissertation lays the groundwork for designing more robust and efficient learning algorithms. These findings have broad implications for machine learning, signal processing, and optimization, paving the way for future research in scalable, principled methods for real-world data analysis.

BIBLIOGRAPHY

Zeyuan Allen-Zhu and Yuanzhi Li. Backward feature correction: How deep learning performs deep learning. *arXiv preprint arXiv:2001.04413*, 2020.

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in neural information processing systems*, 32, 2019a.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019b.

Animashree Anandkumar, Rong Ge, and Majid Janzamin. Analyzing tensor power method dynamics in overcomplete regime. *Journal of Machine Learning Research*, 18(22):1–40, 2017.

Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. *arXiv preprint arXiv:1810.02281*, 2018.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019a.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019b.

Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28 (3):253–263, 2008.

Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of complexity*, 23(1):52–72, 2007.

Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *The annals of statistics*, 44(2):813–852, 2016.

Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. *Advances in Neural Information Processing Systems*, 29, 2016.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of statistics*, 37(4):1705–1732, 2009.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Thierry Bouwmans and El Hadi Zahzah. Robust pca via principal component pursuit: A review for a comparative evaluation in video surveillance. *Computer Vision and Image Understanding*, 122:22–34, 2014.

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

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

Emmanuel J Candes and Yaniv 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.

Joshua Cape, Minh Tang, and Carey E Priebe. The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. 2019.

Vasileios Charisopoulos, Yudong Chen, Damek Davis, Mateo Díaz, Lijun Ding, and Dmitriy Drusvyatskiy. Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. *Foundations of Computational Mathematics*, pages 1–89, 2021.

Ji Chen and Xiaodong Li. Model-free nonconvex matrix completion: Local minima analysis and applications in memory-efficient kernel pca. *J. Mach. Learn. Res.*, 20(142):1–39, 2019.

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

Yan Mei Chen, Xiao Shan Chen, and Wen Li. On perturbation bounds for orthogonal projections. *Numerical Algorithms*, 73(2):433–444, 2016.

Yudong Chen. Incoherence-optimal matrix completion. *IEEE Transactions on Information Theory*, 61(5):2909–2923, 2015.

Yudong Chen and Martin J Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

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

Hung-Hsu Chou, Carsten Gieshoff, Johannes Maly, and Holger Rauhut. Gradient descent for deep matrix factorization: Dynamics and implicit bias towards low rank. *arXiv preprint arXiv:2011.13772*, 2020.

Hung-Hsu Chou, Johannes Maly, and Holger Rauhut. More is less: Inducing sparsity via overparameterization. *arXiv preprint arXiv:2112.11027*, 2021.

Fan Chung and Linyuan Lu. Concentration inequalities and martingale inequalities: a survey. *Internet mathematics*, 3(1):79–127, 2006.

Frank H Clarke. Generalized gradients and applications. *Transactions of the American Mathematical Society*, 205:247–262, 1975.

Frank H Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

Jeremy M Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. *arXiv preprint arXiv:2103.00065*, 2021.

Hadi Daneshmand, Jonas Kohler, Aurelien Lucchi, and Thomas Hofmann. Escaping saddles with stochastic gradients. In *International Conference on Machine Learning*, pages 1155–1164. PMLR, 2018.

Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. *Advances in neural information processing systems*, 29, 2016.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic subgradient method converges at the rate $o(k^{-1/4})$ on weakly convex functions. *arXiv preprint arXiv:1802.02988*, 2018.

Damek Davis and Dmitriy Drusvyatskiy. Proximal methods avoid active strict saddles of weakly convex functions. *Foundations of Computational Mathematics*, 22(2):561–606, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31. Springer-Verlag, 1996.

Lijun Ding, Liwei Jiang, Yudong Chen, Qing Qu, and Zhihui Zhu. Rank overspecified robust matrix recovery: Subgradient method and exact recovery. *arXiv preprint arXiv:2109.11154*, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Simon Du and Wei Hu. Width provably matters in optimization for deep linear neural networks. In *International Conference on Machine Learning*, pages 1655–1664. PMLR, 2019.

Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. *arXiv preprint arXiv:1806.00900*, 2018.

Stanley C Eisenstat and Ilse CF Ipsen. Relative perturbation techniques for singular value problems. *SIAM Journal on Numerical Analysis*, 32(6):1972–1988, 1995.

Stanley C Eisenstat and Ilse CF Ipsen. Relative perturbation results for eigenvalues and eigenvectors of diagonalisable matrices. *BIT Numerical Mathematics*, 38(3):502–509, 1998.

Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *Conference on learning theory*, pages 907–940. PMLR, 2016.

Zhou Fan and Zhichao Wang. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. *Advances in neural information processing systems*, 33: 7710–7721, 2020.

Cong Fang, Zhouchen Lin, and Tong Zhang. Sharp analysis for nonconvex sgd escaping from saddle points. In *Conference on Learning Theory*, pages 1192–1234. PMLR, 2019.

Salar Fattahi and Somayeh Sojoudi. Exact guarantees on the absence of spurious local minima for non-negative rank-1 robust principal component analysis. *Journal of machine learning research*, 2020.

Stanislav Fort, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel M Roy, and Surya Ganguli. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. *Advances in Neural Information Processing Systems*, 33:5850–5861, 2020.

Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. *Advances in neural information processing systems*, 29, 2016.

Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International Conference on Machine Learning*, pages 1233–1242. PMLR, 2017.

Rong Ge, Yunwei Ren, Xiang Wang, and Mo Zhou. Understanding deflation process in over-parametrized tensor decomposition. *Advances in Neural Information Processing Systems*, 34:1299–1311, 2021.

Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. *arXiv preprint arXiv:1904.13262*, 2019.

Daniel Gissin, Shai Shalev-Shwartz, and Amit Daniely. The implicit bias of depth: How incremental learning drives generalization. *arXiv preprint arXiv:1909.12051*, 2019.

David F Gleich and Lek-heng Lim. Rank aggregation via nuclear norm minimization. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 60–68, 2011.

Thore Graepel. Kernel matrix completion by semidefinite programming. In *International Conference on Artificial Neural Networks*, pages 694–699. Springer, 2002.

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–10. IEEE, 2018.

Jeff Z HaoChen, Colin Wei, Jason Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance. In *Conference on Learning Theory*, pages 2315–2357. PMLR, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Yue Hu, Xiaohan Liu, and Mathews Jacob. A generalized structured low-rank matrix completion algorithm for mr image recovery. *IEEE transactions on medical imaging*, 38(8): 1841–1851, 2018.

Zhengmian Hu and Heng Huang. On the random conjugate kernel and neural tangent kernel. In *International Conference on Machine Learning*, pages 4359–4368. PMLR, 2021.

Like Hui and Mikhail Belkin. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. *arXiv preprint arXiv:2006.07322*, 2020.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674, 2013.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International Conference on Machine Learning*, pages 1724–1732. PMLR, 2017.

Cedric Josz, Yi Ouyang, Richard Y Zhang, Javad Lavaei, and Somayeh Sojoudi. A theory on the absence of spurious solutions for nonconvex and nonsmooth optimization. *arXiv preprint arXiv:1805.08204*, 2018.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 795–811. Springer, 2016.

Kenji Kawaguchi. Deep learning without poor local minima. *Advances in neural information processing systems*, 29, 2016.

Daesung Kim and Hye Won Chung. Rank-1 matrix completion with gradient descent and small random initialization. *arXiv preprint arXiv:2212.09396*, 2022.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer Science & Business Media, 1991.

Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid strict saddle points. *Mathematical programming*, 176(1):311–337, 2019.

Nayoung Lee, Kartik Sreenivasan, Jason D Lee, Kangwook Lee, and Dimitris Papailiopoulos. Teaching arithmetic to small transformers. *arXiv preprint arXiv:2307.03381*, 2023.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. In *Journal of Machine Learning Research*, pages 1334–1373, 2016.

Sergey Levine, Peter Pastor, Alex Krizhevsky, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. In *International Symposium on Experimental Robotics*, pages 173–184, 2018.

Jiajin Li, Anthony Man Cho So, and Wing Kin Ma. Understanding notions of stationarity in nonsmooth optimization: A guided tour of various constructions of subdifferential for nonsmooth functions. *IEEE Signal Processing Magazine*, 37(5):18–31, 2020a.

Jiangyuan Li, Thanh Nguyen, Chinmay Hegde, and Ka Wai Wong. Implicit sparse regularization: The impact of depth and early stopping. *Advances in Neural Information Processing Systems*, 34, 2021.

Xiao Li, Zhihui Zhu, Anthony Man-Cho So, and Rene Vidal. Nonconvex robust low-rank matrix recovery. *SIAM Journal on Optimization*, 30(1):660–686, 2020b.

Yuanxin Li, Cong Ma, Yuxin Chen, and Yuejie Chi. Nonconvex matrix factorization from rank-one measurements. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1496–1505. PMLR, 2019.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47. PMLR, 2018.

Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. *arXiv preprint arXiv:2012.09839*, 2020c.

Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye. Tensor completion for estimating missing values in visual data. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):208–220, 2012.

Philip M Long. Properties of the after kernel. *arXiv preprint arXiv:2105.10585*, 2021.

Xiao Luan, Bin Fang, Linghui Liu, Weibin Yang, and Jiye Qian. Extracting sparse error of robust pca for face recognition in the presence of varying illumination and occlusion. *Pattern Recognition*, 47(2):495–508, 2014.

Xin Luo, Mengchu Zhou, Yunni Xia, and Qingsheng Zhu. An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems. *IEEE Transactions on Industrial Informatics*, 10(2):1273–1284, 2014.

Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *arXiv preprint arXiv:1711.10467*, 2017.

Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning*, pages 3345–3354. PMLR, 2018.

Jianhao Ma and Salar Fattahi. Implicit regularization of sub-gradient method in robust matrix recovery: Don't be afraid of outliers. *arXiv preprint arXiv:2102.02969*, 2021a.

Jianhao Ma and Salar Fattahi. Sign-rip: A robust restricted isometry property for low-rank matrix recovery. *arXiv preprint arXiv:2102.02969*, 2021b.

Jianhao Ma and Salar Fattahi. Blessing of depth in linear regression: Deeper models have flatter landscape around the true solution. *Advances in Neural Information Processing Systems*, 35:34334–34346, 2022a.

Jianhao Ma and Salar Fattahi. Blessing of depth in linear regression: Deeper models have flatter landscape around the true solution. In *Advances in Neural Information Processing Systems*, 2022b.

Jianhao Ma and Salar Fattahi. Global convergence of sub-gradient method for robust matrix recovery: Small initialization, noisy measurements, and over-parameterization. *Journal of Machine Learning Research*, 24(96):1–84, 2023a.

Jianhao Ma and Salar Fattahi. Can learning be explained by local optimality in low-rank matrix recovery? *arXiv preprint arXiv:2302.10963*, 2023b.

Jianhao Ma and Salar Fattahi. Convergence of gradient descent with small initialization for unregularized matrix completion. *The Thirty Seventh Annual Conference on Learning Theory*, 2024.

Jianhao Ma, Lingjun Guo, and Salar Fattahi. Behind the scenes of gradient descent: A trajectory analysis via basis function decomposition. *The Eleventh International Conference on Learning Representations*, 2023.

Rahul Mazumder, Peter Radchenko, and Antoine Dedieu. Subset selection with shrinkage: Sparse linear modeling when the snr is low. *arXiv preprint arXiv:1708.03288*, 2017.

Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

Yu Nesterov. Introductory lectures on convex programming, 1998.

Yurii Evgen'evich Nesterov. A method of solving a convex programming problem with convergence rate o\bigl(k^2\bigr). In *Doklady Akademii Nauk*, volume 269, pages 543–547. Russian Academy of Sciences, 1983.

Behnam Neyshabur, Srinadh Bhojanapalli, and Ayan Chakrabarti. Implicit regularization in deep learning. *arXiv preprint arXiv:1709.01953*, 2017.

Roman Novak, Yasaman Bahri, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. *arXiv preprint arXiv:1802.08760*, 2018.

Frank WJ Olver, Daniel W Lozier, Ronald F Boisvert, and Charles W Clark. *NIST handbook of mathematical functions hardback and CD-ROM*. Cambridge university press, 2010.

John Paisley and Lawrence Carin. A nonparametric bayesian model for kernel matrix completion. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2090–2093. IEEE, 2010.

Ioannis Panageas, Georgios Piliouras, and Xiao Wang. First-order methods almost always avoid saddle points: The case of vanishing step-sizes. *Advances in Neural Information Processing Systems*, 32, 2019.

Vivak Patel and Albert S Berahas. Gradient descent in the absence of global lipschitz continuity of the gradients: Convergence, divergence and limitations of its continuous approximation. *arXiv preprint arXiv:2210.02418*, 2022.

Qing Qu, Yuexiang Zhai, Xiao Li, Yuqian Zhang, and Zhihui Zhu. Analysis of the optimization landscapes for overcomplete representation learning. *arXiv preprint arXiv:1912.02427*, 2019.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on information theory*, 57 (10):6976–6994, 2011.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *The Journal of Machine Learning Research*, 15(1):335–366, 2014.

Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. *arXiv preprint arXiv:2005.06398*, 2020.

Noam Razin, Asaf Maman, and Nadav Cohen. Implicit regularization in tensor factorization. In *International Conference on Machine Learning*, pages 8913–8924. PMLR, 2021.

Noam Razin, Asaf Maman, and Nadav Cohen. Implicit regularization in hierarchical tensor factorization and deep convolutional neural networks. *arXiv preprint arXiv:2201.11729*, 2022.

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

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

Itay Safran and Jason D Lee. Optimization-based separations for neural networks. *arXiv preprint arXiv:2112.02393*, 2021.

Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. In *International conference on machine learning*, pages 4433–4441. PMLR, 2018.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Dominik Stöger and Mahdi Soltanolkotabi. Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. *Advances in Neural Information Processing Systems*, 34, 2021.

Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere i: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2016.

Ruoyu Sun. *Matrix completion via nonconvex factorization: Algorithms and theory*. PhD thesis, University of Minnesota, 2015.

Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

Stanislaw J Szarek. Nets of grassmann manifold and orthogonal group. In *Proceedings of research workshop on Banach space theory (Iowa City, Iowa, 1981)*, volume 169, page 185. University of Iowa Iowa City, IA, 1982.

Matus Telgarsky. Representation benefits of deep feedforward networks. *arXiv preprint arXiv:1509.08101*, 2015.

Matus Telgarsky. Benefits of depth in neural networks. In *Conference on learning theory*, pages 1517–1539. PMLR, 2016.

Tian Tong, Cong Ma, and Yuejie Chi. Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *Journal of Machine Learning Research*, 22(150):1–63, 2021.

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

Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Ben Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *International Conference on Machine Learning*, pages 964–973. PMLR, 2016.

Sara Van de Geer. The deterministic lasso. Seminar für Statistik, Eidgenössische Technische Hochschule (ETH) Zürich, 2007.

Aad W Van Der Vaart, Adrianus Willem van der Vaart, Aad van der Vaart, and Jon Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 1996.

Ramon Van Handel. Probability in high dimension. Technical report, PRINCETON UNIV NJ, 2014.

Tomas Vaskevicius, Varun Kanade, and Patrick Rebeschini. Implicit regularization for optimal sparse recovery. *Advances in Neural Information Processing Systems*, 32, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Van Vu. A simple svd algorithm for finding hidden partitions. *Combinatorics, Probability and Computing*, 27(1):124–140, 2018.

Nikhil Vyas, Yamini Bansal, and Preetum Nakkiran. Limitations of the ntk for understanding generalization in deep learning. *arXiv preprint arXiv:2206.10012*, 2022.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Xiang Wang, Chenwei Wu, Jason D Lee, Tengyu Ma, and Rong Ge. Beyond lazy training for over-parameterized tensor decomposition. *Advances in Neural Information Processing Systems*, 33:21934–21944, 2020.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR, 2020.

Tian Ye and Simon S Du. Global convergence of gradient descent for asymmetric low-rank matrix factorization. *Advances in Neural Information Processing Systems*, 34, 2021.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

Gavin Zhang, Salar Fattahi, and Richard Y Zhang. Preconditioned gradient descent for overparameterized nonconvex burer–monteiro factorization with global optimality certification. *arXiv preprint arXiv:2206.03345*, 2022.

Jialun Zhang, Salar Fattahi, and Richard Y Zhang. Preconditioned gradient descent for overparameterized nonconvex matrix factorization. *Advances in Neural Information Processing Systems*, 34:5985–5996, 2021.

Richard Y Zhang. Sharp global guarantees for nonconvex low-rank matrix recovery in the overparameterized regime. *arXiv preprint arXiv:2104.10790*, 2021.

Richard Y Zhang, Somayeh Sojoudi, and Javad Lavaei. Sharp restricted isometry bounds for the inexistence of spurious local minima in nonconvex matrix recovery. *J. Mach. Learn. Res.*, 20(114):1–34, 2019.

Yuqian Zhang, Qing Qu, and John Wright. From symmetry to geometry: Tractable nonconvex problems. *arXiv preprint arXiv:2007.06753*, 2020.

Peng Zhao, Yun Yang, and Qiao-Chu He. Implicit regularization via hadamard product over-parametrization in high-dimensional linear regression. *arXiv preprint arXiv:1903.09367*, 2019.

Qinqing Zheng and John Lafferty. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. *arXiv preprint arXiv:1506.06081*, 2015.

Qinqing Zheng and John Lafferty. Convergence analysis for rectangular matrix completion using burer-monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*, 2016.

Jiacheng Zhuo, Jeongyeol Kwon, Nhat Ho, and Constantine Caramanis. On the computational and statistical complexity of over-parameterized matrix sensing. *arXiv preprint arXiv:2102.02756*, 2021.