Non-Convex and Non-Uniform Approaches to the Euclidean Distance Matrix Completion Problem

A dissertation submitted by

Chandler M. Smith

in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

in

Mathematics

TUFTS UNIVERSITY

May 2026

Advisor: Dr. Abiy F. Tasissa

Abstract

The Euclidean Distance Matrix Completion (EDMC) problem is a foundational problem in engineering, data science, and machine learning. This problem can be simply described by the following question: given partial access to a set of pairwise Euclidean distances between n points in r dimensions, is it possible to reconstruct the set of n points, up to rigid transformations, that generated the pairwise distances? This problem traces back to the 1950s, and its variations are still actively studied in the literature today. Much of the recent research on the EDMC problem relies on low-rank matrix completion techniques, with theoretical guarantees existing for nuclear-norm minimization over the cone of positive semidefinite matrices. These techniques scale poorly for large sets of points, however, so investigation into faster, non-convex surrogates is needed. This thesis provides several state-of-the-art approaches to provably solve this problem under uniform random sampling of pairwise distances using first-order Riemannian optimization techniques. In addition to this, we describe geometric conditions for recovery, and provide a characterization of easy- and hard-to-recover point clouds. To solve the problem for ill-conditioned geometries, we provide a geometrically-aware sampling scheme that provably recovers any point cloud with state-of-the-art sample complexity.

Acknowledgments

This has been an incredibly fun journey. Doing my Ph.D. at Tufts has truly been the most intellectually enriching time of my life, and I have so many people to thank for helping me through the last 5 years; however, I would first like to thank the NSF for supporting my work through DMS-2304489.

To start the more personal note, I would like to thank my advisor, Dr. Abiy Tasissa. Your mentorship and friendship throughout this process have been invaluable, and I look forward to working together beyond the completion of the Ph.D. Thank you for introducing me to this problem, and being patient with the progress throughout. It was far harder than we both anticipated, but the joy is in the journey and I think that we both learned a tremendous amount through the writing of the thesis. Thank you for being so generous with your time and energy.

Next, I would like to thank my committee, Dr. Abiy Tasissa, Dr. HanQin Cai, Dr. Kasso Okoudjou, and Dr. Xiaozhe Hu, for their invaluable feedback on my thesis. Your recommendations have helped me strengthen this document.

Thank you to the Tufts Math Department as a whole. The faculty here have been tremendously supportive over my time here, and have provided a fantastic place for me to grow as a researcher. I want to give a particular thanks to my letter writers Dr. James Murphy, Dr. Xiaozhe Hu, and Dr. Kim Ruane for saying nice things about me and helping me get a job. I also want to thank Dr. James Adler for organizing the summer softball team, and again to Dr. Kim Ruane for fielding my wild throws to 1st (and occasionally wearing one on the shin...).

Next, I would like to thank my friends in the department for their camaraderie throughout the last several years. Thank you Satchel Lefebvre, Hikaru Jitsukawa, and Logo Lou for a welcome relief through basketball, thank you Eoghan O'Keefe for your tremendous DMing

abilities, and thank you to Kate Wall, Danny Riley, Bryan Rust, Sam Gomez, Brian Parkes, and all of the other friends I've made throughout my time at Tufts. The community in the math department at Tufts has been just as important as the research.

Next, I would like to thank my reading group friends, specifically David Cohen, David Gentile, Martin Buck, and Brendan Mallery, for their continual mathematical curiosity and general friendship. I've greatly enjoyed struggling through textbooks that aren't that related to any of our research with you all, and I look forward to celebrating our collective graduation together.

I truly would have never made it through the Ph.D. without the group chat. To Brendan Mallery and Ian McPherson, you both mean the world to me and I largely attribute my love of mathematics to our friendship (and the sloppy chalkies). Having friends to weather the ups and downs of the Ph.D., to share in the excitement of a new idea, and to commiserate over the doldrums of everyday life as a grad student has been so important for our collective sanity. To Brendan, thank you for sharing an office with me throughout the Ph.D. and for the time spent BS'ing instead of working. Some of my favorite memories of my time in New England will be of hanging out in a fluorescently lit, windowless office (sometimes with the Davids). To Ian, thank you for pulling me away from physics and more towards math starting in undergrad, and for continually inspiring me with your passion and work ethic. It's crazy to think how far we've both come. I finished the Ph.D. first though, and don't forget it.

Next, I want to thank both my parents, Kelly and Sheri, my sister Teagan, my Nana Judi, my cousin James, and my friend Yonatan. The support I have received from all of you throughout the years to pursue my passion, even if you don't understand it, has been so important to me. I love fielding your questions about my work, and I deeply appreciate the genuine interest in how I spend my time. Thank you and I love you all.

Lastly, I want to thank my partner Zoë. You have been such a rock for me throughout my Ph.D. For every "it's so over", you always helped me find a "we're so back" inside myself. Thank you for always being there. Moving to the Boston area to come to Tufts was the best decision I've ever made for many reasons, but meeting you is chief among them. I love you very much.

Contents

1 Introduction **2**

1.1 The Euclidean Distance Geometry Problem 2

1.1.1 Classical Multidimensional Scaling 3

1.2 Low-Rank Matrix Completion (LRMC) 7

1.2.1 Incoherence of Low-Rank Matrices 8

1.2.2 Theoretical Recovery through LRMC 9

1.3 The Dual-Basis Approach to EDMC 12

1.4 First-Order Optimization on Manifolds 16

1.4.1 Geometric preliminaries 17

1.4.2 Riemannian Gradient Descent 19

1.5 A Riemannian Approach to LRMC 20

1.6 Incoherence and Geometry 24

1.7 Useful Inequalities from Random Matrix Theory 31

1.8 A Collection of Useful Theorems 34

1.A Further Incoherence Results 35

2 Preliminary Investigations into a Non-convex Riemannian Approach to the EDMC Problem **38**

2.1 Algorithm 38

2.1.1 One Additional Surrogate of Interest 42

2.2 Theoretical Analysis 43

2.2.1 Initialization 47

2.3 Numerical Results . . . 50
2.3.1 Synthetic Data Experiments . . . 50
2.A Properties of the Dual Bases . . . 51
2.B Restricted Isometry Results . . . 54
2.B.1 Proof of Theorem 2.2.1 . . . 54
2.C Proof of Local Convergence (Theorem 2.2.2) . . . 65
2.C.1 Proof of Theorem 2.2.2 . . . 67
2.D Initialization Results . . . 68
2.D.1 Proof of Lemma 2.2.3 . . . 68
2.D.2 Proof of Lemma 2.2.5 . . . 71

3 An Improved Non-convex Riemannian Approach to the EDMC Problem 74
3.1 An Improved Operator . . . 74
3.1.1 Implementation Efficiency . . . 77
3.2 Theoretical Analysis . . . 79
3.2.1 Local Convergence Analysis . . . 80
3.2.2 Initialization Results . . . 83
3.3 Robustness Guarantees . . . 86
3.4 Numerical Results . . . 89
3.4.1 Synthetic Data Experiment . . . 89
3.4.2 Comparison to Existing Methods . . . 89
3.4.3 Experiments on Noisy Distance Measurements . . . 92
3.5 Conclusion . . . 93
3.A Further Properties of the Dual Bases . . . 94
3.B Restricted Isometry Results . . . 95
3.B.1 Supporting Concentration Results . . . 95
3.B.2 Proof of Theorem 3.2.1 . . . 111
3.C Local Convergence Results . . . 115
3.C.1 Proof of Theorem 3.2.2 . . . 117
3.D Initialization Results (Proof of Lemma 3.2.3) . . . 118

3.E Robustness Guarantees . . . 120

4 Non-Uniform Sampling Approaches to EDMC **124**

4.1 A Leverage-Score based approach to LRMC . . . 124

4.2 Motivating Work in LRMC . . . 126

4.3 Geometric Leverage Scores and Sampling Scheme . . . 127

4.3.1 Geometric Interpretation of Dual-Basis Geometric Leverage Scores . . . 130

4.3.2 Geometric Interpretation of Standard Leverage Scores . . . 132

4.4 Algorithm . . . 139

4.5 Theoretical Guarantees . . . 140

4.5.1 Restricted Isometry Property . . . 140

4.6 Numerical Results . . . 142

4.6.1 Approximate Leverage Scores . . . 143

4.6.2 Approximating W-DBRE and Other Leverage Scores . . . 145

4.7 Conclusion . . . 145

4.A Useful Lemmas . . . 146

4.B Restricted Isometry Results . . . 147

4.B.1 Supporting Concentration Results . . . 147

4.B.2 Proof of Theorem 4.5.1 . . . 157

4.C Local Convergence Results . . . 160

4.C.1 Proof of Theorem 4.5.2 . . . 161

5 Conclusion and Future Outlook **164**

5.1 Summary of Results . . . 164

5.2 Future Work . . . 165

5.2.1 Sampling Models . . . 165

5.2.2 Other Algorithms . . . 167

5.2.3 Connections to Rigidity Theory . . . 169

5.3 Conclusion . . . 170

A Appendix **171**

A.1 A Proof Sketch for the Convergence of Algorithm 3 (RF-RGD) 171

A.1.1 Theoretical Analysis 171

A.1.2 Supplementary Results 178

A.2 Dual Bases 179

A.3 Geometry of Low-Rank Matrices 180

List of Tables

2.1 Relative recovery error $\|\boldsymbol{X} - \boldsymbol{X}_{\text{rec}}\|_{\text{F}} / \|\boldsymbol{X}\|_{\text{F}}$ between the recovered Gram matrix and the true Gram matrix averaged over 25 trials using Algorithms 3, 4, and the non-convex algorithm in [TL18]. . . . 51

3.1 Relative recovery error $\|\mathbf{X} - \mathbf{X}_{\text{rec}}\|_{\text{F}}/\|\mathbf{X}\|_{\text{F}}$ averaged over 25 trials. . 89

3.2 Relative recovery error $\|\boldsymbol{X} - \boldsymbol{X}_{\text{rec}}\|_{\text{F}} / \|\boldsymbol{X}\|_{\text{F}}$ between the recovered Gram matrix and the true Gram matrix averaged over 25 trials comparing DBRE, MatrixIRLS, and the Augmented Lagrangian approach. 91

3.3 Relative recovery error $\left\|\boldsymbol{X} - \hat{\boldsymbol{X}}_{\text{rec}}\right\|_{\text{F}} / \|\boldsymbol{X}\|_{\text{F}}$ between the recovered perturbed Gram matrix and the true Gram matrix averaged over 10 trials using DBRE for the Sphere dataset referenced in Table 3.1 and Figure 3.2. σ is the standard deviation of the perturbing noise, and p is the sampling rate. 93

4.1 Comparative mean over 10 trials for relative Gram matrix error results by algorithm, dataset, and percentage of samples. The incoherence constant for the ellipse with outliers is $\nu \approx 38 \approx \frac{n}{3}$, and the incoherence constant for the standard ellipse is $\nu \approx 2$. 143

4.2 Mean computational time across sampling rates for W-DBRE and [CBSW14] using the SDPT3 solver, sorted by sampling rate. 143

4.3 Mean relative Gram matrix recovery error using W-DBRE with varying leverage scores. For this experiment, $\nu \approx 15 \approx \frac{n}{3}$. 144

4.4 Mean relative Gram matrix error across varying sampling rates via approximate leverage scores. The incoherence constant is $\nu \approx 11 \approx \frac{n}{3}$.145

List of Figures

1.1 Geometric interpretation of subspace incoherence in matrix completion. 10

1.2 A diagram of a simple first-order retraction method on $\mathcal{N}$. Again, $\nabla f(\boldsymbol{x}_l)$ is the Euclidean gradient of f at $\boldsymbol{x}_l$, grad $f(\boldsymbol{x}_l)$ is the Riemannian gradient at $\boldsymbol{x}_l$, and $\boldsymbol{x}_{l+1} = \boldsymbol{R}_{\boldsymbol{x}_l}(\boldsymbol{x}_l - \alpha_l \text{grad}\, f(\boldsymbol{x}_l))$ for some step size $\alpha_l > 0$, as in Algorithm 1. . . . 21

1.3 Visualizing the rows of $\boldsymbol{U}$ that lead to a minimal incoherence parameter. 26

1.4 Example of a set of points with a maximal incoherence parameter. 27

3.1 This diagram is a schematic of the overall proof of convergence for DBRE. Arrows indicate how results depend on one another, and how they link together to form the overall proof of convergence. Not every exact dependency is shown in this figure for legibility purposes, instead focusing on the key pieces of the overall flow of the argument. 84

3.2 Reconstruction of the synthetic datasets referenced in Table 3.1. From left to right, the Bernoulli parameter is 0.03, 0.02, and 0.01. 90

3.3 Oversampling ratio ρ versus dimension r for 100 points on the uniform distribution on S^{r-1}. Each parameter was tested 100 times. 91

3.4 Oversampling ratio ρ versus noise level 10^{γ} for 100 points drawn i.i.d. from Unif(S^2). Each parameter was tested 500 times. 92

4.1 Incoherent and Coherent Datasets 125

4.2 Phase Transition plot on the ellipse with outlier datapoints as a function of the size of the dataset and the total percentage of revealed samples. All-white indicates 100% success rate (classified as a relative error $< 10^{-3}$, and all-black indicates 0% success. Each square was averaged over 25 trials. 144

5.1 Comparison between a rigid and non-rigid graph 169

Examining Committee

Chair of committee:	Dr. Abiy Tasissa Assistant Professor Department of Mathematics Tufts University
Internal committee member:	Dr. Xiaozhe Hu Professor Department of Mathematics Tufts University
Internal committee member:	Dr. Kasso Okoudjou Professor Department of Mathematics Tufts University
External committee member:	Dr. HanQin Cai Assistant Professor School of Data, Mathematical, and Statistical Sciences, Department of Computer Science University of Central Florida

Notation

Matrices, Vectors, and Operators

- $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{v}$: Matrices (uppercase bold) and vectors (lowercase bold). $\mathbf{0}, \mathbf{1}$: Zero/ones vectors or matrices. $\boldsymbol{I}$: Identity matrix. $\boldsymbol{e}_i, \boldsymbol{e}_{ij}$: Standard basis vectors/matrices.
- $\mathcal{A}, \mathcal{I}$: Linear/identity operators on matrices. $\mathcal{A}^*$: Adjoint of $\mathcal{A}$. $\mathbb{V}$: Vector spaces/subspaces.
- $\boldsymbol{X}^\top, \mathrm{Trace}(\boldsymbol{X})$: Transpose and trace of $\boldsymbol{X}$. X_{ij}: (i,j)-th entry. $\mathrm{vec}(\boldsymbol{Y})$: Column stack into $\mathbb{R}^{n^2}$.
- $\langle \boldsymbol{A}, \boldsymbol{B} \rangle$: Trace inner product $\mathrm{Trace}(\boldsymbol{A}^\top \boldsymbol{B})$. $\odot$: Hadamard product. δ_{ij}: Kronecker delta.
- $\boldsymbol{A} \succeq \boldsymbol{B}$: Loewner ordering. $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top$: Thin spectral decomposition.
- $H_{\boldsymbol{\alpha}\boldsymbol{\beta}} = \langle \boldsymbol{w}_{\boldsymbol{\alpha}}, \boldsymbol{w}_{\boldsymbol{\beta}} \rangle$, $H^{\boldsymbol{\alpha}\boldsymbol{\beta}} = \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}} \rangle$.

Norms and Spectral Quantities

- $\|\cdot\|_2, \|\cdot\|_{\mathrm{F}}, \|\cdot\|, \|\cdot\|_\infty, \|\cdot\|_*, \|\cdot\|_{2\to\infty}$: Euclidean, Frobenius, operator, max absolute entry, nuclear ($\sum_i \sigma_i$), and $2 \to \infty$ norms.
- $\|\mathcal{A}\|$: Operator norm of $\mathcal{A}$, defined as $\sup_{\|\boldsymbol{X}\|_{\mathrm{F}}=1} \|\mathcal{A}(\boldsymbol{X})\|_{\mathrm{F}}$.
- $\lambda_{\max}(\boldsymbol{X}), \lambda_{\min}(\boldsymbol{X})$: Max/min eigenvalues. $\lambda_1 \geq \cdots \geq \lambda_r$: Ordered non-zero eigenvalues.
- $\sigma_r(\boldsymbol{Y})$: r-th singular value. κ: Condition number $\|\boldsymbol{X}\|/\sigma_r(\boldsymbol{X})$.

Sets, Indexing, Manifolds, and Geometry

- $\mathbb{I}$: Universal index set $\{(i,j) : 1 \leq i < j \leq n\}$. Ω: Random subsets of $\mathbb{I}$. $\emptyset$: Empty set.
- $\boldsymbol{x}_i, \boldsymbol{x}^i$: i-th row and column of $\boldsymbol{X}$, respectively, represented as column vectors.
- $\mathcal{N}_r, \mathcal{N}$: Manifold of rank-$r$ matrices and general smooth manifolds.
- $\mathbb{T}, \mathbb{T}_l$: Tangent space at $\boldsymbol{X} \in \mathcal{N}_r$ and at l-th iterate $\boldsymbol{X}_l \in \mathcal{N}_r$.
- $\nabla f, \mathrm{grad}\, f$: Euclidean and Riemannian gradients of f.

Non-Convex and Non-Uniform Approaches to the Euclidean Distance Matrix Completion Problem

Chapter 1

Introduction

1.1 The Euclidean Distance Geometry Problem

We start this thesis by asking a simple question:

Question 1.1.1

Given a set of pairwise distances alone, can I faithfully reconstruct a unique set of points embedded in $\mathbb{R}^d$?

At first blush, this seems like enough information to reconstruct a set of points in a finite-dimensional vector space, provided sufficiently many points. If one has access to every pairwise distance, i.e., if they are looking to reconstruct a set of points $\{\boldsymbol{p}_i\}_{i=1}^n \subset \mathbb{R}^d$, is access to every $\|\boldsymbol{p}_i - \boldsymbol{p}_j\|_2$ enough? After a second bit of thought, one might notice that the set $\{\|\boldsymbol{p}_i - \boldsymbol{p}_j\|_2\}_{i,j=1}^n$ would be invariant to rotations and translations. Given this observation, one might be better suited to ask the following:

Question 1.1.2: EDG Motivating Question

Given a set $\{\|\boldsymbol{p}_i - \boldsymbol{p}_j\|_2\}_{i,j=1}^n$, is it possible to reconstruct a set of points $\{\boldsymbol{p}_i\}_{i=1}^n \subset \mathbb{R}^d$ that exactly correspond to the distances known?

The answer is **yes**, this is a feasible task. In fact, the solution to this problem has been known for decades [Tor52, Tor58, Gow85]. We briefly provide an explanation, giving us a chance

to set notation as well.

1.1.1 Classical Multidimensional Scaling

We first note that finding the set of points above is equivalent to looking to construct a matrix $\boldsymbol{P} = [\boldsymbol{p}_1 \ \cdots \ \boldsymbol{p}_n]^\top \in \mathbb{R}^{n\times d}$ from the set of pairwise distances $\{\|\boldsymbol{p}_i - \boldsymbol{p}_j\|_2\}_{i,j=1}^n$. We now define the squared Euclidean distance matrix (EDM) $\boldsymbol{D} \in \mathbb{R}^{n\times n}$ entrywise as

$$D_{ij} = \|\boldsymbol{p}_i - \boldsymbol{p}_j\|_2^2.$$

We note that this matrix is symmetric and hollow, meaning that the diagonal of this matrix is all zero. We consider the squared EDM for simplicity, which will become clear shortly. The goal stated in Question 1.1.2 can be better thought of as trying to construct a linear map $\boldsymbol{D} \mapsto \boldsymbol{P}$.

We now introduce the Gram matrix of $\{\boldsymbol{p}_i\}_{i=1}^n$, denoted $\boldsymbol{X} \in \mathbb{R}^{n\times n}$, defined entrywise as

$$X_{ij} = \boldsymbol{p}_i^\top \boldsymbol{p}_j.$$

It is straightforward to see that $\boldsymbol{X} = \boldsymbol{P}\boldsymbol{P}^\top$, which implies that $\boldsymbol{X} \succeq \boldsymbol{0}$, where $\succeq$ denotes the Loewner ordering on matrices. As stated previously, it is not possible to uniquely recover $\boldsymbol{P}$ from $\boldsymbol{D}$ alone, so we make the following simplifying assumption that the point cloud $\boldsymbol{P}$ is centered.

Assumption 1.1.1: Centering and Rank Assumption

Let $\boldsymbol{1} = [1 \ \cdots \ 1]^\top \in \mathbb{R}^n$. Throughout this work, we assume that, for $\boldsymbol{P} \in \mathbb{R}^{n\times r}$

$$\boldsymbol{P}^\top \boldsymbol{1} = \boldsymbol{0}.$$

It follows that

$$\boldsymbol{X}\boldsymbol{1} = \boldsymbol{0}.$$

We also assume that $\operatorname{rank}(\boldsymbol{P}) = \operatorname{rank}(\boldsymbol{X}) = r$.

This assumption indicates that $\sum_{i=1}^n \boldsymbol{p}_i = \boldsymbol{0}$. This assumption will simplify much of our anal-

ysis in this document, and does not drastically change the outcome of answering Question 1.1.2. We now seek to construct a sequence of maps $\boldsymbol{D} \mapsto \boldsymbol{X} \mapsto \boldsymbol{P}$ in order to solve Question 1.1.2. To solve the first leg of this task, we first notice the following entrywise relationship between $\boldsymbol{X}$ and $\boldsymbol{D}$, along with a full-rank assumption.

$$\begin{aligned} D_{ij} &= \|\boldsymbol{p}_i - \boldsymbol{p}_j\|_2^2 \\ &= (\boldsymbol{p}_i - \boldsymbol{p}_j)^T (\boldsymbol{p}_i - \boldsymbol{p}_j)^T \\ &= \boldsymbol{p}_i^T \boldsymbol{p}_i - \boldsymbol{p}_i^T \boldsymbol{p}_j - \boldsymbol{p}_j^T \boldsymbol{p}_i + \boldsymbol{p}_j^T \boldsymbol{p}_j \\ &= \|\boldsymbol{p}_i\|_2^2 + \|\boldsymbol{p}_j\|_2^2 - 2\boldsymbol{p}_i^T \boldsymbol{p}_j \\ &= X_{ii} + X_{jj} - 2X_{ij}. \end{aligned}$$

The goal will now be to find the relationship between linear combinations of various elements D_{ij} to get a closed form expression for each of $X_{ii} + X_{jj} - 2X_{ij}$. We will use Assumption 1.1.1 in this approach. First, we sum across the rows of the pairwise distance matrices, denoting the i-th row average as $\bar{D}_{i,\cdot}$, we see that

$$\begin{aligned} \bar{D}_{i,\cdot} &= \frac{1}{n} \sum_{k=1}^{n} D_{ik} \\ &= \frac{1}{n} \sum_{k=1}^{n} X_{ii} + X_{kk} - 2X_{ik} \\ &= \frac{1}{n} \sum_{k} X_{ii} + \frac{1}{n} \sum_{k} X_{kk} - \frac{2}{n} \sum_{k} X_{ik} \\ &= X_{ii} + \frac{1}{n} \sum_{k} X_{kk} \\ &= X_{ii} + \frac{1}{n} \mathrm{Tr}(X). \end{aligned}$$

By symmetry, we can see that

$$\bar{D}_{\cdot,j} = X_{jj} + \frac{1}{n} \sum_{k} X_{kk} = X_{jj} + \frac{1}{n} \mathrm{Tr}(X),$$

where the j-th column average is $\bar{D}_{\cdot,j}$. Lastly, defining the whole average of entries as $\bar{D}_{\cdot,\cdot}$, we

see that

$$\begin{aligned}\bar{D}_{\cdot,\cdot} &= \frac{1}{n^2}\sum_{k,l=1}^{n} D_{kl} \\ &= \frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n} X_{kk} + X_{ll} - 2X_{kl} \\ &= \frac{1}{n}\sum_{k=1}^{n} X_{kk} + \frac{1}{n}\sum_{l=1}^{n} X_{ll} \\ &= \frac{2}{n}\mathrm{Tr}(X).\end{aligned}$$

From the above computations, it is clear that

$$\begin{aligned}D_{ij} - \bar{D}_{i,\cdot} - \bar{D}_{\cdot,j} + \bar{D}_{\cdot,\cdot} &= X_{ii} + X_{jj} - 2X_{ij} - X_{ii} - \frac{1}{n}\mathrm{Tr}(\boldsymbol{X}) - X_{jj} - \frac{1}{n}\mathrm{Tr}(X) + \frac{2}{n}\mathrm{Tr}(\boldsymbol{X}) \\ &= -2X_{ij}.\end{aligned}$$

This computation provides a clear linear relationship between the entries of $\boldsymbol{D}$ and the entries of $\boldsymbol{X}$. We can see now that

$$\begin{aligned}X_{ij} &= -\frac{1}{2}\left(D_{ij} - \bar{D}_{i,\cdot} - \bar{D}_{\cdot,j} + \bar{D}_{\cdot,\cdot}\right) \\ &= -\frac{1}{2}\left(D_{ij} - \left[\frac{1}{n}\boldsymbol{D}\mathbf{1}\mathbf{1}^\top\right]_{ij} - \left[\frac{1}{n}\mathbf{1}\mathbf{1}^\top\boldsymbol{D}\right]_{ij} + \left[\frac{1}{n}\mathbf{1}\mathbf{1}^\top\boldsymbol{D}\mathbf{1}\mathbf{1}^\top\right]_{ij}\right) \\ &= \left[\underbrace{\left(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right)}_{\boldsymbol{J}}\boldsymbol{D}\left(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right)\right]_{ij} \\ &= -\frac{1}{2}\left[\boldsymbol{J}\boldsymbol{D}\boldsymbol{J}\right]_{ij},\end{aligned}$$

indicating that

$$\boldsymbol{X} = -\frac{1}{2}\boldsymbol{J}\boldsymbol{D}\boldsymbol{J}. \tag{1.1}$$

We note that $\boldsymbol{J}$ is a projection matrix onto the orthogonal complement to the subspace spanned by the vector $\mathbf{1}$.

Now that a linear map from $\boldsymbol{D} \mapsto \boldsymbol{X}$ has been established, we briefly note that this trans-

formation is bijective for points satisfying Assumption 1.1.1. Defining the $\mathrm{diag} : \mathbb{R}^{n\times n} \to \boldsymbol{R}^n$ operation as

$$\begin{pmatrix} A_{11} & A_{12} & \cdots & \\ A_{21} & A_{22} & & \\ \vdots & & \ddots & \\ & & & A_{nn} \end{pmatrix} \mapsto \begin{pmatrix} A_{11} \\ A_{22} \\ \vdots \\ A_{nn} \end{pmatrix},$$

we can see that

$$\boldsymbol{D} = \mathrm{diag}(\boldsymbol{X})\mathbf{1}^\top + \mathbf{1}\mathrm{diag}(\boldsymbol{X})^\top - 2\boldsymbol{X}. \tag{1.2}$$

To construct $\boldsymbol{P}$ from $\boldsymbol{X}$, we look to the eigenvalue decomposition of $\boldsymbol{X}$. Taking a reduced eigenvalue decomposition, we have that $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ for $\boldsymbol{U} \in \mathbb{R}^{n\times d}$ and $\boldsymbol{\Lambda} \in \mathbb{R}^{d\times d}$. As $\boldsymbol{\Lambda} \succ \mathbf{0}$, it has a unique square root computable as

$$\boldsymbol{\Lambda}^{1/2} = \begin{pmatrix} \lambda_1^{1/2} & & \\ & \ddots & \\ & & \lambda_d^{1/2} \end{pmatrix}.$$

It follows then that taking

$$\boldsymbol{P} = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}, \tag{1.3}$$

that

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top \\ &= \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^\top \\ &= \boldsymbol{P}\boldsymbol{P}^\top, \end{aligned}$$

thus giving our desired transformation. This procedure is known as **Classical Multidimensional Scaling (CMDS)** and is a foundational technique in data science, machine learning, and various other disciplines [BLWY06, DKQW10, Roj12, PRT18, TDSL00, GHR93, Tro97, FT13, LLM08, EKS23].

Now that CMDS has been described, this leads us to want to investigate generalizations of this problem. As every pairwise distance is quite a strong assumption (not to say anything

about noisy distance measurements), we want to consider the case where distances are missing.

This assumption is of practical importance as well. Often there are limitations to measurements that can lead to incomplete data. This can be caused by geographic, climatic, or other factors that determine whether a measurement between two points can be obtained, and as such some data may be missing [ASY16, MSMS13]. For instance, in protein structure prediction, nuclear magnetic resonance (NMR) spectroscopy experiments yield spectra for protons that are close together, resulting in incomplete known distance information[CRG93]. Similarly, in sensor networks, we may have mobile nodes with known distances only from fixed anchors [BONL07, KJVRK14].

As such, we are frequently in scenarios where the distance matrix $\boldsymbol{D}$ is incomplete, rendering CMDS an inapplicable technique. The main motivating question of this thesis can be summarized as follows:

Question 1.1.3: Euclidean Distance Matrix Completion (EDMC) Question

Given partial access to entries of a squared EDM $\boldsymbol{D}$, is it possible to algorithmically reconstruct a matrix $\boldsymbol{X}$ such that $\boldsymbol{X} = -\frac{1}{2}\boldsymbol{JDJ}$?

We note that the focus on the reconstruction of $\boldsymbol{X}$ would then allow us to use the eigenvalue decomposition technique described previously. We will use the phrase Euclidean Distance Matrix Completion (EDMC) problem to describe the problem of completing a matrix $\boldsymbol{P}$ from only partial access to $\boldsymbol{D}$. Before describing how we approach this problem in this thesis, however, we will provide a brief detour into the discussion of a field known as Low-Rank Matrix Completion (LRMC).

1.2 Low-Rank Matrix Completion (LRMC)

This work is related to low-rank matrix completion, where a subset of the entries of a low-rank ground truth matrix $\boldsymbol{X}$ are observed, with the goal of algorithmically reconstructing the matrix from this partial observation. This is intuitively possible just from the fact that the "low-rankness" of the matrix lends itself to many repeated entries—or linear combinations thereof—and as such full knowledge of the underlying matrix is unnecessary.

To codify this problem mathematically, let us consider $\boldsymbol{X}$ as an $n \times n$ matrix for simplicity, with $\Omega \subset [n] \times [n]$ representing the set of observed indices. We note that this technique will generalize to rectangular matrices, but as our goal is the adaptation of this technique to positive semi-definite matrices the above assumption suffices. We note that the observed information is of the form $\{X_{ij}\}_{(i,j)\in\Omega}$, which can be re-interpreted through defining the following sampling operator. Let $\mathcal{P}_\Omega : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ be an operator which aggregates the observed entries of $\boldsymbol{X}$ projected onto specific basis elements $\boldsymbol{e}_{ij} = \boldsymbol{e}_i \boldsymbol{e}_j^\top$:

$$\mathcal{P}_\Omega(\boldsymbol{X}) = \sum_{(i,j)\in\Omega} \langle \boldsymbol{X}, \boldsymbol{e}_{ij} \rangle \boldsymbol{e}_{ij}. \tag{1.4}$$

It is clear that knowing the collection $\{X_{ij}\}_{(i,j)\in\Omega}$ is equivalent to knowing $\mathcal{P}_\Omega(\boldsymbol{X})$. In accordance with the observation that low-rank matrices contain many repeated elements, we can construct the following optimization program to recover $\boldsymbol{X}$ from $\mathcal{P}_\Omega(\boldsymbol{X})$:

$$\underset{\boldsymbol{Y}\in\mathbb{R}^{n\times n}}{\text{minimize}} \ \operatorname{rank}(\boldsymbol{Y}) \ \text{ subject to } \ \mathcal{P}_\Omega(\boldsymbol{Y}) = \mathcal{P}_\Omega(\boldsymbol{X}).$$

As minimizing the rank directly is generally a challenging problem [CR09, MJCD08, BCWP22], relaxations of this problem are often considered. In particular, the convex relaxation of the rank to the nuclear norm $\|\cdot\|_*$ has seen tremendous theoretical and computational success [CR09, CP10, Rec11a, Gro11]. We remind readers that the nuclear norm of a matrix $\boldsymbol{X}$ is defined as $\|\boldsymbol{X}\|_* = \sum_{i=1}^n \sigma_i(\boldsymbol{X})$ where $\sigma_i(\boldsymbol{X})$ is the i-th singular value of a matrix $\boldsymbol{X}$. The relaxation is thus explicitly given by

$$\underset{\boldsymbol{Y}\in\mathbb{R}^{n\times n}}{\text{minimize}} \ \|\boldsymbol{Y}\|_* \ \text{ subject to } \ \mathcal{P}_\Omega(\boldsymbol{Y}) = \mathcal{P}_\Omega(\boldsymbol{X}). \tag{1.5}$$

1.2.1 Incoherence of Low-Rank Matrices

A critical condition in the theory of LRMC is that of *incoherence*. This notion can be captured in a rather simple example. Consider the problem of two rank-1 matrices: $\boldsymbol{e}_1\boldsymbol{e}_1^\top$ and $\mathbf{1}\mathbf{1}^\top$, where $\mathbf{1} = [1 \ \cdots \ 1]^\top$. Both of these matrices are clearly rank-1; however, their recoverability is drastically different when accessing both of these matrices from random samples.

Consider the case of $\mathbf{1}\mathbf{1}^\top$ first: all the requisite information that we need to recover this matrix is one sample in each row and column to reconstruct the whole matrix. From a simple application of the coupon collector problem, we see that this can be done uniformly at random in $\mathcal{O}(n \log n)$ entries with high probability. In essence, each entry contains the same "amount" of information about the global structure. We say that this matrix is *incoherent* with respect to the standard basis $\{\boldsymbol{e}_i\}_{i=1}^n$.

By contrast, consider the matrix $\boldsymbol{e}_1\boldsymbol{e}_1^\top$. This matrix has a 1 in the top-left entry, and is entirely empty otherwise. The probability of sampling the one entry that contains any information in this matrix is $\frac{1}{n^2}$, requiring $\mathcal{O}(n^2 \log n)$ samples with replacement to finally gain access to the relevant matrix entry. This matrix, by contrast to above, is *coherent* with respect to the standard basis $\{\boldsymbol{e}_i\}_{i=1}^n$.

We can mathematically codify this as follows: let $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ with reduced SVD $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$. We say that $\boldsymbol{X}$ is μ-incoherent if the following pair of equations hold for some fixed constant $\mu \in \left[1, \frac{\max\{n,m\}}{r}\right]$:

$$\begin{aligned} \max_{i\in[m]} \|\boldsymbol{U}^\top \boldsymbol{e}_i\|_2^2 &\leq \frac{\mu r}{m}, \\ \max_{j\in[n]} \|\boldsymbol{V}^\top \boldsymbol{e}_j\|_2^2 &\leq \frac{\mu r}{n}. \end{aligned} \tag{1.6}$$

We note that, letting the i-th row of $\boldsymbol{U}$ be $\boldsymbol{u}_i$ and similarly for $\boldsymbol{v}_i$, that Equation (1.6) are equivalent to $\|\boldsymbol{u}_i\|_2^2 \leq \frac{\mu r}{m}$ and $\|\boldsymbol{v}_i\|_2^2 \leq \frac{\mu r}{n}$. See Figure 1.1 for a geometric visualization of incoherent versus coherent subspaces.

From this perspective, it becomes clear that the algorithmic recovery will be, in some way, a function of the incoherence of the ground truth matrix. Adapting this perspective to the EDMC problem will be critical, and will require an analogous definition of incoherence as in Section 1.6.

1.2.2 Theoretical Recovery through LRMC

The nuclear norm objective in (1.5) is very neatly characterized and has produced some very strong results. The primary result that we will discuss here is from [Rec11b]. This elegant

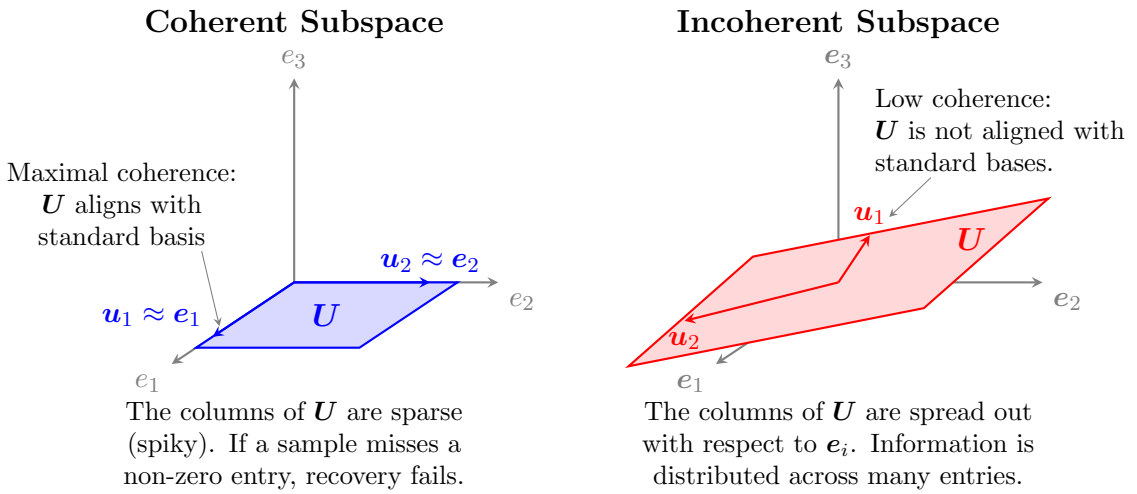


Figure 1.1: Geometric interpretation of subspace incoherence in matrix completion.

approach to matrix completion leveraged some fascinating tools in random matrix theory, first courtesy of [AW02]. These will use the theory of concentration of independent random matrices, which we discuss in further detail in Section 1.7.

To understand the theoretical guarantees for matrix completion, it is helpful to contrast them with classical results from sparse recovery [WM22]. In sparse recovery, the goal is to reconstruct a sparse vector $\boldsymbol{x}$ given access to linear measurements $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}$. The standard Restricted Isometry Property (RIP) states that for some $\delta_k \in (0, 1)$, all k-sparse vectors $\boldsymbol{x} \in \mathbb{R}^n$ satisfy

$$(1 - \delta_k)\|\boldsymbol{x}\|_2^2 \leq \|\boldsymbol{A}\boldsymbol{x}\|_2^2 \leq (1 + \delta_k)\|\boldsymbol{x}\|_2^2.$$

To construct a matrix analogue to the sparse vector recovery problem, known as the affine rank minimization problem, one observes $\boldsymbol{y} = \mathcal{A}(\boldsymbol{M})$ where $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$. The standard rank-$r$ RIP condition requires that for a constant $\delta_r \in (0, 1)$, all matrices $\boldsymbol{X}$ with $\operatorname{rank}(\boldsymbol{X}) \leq r$ satisfy

$$(1 - \delta_r)\|\boldsymbol{X}\|_{\mathrm{F}}^2 \leq \|\mathcal{A}(\boldsymbol{X})\|_2^2 \leq (1 + \delta_r)\|\boldsymbol{X}\|_{\mathrm{F}}^2.$$

This condition holds with high probability if $\mathcal{A}$ is constructed from certain types of random ensembles, such as Gaussian matrices, as these globally mix the entries of $\boldsymbol{X}$ upon measurement.

However, as discussed in Section 1.2.1, the sampling operator $\mathcal{P}_\Omega$ in matrix completion does not globally mix entries. It is easy to construct a coherent, low-rank matrix $\boldsymbol{Y}$ (such as $\boldsymbol{e}_1\boldsymbol{e}_1^\top$)

such that $\mathcal{P}_\Omega(\boldsymbol{Y}) \approx \mathbf{0}$ unless the size of Ω is trivially large. As such, the standard RIP condition is insufficient to capture this measurement inefficiency, as it fails for highly coherent matrices. Instead, a necessary condition for recovery of a matrix through (1.5) relies on restricting the operator to the tangent space of the true matrix:

Definition 1.2.1: Restricted Isometry Property (RIP)

Let $\boldsymbol{X}$ be a rank-r matrix with tangent space $\mathbb{T}$ on the manifold of rank-r matrices, $\mathcal{N}_r$. We say that a random sampling operator $\mathcal{P}_\Omega$ satisfies RIP if the following holds with high probability for some small $\varepsilon > 0$:

$$\|\mathcal{P}_\mathbb{T}\mathcal{P}_\Omega\mathcal{P}_\mathbb{T} - c\mathcal{P}_\mathbb{T}\| \leq \varepsilon,$$

where $c > 0$ is a constant.

In order to prove RIP-type statements, we leverage tools from random matrix theory. In particular, if we define Ω as a subset of $[m] \times [n]$ that is sampled uniformly at random with replacement, we can consider $\mathcal{P}_\Omega(\cdot) = \sum_{(i,j)\in\Omega}\langle\cdot, \boldsymbol{e}_{ij}\rangle\boldsymbol{e}_{ij}$ to be a sum of i.i.d. random operators. This opens up the use of the Operator Bernstein inequality, discussed in greater detail in Section 1.7. Proving concentration statements of this form constitutes the bulk of the theoretical work in this thesis.

By combining the RIP condition with the incoherence assumption, foundational papers establish that if $|\Omega| \geq C\mu nr\log^2 n$ for a sufficiently large constant C, the convex program in (1.5) exactly recovers the ground truth matrix $\boldsymbol{X}$ with high probability. The proof of this recovery heavily leverages the decomposition of the sampling operator $\mathcal{P}_\Omega$ into a sum of independent rank-one random operators, followed by an application of the Matrix Bernstein inequality (reproduced in Theorem 1.7.2). While other mathematical techniques for proving recovery were pioneered in early works like [CR09], since the introduction of the framework in [Rec11a], the Matrix Bernstein approach has become the standard analytical tool for convergence guarantees. However, solving this convex program is computationally expensive for large matrices, which has naturally led to the widespread consideration of non-convex methodologies. We will not provide a comprehensive overview of the various non-convex approaches to standard matrix completion

here, but instead point interested readers to excellent surveys such as [Wal21, WM22].

1.3 The Dual-Basis Approach to EDMC

To adapt the work described in Section 1.2.2, we need to find a way to cast the EDMC problem as a matrix completion problem. The first way to do this, and the most obvious, is to just directly complete $\boldsymbol{D}$ from observations $\{D_{ij}\}_{(i,j)\in\Omega}$ and to use a nuclear-norm minimization from (1.5). This approach does technically work, although it is suboptimal for a couple of reasons.

The first reason is that, for a d-dimensional set of points $\boldsymbol{P} \in \mathbb{R}^{n\times d}$, $\operatorname{rank}(\boldsymbol{D}) = d + 2$. This observation follows directly from (1.2), as $\operatorname{rank}(\boldsymbol{X}) = d$ and we are adding two additional rank-one vectors. A completion algorithm that operates directly on the Gram matrix $\boldsymbol{X}$ will thereby have a lower sample complexity than directly completing the distance matrix. This observation is the basis for the development of some of the original EDMC theory using the framework of low-rank matrix completion, seen in [LL17, TL18].

Another more relevant reason that considering completion on $\boldsymbol{X}$ is numerically preferable to $\boldsymbol{D}$ is specifically relevant for non-convex solves to the EDMC problem. Consider the case of Algorithm 2 applied on $\boldsymbol{D}$ alone. As this is an unconstrained optimization problem, there is no guarantee that the iterates $\boldsymbol{D}_l$ are going to be pairwise distance matrices. As discussed previously, pairwise distance matrices need to satisfy an entrywise triangle inequality, and without explicitly enforcing this condition in a convex program this is a difficult condition to match. As such, algorithms like Algorithm 2 perform poorly when attempting to directly complete pairwise distance matrices algorithmically. The goal of this section is to outline how to instead consider the completion problem not as a completion problem on $\boldsymbol{D}$, but rather a completion problem on $\boldsymbol{X}$ given access to linear combinations of entries of $\boldsymbol{X}$, and not the raw entries of $\boldsymbol{X}$ themselves.

From MDS theory discussed in Section 1.1 (and the simple expansion of the ℓ_2 norm), we know that there is a clear relationship between each entry of the squared distance matrix to the Gram matrix given by

$$D_{ij} = X_{ii} + X_{jj} - X_{ij} - X_{ji}.$$

In [TL18], an approach for re-casting completion on $\boldsymbol{D}$ to completion on $\boldsymbol{X}$ was constructed as follows: given $\boldsymbol{\alpha} = (i, j), i < j$, we define the matrix $\boldsymbol{w}_{\boldsymbol{\alpha}}$ as

$$\begin{aligned}\boldsymbol{w}_{\boldsymbol{\alpha}} &= \boldsymbol{e}_{ii} + \boldsymbol{e}_{jj} - \boldsymbol{e}_{ij} - \boldsymbol{e}_{ji} \\ &= (\boldsymbol{e}_i - \boldsymbol{e}_j)(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top.\end{aligned}$$

Using the collection of matrices $\{\boldsymbol{w}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$, where $\mathbb{I} = \{(i, j), 1 \leq i < j \leq n\}$, we can realize each entry of the squared distance matrix as the trace inner product of the Gram matrix with the basis as for $\boldsymbol{\alpha} = (i, j)$

$$\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle = X_{ii} + X_{jj} - X_{ji} - X_{ij} = D_{ij}.$$

Furthermore, it turns out that $\{\boldsymbol{w}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$ is a non-orthogonal basis for the subspace of symmetric matrices with zero row sum, denoted $\mathbb{S} = \{\boldsymbol{Y} \in \mathbb{R}^{n\times n} \mid \boldsymbol{Y} = \boldsymbol{Y}^\top, \boldsymbol{Y}\boldsymbol{1} = \boldsymbol{0}\}$, where

$$\dim(\mathbb{S}) = \frac{n(n-1)}{2} =: L. \tag{1.7}$$

In fact, for any two pairs of indices $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{I}$ with $\boldsymbol{\alpha} = (i, j)$ and $\boldsymbol{\beta} = (k, l)$, we have:

$$\begin{aligned}\langle \boldsymbol{w}_{\boldsymbol{\alpha}}, \boldsymbol{w}_{\boldsymbol{\beta}} \rangle &= \mathrm{Tr}(\boldsymbol{w}_{\boldsymbol{\alpha}}\boldsymbol{w}_{\boldsymbol{\beta}}) \\ &= \mathrm{Tr}\left((\boldsymbol{e}_i - \boldsymbol{e}_j)(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top(\boldsymbol{e}_k - \boldsymbol{e}_l)(\boldsymbol{e}_k - \boldsymbol{e}_l)^\top\right) \\ &= \mathrm{Tr}\left((\boldsymbol{e}_i - \boldsymbol{e}_j)(\delta_{ik} - \delta_{il} - \delta_{jk} + \delta_{jl})(\boldsymbol{e}_k - \boldsymbol{e}_l)^\top\right) \\ &= \mathrm{Tr}\left((\boldsymbol{e}_k - \boldsymbol{e}_l)^\top(\boldsymbol{e}_i - \boldsymbol{e}_j)(\delta_{ik} - \delta_{il} - \delta_{jk} + \delta_{jl})\right) \\ &= (\delta_{ki} - \delta_{kj} - \delta_{li} + \delta_{lj})(\delta_{ik} - \delta_{il} - \delta_{jk} + \delta_{jl}) \\ &= \begin{cases} 4 & \boldsymbol{\alpha} = \boldsymbol{\beta}; \\ 1 & \boldsymbol{\alpha} \neq \boldsymbol{\beta},\ \boldsymbol{\alpha} \cap \boldsymbol{\beta} \neq \emptyset; \\ 0 & \boldsymbol{\alpha} \cap \boldsymbol{\beta} = \emptyset, \end{cases}\end{aligned}$$

where δ_{ij} is the Kronecker delta.

As $\{\boldsymbol{w}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$ does not form an orthonormal basis over $\mathbb{S}$, we consider its bi-orthogonal, or dual, basis for both theoretical and computational purposes. We provide a brief description of

dual bases in Section A.2 for posterity. We denote this basis as $\{\boldsymbol{v}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$, and represent any centered Gram matrix $\boldsymbol{X}$ using the following expansion:

$$\boldsymbol{X} = \sum_{\boldsymbol{\alpha}} \langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \boldsymbol{v}_{\boldsymbol{\alpha}}.$$

The advantage of the dual basis representation is that it allows us to recast the EDMC problem as a low-rank matrix recovery problem where we observe a subset of the expansion coefficients. In [TL18], this dual basis formulation has been used to provide theoretical guarantees for the convex program given in (1.5). To make use of the dual basis approach both in theory and applications, one of the first steps is to have a representation of the dual basis that is easier to use. Let $\boldsymbol{H} \in \mathbb{R}^{L\times L}$ be the Gram matrix with

$$H_{\boldsymbol{\alpha}\boldsymbol{\beta}} = \langle \boldsymbol{w}_{\boldsymbol{\alpha}}, \boldsymbol{w}_{\boldsymbol{\beta}} \rangle.$$

Furthermore, let

$$H^{\boldsymbol{\alpha}\boldsymbol{\beta}} = \left(\boldsymbol{H}^{-1}\right)_{\boldsymbol{\alpha}\boldsymbol{\beta}} = \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{w}_{\boldsymbol{\beta}} \rangle.$$

From Section A.2, we have a closed form expression for $\boldsymbol{v}_{\boldsymbol{\alpha}}$:

$$\boldsymbol{v}_{\boldsymbol{\alpha}} = \sum_{\boldsymbol{\beta}\in\mathbb{I}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \boldsymbol{w}_{\boldsymbol{\beta}}. \tag{1.8}$$

To compute this explicitly from just knowledge of $\boldsymbol{H}$ relies on an inverse of an $L \times L$ matrix, which requires the solution of a large linear system. Thankfully, in [LT23] it was shown that the dual basis admits a simple explicit form

$$\boldsymbol{v}_{\boldsymbol{\alpha}} = -\frac{1}{2}\left(\boldsymbol{a}\boldsymbol{b}^\top + \boldsymbol{b}\boldsymbol{a}^\top\right), \tag{1.9}$$

where $\boldsymbol{a} = \boldsymbol{e}_i - \frac{1}{n}\mathbf{1}$ and $\boldsymbol{b} = \boldsymbol{e}_j - \frac{1}{n}\mathbf{1}$ for $\boldsymbol{\alpha} = (i, j)$. We now highlight a few operators that are related to the dual basis approach. The first one is the sampling operator $\mathcal{R}_\Omega : \mathbb{S} \to \mathbb{S}$ defined as follows:

$$\mathcal{R}_\Omega(\cdot) = \sum_{\boldsymbol{\alpha}\in\Omega} \langle \cdot, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \boldsymbol{v}_{\boldsymbol{\alpha}}.$$

The bi-orthogonality relationship of the dual basis gives that $\mathcal{R}_\Omega^2 = \mathcal{R}_\Omega$ if Ω does not have repeated indices, and that

$$\mathcal{R}_\Omega^\star(\cdot) = \sum_{\boldsymbol{\alpha}\in\Omega} \langle \cdot, \boldsymbol{v}_{\boldsymbol{\alpha}} \rangle \boldsymbol{w}_{\boldsymbol{\alpha}}.$$

Due to the lack of self-adjointness, $\mathcal{R}_\Omega$ without repeated indices in Ω is not an orthogonal projection operator, and is instead an oblique projection operator. In [SLCT23], $\mathcal{R}_\Omega(\boldsymbol{X})$ is related to the sampling operator $\mathcal{P}_\Omega(\boldsymbol{D})$ as follows:

$$\mathcal{R}_\Omega(\boldsymbol{X}) = -\frac{1}{2}\boldsymbol{J}\mathcal{P}_\Omega(\boldsymbol{D})\boldsymbol{J}, \tag{1.10}$$

where $\boldsymbol{J}$ is as defined in (1.1). The next operator is the restricted frame operator $\mathcal{F}_\Omega : \mathbb{S} \to \mathbb{S}$, first studied in [TL18], and defined as

$$\mathcal{F}_\Omega(\cdot) = \sum_{\boldsymbol{\alpha}\in\Omega} \langle \cdot, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \boldsymbol{w}_{\boldsymbol{\alpha}}. \tag{1.11}$$

This operator is self-adjoint, positive semi-definite, but unlike $\mathcal{R}_\Omega$, does not reference the dual basis. We note that this operator under a different name was critical to the analysis of the algorithm in [LS24].

We briefly describe the results of [TL18], which inspired the direction of much of this research. In particular, [TL18] established that the exact recovery of Gram matrices from incomplete pairwise distance measurements is theoretically guaranteed via nuclear-norm minimization. This is done using a dual-basis formulation by considering

$$\min_{\boldsymbol{Y}\in\mathbb{S},\boldsymbol{Y}\succeq\boldsymbol{0}} \|\boldsymbol{Y}\|_* \quad \text{subject to } \mathcal{R}_\Omega(\boldsymbol{Y}) = \mathcal{R}_\Omega(\boldsymbol{X}). \tag{1.12}$$

Similar to (1.5), (1.12) will recover the ground truth matrix $\boldsymbol{X}$ with high probability if $|\Omega| \geq C\nu^2 r^2 n \log^2 n$, where ν is an EDMC-specific incoherence constant discussed in Section 1.6. This strategy suffers from the same computational issues as (1.5), and as such the investigation into provably convergent non-convex surrogates is the main direction of interest in this thesis. We summarize this in the following question:

Question 1.3.1: Non-Convex EDMC Question

Can we construct a provably convergent, fast, non-convex algorithm for the EDMC problem under a uniform sampling scheme?

Chapter 2, Chapter 3, and Section A.1 will be dedicated to answering this question in the affirmative. The approach taken in this thesis will be a first-order Riemannian optimization scheme on the manifold of rank-r matrices for some fixed r. We dedicate the next two sections to background information on Riemannian optimization, and a Riemannian optimization scheme for the LRMC problem.

1.4 First-Order Optimization on Manifolds

Traditional optimization techniques primarily focus on the minimization of functionals defined in Euclidean space, and subsets thereof. That is, algorithms are mainly designed to find the optima of functions that map from convex subsets of $\mathbb{R}^n$ to $\mathbb{R}$, with an incredibly rich theory of convergence across the algorithmic zoo [NW06]. However, this perspective ignores the underlying geometry that a particular problem might possess. Consider the following concrete example:

$$\underset{\boldsymbol{x}\in\mathbb{R}^d}{\text{minimize}} \quad f(\boldsymbol{x}) \quad \text{subject to}\, \boldsymbol{x}^\top \boldsymbol{x} = 1.$$

This objective function is a constrained optimization program. Now, this constraint is rather simple and not difficult to consider how to solve in and of itself, but it might still be simpler to recognize that this objective function is equivalent to

$$\underset{\boldsymbol{x}\in S^{d-1}}{\text{minimize}} \; f(x),$$

where S^{d-1} is the sphere in d-dimensions. We notice now that this problem is a simple, unconstrained optimization problem whose input set is a manifold. This approach motivates the broader field known as Riemannian optimization, which has emerged in recent years as a powerful technique for solving optimization problems for non-convex functions defined on manifolds. By taking into account the curvature and other geometric properties of the manifold, Rie-

mannian optimization algorithms can provide better convergence rates and improved accuracy compared to standard Euclidean optimization methods [Bou23, Sat21].

We will provide a brief description of how first-order retraction methodologies defined on manifolds work.

1.4.1 Geometric preliminaries

Throughout this report, we will use $\mathcal{N}$ to denote a smooth, connected manifold that is embedded in a Euclidean vector space, i.e. $\mathcal{N} \subset \mathbb{R}^n$. By smooth, we mean that the manifold can be defined as the zero-set of some function $f \in C^\infty(\mathbb{R}^n)$. We will refer to tangent spaces at a point $\boldsymbol{x} \in \mathcal{N}$ as

$$T_{\boldsymbol{x}}\mathcal{N} = \{\boldsymbol{v} \in \mathbb{R}^n : \boldsymbol{x}^\top \boldsymbol{v} = 0\},$$

and the tangent bundle as

$$T\mathcal{N} = \{(\boldsymbol{x}, \boldsymbol{v}) : x \in \mathcal{N}, \boldsymbol{v} \in T_{\boldsymbol{x}}\mathcal{N}\}.$$

Elements in tangent spaces are vectors, and vector fields are smooth, vector-valued functions that live along a manifold. More specifically for a given vector field X, at each point $\boldsymbol{p} \in \mathcal{N}$, $X(p) \in T_{\boldsymbol{p}}\mathcal{N}$. The set of smooth vector fields on $\mathcal{N}$ is denoted by $\mathfrak{X}(\mathcal{N})$.

We denote curves c on a manifold $\mathcal{N}$ via $c : I \to \mathcal{N}$ where $I = [a, b]$ for some $a < b \in \mathbb{R}$. Additionally, we define a retraction map at each point $R_{\boldsymbol{x}} : T_{\boldsymbol{x}}\mathcal{N} \to \mathcal{N}$ as a smooth map from a tangent space back to the manifold. This map defines a curve on the manifold from vectors in the tangent space. To define a retraction independent of an individual point, we say a retraction R on $\mathcal{N}$ is a smooth map that picks out a particular curve for each $(\boldsymbol{x}, \boldsymbol{v}) \in T_{\boldsymbol{x}}\mathcal{N}$ such that

$$R : T\mathcal{N} \to \mathcal{N} : (\boldsymbol{x}, \boldsymbol{v}) \to R_{\boldsymbol{x}}(\boldsymbol{v}),$$

where each curve $c(t) = R_{\boldsymbol{x}}(t\boldsymbol{v})$ and $c(0) = \boldsymbol{x}$, $c'(0) = \boldsymbol{v}$. One specific type of retraction map is the exponential map. The exponential map goes from the tangent space to a geodesic on the manifold, where a geodesic is a constant-speed, locally distance-minimizing curve.

Now, let $F : \mathcal{N} \to \mathcal{N}'$ be a smooth map. We define the differential of F, denoted by DF,

as a linear map defined by $DF : T_{\boldsymbol{x}}\mathcal{N} \to T_{F(\boldsymbol{x})}\mathcal{N}$ via

$$DF(\boldsymbol{x})[\boldsymbol{v}] = \frac{d}{dt}F(c(t))\bigg|_{t=0} = (F \circ c)'(0),$$

where $c(0) = \boldsymbol{x}$ and $c'(0) = \boldsymbol{v}$. To make a manifold a Riemannian manifold, the manifold must be endowed with an inner product structure for each tangent space that smoothly varies with $x \in \mathcal{N}$. That is, the metric at a point $\boldsymbol{x} \in \mathcal{N}$ is given as, for two vectors $\boldsymbol{u}, \boldsymbol{v} \in T_{\boldsymbol{x}}\mathcal{N}$, as $\langle \boldsymbol{u}, \boldsymbol{v} \rangle_x = \boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{v}$ for some symmetric $\boldsymbol{A} \succ 0$. To define the Riemannian gradient of a smooth function $f : \mathcal{N} \to \mathbb{R}$, we can employ the differential and define the gradient of f, denoted $\mathrm{grad} f(x) \in T_{\boldsymbol{x}}\mathcal{N}$, as the unique tangent vector at $\boldsymbol{x} \in \mathcal{N}$ such that for any $\boldsymbol{v} \in T_{\boldsymbol{x}}\mathcal{N}$,

$$Df(\boldsymbol{x})[\boldsymbol{v}] = \langle \boldsymbol{v}, \mathrm{grad} f \rangle_{\boldsymbol{x}}.$$

In the case where $\mathcal{N}$ is a Riemannian manifold embedded in some larger Euclidean space, one can show that $\mathrm{grad} f$ is the orthogonal projection of the usual gradient onto the tangent space in the sense that f has been smoothly extended into all of $\mathbb{R}^n$ and its classical gradient is computed in the standard way.

Up to now, we have only considered notions of a first derivative on a manifold in the form of the differential. For the proof of convergence of the Riemannian gradient descent algorithm, we will need to define more clearly notions of a second derivative on a manifold, namely the Riemannian Hessian of our objective function, and to do that we need to discuss connections on manifolds. Intuitively, a connection at a point $\boldsymbol{x} \in \mathcal{N}$ is a projection of the differential of a smooth extension of a vector field onto $T_{\boldsymbol{x}}\mathcal{N}$. This projection is needed to establish a notion of directional derivatives of vector fields that stay tangent to the manifold. More formally, a connection ∇ is an $\mathbb{R}$-bilinear map from $\mathfrak{X}(\mathcal{N}) \times \mathfrak{X}(\mathcal{N}) \to \mathfrak{X}(\mathcal{N})$. More properties of the connection can be found in [Tu10] but an intuitive grasp is adequate for this thesis.

Now, the Riemannian Hessian $\mathrm{Hess} f(\boldsymbol{x})$ of a smooth function $f : \mathcal{N} \to \mathbb{R}$ is a linear map from $T_{\boldsymbol{x}}\mathcal{N} \to T_{\boldsymbol{x}}\mathcal{N}$ defined as

$$\mathrm{Hess} f(\boldsymbol{x})[\boldsymbol{v}] = \nabla_{\boldsymbol{v}} \mathrm{grad} f.$$

One can show using the properties of the Riemannian connection and metric that this

operator Hess$f$ is self-adjoint.

The final concept to introduce here is that of vector transport. For $\boldsymbol{\eta} \in T_{\boldsymbol{x}}\mathcal{N}$, a vector transport map $\boldsymbol{T}_\eta$ is a linear map takes a vector $\boldsymbol{\xi} \in T_{\boldsymbol{x}}\mathcal{N}$ to $\boldsymbol{T}_{\boldsymbol{\eta}}(\boldsymbol{\xi}) \in T_{R_{\boldsymbol{x}}(\boldsymbol{\eta})}\mathcal{N}$. In words, this map can be thought of as moving $\boldsymbol{\xi}$ into the tangent space determined by some retraction map $R_{\boldsymbol{x}}$ in the direction of $\boldsymbol{\eta}$. A specific type of vector transport map is a parallel transport map, which uniquely transports vectors along curves in such a way that the vectors stay parallel to each other with respect to the connection, but this can be difficult to compute [SKM19].

1.4.2 Riemannian Gradient Descent

The following section has drawn upon Chapter 4 of [Bou23]. In this section, we will briefly investigate a first-order retraction-based algorithm for minimizing a smooth objective function $f : \mathcal{N} \to \mathbb{R}$. We define this Riemannian Gradient Descent algorithm in Algorithm 1.

Algorithm 1 Riemannian Gradient Descent

Input: $\boldsymbol{x}_0 \in \boldsymbol{M}$
for $k = 0, 1, 2, ...$ **do**
  Pick $\alpha_k > 0$
  $\boldsymbol{s}_k = -\alpha_k \mathbf{grad} f(\boldsymbol{x}_k)$
  $\boldsymbol{x}_{k+1} = R_{\boldsymbol{x}_k}(\boldsymbol{s}_k)$
end for

The intuition behind this algorithm is simple: For each $\boldsymbol{x}_k \in \boldsymbol{M}$, examine $T_{\boldsymbol{x}}\mathcal{N}$ and find the steepest descent direction in the tangent space. Pick a step size, then retract back to the manifold for the next iteration. One thing to note about this algorithm is that the retraction map is a first-order approximation to the exponential map back to the manifold. This can be seen from a Taylor-series expansion. A generic retraction map in many cases is easier to compute than the exponential map. We will also impose the following assumptions:

1. There exists $f_{\text{low}} \in \mathbb{R}$ such that $f(\boldsymbol{x}) \geq f_{\text{low}}$ $\forall \boldsymbol{x} \in \mathcal{N}$.

2. At each iteration, the algorithm achieves sufficient decrease for f in the sense that, for some $c > 0$,

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \geq c\|\text{grad} f(\boldsymbol{x}_k)\|^2.$$

3. For a given subset $S \subset T\mathcal{N}$, there exists a constant $L > 0$ such that, for all $(\boldsymbol{x}, \boldsymbol{s}) \in S$

$$f(R_{\boldsymbol{x}}(s)) \leq f(\boldsymbol{x}) + \langle \text{grad} f(\boldsymbol{x}), \boldsymbol{s} \rangle_{\boldsymbol{x}} + \frac{L}{2}\|\boldsymbol{s}\|^2.$$

The first assumption of a global lower bound is a common one in optimization and not an unreasonable requirement. The third requirement is the Riemannian generalization of L-smoothness of a convex function and is itself reasonable. The second assumption on its own seems a bit unmotivated, but is an important condition to ensure that the limit of the sequence $\lim_{k\to\infty} \|\text{grad} f(\boldsymbol{x}_k)\| \to 0$. Once these assumptions are satisfied, local convergence of Algorithm 1 can be established as follows:

Theorem 1.4.1 (Local Convergence of Algorithm 1 [Bou23])**.** *Let $\mathcal{N}$ be a Riemannian manifold with a retraction R, and let $f : \mathcal{N} \to \mathbb{R}$ be a smooth function. Assume $\boldsymbol{x}^\star \in \mathcal{N}$ satisfies $gradf(\boldsymbol{x}^\star) = 0$ and $Hessf(\boldsymbol{x}^\star) \succ 0$. Let $\kappa = \frac{\lambda_{max}}{\lambda_{min}}$ be the condition number of $Hessf(\boldsymbol{x}^\star)$ with λ_{max} and λ_{min} being the largest and smallest eigenvalues, respectively. With $L > \frac{1}{2}\lambda_{max}$, set $\alpha_k = \alpha = \frac{1}{L}$. Given $F(\boldsymbol{x}) = R_{\boldsymbol{x}}(-\alpha gradf(x))$ and $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k$, there exists a neighborhood of $\boldsymbol{x}^\star$ such that if the sequence $\{\boldsymbol{x}_k\}$ enters the neighborhood, it converges to $\boldsymbol{x}^\star$ linearly. If $L = \lambda_{max}$, the convergence factor is at most $\frac{\kappa-1}{\kappa}$.*

By this theorem, Algorithm 1 is an algorithm with local linear convergence. The main difficulties in the computation of Algorithm 1 is the computation of grad$f$ and the retraction map at each point $R_{\boldsymbol{x}_k}$. We sketch out a step of Algorithm 1 in Figure 1.2.

1.5 A Riemannian Approach to LRMC

One prior that we can leverage in constructing non-convex solutions to the LRMC problem is the idea that a ground-truth matrix $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ has a known rank of r. This allows us to more naturally consider the low-rank matrix completion problem as a least-squares problem given via

$$\underset{\boldsymbol{Y}\in\mathbb{R}^{m\times n}}{\text{minimize}} \ \|\mathcal{P}_\Omega(\boldsymbol{Y} - \boldsymbol{X})\|_{\text{F}}^2 \quad \text{subject to} \quad \text{rank}(\boldsymbol{Y}) = r. \tag{1.13}$$

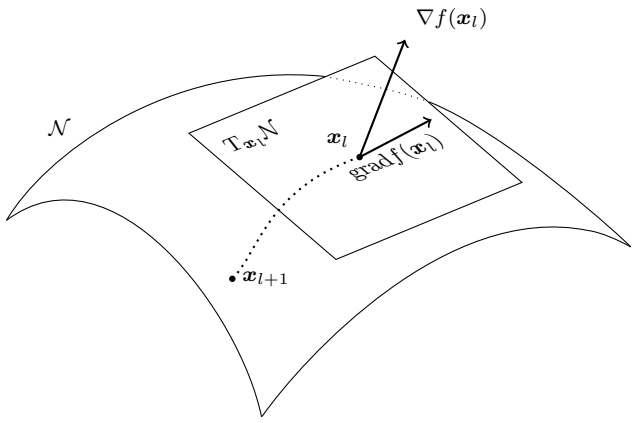


Figure 1.2: A diagram of a simple first-order retraction method on $\mathcal{N}$. Again, $\nabla f(\boldsymbol{x}_l)$ is the Euclidean gradient of f at $\boldsymbol{x}_l$, $\mathrm{grad}\, f(\boldsymbol{x}_l)$ is the Riemannian gradient at $\boldsymbol{x}_l$, and $\boldsymbol{x}_{l+1} = \boldsymbol{R}_{\boldsymbol{x}_l}(\boldsymbol{x}_l - \alpha_l \mathrm{grad}\, f(\boldsymbol{x}_l))$ for some step size $\alpha_l > 0$, as in Algorithm 1.

This was studied in [Van12], in which the author showed that the algorithm does converge to local minima. It does not, however, provide any guarantees on the optimality of recovered solutions.

In [WCCL20], a variation on this theme was considered. If we take now the objective

$$\underset{\boldsymbol{Y}\in\mathbb{R}^{m\times n}}{\text{minimize}}\ \langle \boldsymbol{Y}-\boldsymbol{X}, \mathcal{P}_\Omega(\boldsymbol{Y}-\boldsymbol{X})\rangle \quad \text{subject to} \quad \mathrm{rank}(\boldsymbol{Y}) = r, \tag{1.14}$$

more can be done algorithmically to guarantee convergence. Now, as in Section 1.2.2, it is often beneficial to consider $\mathcal{P}_\Omega$ as a sum of random rank-one operators sampled with replacement from $[m] \times [n]$; as such, it is overwhelmingly likely that there are repeated entries of $\mathcal{P}_\Omega$ and in that case $\mathcal{P}_\Omega^2 \neq \mathcal{P}_\Omega$. However, if there are no repeated entries we note that (1.14) is equivalent to (1.13).

Following Section 1.4.2 and Section A.3, we can now construct an algorithm to minimize the objective function in (1.14). In order to do so, we need to compute the Riemannian gradient of $\langle \boldsymbol{Y}-\boldsymbol{X}, \mathcal{P}_\Omega(\boldsymbol{Y}-\boldsymbol{X})\rangle$ at any $\boldsymbol{Y} \in \mathcal{N}_r$ first. First, we can compute the Euclidean gradient as $\nabla_{\boldsymbol{Y}}\langle \boldsymbol{Y}-\boldsymbol{X}, \mathcal{P}_\Omega(\boldsymbol{Y}-\boldsymbol{X})\rangle = 2\mathcal{P}_\Omega(\boldsymbol{Y}-\boldsymbol{X})$. To compute the Riemannian gradient at a point $\boldsymbol{Y}$, we just need to consider the projection onto $T_{\boldsymbol{Y}}\mathcal{N}_r$, with the projection formula seen in (1.16). Let the tangent space at the ground-truth matrix $\boldsymbol{X}$ be denoted $\mathbb{T}$, and let the tangent space at each iterate $\boldsymbol{X}_l$ be denoted $\mathbb{T}_l$, with corresponding projection operators $\mathcal{P}_{\mathbb{T}}$ and $\mathcal{P}_{\mathbb{T}_l}$, respectively. One of the benefits of working on $\mathcal{N}_r$ is the well-defined characterization of the

manifold of rank-r matrices.

Statement 1.5.1: Characterization of the Tangent Spaces of $\mathcal{N}_r$

Let $\boldsymbol{X} \in \mathcal{N}_r \subset \mathbb{R}^{m\times n}$ have a reduced SVD given by $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, where $\boldsymbol{U} \in \mathbb{R}^{m\times r}$ and $\boldsymbol{V} \in \mathbb{R}^{n\times r}$. Then

$$T_{\boldsymbol{X}}\mathcal{N}_r = \left\{ \boldsymbol{U}\boldsymbol{Z}_1^\top + \boldsymbol{Z}_2\boldsymbol{V}^\top \mid \boldsymbol{Z}_1 \in \mathbb{R}^{n\times r}, \boldsymbol{Z}_2 \in \mathbb{R}^{m\times r} \right\}. \tag{1.15}$$

Furthermore, the projection operation onto $T_{\boldsymbol{X}}\mathcal{N}_r$ can be neatly characterized as

$$\mathcal{P}_{\mathbb{T}}\boldsymbol{Y} = \mathcal{P}_U\boldsymbol{Y} + \boldsymbol{Y}\mathcal{P}_V - \mathcal{P}_U\boldsymbol{Y}\mathcal{P}_V, \tag{1.16}$$

where $\mathcal{P}_U = \boldsymbol{U}\boldsymbol{U}^\top$ and $\mathcal{P}_V = \boldsymbol{V}\boldsymbol{V}^\top$ are the projections onto the column/row spaces.

We provide more details on the geometry of $\mathcal{N}_r$ in Section A.3. The last thing needed to define a first-order retraction algorithm is to define the retraction map. When working on $\mathcal{N}_r$, the defining constraint is that the matrices in $\mathcal{N}_r$ are, of course, rank-r. To build a good retraction, one only needs to consider the hard-thresholding operator $\mathcal{H}_r : \mathbb{R}^{m\times n} \to \mathcal{N}_r$ as

$$\mathcal{H}_r(\boldsymbol{Y}) = \sum_{i=1}^{r} \sigma_i \boldsymbol{U}_i \boldsymbol{V}_i^\top, \tag{1.17}$$

where $\boldsymbol{U}_i$, $\boldsymbol{V}_i$, and σ_i are the i-th left singular vector, right singular vector, and singular value, respectively. As such, we can define the following algorithm to solve the problem posed in (1.14):

Algorithm 2 Riemannian Gradient Descent for LRMC [WCCL20]

Input: Ω, $\{X_{ij}\}_{(i,j)\in\Omega}$, $\operatorname{rank}(\boldsymbol{X}) = r$
Initialization: $\boldsymbol{X}_0 = \boldsymbol{U}_0\boldsymbol{D}_0\boldsymbol{V}_0^\top$
for $l = 0, 1, \cdots$ **do**
  $\boldsymbol{G}_l = \mathcal{P}_\Omega(\boldsymbol{X} - \boldsymbol{X}_0)$
  $\alpha_l = \frac{\left\|\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l\right\|_{\mathrm{F}}^2}{\langle \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l, \mathcal{M}_\Omega\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l\rangle}$
  $\boldsymbol{W}_l = \boldsymbol{X}_l + \alpha_l\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l$
  $\boldsymbol{X}_{l+1} = \mathcal{H}_r(\boldsymbol{W}_l)$
end for
Output: $\boldsymbol{X}_{\mathrm{rec}}$: Estimated matrix.

The step size seen in Algorithm 2 is the solution to a Euclidean exact line search in the tangent space. Other choices of step size are possible, but this one provides nice theoretical guarantees. In addition to this approach, the authors considered two initialization schemes. One is simple, and only involves setting $\boldsymbol{X}_0 = \frac{n^2}{m}\mathcal{H}_r(\mathcal{P}_\Omega(\boldsymbol{X}))$. This one-step hard thresholding is, in some sense, the closest representation of $\boldsymbol{X}$ in $\mathcal{N}_r$ based on the observed entries, and is in fact an unbiased estimator of $\boldsymbol{X}$. A more delicate initialization can be considered as well, however, by partitioning the set Ω into S equally sized subsets, and performing one Riemannian gradient descent step for each subset. This more delicate initialization breaks the dependence on each iterate from the previous, and provides a more reliable initialization for large enough sample sizes. We summarize their results here:

Theorem 1.5.1 (Convergence guarantees of Algorithm 2 [WCCL20]). *Let $|\Omega| = m$, and let $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ be the μ-incoherent rank-r ground truth solution. If*

$$m \geq C\beta^{3/2}\mu r n^{3/2}\log^{3/2}(n),$$

for some sufficiently large constant C, with probability at least $1 - 2n^{2-2\beta} - 2n^{1-\beta}$ when considering a one-step hard thresholding initialization we have that Algorithm 2 recovers the ground truth $\boldsymbol{X}$. For the Riemannian resampling scheme, the sample complexity result improves to

$$m \geq C'\mu r\beta n\log(n),$$

for some matrix specific constant C' with probability at least $1 - 2n^{1-\beta} - cn^{2-2\beta}$ for some $c > 0$.

In order to prove this statement, the authors show that in an attractive basin around the ground truth matrix, Algorithm 2 converges linearly to the true solution, an improvement over existing Riemannian methods such as [Van12]. They then show that the aforementioned initialization methods with high enough sample complexity are, with high probability, in this attractive basin.

1.6 Incoherence and Geometry

As we have seen in Section 1.2.1, matrix incoherence exhibits strong influence on the sample complexity guarantees of LRMC algorithms. In this section, we investigate the dual-basis analogue to (1.6) and the geometric implications of a similar condition. We note that this section was heavily adapted from [SCT25].

In pathological cases, the ground truth matrix $\boldsymbol{X}$ may exhibit a sparse representation in the basis $\{\boldsymbol{w}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$, which could lead to challenges in its recovery from sampled measurements. While the concept of incoherence is well-established in the standard matrix completion literature, the condition specific to the EDMC problem slightly differs in structure and admits a natural geometric interpretation. This section is devoted to a detailed examination of this geometric perspective. We will state more formally the incoherence assumptions in Section 2.2 and Section 3.2, but we will first introduce one of the conditions below. We say that a rank-r Gram matrix $\boldsymbol{X} \in \mathbb{R}^{n\times n}$ is ν-incoherent with respect to $\{\boldsymbol{w}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$ if the following statement holds:

$$\max_{\boldsymbol{\alpha}\in\mathbb{I}} \| \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \leq \sqrt{\frac{4\nu r}{n}}. \tag{1.18}$$

We remark that the above is inspired by the standard incoherence condition, which up to a scaling factor, states that

$$\max_{i\in[n]} \| \mathcal{P}_U \boldsymbol{e}_i\|_2 \leq \sqrt{\frac{4\nu r}{n}}. \tag{1.19}$$

The standard incoherence assumption, shown in (1.19), is prevalent throughout matrix completion literature and is a measure of "entrywise diffuseness" in the ground truth matrix. Further discussion of standard matrix incoherence can be seen in [CR09].

The incoherence condition introduced in (1.18) can be interpreted in terms of the underlying point cloud data. For the specific case of the EDMC problem, (1.18) can be expanded as follows for $\boldsymbol{\alpha} = (i, j)$ with $i < j$:

$$\begin{aligned}
\|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 &= \langle \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}, \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \\
&= \mathrm{Tr}\left(\boldsymbol{w}_{\boldsymbol{\alpha}}\boldsymbol{w}_{\boldsymbol{\alpha}}\boldsymbol{U}\boldsymbol{U}^\top\right) \\
&= \mathrm{Tr}\left(2\boldsymbol{w}_{\boldsymbol{\alpha}}\boldsymbol{U}\boldsymbol{U}^\top\right)
\end{aligned}$$

$$
\begin{aligned}
&= 2\left\langle \boldsymbol{e}_{ii} - \boldsymbol{e}_{ij} + \boldsymbol{e}_{jj} - \boldsymbol{e}_{ji}, \boldsymbol{U}\boldsymbol{U}^\top \right\rangle \\
&= 2\left((\boldsymbol{U}\boldsymbol{U}^\top)_{ii} + (\boldsymbol{U}\boldsymbol{U}^\top)_{jj} - (\boldsymbol{U}\boldsymbol{U}^\top)_{ij} - (\boldsymbol{U}\boldsymbol{U}^\top)_{ji} \right) \\
&= 2\left(\boldsymbol{u}_i{}^\top \boldsymbol{u}_i + \boldsymbol{u}_j{}^\top \boldsymbol{u}_j - \boldsymbol{u}_j{}^\top \boldsymbol{u}_i - \boldsymbol{u}_i{}^\top \boldsymbol{u}_j \right) \\
&= 2\left(\boldsymbol{u}_i{}^\top (\boldsymbol{u}_i - \boldsymbol{u}_j) + \boldsymbol{u}_j{}^\top (\boldsymbol{u}_j - \boldsymbol{u}_i) \right) \\
&= 2\,(\boldsymbol{u}_i - \boldsymbol{u}_j)^\top (\boldsymbol{u}_i - \boldsymbol{u}_j)\,. \qquad (1.20)
\end{aligned}
$$

The incoherence condition can then equivalently be stated as

$$\max_{(i,j):i<j} (\boldsymbol{u}_i - \boldsymbol{u}_j)^\top (\boldsymbol{u}_i - \boldsymbol{u}_j) \leq 2\frac{\nu r}{n}. \qquad (1.21)$$

The next Lemma provides the lower and upper bounds for ν.

Lemma 1.6.1. *For the incoherence condition in* (1.21)*, ν is bounded below by $1 + \frac{1}{n-1}$ and above by $2\frac{n}{r}$.*

Proof. We consider $\sum_{(i,j):i<j} (\boldsymbol{u}_i - \boldsymbol{u}_j)^\top (\boldsymbol{u}_i - \boldsymbol{u}_j)$. Note that

$$\sum_i (\boldsymbol{u}_i)^\top \boldsymbol{u}_i = \mathrm{Tr}(\boldsymbol{U}\boldsymbol{U}^\top) = \mathrm{Tr}(\boldsymbol{U}^\top \boldsymbol{U}) = r.$$

Since we assume centered configurations, $\boldsymbol{U}^\top \mathbf{1} = \mathbf{0}$. It then follows that, $\sum_{i,j} (\boldsymbol{u}_i)^\top \boldsymbol{u}_j = (\sum_i \boldsymbol{u}_i)^\top \sum_j \boldsymbol{u}_j = 0$. Using these two relations, we obtain:

$$\sum_{(i,j):i<j} (\boldsymbol{u}_i - \boldsymbol{u}_j)^\top (\boldsymbol{u}_i - \boldsymbol{u}_j) = \frac{1}{2} \sum_{(i,j)} (\boldsymbol{u}_i - \boldsymbol{u}_j)^\top (\boldsymbol{u}_i - \boldsymbol{u}_j) = \frac{1}{2} 2nr = nr.$$

The above equality shows that the sum of L terms is nr. Therefore, the maximum summand must be at least $\frac{nr}{L}$. In particular, we have:

$$\max_{(i,j):i<j} (\boldsymbol{u}_i - \boldsymbol{u}_j)^\top (\boldsymbol{u}_i - \boldsymbol{u}_j) \geq \frac{nr}{L} = 2\frac{r}{n}\left(1 + \frac{1}{n-1}\right).$$

Therefore, the minimum value of the incoherence parameter ν is $1 + \frac{1}{n-1}$. To find the maximum value of the incoherence, using the parallelogram inequality, $(\boldsymbol{u}_i - \boldsymbol{u}_j)^\top (\boldsymbol{u}_i - \boldsymbol{u}_j) \leq 2||\boldsymbol{u}_i||^2 + 2||\boldsymbol{u}_j||^2 \leq 4$. Therefore, the upper bound for $\boldsymbol{v}$ is $2\frac{n}{r}$. □

Remark 1.6.1: EDMC Incoherence Lower Bound

To show that the lower bound for the incoherence can be attained, we consider the following example:

$$\boldsymbol{U} = \sqrt{\frac{2}{3}} \begin{bmatrix} 1 & 0 \\ -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{1}{2} & -\frac{\sqrt{3}}{2} \end{bmatrix}.$$

Up to the scaling factor of $\sqrt{\frac{2}{3}}$, the rows of $\boldsymbol{U}$ correspond to the vertices of an equilateral triangle inscribed in the unit circle. It can be easily verified that this attains the lower bound on incoherence. For the upper bound, a simple example is the matrix $\boldsymbol{U} \in \mathbb{R}^{n\times 3}$, where the first two columns are the standard basis vectors $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$, respectively, and the third column is a unit vector which is zero in its first two entries. Any set of points generated from this $\boldsymbol{U}$ lies entirely along the z-axis, except for two points, which lie on the x- and y-axes, respectively. Figures 1.3 and 1.4 provide a visual illustration of these examples.

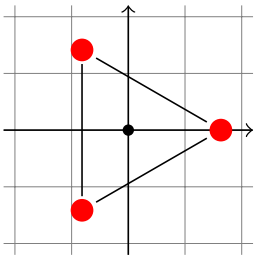


Figure 1.3: Visualizing the rows of $\boldsymbol{U}$ that lead to a minimal incoherence parameter.

Next, we aim to state the incoherence condition in terms of the points. Using the fact that $\boldsymbol{U}\boldsymbol{U}^\top$ is a projection operator onto the column space of $\boldsymbol{P}$, along with the fact that $\boldsymbol{P}$ is full-rank from Assumption 1.1.1, it follows that the unique orthogonal projector onto this range is defined as:

$$\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{P}(\boldsymbol{P}^\top\boldsymbol{P})^{-1}\boldsymbol{P}^\top.$$

By substituting this definition directly into the expanded trace formulation, we notice the

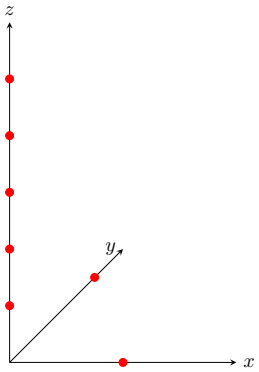


Figure 1.4: Example of a set of points with a maximal incoherence parameter.

following:

$$\begin{aligned}
&\left((\boldsymbol{U}\boldsymbol{U}^\top)_{ii} + (\boldsymbol{U}\boldsymbol{U}^\top)_{jj} - (\boldsymbol{U}\boldsymbol{U}^\top)_{ij} - (\boldsymbol{U}\boldsymbol{U}^\top)_{ji}\right) \\
&= \left(\boldsymbol{e}_i^\top \boldsymbol{P}(\boldsymbol{P}^\top\boldsymbol{P})^{-1}\boldsymbol{P}^\top\boldsymbol{e}_i + \boldsymbol{e}_j^\top \boldsymbol{P}(\boldsymbol{P}^\top\boldsymbol{P})^{-1}\boldsymbol{P}^\top\boldsymbol{e}_j - 2\boldsymbol{e}_i^\top \boldsymbol{P}(\boldsymbol{P}^\top\boldsymbol{P})^{-1}\boldsymbol{P}^\top\boldsymbol{e}_j\right) \\
&= \left(\boldsymbol{p}_i^\top (\boldsymbol{P}^\top\boldsymbol{P})^{-1}\boldsymbol{p}_i + \boldsymbol{p}_j^\top (\boldsymbol{P}^\top\boldsymbol{P})^{-1}\boldsymbol{p}_j - 2\boldsymbol{p}_i^\top (\boldsymbol{P}^\top\boldsymbol{P})^{-1}\boldsymbol{p}_j\right) \\
&= (\boldsymbol{p}_i - \boldsymbol{p}_j)^\top (\boldsymbol{P}^\top\boldsymbol{P})^{-1}(\boldsymbol{p}_i - \boldsymbol{p}_j).
\end{aligned}$$

The norm generated by $(\boldsymbol{P}^\top\boldsymbol{P})^{-1}$ is known in the statistics literature as the Mahalanobis distance [Mah18]. As we can see above, the Mahalanobis distance emerges algebraically in this expression, establishing the direct equivalence between the projected Frobenius norm of the measurement operator and the scaled pairwise Mahalanobis distance of the spatial coordinates.

This indicates that our incoherence condition can be reinterpreted as

$$\max_{(i,j):i<j} \left(\boldsymbol{p}_i - \boldsymbol{p}_j\right)^\top \left(\boldsymbol{P}^\top\boldsymbol{P}\right)^{-1} \left(\boldsymbol{p}_i - \boldsymbol{p}_j\right) \leq 2\frac{\nu r}{n}. \tag{1.22}$$

We first start our interpretation for the case where $(\boldsymbol{P}^\top\boldsymbol{P})^{-1}$ is the identity matrix. In this setting, for any pair (i,j), the expression $\left(\boldsymbol{p}_i - \boldsymbol{p}_j\right)^\top \left(\boldsymbol{p}_i - \boldsymbol{p}_j\right)$ is the squared Euclidean distance between the points $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$. Hence, the incoherence can be directly linked to the maximum distance among the points. In the general case, the above quantity suggests that incoherence serves as a measure of how the displacement vectors $\boldsymbol{p}_i - \boldsymbol{p}_j$ align with the principal components of the embedding. In particular, for a fixed choice of $(\boldsymbol{P}^\top\boldsymbol{P})^{-1}$, varying the matrix $\boldsymbol{U}$ leads to

different sets of points. If the displacement vectors tend to align with directions corresponding to the smallest principal components (i.e., those with the lowest variance), the incoherence is expected to be high. Conversely, if they align more with the dominant components (those with the highest variance), the incoherence tends to be low. In essence, high incoherence indicates that certain pairs of points are stretching significantly in directions where the embedding space has low variance.

Using the variational characterization, note that

$$(\boldsymbol{p}_i - \boldsymbol{p}_j)^\top (\boldsymbol{P}^\top \boldsymbol{P})^{-1} (\boldsymbol{p}_i - \boldsymbol{p}_j) \leq \lambda_1((\boldsymbol{P}^\top \boldsymbol{P})^{-1}) (\boldsymbol{p}_i - \boldsymbol{p}_j)^\top (\boldsymbol{p}_i - \boldsymbol{p}_j).$$

Noting that $\lambda_1((\boldsymbol{P}^\top \boldsymbol{P})^{-1}) = \frac{1}{\lambda_r}$, we can also state the incoherence condition in (1.22) as:

$$(\boldsymbol{p}_i - \boldsymbol{p}_j)^\top (\boldsymbol{p}_i - \boldsymbol{p}_j) \leq 2\frac{\nu r}{n}\lambda_r. \tag{1.23}$$

We note that these statements are not equivalent, merely that this simpler statement implies the original incoherence condition. Continuing with the simplified incoherence condition in (1.23), we seek to derive an upper bound on ν in terms of other geometric properties of $\boldsymbol{P}$, or spectral properties of $\boldsymbol{X}$. First, notice that

$$(\boldsymbol{p}_i - \boldsymbol{p}_j)^\top (\boldsymbol{p}_i - \boldsymbol{p}_j) \leq \max_{ij} \|\boldsymbol{p}_i - \boldsymbol{p}_j\|_2^2 = \|\boldsymbol{D}\|_\infty.$$

As we seek a constant ν such that (1.23) is satisfied for all $(i, j) \in \mathbb{I}$, we can see that this will be satisfied if

$$2\frac{\nu r}{n}\lambda_r(\boldsymbol{X}) \leq \|\boldsymbol{D}\|_\infty.$$

This yields the following upper bound for ν, in terms of a geometric constant and a spectral constant:

$$\nu \leq \frac{n}{2r}\frac{\|\boldsymbol{D}\|_\infty}{\lambda_r(\boldsymbol{X})}. \tag{1.24}$$

In the following lemmas, we show that data drawn from subgaussian isotropic distributions exhibit the property that $\nu \ll \mathcal{O}(n)$ with high probability, indicating that the recovery results derived in Section 2.2 and Section 3.2 will hold with reasonable sample complexity for a wide

category of data distributions.

Lemma 1.6.2. *[Ver12, Page 31] Let $\{\boldsymbol{p}_i\}_{i=1}^n \sim \mu$ where μ is a subgaussian probability measure defined on $\mathbb{R}^r$, and let $\boldsymbol{P} = [\boldsymbol{p}_1 \cdots \boldsymbol{p}_n]^\top \in \mathbb{R}^{n\times r}$. Define the covariance matrix of μ as $\boldsymbol{\Sigma}$. If $n \geq C(t/\varepsilon)^2 r$ for some constant $C > 0$, then with probability at least $1 - 2\exp(-t^2 n)$*

$$\left\| \frac{1}{n}\boldsymbol{P}^\top \boldsymbol{P} - \boldsymbol{\Sigma} \right\| \leq \varepsilon \|\boldsymbol{\Sigma}\|.$$

Let us now assume that μ is both subgaussian and isotropic, that is $\boldsymbol{\Sigma} = \boldsymbol{I}$. Furthermore, as we are interested in point clouds satisfying $\boldsymbol{P}^\top \mathbf{1} = \mathbf{0}$, we consider mean zero distributions. As such, we provide the following lemma:

Lemma 1.6.3. *Let $\{\boldsymbol{p}_i\}_{i=1}^n \subset \mathbb{R}^r$ be a collection of points drawn i.i.d. from an isotropic subgaussian distribution μ. Furthermore, let $\mathbb{E}[\boldsymbol{p}_i] = \mathbf{0}$, and assume each coordinate of $\boldsymbol{p}_i$ is independent. Let $\|\boldsymbol{p}_i\|_{\psi_2} \leq K$, where $\|\cdot\|_{\psi_2}$ is the subgaussian norm. Then for every $i, j \in [n] \times [n]$ with $i \neq j$, with probability at least $1 - Cn^{-2}$,*

$$\left| \|\boldsymbol{p}_i - \boldsymbol{p}_j\|_2^2 - 2r \right| \leq 4K^2 \sqrt{r} \log n,$$

where $C > 0$ is an absolute constant.

Proof. We first notice that, as $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$ are independent and $\mathbb{E}[\boldsymbol{p}_i] = \mathbf{0}$,

$$\begin{aligned}
\mathbb{E}\left[(\boldsymbol{p}_i - \boldsymbol{p}_j)^\top (\boldsymbol{p}_i - \boldsymbol{p}_j) \right] &= \mathbb{E}\left[\|\boldsymbol{p}_i\|_2^2 \right] - \mathbb{E}\left[\boldsymbol{p}_j{}^\top \boldsymbol{p}_i \right] - \mathbb{E}\left[\boldsymbol{p}_j{}^\top \boldsymbol{p}_i \right] + \mathbb{E}\left[\boldsymbol{p}_j{}^\top \boldsymbol{p}_j \right] \\
&= \mathbb{E}\left[\|\boldsymbol{p}_i\|_2^2 \right] + \mathbb{E}\left[\|\boldsymbol{p}_j\|_2^2 \right] - 2\mathbb{E}\left[\boldsymbol{p}_i\right]^\top \mathbb{E}\left[\boldsymbol{p}_j\right] \\
&= \mathbb{E}\left[\|\boldsymbol{p}_i\|_2^2 \right] + \mathbb{E}\left[\|\boldsymbol{p}_j\|_2^2 \right] \\
&= 2\mathbb{E}\left[\|\boldsymbol{p}_i\|_2^2 \right] \\
&= 2\mathbb{E}\left[\mathrm{Tr}\left(\boldsymbol{p}_i \boldsymbol{p}_i{}^\top \right) \right] \\
&= 2\ \mathrm{Tr}\left(\mathbb{E}\left[\boldsymbol{p}_i \boldsymbol{p}_i{}^\top \right] \right) \\
&= 2\ \mathrm{Tr}(\boldsymbol{I}) = 2r,
\end{aligned}$$

where the second and fourth lines follow from the independence of $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$, the third line

follows from the fact that $\mathbb{E}[\mu] = 0$, and the seventh line follows from the fact that μ is isotropic, i.e. $\boldsymbol{\Sigma} = \boldsymbol{I}$. The lemma statement follows from a simple application of the Hanson-Wright inequality, seen in Theorem 1.7.4. First, given vectors $\boldsymbol{p}_i, \boldsymbol{p}_j$, notice that

$$\begin{pmatrix} \boldsymbol{p}_i & \boldsymbol{p}_j \end{pmatrix} \underbrace{\begin{pmatrix} \boldsymbol{I} & -\boldsymbol{I} \\ -\boldsymbol{I} & \boldsymbol{I} \end{pmatrix}}_{\boldsymbol{A}} \begin{pmatrix} \boldsymbol{p}_i \\ \boldsymbol{p}_j \end{pmatrix} = (\boldsymbol{p}_i - \boldsymbol{p}_j)^\top (\boldsymbol{p}_i - \boldsymbol{p}_j) .$$

Previously, we have shown that $\mathbb{E}[(\boldsymbol{p}_i - \boldsymbol{p}_j)^\top (\boldsymbol{p}_i - \boldsymbol{p}_j)] = 2r$. Furthermore, $\|\boldsymbol{A}\|_{\mathrm{F}}^2 = 4r$, and $\|\boldsymbol{A}\| \leq 2$ by Gershgorin circle theorem. The result follows from an application of Theorem 1.7.4.

□

We can use the above lemma to prove an upper bound on the incoherence of an isotropic, subgaussian distribution, presented in the following lemma:

Lemma 1.6.4. *Let $\{\boldsymbol{p}_i\}_{i=1}^n \sim \mu$ where μ is an isotropic subgaussian probability measure defined on $\mathbb{R}^r$ with subgaussian norm $\|\mu\|_{\psi_2} \leq K$, and let $\boldsymbol{P} = [\boldsymbol{p}_1 \cdots \boldsymbol{p}_n]^\top \in \mathbb{R}^{n \times r}$. Then with probability at least $1 - Dn^{-2}$ for an absolute constant $D > 0$, the incoherence parameter ν of $\boldsymbol{X} = \boldsymbol{P}\boldsymbol{P}^\top$ can be upper bounded by*

$$\nu \leq \mathcal{O}\left(K^2 \frac{\log n}{\sqrt{r}} \right).$$

Proof. From Theorem 1.6.3, we have that $\|\boldsymbol{D}\|_\infty \leq 2r + 4K^2\sqrt{r}\log n$ with high probability for a subgaussian isotropic distribution. Furthermore, from Theorem 1.6.2, $\lambda_r(\boldsymbol{X}) \geq cn$ for some $\mathcal{O}(1)$ constant $c > 0$ with high probability. As such, we can see that the incoherence constant can be upper-bounded using (1.24) by

$$\begin{aligned} \nu &\leq \frac{n}{2r} \frac{2r + 4K^2\sqrt{r}\log n}{cn} \\ &\leq 1 + \frac{2K^2 \log n}{c\sqrt{r}} \\ &= \mathcal{O}\left(K^2 \frac{\log n}{\sqrt{r}} \right), \end{aligned}$$

thus concluding the proof. □

We note that Theorem 1.6.4 shows, with high probability, that the incoherence constant remains in a regime where it does not degrade the recovery guarantees established in Section 2.2 and Section 3.2 for data generated from subgaussian distributions. We note that this result is very similar to the condition derived in [CR09] for the incoherence of matrices in the random orthogonal model.

We comment that the analysis in this section exclusively pertained to data generated from subgaussian isotropic measures for expositional simplicity. These techniques can be extended to anisotropic subgaussian measures, and one can show the resulting bound for $\nu = \mathcal{O}(\kappa + K^2 \log n/\sqrt{r})$, where κ is the condition number of $\boldsymbol{X}$. We provide a proof of this result in Theorem 1.A.1.

Remark 1.6.2: Geometric Interpretation of Standard Incoherence

We now provide a geometric interpretation of (1.19) for positive semidefinite matrices. Expanding (1.19), we obtain:

$$\begin{aligned}\|\mathcal{P}_U \boldsymbol{e}_i\|_2^2 &= \boldsymbol{e}_i^\top \mathcal{P}_U \boldsymbol{e}_i \\ &= \mathrm{Tr}(\boldsymbol{e}_i^\top (\boldsymbol{P}(\boldsymbol{P}^\top \boldsymbol{P})^{-1} \boldsymbol{P}\boldsymbol{e}_i) \\ &= \boldsymbol{p}_i^\top (\boldsymbol{P}^\top \boldsymbol{P})^{-1} \boldsymbol{p}_i.\end{aligned}$$

As such, standard incoherence in the EDMC framework represents the standard Mahalanobis distance of each point, rather than the pairwise Mahalanobis distance between points in the geometric framework.

1.7 Useful Inequalities from Random Matrix Theory

Concentration of measure is a beautiful phenomenon in the study of high-dimensional mathematics, and is well worth writing about in detail. It is extremely relevant in a variety of disciplines. In lieu of an introduction to the the curious reader to the following resources [Ver18, BLM13, Led01]. We now state some important inequalities from this field that we will use frequently throughout this work.

Theorem 1.7.1 (Scalar Bernstein Inequality [Ver18]). *Let $Y_1, \cdots, Y_n$ be independent, mean zero random variables such that $|Y_i| \leq R$ for all i, and let $\sigma^2 = \mathbb{E}\left[\sum_{i=1}^n Y_i^2\right]$. Then*

$$\mathbb{P}\left[\left|\sum_{i=1}^n \boldsymbol{Y}_i\right| \geq t\right] \leq 2\exp\left(-\frac{t^2/2}{\sigma^2 + Rt/3}\right).$$

In particular, if $t \leq \frac{\sigma^2}{R}$ this simplifies to

$$\mathbb{P}\left[\left|\sum_{i=1}^n \boldsymbol{Y}_i\right| \geq t\right] \leq 2\exp\left(-\frac{3t^2}{8\sigma^2}\right).$$

Another set of relevant tools are matrix concentration inequalities. These powerful tools allow for sharp understanding of the tail probabilities of the spectral norms of certain random matrices. We point the interested reader to [Tro12, T$^+$15] for an excellent treatment of the subject. We provide these inequalities below.

Theorem 1.7.2 (Operator Bernstein Inequality[Rec11a, Tro12]). *Let $\boldsymbol{X}_i \in \mathbb{R}^{n\times n}$, $i = 1, \cdots, m$ be independent, zero-mean, matrix-valued random variables, and let*

$$\sigma^2 \geq \max\left\{\left\|\sum_{i=1}^m \mathbb{E}\left(\boldsymbol{X}_i \boldsymbol{X}_i^\top\right)\right\|, \left\|\sum_{i=1}^m \mathbb{E}\left(\boldsymbol{X}_i^\top \boldsymbol{X}_i\right)\right\|\right\}.$$

Assume there exists a $c \in \mathbb{R}$ such that $\|\boldsymbol{X}_i\| \leq c$ almost surely. Then for $t > 0$

$$\mathbb{P}\left(\left\|\sum_{i=1}^m \boldsymbol{X}_i\right\| > t\right) \leq 2n\exp\left(-\frac{t^2/2}{\sigma^2 + ct/3}\right).$$

If we assume that $t \leq \frac{\sigma^2}{c}$, this simplifies to

$$\mathbb{P}\left(\left\|\sum_{i=1}^m \boldsymbol{X}_i\right\| > t\right) \leq 2n\exp\left(-\frac{3t^2}{8\sigma^2}\right), \tag{1.25}$$

and if $t > \frac{\sigma^2}{c}$,

$$\mathbb{P}\left(\left\|\sum_{i=1}^m \boldsymbol{X}_i\right\| > t\right) \leq 2n\exp\left(-\frac{3t}{8c}\right). \tag{1.26}$$

Theorem 1.7.3 (Vector Bernstein Inequality [Gro11]). *Let $\{X_i\}_{i=1}^n$ be independent, mean-zero random vectors in $\mathbb{R}^d$ with $\mathbb{E}[X_i] = 0$ for all $i = 1, \ldots, n$. Suppose there exist constants*

$R > 0$ and $\sigma^2 > 0$ such that for every i, $\|X_i\|_2 \leq R$ almost surely and $\sum_{i=1}^n \mathbb{E}\big[\|X_i\|_2^2\big] \leq \sigma^2$. $\mathbb{E}\big[\|X_i\|_2^2\big] \leq \sigma^2$. Then for any $t > 0$, the following tail bound holds for $t \leq \frac{\sigma^2}{R}$:

$$\mathbb{P}\left(\left\|\sum_{i=1}^n X_i\right\|_2 \geq t\right) \leq \exp\left(\frac{-t^2}{4\sigma^2}\right).$$

Theorem 1.7.4 (Hanson-Wright Inequality [RV13])**.** *Let $\boldsymbol{Y} = (Y_1 \cdots Y_n) \in \mathbb{R}^n$ be a random vector with Y_i independent and $\mathbb{E}[Y_i] = 0$, and $\|Y_i\|_{\psi_2} \leq K$ for some $K \geq 0$, where $\|\cdot\|_{\psi_2}$ is the sub-Gaussian norm. Additionally, let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. Then*

$$\mathbb{P}\left[\left|\boldsymbol{Y}^\top \boldsymbol{A}\boldsymbol{Y} - \mathbb{E}\left[\boldsymbol{Y}^\top \boldsymbol{A}\boldsymbol{Y}\right]\right| \geq t\right] \leq 2\exp\left(-c\min\left\{\frac{t^2}{K^4\|\boldsymbol{A}\|_{\mathrm{F}}^2}, \frac{t}{K^2\|\boldsymbol{A}\|}\right\}\right).$$

The last important inequality used in this work is the decoupling inequality of de la Peña [DlPG12]. A partial statement of this remarkable theorem is reproduced below.

Theorem 1.7.5 (Matrix–valued decoupling inequality [DlPG12, Theorem 3.4.1])**.** *Let $\{\boldsymbol{X}_i\}_{i=1}^n$ be a sequence of independent random variables in $\mathbb{R}^m$, and let $\{\boldsymbol{X}_i'\}_{i=1}^n$ be an independent copy of $\{\boldsymbol{X}_i\}_{i=1}^n$ defined on the same probability space. Let $\{h_{ij}\}_{1\leq i\neq j\leq n}$ be a collection of measurable kernels $h_{ij} : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^{m\times m}$, satisfying*

$$\mathbb{E}\big[h_{ij}(\boldsymbol{X}_i, \boldsymbol{X}_j) \mid \boldsymbol{X}_i\big] = \boldsymbol{0}, \qquad \mathbb{E}\big[h_{ij}(\boldsymbol{X}_i, \boldsymbol{X}_j) \mid \boldsymbol{X}_j\big] = \boldsymbol{0},$$

for all $= i \neq j$.

Define the coupled and decoupled chaos sums as

$$S := \sum_{i\neq j} h_{ij}(\boldsymbol{X}_i, \boldsymbol{X}_j), \qquad S' := \sum_{i\neq j} h_{ij}(\boldsymbol{X}_i, \boldsymbol{X}_j').$$

Then there exists a universal constant $C > 0$ such that, for all $t > 0$,

$$\mathbb{P}\big(\|S\| \geq t\big) \;\leq\; C\,\mathbb{P}\big(\|S'\| \geq t/C\big).$$

The constant C is universal and is independent of all h_{ij}, $\{\boldsymbol{X}_i\}_{i=1}^n$, n and m.

We note that the preceding theorem holds for higher than second–order chaoses, and that

the constant C depends on solely on the order of the chaos. Furthermore, the above inequality holds in any Banach space equipped with any norm. However, for the problem that we consider we only need the second–order, matrix–valued version, and are simplifying the statement for expositional purposes.

1.8 A Collection of Useful Theorems

Lastly, we provide a few other theorems that are leveraged throughout this thesis in various forms. We provide them below.

Theorem 1.8.1 (Davis–Kahan $\sin\Theta$ theorem (rank-r case) [DK70], [YWS15])**.** *Let $\boldsymbol{X}, \hat{\boldsymbol{X}} \in \mathbb{R}^{n\times n}$ be symmetric positive semidefinite matrices of rank r. Let $\boldsymbol{V}, \hat{\boldsymbol{V}} \in \mathbb{R}^{n\times r}$ have orthonormal columns spanning the top-r eigenspaces of $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$, respectively. Then*

$$\|\sin\Theta(\boldsymbol{V}, \hat{\boldsymbol{V}})\| \leq \frac{\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|}{\lambda_r(\boldsymbol{X})}.$$

One result that will be used throughout this work is a technique for constructing eigenvalue bounds through a vectorization technique. This result is as follows.

Lemma 1.8.2 (Vectorization Technique)**.** *Let $\{\boldsymbol{Z}_k\}_{k=1}^m$ be a basis for some subspace $\mathbb{V} \subset \mathbb{R}^{n\times n}$ of dimension m, $\boldsymbol{G} = [\langle \boldsymbol{Z}_i, \boldsymbol{Z}_j\rangle] \in \mathbb{R}^{m\times m}$, and $\boldsymbol{Z}_{\mathbb{V}} \in \mathbb{R}^{n^2\times m}$ be the matrix where the k-th column vector is $\mathrm{vec}(\boldsymbol{Z}_k)$. Then for any $\boldsymbol{Y} \in \mathbb{R}^{n\times n}$*

$$\max_{\|\boldsymbol{Y}\|_{\mathrm{F}}=1} \sum_{k=1}^m \langle \boldsymbol{Y}, \boldsymbol{Z}_k\rangle^2 = \lambda_{\max}(\boldsymbol{G}).$$

Proof. We can see that

$$\begin{aligned}
\max_{\|\boldsymbol{Y}\|_{\mathrm{F}}=1} \sum_{k=1}^m \langle \boldsymbol{Y}, \boldsymbol{Z}_k\rangle^2 &= \max_{\|\boldsymbol{Y}\|_{\mathrm{F}}=1} \sum_{k=1}^m \Big(\mathrm{vec}(\boldsymbol{Y})^\top \mathrm{vec}(\boldsymbol{Z}_k)\Big)\Big(\mathrm{vec}(\boldsymbol{Z}_k)^\top \mathrm{vec}(\boldsymbol{Y})\Big) \\
&= \max_{\|\boldsymbol{Y}\|_{\mathrm{F}}=1} \mathrm{vec}(\boldsymbol{Y})^\top \left(\sum_{k=1}^m \mathrm{vec}(\boldsymbol{Z}_k)\mathrm{vec}(\boldsymbol{Z}_k)^\top\right) \mathrm{vec}(\boldsymbol{Y}) \\
&= \max_{\|\boldsymbol{Y}\|_{\mathrm{F}}=1} \mathrm{vec}(\boldsymbol{Y})^\top \boldsymbol{Z}_{\mathbb{V}} \boldsymbol{Z}_{\mathbb{V}}^\top \mathrm{vec}(\boldsymbol{Y}).
\end{aligned}$$

As for any matrix $\boldsymbol{A} \succeq \boldsymbol{0}$, $\max_{\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} = \lambda_{\max}(\boldsymbol{A})$, it follows that

$$\max_{\|\boldsymbol{Y}\|_{\mathrm{F}}=1} \sum_{k=1}^{m} \langle \boldsymbol{Y}, \boldsymbol{Z}_k \rangle^2 = \lambda_{\max}(\boldsymbol{Z}_{\mathbb{V}} \boldsymbol{Z}_{\mathbb{V}}^\top).$$

Now, as for any $\boldsymbol{A} \in \mathbb{R}^{r\times s}$, $\lambda_{\max}(\boldsymbol{A}\boldsymbol{A}^\top) = \lambda_{\max}(\boldsymbol{A}^\top \boldsymbol{A})$, we see that

$$\max_{\|\boldsymbol{Y}\|_{\mathrm{F}}=1} \sum_{k=1}^{m} \langle \boldsymbol{Y}, \boldsymbol{Z}_k \rangle^2 = \lambda_{\max}(\boldsymbol{Z}_{\mathbb{V}} \boldsymbol{Z}_{\mathbb{V}}^\top) = \lambda_{\max}(\boldsymbol{Z}_{\mathbb{V}}^\top \boldsymbol{Z}_{\mathbb{V}}) = \lambda_{\max}(\boldsymbol{G}).$$

This concludes the proof. □

Lemma 1.8.3 (Bounds for Projections [WCCL16, WCCL20])**.** *Let $\boldsymbol{X}_l = \boldsymbol{U}_l \boldsymbol{D}_l \boldsymbol{U}_l^\top$ be a rank-r matrix and $\mathbb{T}_l$ be the tangent space of $\mathcal{N}_r$ at $\boldsymbol{X}_l$. Let $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top$ be another rank-r matrix, and $\mathbb{T}$ be the corresponding tangent space. Then*

$$\|\boldsymbol{U}_l \boldsymbol{U}_l^\top - \boldsymbol{U}\boldsymbol{U}^\top\| \leq \frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\sigma_r(\boldsymbol{X})}, \qquad \|\boldsymbol{U}_l \boldsymbol{U}_l^\top - \boldsymbol{U}\boldsymbol{U}^\top\|_{\mathrm{F}} \leq \frac{\sqrt{2}\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\sigma_r(\boldsymbol{X})},$$
$$\|(\mathcal{I} - \mathcal{P}_{\mathbb{T}_l})\boldsymbol{X}\|_{\mathrm{F}} \leq \frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}^2}{\sigma_r(\boldsymbol{X})}, \qquad \|\mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_{\mathbb{T}}\| \leq \frac{2\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\sigma_r(\boldsymbol{X})}.$$

1.A Further Incoherence Results

Lemma 1.A.1. *Let μ be an a.s. bounded, mean-zero, subgaussian distribution with positive definite covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}$. Let n points $\{\boldsymbol{p}_i\}_{i=1}^n \sim \mu$ be sampled i.i.d., and let $\boldsymbol{P} = [\boldsymbol{p}_1 \dots \boldsymbol{p}_n]^\top \in \mathbb{R}^{n\times d}$ be the corresponding point matrix with Gram matrix $\boldsymbol{X}$, which has condition number κ. Let $\|\boldsymbol{p}_i\|_{\psi_2} \leq K$ for some $K > 0$. Then with probability at least $1 - Cn^{-2}$ for some absolute constant $C > 0$, the incoherence parameter of $\boldsymbol{X}$ is bounded by*

$$\nu \leq \mathcal{O}\left(\kappa + \frac{K^2 \log n}{\sqrt{r}}\right).$$

Proof. This proof is much the same as the proof in Section 1.6. First, we remark that

$$\begin{aligned}\mathbb{E}\left[(\boldsymbol{p}_i - \boldsymbol{p}_j)^\top (\boldsymbol{p}_i - \boldsymbol{p}_j)\right] &= \mathbb{E}\left[\|\boldsymbol{p}_i\|_2^2\right] - \mathbb{E}\left[\boldsymbol{p}_j{}^\top \boldsymbol{p}_i\right] - \mathbb{E}\left[\boldsymbol{p}_j{}^\top \boldsymbol{p}_i\right] + \mathbb{E}\left[\boldsymbol{p}_j{}^\top \boldsymbol{p}_j\right] \\ &= \mathbb{E}\left[\|\boldsymbol{p}_i\|_2^2\right] + \mathbb{E}\left[\|\boldsymbol{p}_j\|_2^2\right] - 2\mathbb{E}\left[\boldsymbol{p}_i\right]^\top \mathbb{E}\left[\boldsymbol{p}_j\right]\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left[\|\boldsymbol{p}_i\|_2^2\right] + \mathbb{E}\left[\|\boldsymbol{p}_j\|_2^2\right] \\
&= 2\mathbb{E}\left[\|\boldsymbol{p}_i\|_2^2\right] \\
&= 2\mathbb{E}\left[\text{Trace}\left(\boldsymbol{p}_i \boldsymbol{p}_i{}^\top\right)\right] \\
&= 2\ \text{Trace}\left(\mathbb{E}\left[\boldsymbol{p}_i \boldsymbol{p}_i{}^\top\right]\right) \\
&= 2\ \text{Trace}(\boldsymbol{\Sigma}) \leq 2r\lambda_1(\boldsymbol{\Sigma}),
\end{aligned}
$$

where the second and fourth lines follow from the independence of $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$, the third line follows from the fact that $\mathbb{E}[\mu] = 0$, and the seventh line follows from the fact that $\boldsymbol{\Sigma}$ has r non-zero eigenvalues.

Next, following the argument of Theorem 1.6.3 but replacing $2r$ with $\mathbb{E}\left[(\boldsymbol{p}_i - \boldsymbol{p}_j)^\top (\boldsymbol{p}_i - \boldsymbol{p}_j)\right]$, we have that, with probability at least $1 - Cn^{-2}$

$$
\|\boldsymbol{p}_i - \boldsymbol{p}_j\|_2^2 \leq 2r\lambda_1(\boldsymbol{\Sigma}) + 4K^2\sqrt{r}\log n.
$$

Now, leveraging that from Theorem 1.6.2 that $\lambda_r(\boldsymbol{X}) \approx n\lambda_r(\boldsymbol{\Sigma})$, we have that for some $c > 0$ with high probability that

$$
\begin{aligned}
\nu &\leq \frac{n}{2r}\frac{2r\lambda_1(\boldsymbol{\Sigma}) + 4K^2\sqrt{r}\log n}{\lambda_r(\boldsymbol{X})} \\
&= \frac{n}{2r}\frac{2r\lambda_1(\boldsymbol{\Sigma}) + 4K^2\sqrt{r}\log n}{cn\lambda_r(\boldsymbol{\Sigma})} \\
&\leq \frac{\kappa}{c} + \frac{2K^2\log n}{c\sqrt{r}} \\
&= \mathcal{O}\left(\kappa + \frac{K^2\log n}{\sqrt{r}}\right).
\end{aligned}
$$

This concludes the proof. □

Lemma 1.A.2. *If $\|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\text{F}} \leq \sqrt{\frac{\nu r}{8n}}$, it follows that $\|\mathcal{P}_U \boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\text{F}} \leq \sqrt{\frac{\nu r}{2n}}$. Similarly, if $\|\mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\text{F}} \leq \sqrt{\frac{\nu r}{8n}}$, it follows that $\|\mathcal{P}_{\mathbb{T}} \boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\text{F}} \leq \sqrt{\frac{\nu r}{2n}}$.*

Proof. To see this result, notice that

$$
\|\mathcal{P}_U \boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\text{F}} = \left\|\mathcal{P}_U\left(\sum_{\boldsymbol{\beta}\in\mathbb{I}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \boldsymbol{w}_{\boldsymbol{\beta}}\right)\right\|_{\text{F}}
$$

$$\leq \sum_{\boldsymbol{\beta}\in\mathbb{I}} \left|H^{\boldsymbol{\alpha}\boldsymbol{\beta}}\right| \|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\beta}}\|_{\mathrm{F}}$$
$$\leq \sqrt{\frac{\nu r}{8n}} \sum_{\boldsymbol{\beta}\in\mathbb{I}} \left|H^{\boldsymbol{\alpha}\boldsymbol{\beta}}\right|,$$

and as $\sum_{\boldsymbol{\alpha}\in\mathbb{I}} |H^{\boldsymbol{\alpha}\boldsymbol{\beta}}| \leq 2$ from Theorem 2.A.3, the claim follows. An identical proof shows the second result, with $\mathcal{P}_{\mathbb{T}}$ in place of $\mathcal{P}_U$. □

Lemma 1.A.3. *Let $\boldsymbol{X} \succeq \mathbf{0}$ be a rank-r, ν-incoherent matrix satisfying* (3.4) *with constant ν. Then*

$$\left(\max_{\boldsymbol{\alpha}\in\mathbb{I}} |\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle|\right) \leq \frac{\nu r}{4n}\|\boldsymbol{X}\|.$$

Proof. To see the above statement, notice that for $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top = \boldsymbol{P}\boldsymbol{P}^\top$,

$$\begin{aligned}
\frac{(\max_{\boldsymbol{\alpha}\in\mathbb{I}} |\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle|)}{\|\boldsymbol{X}\|} &= \frac{1}{\|\boldsymbol{X}\|} \max_{i,j} |X_{ii} + X_{jj} - 2X_{ij}| \\
&= \frac{1}{\|\boldsymbol{X}\|} \max_{i,j} \|\boldsymbol{p}_i - \boldsymbol{p}_j\|_2^2 \\
&= \frac{1}{\|\boldsymbol{X}\|} \max_{i,j} (\boldsymbol{u}_i - \boldsymbol{u}_j)^\top \boldsymbol{\Lambda} (\boldsymbol{u}_i - \boldsymbol{u}_j) \\
&\leq \max_{ij} \|\boldsymbol{u}_i - \boldsymbol{u}_j\|_2^2 \\
&= \max_{\boldsymbol{\alpha}\in\mathbb{I}} \frac{1}{2}\|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 \\
&\leq \frac{\nu r}{4n},
\end{aligned}$$

where the third line follows from (1.3), the fourth line follows from the fact that $\boldsymbol{X} \succeq \mathbf{0}$, the fifth line follows from (1.20), and the final line follows from (3.4), thus concluding the proof. □

Chapter 2

Preliminary Investigations into a Non-convex Riemannian Approach to the EDMC Problem

This chapter constitutes our first attempts at addressing Question 1.3.1. In this section, we provide convergence guarantees for a Riemannian dual-basis algorithm, but note that its numerical properties are worse than state-of-the-art. We also introduce another algorithm that performs comparably to state-of-the-art approaches, but only recently admitted a possible proof of convergence. The sketch of this proof is discussed in Section A.1. Much of this work has been taken from [SCT24].

2.1 Algorithm

With the goal of translating the standard matrix completion problem to Gram matrix completion for EDMC in mind, the most direct adaptation of the work conducted in [WCCL20] would be defining an objective function by analogy to (1.14) as follows:

$$\underset{\boldsymbol{Y}\in\mathbb{S}}{\text{minimize}}\ \langle \boldsymbol{Y}-\boldsymbol{X}, \mathcal{R}_\Omega(\boldsymbol{Y}-\boldsymbol{X})\rangle \ \text{ subject to } \ \text{rank}(\boldsymbol{Y}) = r.$$

However, a notable challenge arises: computing the Euclidean gradient of the objective function necessitates unavailable information in the form $\langle \boldsymbol{X}, \boldsymbol{v}_{\boldsymbol{\alpha}} \rangle$ from $\mathcal{R}_\Omega^*(\boldsymbol{X})$ as

$$\nabla_{\boldsymbol{Y}} \left(\langle \boldsymbol{Y} - \boldsymbol{X}, \mathcal{R}_\Omega(\boldsymbol{Y} - \boldsymbol{X}) \rangle \right) = \mathcal{R}_\Omega(\boldsymbol{Y} - \boldsymbol{X}) + \mathcal{R}_\Omega^*(\boldsymbol{Y} - \boldsymbol{X}),$$

where $\nabla_{\boldsymbol{Y}}$ denotes the gradient with respect to $\boldsymbol{Y}$. To circumvent this difficulty, there has been exploration into self-adjoint alternatives to $\mathcal{R}_\Omega$ [TL21, TL18, SLCT23], one of which we will expand upon shortly. These surrogates allow for the definition of an objective function in analogy to (1.14), but as of now lack the requisite theoretical properties for convergence.

One surrogate of interest in this work is the restricted frame operator $\mathcal{F}_\Omega$, defined in (1.11). This operator is self-adjoint, positive semi-definite, and expresses the same information as $\mathcal{R}_\Omega$ without reference to the dual basis. Additionally, this operator under a different name was critical to the analysis of the algorithm in [LS24]. Using this operator, we can define the following objective function:

$$\underset{\boldsymbol{Y} \in \mathbb{S}}{\text{minimize}} \ \frac{1}{2} \langle \boldsymbol{Y} - \boldsymbol{X}, \mathcal{F}_\Omega(\boldsymbol{Y} - \boldsymbol{X}) \rangle \ \text{ subject to } \ \text{rank}(\boldsymbol{Y}) = r. \tag{2.1}$$

This is a true quadratic form, minimized over $\mathcal{N}_r$, and can be approached in an identical manner algorithmically as (1.14). To perform this first-order retraction method, we need to define the tangent space at each iterate $\boldsymbol{X}_l \in \mathcal{N}_r$. Given $\boldsymbol{X}_l = \boldsymbol{U}_l \boldsymbol{\Lambda}_l \boldsymbol{U}_l^\top$ for $\boldsymbol{U}_l \in \mathbb{R}^{n \times r}$, we can define $\mathbb{T}_l$ as in (1.15). The projection onto $\mathbb{T}_l$, denoted $\mathcal{P}_{\mathbb{T}_l}$ and defined in (1.16) allows for easy computation of the Riemannian gradient of the objective function. The retraction map is the hard thresholding operator, denoted $\mathcal{H}_r : \mathbb{T}_l \to \mathcal{N}_r$ is as defined in (1.17). We note that for matrices $\boldsymbol{Y}$ with $\text{rank}(\boldsymbol{Y}) \geq r$ that $\text{rank}\left(\mathcal{H}_r(\boldsymbol{Y})\right) = r$.

The construction of $\mathcal{F}_\Omega$ motivates Algorithm 3. This algorithm, while effective numerically, has technical weaknesses that prevent an easy proof of convergence. See the discussion in Remark 2.1.1.

To leverage the analytical properties of $\mathcal{R}_\Omega$ while sidestepping the technical challenges of its non-self-adjoint nature, we define an algorithm by analogy to RF-RGD but without any reference to an objective function. As we will show in Section 2.2, this algorithm will give us

Algorithm 3 Restricted Frame Riemannian Gradient Descent for EDMC (RF-RGD)

Initialization: $\boldsymbol{X}_0 = \boldsymbol{U}_0\boldsymbol{D}_0\boldsymbol{U}_0^\top$
for $l = 0, 1, ...$ **do**
1. $\boldsymbol{G}_l = \mathcal{F}_\Omega(\boldsymbol{X} - \boldsymbol{X}_l)$
2. $\alpha_l = \frac{\|\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l\|_{\mathrm{F}}^2}{\langle \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l, \mathcal{F}_\Omega\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l\rangle}$
3. $\boldsymbol{W}_l = \boldsymbol{X}_l + \alpha_l\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l$
4. $\boldsymbol{X}_{l+1} = \mathcal{H}_r(\boldsymbol{W}_l)$

end for
Output: $\boldsymbol{X}_{\mathrm{rec}}$: Estimated Gram matrix.

strong convergence guarantees with reasonable sample complexities at a cost of interpretability. As such, we define Algorithm 4 to reconstruct a ground truth matrix $\boldsymbol{X}$ as follows:

Algorithm 4 Riemannian Pseudogradient Descent

Initialization: $\boldsymbol{X}_0 = \boldsymbol{U}_0\boldsymbol{D}_0\boldsymbol{U}_0^\top$
for $l = 0, 1, ...$ **do**
1. $\boldsymbol{G}_l = \mathcal{R}_\Omega(\boldsymbol{X} - \boldsymbol{X}_l)$
2. $\alpha_l = \max\left\{\frac{\|\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l\|_{\mathrm{F}}^2}{\langle \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l, \mathcal{R}_\Omega\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l\rangle}, 0\right\}$
3. $\boldsymbol{W}_l = \boldsymbol{X}_l + \alpha_l\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l$
4. $\boldsymbol{X}_{l+1} = \mathcal{H}_r(\boldsymbol{W}_l)$

end for
Output: $\boldsymbol{X}_{\mathrm{rec}}$: Estimated Gram matrix.

Unlike in RF-RGD, we cannot compute the steepest descent of an objective function in $\mathbb{T}_l$, so we consider a surrogate modeled after each of the preceding algorithms. The maximum with zero in Step 2 of the algorithm is introduced to avoid divergence, as $\mathcal{R}_\Omega$ is not a positive semi-definite operator and the denominator cannot be guaranteed to be positive for arbitrary points in $\mathcal{N}_r$. When $\alpha_l = 0$ occurs, the algorithm terminates. Positive α_l is required for convergence, and the conditions for this are provided and characterized in Lemma 2.C.1. This condition is satisfied in the high-sample regime, where ε_0 is small.

In the above approaches, as in Algorithm 2, the truncated SVD step in the gradient descent scheme is the most expensive, especially when n is large. The authors in [WCCL20] noticed the following: in Algorithm 2, $\boldsymbol{W}_l$ can be rewritten as, for $\boldsymbol{Z}_l = \alpha_l\boldsymbol{G}_l$,

$$\boldsymbol{W}_l = \boldsymbol{X}_l + \alpha_l\mathcal{P}_{\mathbb{T}_l}(\boldsymbol{Z}_l).$$

In particular, we can see that

$$\begin{aligned}\boldsymbol{W}_l &= \boldsymbol{U}_l\boldsymbol{D}_l\boldsymbol{U}_l^\top + \boldsymbol{U}_l\boldsymbol{U}_l^\top\boldsymbol{Z}_l + \boldsymbol{Z}_l\boldsymbol{U}_l\boldsymbol{U}_l^\top - \boldsymbol{U}_l\boldsymbol{U}_l^\top\boldsymbol{Z}_l\boldsymbol{U}_l\boldsymbol{U}_l^\top \\ &= \boldsymbol{U}_l(\boldsymbol{D}_l + \boldsymbol{U}_l^\top\boldsymbol{Z}_l\boldsymbol{U}_l)\boldsymbol{U}_l^\top + \boldsymbol{U}_l\boldsymbol{U}_l^\top\boldsymbol{Z}_l(\boldsymbol{I} - \boldsymbol{U}_l\boldsymbol{U}_l^\top) + \underbrace{(\boldsymbol{I} - \boldsymbol{U}_l\boldsymbol{U}_l^\top)\boldsymbol{Z}_l\boldsymbol{U}_l}_{\boldsymbol{Y}}\boldsymbol{U}_l^\top \\ &:= \boldsymbol{U}_l(\boldsymbol{D}_l + \boldsymbol{U}_l^\top\boldsymbol{Z}_l\boldsymbol{U}_l)\boldsymbol{U}_l^\top + \boldsymbol{U}_l\boldsymbol{Y}^\top + \boldsymbol{Y}\boldsymbol{U}_l^\top.\end{aligned}$$

The authors then notice that by taking the QR decomposition $\boldsymbol{Y} = \boldsymbol{Q}\boldsymbol{R}$, $\boldsymbol{U}_l^\top\boldsymbol{Q} = \boldsymbol{0}$ so $\boldsymbol{W}_l$ becomes

$$\begin{aligned}\boldsymbol{W}_l &= \begin{bmatrix}\boldsymbol{U}_l & \boldsymbol{Q}\end{bmatrix}\begin{bmatrix}\boldsymbol{D}_l + \boldsymbol{U}_l^\top\boldsymbol{Z}_l\boldsymbol{U}_l & \boldsymbol{R}^\top \\ \boldsymbol{R} & \boldsymbol{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{U}_l^\top \\ \boldsymbol{Q}^\top\end{bmatrix} \\ &:= \begin{bmatrix}\boldsymbol{U}_l & \boldsymbol{Q}\end{bmatrix}\boldsymbol{M}_l\begin{bmatrix}\boldsymbol{U}_l^\top \\ \boldsymbol{Q}^\top\end{bmatrix},\end{aligned}$$

which reduces the retraction map $\mathcal{H}_r$ from a truncated SVD on an $n \times n$ matrix to an SVD on $\boldsymbol{M}_l$, which is $2r \times 2r$. This result generalizes for non-symmetric matrices, but only the symmetric case is considered here for simplicity.

Remark 2.1.1: Discussion of the RIP of $\mathcal{F}_\Omega$ and $\mathcal{R}_\Omega$

Algorithm 3 is easily implementable and gives strong numerical results, provided in Section 2.3, but the proof of local convergence remains an open question due to a lack of proven RIP (described in Definition 1.2.1). In practice, this statement is proven using non-commutative concentration inequalities like Theorem 1.7.2, first introduced in the matrix completion literature in [Rec11a, GN10]. As in Section 1.2.2, this requires the decomposition of the measurement operator as a sum of i.i.d. random operators. It is well-established that $\mathcal{P}_\Omega$ possesses this property, and in this paper we show that $\mathcal{R}_\Omega$ also exhibits RIP. The proof, seen in Theorem 2.2.1, is completed using standard techniques with some specific properties of the dual bases $\{\boldsymbol{w}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$ and $\{\boldsymbol{v}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$, and RIP for $\mathcal{R}_\Omega$ is established with only slightly worse sample complexity than for $\mathcal{P}_\Omega$.

In the case of $\mathcal{F}_\Omega$, the object to consider would be $\mathcal{P}_\mathbb{T}\mathcal{F}_\Omega\mathcal{P}_\mathbb{T} - \frac{m}{L}\mathcal{P}_\mathbb{T}\mathcal{F}_\mathbb{I}\mathcal{P}_\mathbb{T}$, as $\mathbb{E}[\mathcal{F}_\Omega] =$

$\frac{m}{L}\mathcal{F}_{\mathbb{I}} \neq \mathcal{I}$, where $\mathcal{F}_{\mathbb{I}} = \sum_{\boldsymbol{\alpha}\in\mathbb{I}}\langle\cdot, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\boldsymbol{w}_{\boldsymbol{\alpha}}$. Proving a similar statement for $\mathcal{F}_{\Omega}$ is challenging, as $\mathbb{E}[\mathcal{F}_{\Omega}] \neq c\mathcal{I}$. However, numerical evidence strongly indicates that $\|\mathcal{P}_{\mathbb{T}}\mathcal{F}_{\Omega}\mathcal{P}_{\mathbb{T}} - \frac{m}{L}\mathcal{P}_{\mathbb{T}}\|$ is in fact small for random ground truth matrices in $\mathbb{S}$, which is surprising. Chapter 3 outlines an approach that takes the best of both of these operators into consideration, at the cost of some technical complexity.

Computation of $\mathcal{R}_{\Omega}(\boldsymbol{X}_l)$ can be done efficiently, with minimal complexity per iteration. This is because a given iterate $\boldsymbol{X}_l$ can be easily translated to its distance matrix $\boldsymbol{D}_l$ via (1.2), and through (1.10), $\mathcal{R}_{\Omega}(\boldsymbol{X}_l)$ can be computed in $\mathcal{O}(m)$ operations, for $|\Omega| = m$. From [SLCT23], it can be shown that the total cost per loop is approximately $\mathcal{O}(m) + \mathcal{O}(n^2) + \mathcal{O}(mr) + \mathcal{O}(nr^2) + \mathcal{O}(r^3)$ for an $n \times n$ rank-r matrix.

2.1.1 One Additional Surrogate of Interest

One other surrogate was investigated throughout the course of my thesis work, published in [SLCT23]. This surrogate is $\mathcal{R}_{\Omega}^*\mathcal{R}_{\Omega}(\cdot)$, and can be computed as

$$\mathcal{R}_{\Omega}^*\mathcal{R}_{\Omega}(\cdot) = \sum_{\boldsymbol{\alpha}\in\Omega}\langle\cdot, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \sum_{\boldsymbol{\beta}\in\Omega}\langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle \boldsymbol{w}_{\boldsymbol{\beta}}.$$

This operator is symmetric and positive semidefinite, and as such an objective function analogous to (1.14) can be defined below:

$$\underset{\boldsymbol{Y}\in\mathbb{S}}{\text{minimize}} \ \frac{1}{2}\langle \boldsymbol{Y} - \boldsymbol{X}, \mathcal{R}_{\Omega}^*\mathcal{R}_{\Omega}(\boldsymbol{Y} - \boldsymbol{X})\rangle \ \text{ subject to } \ \text{rank}(\boldsymbol{Y}) = r.$$

No theoretical analysis of this surrogate was conducted in [SLCT23], and the operator was superseded by the introduction of the operator in Chapter 3. We mention this object here for posterity as a first pass to this problem, but ultimately it is not particularly relevant to the work conducted in this thesis. More can be seen on this in Remark 3.1.1.

2.2 Theoretical Analysis

In this section, we will provide the main results of this work, which are the local convergence and recovery guarantees for Riemannian Pseudogradient Descent, presented in Theorems 2.2.2, 2.2.4, and 2.2.6. Prior to these guarantees, we will first introduce slightly altered incoherence conditions for the non-orthogonal problem at hand. Pathological cases can arise where a ground truth matrix $\boldsymbol{X}$ has few non-zero coefficients in a dual basis expansion, which can cause issues in the recovery of said matrix from samples. This is well-studied in the standard matrix completion problem, and is captured in the idea of incoherence with respect to the standard basis. Relating incoherence to the underlying geometry of points is an interesting problem, but this is outside the scope of the current work.

Assumption 2.2.1: Incoherence assumption

Let $\boldsymbol{X} \in \mathbb{R}^{n\times n}$ be a rank-r matrix with eigenvalue decomposition $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top$. We assume that $\boldsymbol{X}$ is $\frac{\nu}{4}$-incoherent to the basis $\{\boldsymbol{w}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$, ν-incoherent to its dual basis $\{\boldsymbol{v}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$, and $\frac{\nu}{64}$-incoherent in the standard matrix basis; that is, there exists a constant $\nu \geq 1$ such that for all $\boldsymbol{\alpha} = (i,j) \in \mathbb{I}$:

$$\|\mathcal{P}_U \boldsymbol{e}_{ij}\|_{\mathrm{F}} \leq \sqrt{\frac{\nu r}{128n}}, \qquad \|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \leq \sqrt{\frac{\nu r}{8n}}, \quad \text{and} \quad \|\mathcal{P}_U \boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \leq \sqrt{\frac{\nu r}{2n}}. \tag{2.2}$$

In addition to the above, we require that

$$\|\mathcal{P}_{\mathbb{T}} \boldsymbol{e}_{ij}\|_{\mathrm{F}} \leq \sqrt{\frac{\nu r}{128n}}, \qquad \|\mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \leq \sqrt{\frac{\nu r}{8n}}, \quad \text{and} \quad \|\mathcal{P}_{\mathbb{T}} \boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \leq \sqrt{\frac{\nu r}{2n}}. \tag{2.3}$$

This assumption is in accordance with both the standard definitions of incoherence as $\|\mathcal{P}_U \boldsymbol{e}_{ij}\|_{\mathrm{F}} \leq \|\mathcal{P}_U \boldsymbol{e}_i\|_2$. Additionally, notice that these two definitions are equivalent up to a small constant, as

$$\|\mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} = \|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}} + \boldsymbol{w}_{\boldsymbol{\alpha}} \mathcal{P}_U - \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}} \mathcal{P}_U\|_{\mathrm{F}} \leq 3 \|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}},$$

where the first inequality follows from the triangle inequality and the self-adjointness of $\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}$.

As such, we pick a ν large enough such that the inequalities in (2.2) and (2.3) hold. As in [WCCL20], we note that the first condition above implies the following:

$$\|\mathcal{P}_U \boldsymbol{e}_{ij}\|_{\mathrm{F}}^2 = \langle \mathcal{P}_U \boldsymbol{e}_{ij}, \mathcal{P}_U \boldsymbol{e}_{ij} \rangle = \langle \mathcal{P}_U \boldsymbol{e}_{ij}, \boldsymbol{e}_{ij} \rangle = \mathrm{Trace}\left(\boldsymbol{e}_{ij}\boldsymbol{e}_{ji}\mathcal{P}_U\right) = \langle \mathcal{P}_U, \boldsymbol{e}_{ii} \rangle = \left\|\boldsymbol{U}^{(i)}\right\|_2^2.$$

This indicates that $\left\|\boldsymbol{U}^{(i)}\right\|_2^2 \leq \frac{\nu r}{128n}$, which will be relevant when discussing the initialization of Riemannian Pseudogradient Descent using a trimming step in Algorithm 6.

Remark 2.2.1: Equivalence of incoherence in $\boldsymbol{e}_{ij}$, $\boldsymbol{w}_{\boldsymbol{\alpha}}$, and $\boldsymbol{v}_{\boldsymbol{\alpha}}$

We want to note that the first assumption in (2.2) actually implies the next two. That is to say, if $\|\mathcal{P}_U \boldsymbol{e}_{ij}\|_{\mathrm{F}} \leq \sqrt{\frac{\nu r}{128n}}$, then by the triangle inequality

$$\|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \leq 4 \max_{(i,j)\in\mathbb{I}} \|\mathcal{P}_U \boldsymbol{e}_{ij}\|_{\mathrm{F}} \leq \sqrt{\frac{\nu r}{8n}}.$$

To see the last result, notice that

$$\begin{aligned}
\|\mathcal{P}_U \boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} &= \left\| \mathcal{P}_U \left(\sum_{\boldsymbol{\beta}\in\mathbb{I}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \boldsymbol{w}_{\boldsymbol{\beta}} \right) \right\|_{\mathrm{F}} \\
&\leq \sum_{\boldsymbol{\beta}\in\mathbb{I}} \left| H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \right| \|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\beta}}\|_{\mathrm{F}} \\
&\leq \sqrt{\frac{\nu r}{8n}} \sum_{\boldsymbol{\beta}\in\mathbb{I}} \left| H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \right|,
\end{aligned}$$

and as $\sum_{\boldsymbol{\alpha}\in\mathbb{I}} |H^{\boldsymbol{\alpha}\boldsymbol{\beta}}| \leq 2$ from Lemma 2.A.3, the claim follows. A similar proof shows the same relationship for the equations in (2.3).

Lastly, we are typically interested in large n. Assuming that $n \geq 3$ produces uniform results for several bounds in the appendix, and is formally stated as an assumption.

Assumption 2.2.2: Size of n

For the given ground truth rank-r matrix $\boldsymbol{X} \in \mathbb{R}^{n\times n}$, we assume that $n \geq 3$.

Throughout the remainder of this work, we will assume that our ground truth matrix $\boldsymbol{X} \in \mathbb{S}$ satisfies Assumption 2.2.1 with $\mathcal{O}(1)$ constant factor ν. As in [WCCL20], we identify a

neighborhood in $\mathcal{M}_r$ around which any initial guess in this neighborhood converges linearly to the true solution with high probability.

As mentioned previously, the most critical property for convergence of Riemannian Pseudo-gradient Descent is RIP. This theorem provides the conditions needed for RIP of $\mathcal{R}_\Omega$:

Theorem 2.2.1 (Restricted Isometry Property (RIP) for $\mathcal{P}_\mathbb{T}\mathcal{R}_\Omega\mathcal{P}_\mathbb{T}$)**.** *With probability at least* $1 - 2n^{1-\beta}$,

$$\frac{L}{m}\left\|\mathcal{P}_\mathbb{T}\mathcal{R}_\Omega\mathcal{P}_\mathbb{T} - \frac{m}{L}\mathcal{P}_\mathbb{T}\right\| \leq \sqrt{\frac{8\beta\nu^2r^2n\log(n)}{3m}},$$

for $m \geq \frac{8}{3}\beta\nu^2r^2n\log(n)$*. In particular,*

$$\frac{L}{m}\left\|\mathcal{P}_\mathbb{T}\mathcal{R}_\Omega\mathcal{P}_\mathbb{T} - \frac{m}{L}\mathcal{P}_\mathbb{T}\right\| \leq \varepsilon_0,$$

for any $\varepsilon_0 > 0$ *if* $m \geq \frac{8}{3}\beta\left(\frac{\nu r}{\varepsilon_0}\right)^2 n\log(n)$*. Additionally, under the same conditions as above, we also have*

$$\frac{L}{m}\left\|\mathcal{P}_\mathbb{T}\mathcal{R}_\Omega^*\mathcal{P}_\mathbb{T} - \frac{m}{L}\mathcal{P}_\mathbb{T}\right\| \leq \varepsilon_0.$$

Proof sketch. This result is primarily a consequence of Theorem 1.7.2, the non-commutative Bernstein inequality. As $\mathbb{E}[\mathcal{R}_\Omega] = \frac{m}{L}\mathcal{I}$, we see that $\mathbb{E}[\mathcal{P}_\mathbb{T}\mathcal{R}_\Omega\mathcal{P}_\mathbb{T}] = \frac{m}{L}\mathcal{P}_\mathbb{T}$, and the rest of the proof leverages specific properties of the dual bases $\{\boldsymbol{w}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$ and $\{\boldsymbol{v}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$ to prove the concentration of $\mathcal{P}_\mathbb{T}\mathcal{R}_\Omega\mathcal{P}_\mathbb{T}$ around its expectation. See Appendix 2.B.1 for details. □

This leads us into the following theorem, which is the crux of this work. Theorem 2.2.2, stated with a brief proof outline in the main text, describes a local attractive basin around the ground truth solution, provided that $\mathcal{P}_\mathbb{T}\mathcal{R}_\Omega\mathcal{P}_\mathbb{T}$ exhibits RIP. With high probability, the algorithm will converge to the ground truth from any initialization in this attractive basin. The full proof is delayed to Appendix 2.C.1.

Theorem 2.2.2 (Local Convergence of Algorithm 4)**.** *Let* $\boldsymbol{X} \in \mathbb{R}^{n\times n}$ *be the measured rank-r matrix and let* $\mathbb{T}$ *be the tangent space of* $\mathcal{M}_r$ *at* $\boldsymbol{X}$*. Suppose that*

$$\left\|\frac{L}{m}\mathcal{P}_\mathbb{T}\mathcal{R}_\Omega\mathcal{P}_\mathbb{T} - \mathcal{P}_\mathbb{T}\right\| \leq \varepsilon_0, \tag{2.4}$$

$$\frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_\mathrm{F}}{\sigma_{min}(\boldsymbol{X})} \leq \frac{\sqrt{m}\varepsilon_0}{16n^{5/4}\sqrt{\beta\nu r\log n}}, \tag{2.5}$$

$$\|\mathcal{R}_\Omega\| \leq \frac{m}{L} + 4\sqrt{\frac{8m\log(n)}{n}}, \tag{2.6}$$

$$\|\mathcal{R}_\Omega\mathcal{P}_\mathbb{T}\| \leq \frac{m}{L} + \frac{m\sqrt{n}}{L}\sqrt{\frac{\beta\nu rn\log(n)}{3m}}, \tag{2.7}$$

$$\|\mathcal{P}_\mathbb{T}\mathcal{R}_\Omega\| \leq \frac{m}{L} + \frac{4m\sqrt{n}}{L}\sqrt{\frac{\beta\nu rn\log(n)}{3m}}, \tag{2.8}$$

where ε_0 is a constant satisfying

$$\delta = \frac{18\varepsilon_0}{1-4\varepsilon_0} < 1.$$

Then the algorithm converges linearly as the iterates satisfy

$$\|\boldsymbol{X}_{l+1} - \boldsymbol{X}\|_\mathrm{F} \leq \delta^l \|\boldsymbol{X}_0 - \boldsymbol{X}\|_\mathrm{F}.$$

Proof sketch of Theorem 2.2.2. The theorem begins first by simple linear algebra, as we have

$$\begin{aligned}\|\boldsymbol{X}_{l+1} - \boldsymbol{X}\|_\mathrm{F} &= \|\boldsymbol{X}_{l+1} - \boldsymbol{W}_l - \boldsymbol{X} + \boldsymbol{W}_l\|_\mathrm{F} \\ &\leq \|\boldsymbol{X}_{l+1} - \boldsymbol{W}_l\|_\mathrm{F} + \|\boldsymbol{X} - \boldsymbol{W}_l\|_\mathrm{F} \\ &\leq 2\|\boldsymbol{W}_l - \boldsymbol{X}\|_\mathrm{F},\end{aligned}$$

where the last inequality follows from $\boldsymbol{X}_{l+1}$ being the best rank-r approximation to $\boldsymbol{W}_l$ by Eckart-Young-Mirsky [EY36]. Next, plugging in $\boldsymbol{W}_l = \boldsymbol{X}_l + \alpha_l\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l$, we see that

$$\begin{aligned}\|\boldsymbol{X}_{l+1} - \boldsymbol{X}\|_\mathrm{F} &\leq 2\left\|\boldsymbol{X}_l + \alpha_l\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l - \boldsymbol{X}\right\|_\mathrm{F} \\ &= 2\|\boldsymbol{X}_l - \boldsymbol{X} - \alpha_l\mathcal{P}_{\mathbb{T}_l}\mathcal{R}_\Omega(\boldsymbol{X}_l - \boldsymbol{X})\|_\mathrm{F} \\ &\leq \underbrace{2\|(\mathcal{P}_{\mathbb{T}_l} - \alpha_l\mathcal{P}_{\mathbb{T}_l}\mathcal{R}_\Omega\mathcal{P}_{\mathbb{T}_l})(\boldsymbol{X}_l - \boldsymbol{X})\|_\mathrm{F}}_{I_1} \\ &\quad + \underbrace{2\|(\mathcal{I} - \mathcal{P}_{\mathbb{T}_l})(\boldsymbol{X}_l - \boldsymbol{X})\|_\mathrm{F}}_{I_2} \\ &\quad + \underbrace{2|\alpha_l|\|\mathcal{P}_{\mathbb{T}_l}\mathcal{R}_\Omega(\mathcal{I} - \mathcal{P}_{\mathbb{T}_l})(\boldsymbol{X}_l - \boldsymbol{X})\|_\mathrm{F}}_{I_3}.\end{aligned}$$

The remainder of the proof is in the bounding of I_1, I_2, and I_3. I_1 is proven by showing that in a neighborhood of the solution, defined by (2.5), a local form of RIP for $\mathcal{R}_\Omega$ holds if

(2.4) is true. This proof leverages the assumptions made in (2.6), (2.7), and (2.8). I_2 follows from the neighborhood assumption of (2.5) in tandem with Lemma 1.8.3, and I_3 follows from bounds on the step size (seen in Lemma 2.C.1), the assumption in (2.8), and Lemma 1.8.3. The assumptions in (2.4), (2.6), (2.7), and (2.8) are all proven via high probability guarantees using Theorem 1.7.2. The technical details are deferred to the appendix. See the full proof in 2.C.1. □

In Theorem 2.2.2, the linear convergence rate required that $\varepsilon_0 < \frac{1}{22}$, which is satisfied for $m > 1300\nu^2 r^2 n \log(n)$, a smaller constant factor than that required of the following initialization guarantees. That is to say, the local neighborhood guarantees are stricter in practice than the RIP requirement.

2.2.1 Initialization

Given that the convergence of this algorithm is only local, initialization is important to consider in the context of sample complexity. The simplest initialization, a hard thresholding to $\mathcal{M}_r$ of the measured information, provides a reasonable starting point. The following theorem describes how close such an initialization might be to the ground truth.

Lemma 2.2.3 (Initialization via One Step Hard Thresholding)**.** *Let $\boldsymbol{X} \in \mathbb{R}^{n\times n}$ be the underlying measured rank-r matrix, and let $\boldsymbol{X}_0 = \frac{L}{m}\mathcal{H}_r(\mathcal{R}_\Omega(X))$ with $\Omega = m \geq \frac{32}{3}\beta n \log n$. It follows that with probability at least $1 - 2n^{1-\beta}$ that*

$$\|\boldsymbol{X}_0 - \boldsymbol{X}\|_{\mathrm{F}} \leq \sqrt{\frac{\beta\nu^2 r^3 n \log(n)}{3m}}\|\boldsymbol{X}\|, \tag{2.9}$$

Proof. See Section 2.D.1. □

This result shows that with probability at least $1 - 2n^{1-\beta}$, the neighborhood assumption (2.5), important to Lemma 2.B.3, is satisfied if

$$m \geq \frac{10\kappa\beta\nu^{3/2}r^2}{\varepsilon_0} n^{7/4}\log(n),$$

where $\kappa = \frac{\|\boldsymbol{X}\|}{\sigma_{\min}(\boldsymbol{X})}$ is the condition number of $\boldsymbol{X}$.

This leads into the following theorem, which is one of the primary results of the work.

Theorem 2.2.4 (Recovery Guarantee I)**.** *Suppose $|\Omega| = m$ with the indices sampled uniformly with replacement. Given an initialization of $\boldsymbol{X}_0 = \frac{L}{m}\mathcal{H}_r(\mathcal{R}_\Omega(\boldsymbol{X}))$ for the rank-r, ν-incoherent ground truth matrix $\boldsymbol{X}$ with condition number κ, and given*

$$m \geq \frac{10\kappa\beta\nu^{3/2}r^2}{\varepsilon_0} n^{7/4}\log(n),$$

with $\beta > 1$, then Riemannian Pseudogradient Descent linearly converges to the ground truth $\boldsymbol{X}$ with probability at least $1 - 10n^{1-\beta}$.

Proof. This follows from taking the local neighborhood assumption seen in Theorem 2.2.2. By setting the bound produced in Lemma 2.2.3 to be less than required for local convergence, the result follows after some minor algebra and taking the maximum with the sample complexity requirements seen in Theorem 2.2.1. □

This naive one-step hard thresholding initialization can be improved again following a construction in [WCCL20], using a resampling and trimming algorithm, both defined as follows:

Algorithm 5 Riemannian Resampling for Initialization

Partition Ω into $S+1$ equal groups $\Omega_0, \Omega_1, ..., \Omega_S$, each of size $\hat{m}$
Set $\boldsymbol{Z}_0 = \mathcal{H}_r\left(\frac{L}{\hat{m}}\mathcal{R}_{\Omega_0}(\boldsymbol{X})\right)$
for $l = 0, 1, ..., S-1$ **do**
1. $\hat{\boldsymbol{Z}}_l = \texttt{trim}(\boldsymbol{Z}_l)$
2. $\boldsymbol{Z}_{l+1} = \mathcal{H}_r\left(\hat{\boldsymbol{Z}}_l + \frac{L}{\hat{m}}\mathcal{P}_{\hat{\mathbb{T}}_l}\mathcal{R}_{\Omega_{l+1}}(\boldsymbol{X} - \hat{\boldsymbol{Z}}_l)\right)$

end for
Output: $\boldsymbol{X}_0 = \boldsymbol{Z}_S$.

Algorithm 6 `trim`

Input: $\boldsymbol{Z}_l = \boldsymbol{U}_l\boldsymbol{D}_l\boldsymbol{U}_l^\top$
Output: $\widehat{\boldsymbol{Z}}_l = \boldsymbol{A}_l\boldsymbol{D}_l\boldsymbol{A}_l^\top$, where $\boldsymbol{A}_l^{(i)} = \frac{\boldsymbol{U}_l^{(i)}}{\left\|\boldsymbol{U}_l^{(i)}\right\|_2}\min\left\{\left\|\boldsymbol{U}_l^{(i)}\right\|_2, \sqrt{\frac{\nu r}{n}}\right\}$.

This trimming algorithm is a projection onto the space of matrices that are ν-incoherent with respect to the standard matrix basis, not necessarily with respect to the basis $\{\boldsymbol{w}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$. However, as noted previously, the incoherence parameter differs by at most an $\mathcal{O}(1)$ constant, so this is a reasonable surrogate, especially for large n. Following Section 1.2.1 and Remark 1.6.2,

we can view this trimming algorithm as a projection of the $\boldsymbol{p}_i$ onto a sphere of radius $\frac{\nu r}{n}$ in the Mahalanobis metric, defined as $\boldsymbol{p}_i(\boldsymbol{P}^\top \boldsymbol{P})^{-1}\boldsymbol{p}_i$.

We can analyze Algorithm 5 and get the following result:

Lemma 2.2.5 (Riemannian Resampling Result)**.** *Let $\boldsymbol{X} \in \mathbb{R}^{n\times n}$ be the measured rank-r matrix with condition number κ. Let S be the number of partitions specified in Algorithm 5, and let $\hat{m} = \frac{m}{S+1}$. Then for all $\beta > 1$, with probability at least $1-(2+4S)n^{1-\beta}$, the output of Algorithm 5 satisfies*

$$\|\boldsymbol{X}_0 - \boldsymbol{X}\|_{\mathrm{F}} \leq \left(\frac{5}{6}\right)^S \frac{\sigma_{\min}(\boldsymbol{X})}{256\kappa^2},$$

provided that $\hat{m} \geq (2.2 \times 10^4)\kappa^6\nu^2 r^3 \beta n \log(n)$.

Proof. See 2.D.2. □

Assuming that $m \geq \beta\nu r n \log(n)$, which is relaxed from the requirement for RIP seen in Theorem 2.2.1, Lemma 2.2.5 shows that taking

$$S \geq 6 \log\left(\frac{n^{3/4}}{16\varepsilon_0}\right),$$

the third condition in Lemma 2.B.3 can be satisfied with probability at least

$$1 - \left(6 + 24\log\left(\frac{n^{3/4}}{16\varepsilon_0}\right)\right) n^{1-\beta},$$

for a large enough sample complexity. This leads to the final recovery guarantee, attenuating the n dependence in the sample complexity.

Theorem 2.2.6. *Let $\boldsymbol{X} \in \mathbb{R}^{n\times n}$ be the measured rank-r, ν-incoherent matrix with condition number κ, and suppose that $\Omega \subset \mathbb{I}$ is a set of indices sampled uniformly at random with replacement with $|\Omega| = m$. Let $\boldsymbol{X}_0$ be the output of Algorithm 5. Then for any $\beta > 1$, the iterates of Riemannian Pseudogradient Descent linearly converge to $\boldsymbol{X}$ with probability at least $1 - \left(6 + 24\log\left(\frac{n^{3/4}}{16\varepsilon_0}\right)\right) n^{1-\beta}$ provided that*

$$m \geq (2.4 \times 10^5)\kappa^6\nu^2 r^3 \beta n \log^2(n).$$

Proof. As in Theorem 2.2.4, this proof follows from the local neighborhood conditions of Theorem 2.2.2, combined with the sample complexity results from Lemma 2.2.5 and Theorem 2.2.1. The constants are not optimized, and could be further improved. □

2.3 Numerical Results

In this section, we test the proposed algorithms on synthetic and real data.

2.3.1 Synthetic Data Experiments

To test Algorithms 3 and 4, various two and three dimensional datasets were used, and are referred to in Table 2.1 with their corresponding sizes. The goal of Algorithms 3 and 4 is to recover the full set of points $\boldsymbol{P}$ up to orthogonal transformation by sampling the entries above the diagonal of $\boldsymbol{D}$ uniformly with replacement, with a total of γL entries chosen for $\gamma \in [0, 1]$. Riemannian Pseudogradient Descent reconstructs the Gram matrix $\boldsymbol{X} = \boldsymbol{P}^\top \boldsymbol{P}$, from which $\boldsymbol{P}$ can be recovered. The comparison referenced in Table 2.1 is the relative error between the recovered matrix $\boldsymbol{X}_{\text{rec}}$ and the ground truth matrix $\boldsymbol{X}$ in Frobenius norm. Each run was terminated at either 1000 iterations or when a relative Frobenius norm difference between iterates of 10^{-5} was achieved.

As indicated by these experiments, RF-RGD more reliably reconstructs the underlying datasets from distance samples than Riemannian Pseudogradient Descent, but both are outperformed by the non-convex algorithm in [TL18]. However, with the non-convex algorithm in [TL18] there is little hope of ever conducting local convergence analysis of this algorithm, whereas Riemannian Pseudogradient Descent has been proven to exhibit local convergence. The success of $\mathcal{F}_\Omega$ in this algorithm indicates that there might be a different method for local convergence that does not use RIP explicitly. Future work in this direction would be interesting, particularly as this approach is more theoretically straightforward than the approach outlined in Chapter 3.

It remains to be seen if RF-RGD will be provably locally convergent, as determining a high-probability bound on $\|\mathcal{P}_{\mathbb{T}}\mathcal{F}_\Omega\mathcal{P}_{\mathbb{T}} - c\mathcal{P}_{\mathbb{T}}\|$ for some $c > 0$ has proven challenging. As such, more practical utility lies in RF-RGD and the algorithm in [TL18] than in Riemannian Pseudogradient

Table 2.1: Relative recovery error $\|\boldsymbol{X} - \boldsymbol{X}_{\text{rec}}\|_{\text{F}} / \|\boldsymbol{X}\|_{\text{F}}$ between the recovered Gram matrix and the true Gram matrix averaged over 25 trials using Algorithms 3, 4, and the non-convex algorithm in [TL18].

Dataset \ γ	7%	5%	3%	2%	1%
RF-RGD					
Sphere (3D, $n = 1002$)	3.69×10^{-5}	6.01×10^{-5}	1.28×10^{-4}	6.82×10^{-3}	9.11×10^{-1}
Cow (3D, $n = 2601$)	4.06×10^{-5}	4.89×10^{-5}	7.61×10^{-5}	1.07×10^{-4}	8.60×10^{-3}
Swiss Roll (3D, $n = 2048$)	3.97×10^{-5}	5.06×10^{-5}	7.71×10^{-5}	1.21×10^{-4}	5.52×10^{-2}
U.S. Cities (2D, $n = 2920$)	4.67×10^{-5}	2.01×10^{-3}	6.94×10^{-3}	1.63×10^{-2}	6.08×10^{-2}
Riemannian Pseudogradient Descent					
Sphere (3D, $n = 1002$)	3.86×10^{-5}	2.11×10^{-4}	7.88×10^{-2}	2.29×10^{-1}	1.78×10^{0}
Cow (3D, $n = 2601$)	3.40×10^{-2}	6.28×10^{-2}	1.76×10^{-1}	3.60×10^{0}	7.73×10^{-1}
Swiss Roll (3D, $n = 2048$)	1.46×10^{-5}	3.00×10^{-5}	1.52×10^{-3}	2.82×10^{-1}	9.73×10^{-1}
U.S. Cities (2D, $n = 2920$)	1.23×10^{-1}	1.46×10^{-1}	1.86×10^{-1}	3.13×10^{-1}	8.35×10^{-1}
Non-convex algorithm in [TL18]					
Sphere (3D, $n = 1002$)	9.86×10^{-6}	1.36×10^{-5}	3.04×10^{-5}	6.18×10^{-5}	1.00×10^{-1}
Cow (3D, $n = 2601$)	7.66×10^{-6}	1.06×10^{-5}	1.65×10^{-5}	2.11×10^{-5}	4.46×10^{-5}
Swiss Roll (3D, $n = 2048$)	1.22×10^{-6}	1.01×10^{-6}	1.87×10^{-6}	1.06×10^{-6}	3.34×10^{-5}
U.S. Cities (2D, $n = 2920$)	6.09×10^{-7}	8.19×10^{-7}	1.32×10^{-6}	2.30×10^{-6}	4.69×10^{-6}

Descent, but the strong theoretical results provided by Riemannian Pseudogradient Descent will guide future work for the convergence analysis of RF-RGD and other self-adjoint surrogates for $\mathcal{R}_\Omega$.

2.A Properties of the Dual Bases

In this section, we provide a few properties of the dual bases $\{\boldsymbol{w}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$ and $\{\boldsymbol{v}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$ that will be relevant through the remainder of this work.

Lemma 2.A.1 ($\lambda_{\max}(\tilde{\boldsymbol{H}})$ bound)**.** *Let $\tilde{\boldsymbol{H}} = [\langle \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}, \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\beta}} \rangle] \in \mathbb{R}^{L\times L}$, where U is the row/column space of the true solution $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top$, which is rank-$r$, and where $\mathcal{P}_U$ is the projection operator onto U. It follows that*

$$\lambda_{\max}(\tilde{\boldsymbol{H}}) \leq \nu r.$$

Proof. First, by incoherence we have that

$$|\langle \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}, \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\beta}} \rangle| \leq \|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\text{F}} \|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\beta}}\|_{\text{F}} \leq \frac{\nu r}{2n}.$$

Next, as $\mathcal{P}_U = \boldsymbol{U}\boldsymbol{U}^\top$, for $\boldsymbol{\alpha} \cap \boldsymbol{\beta} = \emptyset$

$$\langle \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}, \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\beta}} \rangle = \mathrm{Trace}(\boldsymbol{w}_{\boldsymbol{\alpha}} \mathcal{P}_U \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\beta}}) = \mathrm{Trace}(\boldsymbol{w}_{\boldsymbol{\beta}} \boldsymbol{w}_{\boldsymbol{\alpha}} \mathcal{P}_U) = \mathrm{Trace}(\boldsymbol{0}\mathcal{P}_U) = 0,$$

as $\boldsymbol{w}_{\boldsymbol{\alpha}} \boldsymbol{w}_{\boldsymbol{\beta}} = \boldsymbol{w}_{\boldsymbol{\beta}} \boldsymbol{w}_{\boldsymbol{\alpha}} = \boldsymbol{0}$, where $\boldsymbol{0}$ is the zero matrix. Thus $\tilde{\boldsymbol{H}}$ is sparse, with each row having at most $2n-3$ non-zero entries. The result follows from a Gershgorin argument and the entrywise bound derived from the incoherence condition above. □

Lemma 2.A.2. *For any $\boldsymbol{X} \in \mathbb{R}^{n \times n}$, $\boldsymbol{X} = \boldsymbol{X}^\top$, and any $\boldsymbol{w}_{\boldsymbol{\alpha}} \in \{\boldsymbol{w}_{\boldsymbol{\beta}}\}_{\boldsymbol{\beta} \in \mathbb{I}}$,*

$$\langle \mathcal{P}_{\mathbb{T}} \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle = \langle \boldsymbol{X}, \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle + \langle \boldsymbol{X} \mathcal{P}_{U^\perp}, \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle.$$

Additionally for $\|\boldsymbol{X}\|_{\mathrm{F}} = 1$,

$$\sum_{\boldsymbol{\alpha} \in \mathbb{I}} \langle \boldsymbol{X}, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle^2 \leq 4\lambda_{\max}(\tilde{\boldsymbol{H}}) = 4\nu r.$$

Proof. First, notice that $\langle \boldsymbol{X}\mathcal{P}_U, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle = \langle \mathcal{P}_U \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle$ due to cyclicity of the trace and symmetry of $\boldsymbol{X}$, $\mathcal{P}_U$, and $\boldsymbol{w}_{\boldsymbol{\alpha}}$. It follows then that

$$\begin{aligned}
\langle \mathcal{P}_{\mathbb{T}} \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle &= \langle \mathcal{P}_U \boldsymbol{X} + \boldsymbol{X}\mathcal{P}_U - \mathcal{P}_U \boldsymbol{X} \mathcal{P}_U, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \\
&= 2\langle \mathcal{P}_U \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle - \langle \mathcal{P}_U \boldsymbol{X} \mathcal{P}_U, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \\
&= \langle \mathcal{P}_U \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle + \langle \mathcal{P}_U \boldsymbol{X} - \mathcal{P}_U \boldsymbol{X} \mathcal{P}_U, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \\
&= \langle \mathcal{P}_U \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle + \langle \mathcal{P}_U \boldsymbol{X} \mathcal{P}_{U^\perp}, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \\
&= \langle \boldsymbol{X}, \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle + \langle \boldsymbol{X} \mathcal{P}_{U^\perp}, \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle.
\end{aligned}$$

To see the second statement, notice that from Young's inequality we have

$$\begin{aligned}
\sum_{\boldsymbol{\alpha} \in \mathbb{I}} \langle \boldsymbol{X}, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle^2 &= \sum_{\boldsymbol{\alpha} \in \mathbb{I}} \left(\langle \boldsymbol{X}, \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle + \langle \boldsymbol{X} \mathcal{P}_{U^\perp}, \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \right)^2 \\
&\leq 2 \sum_{\boldsymbol{\alpha} \in \mathbb{I}} \langle \boldsymbol{X}, \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle^2 + 2 \sum_{\boldsymbol{\alpha} \in \mathbb{I}} \langle \boldsymbol{X} \mathcal{P}_{U^\perp}, \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle^2.
\end{aligned}$$

As such,

$$\begin{aligned}
\sum_{\boldsymbol{\alpha}\in\mathbb{I}}\langle \boldsymbol{X}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 &\leq \max_{\|\boldsymbol{X}\|_{\mathrm{F}}=1}\sum_{\boldsymbol{\alpha}\in\mathbb{I}}\langle \boldsymbol{X}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \\
&\leq \max_{\|\boldsymbol{X}\|_{\mathrm{F}}=1}\sum_{\boldsymbol{\alpha}\in\mathbb{I}} 2\langle \boldsymbol{X}, \mathcal{P}_{U}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 + 2\langle \boldsymbol{X}\mathcal{P}_{U^\perp}, \mathcal{P}_{U}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \\
&\leq 2\max_{\|\boldsymbol{X}\|_{\mathrm{F}}=1}\sum_{\boldsymbol{\alpha}\in\mathbb{I}}\langle \boldsymbol{X}, \mathcal{P}_{U}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 + 2\max_{\|\boldsymbol{X}\|_{\mathrm{F}}=1}\sum_{\boldsymbol{\alpha}\in\mathbb{I}}\langle \boldsymbol{X}\mathcal{P}_{U^\perp}, \mathcal{P}_{U}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \\
&\leq 4\max_{\|\boldsymbol{X}\|_{\mathrm{F}}=1}\sum_{\boldsymbol{\alpha}\in\mathbb{I}}\langle \boldsymbol{X}, \mathcal{P}_{U}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \\
&\leq 4\nu r,
\end{aligned}$$

with the second to last inequality following from the fact that $\mathcal{P}_{U^\perp}$ is an orthogonal projection operator and the last line following from Theorem 1.8.2 and Theorem 2.A.1. This concludes the proof. □

Lemma 2.A.3 (Eigenvalues of $\boldsymbol{H}$ and $\boldsymbol{H}^{-1}$, entries of $\boldsymbol{H}^{-1}$, and spectral norms of $\boldsymbol{w}_{\boldsymbol{\alpha}}$ and $\boldsymbol{v}_{\boldsymbol{\alpha}}$ [LT23])**.** *Let $\boldsymbol{H} \in \mathbb{R}^{L\times L}$ be the Gram matrix for $\{\boldsymbol{w}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$ defined by $H_{\boldsymbol{\alpha}\boldsymbol{\beta}} = \langle \boldsymbol{w}_{\boldsymbol{\alpha}}, \boldsymbol{w}_{\boldsymbol{\beta}}\rangle$, and let $\boldsymbol{H}^{-1}$ be its inverse. Then*

$$\lambda_{\max}(\boldsymbol{H}) = 2n, \qquad \lambda_{\max}(\boldsymbol{H}^{-1}) = \frac{1}{2}.$$

Additionally,

$$H^{\boldsymbol{\alpha}\boldsymbol{\beta}} = \begin{cases} \frac{1}{n^2} & \boldsymbol{\alpha}\cap\boldsymbol{\beta} = \emptyset; \\ -\frac{1}{2n} + \frac{1}{n^2} & \boldsymbol{\alpha}\cap\boldsymbol{\beta} \neq \emptyset, \boldsymbol{\alpha} \neq \boldsymbol{\beta}; \\ \frac{1}{2}\left(1 - \frac{2}{n} + \frac{2}{n^2}\right) & \boldsymbol{\alpha} = \boldsymbol{\beta}, \end{cases}$$

and,

$$\|\boldsymbol{w}_{\boldsymbol{\alpha}}\| = 2, \qquad \|\boldsymbol{v}_{\boldsymbol{\alpha}}\| = \frac{1}{2}.$$

Lemma 2.A.4. *Let $\{\boldsymbol{v}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$ be the dual basis to $\{\boldsymbol{w}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$. It follows that*

$$\sum_{\boldsymbol{\alpha}\in\mathbb{I}} \boldsymbol{v}_{\boldsymbol{\alpha}}^2 = \frac{n^2 - 2n + 2}{4n}\boldsymbol{J}.$$

Proof. Recall that $\boldsymbol{v}_{\boldsymbol{\alpha}} = -\frac{1}{2}\left(\boldsymbol{a}\boldsymbol{b}^\top + \boldsymbol{b}\boldsymbol{a}^\top\right)$ where $\boldsymbol{a} = \boldsymbol{e}_i - \frac{1}{n}\mathbf{1}$ and $\boldsymbol{b} = \boldsymbol{e}_j - \frac{1}{n}\mathbf{1}$ for $\boldsymbol{\alpha} = (i,j)$. It follows that

$$4\boldsymbol{v}_{\boldsymbol{\alpha}}^2 = \boldsymbol{a}\boldsymbol{b}^\top\boldsymbol{a}\boldsymbol{b}^\top + \boldsymbol{a}\boldsymbol{b}^\top\boldsymbol{b}\boldsymbol{a}^\top + \boldsymbol{b}\boldsymbol{a}^\top\boldsymbol{a}\boldsymbol{b}^\top + \boldsymbol{b}\boldsymbol{a}^\top\boldsymbol{b}\boldsymbol{a}^\top,$$

and as $\boldsymbol{b}^\top\boldsymbol{b} = \boldsymbol{a}^\top\boldsymbol{a} = \frac{n-1}{n}$ and $\boldsymbol{a}^\top\boldsymbol{b} = -\frac{1}{n}$, we see that

$$\begin{aligned}
4\boldsymbol{v}_{\boldsymbol{\alpha}}^2 &= \frac{n-1}{n}\left[\left(\boldsymbol{e}_{ii} - \frac{1}{n}\boldsymbol{e}_i\mathbf{1}^\top - \frac{1}{n}\mathbf{1}\boldsymbol{e}_i^\top + \frac{1}{n^2}\mathbf{1}\mathbf{1}^\top\right) + \left(\boldsymbol{e}_{jj} - \frac{1}{n}\boldsymbol{e}_j\mathbf{1}^\top - \frac{1}{n}\mathbf{1}\boldsymbol{e}_j^\top + \frac{1}{n^2}\mathbf{1}\mathbf{1}^\top\right)\right] \\
&\quad - \frac{1}{n}\left[\left(\boldsymbol{e}_{ij} - \frac{1}{n}\boldsymbol{e}_i\mathbf{1}^\top - \frac{1}{n}\mathbf{1}\boldsymbol{e}_j + \frac{1}{n^2}\mathbf{1}\mathbf{1}^\top\right) + \left(\boldsymbol{e}_{ji} - \frac{1}{n}\boldsymbol{e}_j\mathbf{1}^\top - \frac{1}{n}\mathbf{1}\boldsymbol{e}_i^\top + \frac{1}{n^2}\mathbf{1}\mathbf{1}^\top\right)\right] \\
&= \frac{n-1}{n}\left(\boldsymbol{e}_{ii} + \boldsymbol{e}_{jj}\right) + \frac{2-n}{n^2}\left(\boldsymbol{e}_i\mathbf{1}^\top + \mathbf{1}\boldsymbol{e}_i^\top + \boldsymbol{e}_j\mathbf{1}^\top + \mathbf{1}\boldsymbol{e}_j^\top\right) + \frac{2(n-2)}{n^3}\mathbf{1}\mathbf{1}^\top - \frac{1}{n}\left(\boldsymbol{e}_{ij} + \boldsymbol{e}_{ji}\right).
\end{aligned}$$

So it follows that

$$\begin{aligned}
\sum_{\boldsymbol{\alpha}\in\mathbb{I}} 4\boldsymbol{v}_{\boldsymbol{\alpha}}^2 &= \frac{(n-1)^2}{n}\boldsymbol{I} + \frac{2(2-n)(n-1)}{n^2}\mathbf{1}\mathbf{1}^\top + \frac{(n-1)(n-2)}{n^2}\mathbf{1}\mathbf{1}^\top - \frac{1}{n}(\mathbf{1}\mathbf{1}^\top - \boldsymbol{I}), \\
&= \frac{n^2-2n+2}{n}\boldsymbol{I} - \frac{n^2-2n+2}{n^2}\mathbf{1}\mathbf{1}^\top,
\end{aligned}$$

yielding the desired result as $\boldsymbol{J} = \boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$. □

2.B Restricted Isometry Results

As RIP and its variants are critical to the analysis of Riemannian Pseudogradient Descent, this section is dedicated to the proofs of RIP and similar results.

2.B.1 Proof of Theorem 2.2.1

Proof. First, notice that for any dual basis pair $\{\boldsymbol{w}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$ and $\{\boldsymbol{v}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$, we can decompose any $\boldsymbol{X} \in \mathbb{S}$ as

$$\boldsymbol{X} = \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \boldsymbol{v}_{\boldsymbol{\alpha}}.$$

It follows then that

$$\mathbb{E}\left(\mathcal{R}_\Omega\right) = \frac{m}{L}\mathcal{I},$$

where $|\Omega| = m$ and $\mathcal{I}$ is the identity operator on $\mathbb{S}$. From this, we can see that

$$\begin{aligned}\mathcal{P}_{\mathbb{T}}\boldsymbol{X} &= \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \boldsymbol{X}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \boldsymbol{v}_{\boldsymbol{\alpha}},\\ \mathcal{R}_{\Omega}\mathcal{P}_{\mathbb{T}}\boldsymbol{X} &= \sum_{\boldsymbol{\alpha}\in\Omega} \langle \boldsymbol{X}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \boldsymbol{v}_{\boldsymbol{\alpha}},\\ \mathcal{P}_{\mathbb{T}}\mathcal{R}_{\Omega}\mathcal{P}_{\mathbb{T}}\boldsymbol{X} &= \sum_{\boldsymbol{\alpha}\in\Omega} \langle \boldsymbol{X}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \mathcal{P}_{\mathbb{T}}\boldsymbol{v}_{\boldsymbol{\alpha}}.\end{aligned}$$

Therefore it follows that $\mathbb{E}(\mathcal{P}_{\mathbb{T}}\mathcal{R}_{\Omega}\mathcal{P}_{\mathbb{T}}) = \frac{m}{L}\mathcal{P}_{\mathbb{T}}$. We can now use Theorem 1.7.2 to bound the probability that $\mathcal{P}_{\mathbb{T}}\mathcal{R}_{\Omega}\mathcal{P}_{\mathbb{T}}$ deviates from its expected spectral norm. To do this, we first define an operator $\mathcal{T}_{\boldsymbol{\alpha}} = \langle \cdot, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \mathcal{P}_{\mathbb{T}}\boldsymbol{v}_{\boldsymbol{\alpha}} - \frac{1}{L}\mathcal{P}_{\mathbb{T}}$. First, observe the following coherence conditions outlined in Assumption 2.2.1,

$$\|\langle \cdot, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \mathcal{P}_{\mathbb{T}}\boldsymbol{v}_{\boldsymbol{\alpha}}\| \leq \|\mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \|\mathcal{P}_{\mathbb{T}}\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \leq \sqrt{\frac{\nu r}{8n}}\sqrt{\frac{\nu r}{2n}} \leq \frac{\nu r}{2n}.$$

Additionally, $\mathbb{E}(\mathcal{T}_{\alpha}) = \frac{1}{L}\mathcal{I}$. It follows then that

$$\begin{aligned}\|\mathcal{T}_{\alpha}\| &\leq \|\langle \cdot, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \mathcal{P}_{\mathbb{T}}\boldsymbol{v}_{\boldsymbol{\alpha}}\| + \frac{1}{L}\\ &\leq \frac{\nu r}{2n} + \frac{1}{L}\\ &\leq \frac{\nu r}{n} \leq \frac{\nu^2 r^2}{n} =: c,\end{aligned}$$

as $\nu, r \geq 1$, which gives us our almost sure estimate on the spectral norm of each term $\mathcal{T}_{\alpha}$. Next, to estimate the variance we notice first that

$$\begin{aligned}\mathbb{E}\left(\mathcal{T}_{\boldsymbol{\alpha}}\mathcal{T}_{\boldsymbol{\alpha}}^*\right) &= \frac{1}{L}\sum_{\boldsymbol{\alpha}\in\mathbb{I}} \mathcal{T}_{\alpha}\mathcal{T}_{\alpha}^* = \frac{1}{L}\sum_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \cdot, \mathcal{P}_{\mathbb{T}}\boldsymbol{v}_{\boldsymbol{\alpha}}\rangle \langle \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \mathcal{P}_{\mathbb{T}}\boldsymbol{v}_{\boldsymbol{\alpha}} - \frac{1}{L^2}\mathcal{P}_{\mathbb{T}},\\ \mathbb{E}\left(\mathcal{T}_{\boldsymbol{\alpha}}^*\mathcal{T}_{\boldsymbol{\alpha}}\right) &= \frac{1}{L}\sum_{\boldsymbol{\alpha}\in\mathbb{I}} \mathcal{T}_{\alpha}^*\mathcal{T}_{\alpha} = \frac{1}{L}\sum_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \cdot, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \langle \mathcal{P}_{\mathbb{T}}\boldsymbol{v}_{\boldsymbol{\alpha}}, \mathcal{P}_{\mathbb{T}}\boldsymbol{v}_{\boldsymbol{\alpha}}\rangle \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}} - \frac{1}{L^2}\mathcal{P}_{\mathbb{T}}.\end{aligned}$$

To bound the maximum spectral norm of the above two terms, notice that for $\sum_{\boldsymbol{\alpha}\in\mathbb{I}} \mathcal{T}_{\boldsymbol{\alpha}}\mathcal{T}_{\boldsymbol{\alpha}}^*$,

$$\left\|\frac{1}{L}\sum_{\boldsymbol{\alpha}\in\mathbb{I}} \mathcal{T}_{\boldsymbol{\alpha}}\mathcal{T}_{\boldsymbol{\alpha}}^*\right\| \leq \max_{\|\boldsymbol{X}\|_{\mathrm{F}}=1} \frac{1}{L}\sum_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \langle \boldsymbol{X}, \mathcal{P}_{\mathbb{T}}\boldsymbol{v}_{\boldsymbol{\alpha}}\rangle^2 + \frac{1}{L^2}$$

$$
\begin{aligned}
&\leq \max_{\|\boldsymbol{X}\|_{\mathrm{F}}=1} \frac{\nu r}{2nL} \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \boldsymbol{X}, \mathcal{P}_{\mathbb{T}} \boldsymbol{v}_{\boldsymbol{\alpha}} \rangle^2 + \frac{1}{L^2} \\
&= \max_{\boldsymbol{X}\in\mathbb{T}, \|\boldsymbol{X}\|_{\mathrm{F}}=1} \frac{\nu r}{2nL} \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \boldsymbol{X}, \boldsymbol{v}_{\boldsymbol{\alpha}} \rangle^2 + \frac{1}{L^2} \\
&\leq \max_{\|\boldsymbol{X}\|_{\mathrm{F}}=1} \frac{\nu r}{2nL} \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \boldsymbol{X}, \boldsymbol{v}_{\boldsymbol{\alpha}} \rangle^2 + \frac{1}{L^2} \\
&= \frac{\nu r}{2nL} \lambda_{\max}(\boldsymbol{H}^{-1}) + \frac{1}{L^2} \\
&= \frac{\nu r}{4nL},
\end{aligned}
$$

where the first inequality follows from the triangle inequality, the second comes from the coherence conditions in Assumption 2.2.1, the third line comes from the self-adjointness of $\mathcal{P}_{\mathbb{T}}$, the fourth comes from the definition of the max, the fifth comes from an application of Lemma 1.8.2, and the sixth comes from the $\lambda_{\max}(\boldsymbol{H}^{-1})$ bound from Lemma 2.A.3. Next for $\sum_{\boldsymbol{\alpha}\in\mathbb{I}} \mathcal{T}_{\boldsymbol{\alpha}}^* \mathcal{T}_{\boldsymbol{\alpha}}$, we have that

$$
\begin{aligned}
\left\| \frac{1}{L} \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \mathcal{T}_{\boldsymbol{\alpha}}^* \mathcal{T}_{\boldsymbol{\alpha}} \right\| &\leq \max_{\|\boldsymbol{X}\|_{\mathrm{F}}=1} \frac{1}{L} \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \mathcal{P}_{\mathbb{T}} \boldsymbol{v}_{\boldsymbol{\alpha}}, \mathcal{P}_{\mathbb{T}} \boldsymbol{v}_{\boldsymbol{\alpha}} \rangle \langle \boldsymbol{X}, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle^2 + \frac{1}{L^2} \\
&\leq \max_{\|\boldsymbol{X}\|_{\mathrm{F}}=1} \frac{\nu r}{2nL} \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \boldsymbol{X}, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle^2 + \frac{1}{L^2} \\
&= \max_{\boldsymbol{Z}=\boldsymbol{X}\mathcal{P}_U, \|\boldsymbol{X}\|_{\mathrm{F}}=1} \frac{\nu r}{2nL} \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \boldsymbol{Z}, \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle^2 + \frac{1}{L^2} \\
&\leq \max_{\|\boldsymbol{X}\|_{\mathrm{F}}=1} \frac{\nu r}{2nL} \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \boldsymbol{X}, \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle^2 + \frac{1}{L^2} \\
&= \frac{\nu r}{2nL} \lambda_{\max}(\tilde{\boldsymbol{H}}) + \frac{1}{L^2} \\
&\leq \frac{\nu^2 r^2}{2nL},
\end{aligned}
$$

where the first inequality follows from the triangle inequality, the second comes from the coherence conditions in Assumption 2.2.1, the third line comes from the fact that for any symmetric $\boldsymbol{Y} \in \mathbb{R}^{n\times n}$, $\langle \mathcal{P}_{\mathbb{T}} \boldsymbol{Y}, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle = \langle \boldsymbol{Y}\mathcal{P}_U, \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle$ from Lemma 2.A.2, the fourth comes from the definition of the max, the fifth comes from an application of Lemma 1.8.2, and the sixth comes from the bound on $\lambda_{\max}(\tilde{\boldsymbol{H}})$ in Lemma 2.A.1.

As the latter term is larger, we get a variance estimate $V_0 = \frac{\nu^2 r^2}{nL}$. Now, for $t < \frac{mV_0}{c} = \frac{m}{L}$,

we can use (1.25). It follows that for $m \geq \frac{8}{3}\beta\nu^2 r^2 n \log(n)$,

$$\mathbb{P}\left(\left\|\mathcal{P}_{\mathbb{T}}\mathcal{R}_{\Omega}\mathcal{P}_{\mathbb{T}} - \frac{m}{L}\mathcal{P}_{\mathbb{T}}\right\| \geq \frac{m}{L}\sqrt{\frac{8\beta\nu^2 r^2 n\log(n)}{3m}}\right) \leq 2n\exp\left(-\beta\log(n)\right) = 2n^{1-\beta},$$

yielding the desired result.

Additionally, as $\mathbb{E}(\mathcal{R}_{\Omega}^*) = \frac{m}{L}\mathcal{I}$, the same proof can be repeated to show RIP for $\mathcal{R}_{\Omega}^*$ with the same constants provided. □

The proof of RIP for $\mathcal{R}_{\Omega}$ allows us to define a neighborhood around the ground truth solution where a slightly weakened version of RIP holds. First, however, we will introduce some technical lemmas:

Lemma 2.B.1 (Spectral norm bounds for $\mathcal{P}_{\mathbb{T}}\mathcal{R}_{\Omega}$ and $\mathcal{R}_{\Omega}\mathcal{P}_{\mathbb{T}}$)**.** *Let $\boldsymbol{X}$ be a ν-incoherent rank-r ground truth matrix. For $m \geq \frac{16}{3}\nu r n\log(n)$, both results hold, each with probability $1-2n^{1-\beta}$:*

$$\|\mathcal{R}_{\Omega}\mathcal{P}_{\mathbb{T}}\| \leq \frac{m}{L} + \frac{m\sqrt{n}}{L}\sqrt{\frac{8\beta\nu r n\log(n)}{3m}} \quad \text{and} \quad \|\mathcal{P}_{\mathbb{T}}\mathcal{R}_{\Omega}\| \leq \frac{m}{L} + \frac{4m\sqrt{n}}{L}\sqrt{\frac{\beta\nu r n\log(n)}{3m}}.$$

Proof of 2.B.1. We will start by proving the bound for $\|\mathcal{R}_{\Omega}\mathcal{P}_{\mathbb{T}}\|$.

First, notice that

$$\|\mathcal{R}_{\Omega}\mathcal{P}_{\mathbb{T}}\| \leq \left\|\mathcal{R}_{\Omega}\mathcal{P}_{\mathbb{T}} - \frac{m}{L}\mathcal{P}_{\mathbb{T}}\right\| + \frac{m}{L},$$

following from the triangle inequality, and notice that the middle term can be decomposed as the sum of i.i.d. operators as follows. Let $\mathcal{T}_{\alpha} = \langle\cdot, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\boldsymbol{v}_{\boldsymbol{\alpha}} - \frac{1}{L}\mathcal{P}_{\mathbb{T}}$. Much in the same vein as in Theorem 2.2.1 and Lemmas 2.B.4 and 2.2.3, we will use Theorem 1.7.2 to prove a concentration result. For this, we must get a spectral norm and variance estimate.

For the spectral norm estimate, notice that

$$\begin{aligned}
\|\mathcal{T}_{\alpha}\| &= \left\|\langle\cdot, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\boldsymbol{v}_{\boldsymbol{\alpha}} - \frac{1}{L}\mathcal{P}_{\mathbb{T}}\right\| \\
&\leq \|\mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}\|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} + \frac{1}{L} \\
&\leq \sqrt{\frac{\nu r}{2n}} + \frac{1}{L} \\
&\leq \frac{2\nu r}{\sqrt{n}} =: c.
\end{aligned}$$

To get the variance bounds, notice that

$$\begin{aligned}\|\boldsymbol{E}[\mathcal{T}_\alpha \mathcal{T}_\alpha^*]\| &= \left\| \frac{1}{L} \sum_{\boldsymbol{\alpha} \in \mathbb{I}} \langle \cdot, \boldsymbol{v}_{\boldsymbol{\alpha}} \rangle \langle \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}}, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \boldsymbol{v}_{\boldsymbol{\alpha}} - \frac{1}{L^2} \mathcal{P}_{\mathbb{T}} \right\| \\ &\leq \frac{\nu r}{2nL} \lambda_{\max}(\boldsymbol{H}^{-1}) + \frac{1}{L^2} \\ &\leq \frac{\nu r}{nL},\end{aligned}$$

where the first inequality follows from Assumption 2.2.1, the triangle inequality, and Lemma 1.8.2. For the other term, we see that

$$\begin{aligned}\|\boldsymbol{E}[\mathcal{T}_\alpha^* \mathcal{T}_\alpha]\| &= \left\| \frac{1}{L} \sum_{\boldsymbol{\alpha} \in \mathbb{I}} \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\alpha}} \rangle \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} - \frac{1}{L^2} \mathcal{P}_{\mathbb{T}} \right\| \\ &\leq \left\| \frac{1}{L} \sum_{\boldsymbol{\alpha} \in \mathbb{I}} \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\alpha}} \rangle \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \right\| + \frac{1}{L^2} \\ &\leq \frac{1}{L^2} + \frac{1}{2L} \max_{\|\boldsymbol{X}\|_{\mathrm{F}}=1} \sum_{\boldsymbol{\alpha} \in \mathbb{I}} \langle \boldsymbol{X}, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle^2 \\ &\leq \frac{1}{L^2} + \frac{1}{2L} \max_{\|\boldsymbol{X}\|_{\mathrm{F}}=1} \sum_{\boldsymbol{\alpha} \in \mathbb{I}} \langle \boldsymbol{X}, \mathcal{P}_{U} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle^2 \\ &\leq \frac{1}{L^2} + \frac{1}{2L} \lambda_{\max}(\tilde{\boldsymbol{H}}) \\ &\leq \frac{1}{L^2} + \frac{\nu r}{2L} \\ &\leq \frac{\nu r}{L},\end{aligned}$$

where the third inequality follows from $\|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} < 1$, the fourth inequality follows from Lemma 2.A.2, and fifth inequality from Lemma 2.A.1. As this latter term is the larger of the two variance terms, we set $V = \frac{\nu r m}{L}$ as our variance estimate. This allows us to state the following result using Theorem 1.7.2 for $m \geq \frac{8}{3} \nu r \beta n \log(n)$:

$$\mathbb{P}\left(\left\| \mathcal{R}_\Omega \mathcal{P}_{\mathbb{T}} - \frac{m}{L} \mathcal{P}_{\mathbb{T}} \right\| \geq \frac{m\sqrt{n}}{L} \sqrt{\frac{8\beta\nu r n \log(n)}{3m}} \right) \leq 2n \exp\left(-\beta \log(n)\right) = 2n^{1-\beta},$$

giving a bound on $\mathcal{R}_\Omega \mathcal{P}_{\mathbb{T}}$ of

$$\|\mathcal{R}_\Omega \mathcal{P}_{\mathbb{T}}\| \leq \frac{m}{L} + \frac{m\sqrt{n}}{L} \sqrt{\frac{8\beta\nu r n \log(n)}{3m}},$$

with probability at least $1 - 2n^{1-\beta}$.

The next step is to produce a similar bound for $\|\mathcal{P}_{\mathbb{T}}\mathcal{R}_{\Omega}\|$. The analysis is much the same, with $c = \frac{2\nu r}{\sqrt{n}}$ and $V = \frac{2\nu r}{L}$. Following the same steps, we see that for $m \geq \frac{16}{3}\nu r \beta n \log(n)$,

$$\|\mathcal{P}_{\mathbb{T}}\mathcal{R}_{\Omega}\| \leq \frac{m}{L} + \frac{4m\sqrt{n}}{L}\sqrt{\frac{\beta\nu r n \log(n)}{3m}},$$

with probability at least $1 - 2n^{1-\beta}$. □

Lemma 2.B.2 (Spectral Bound on $\mathcal{P}_{\mathbb{T}_l}\mathcal{R}_{\Omega}$)**.** *Assume that*

$$\|\mathcal{R}_{\Omega}\| \leq \frac{m}{L} + 4\sqrt{\frac{8m\log(n)}{3n}} \quad \text{and} \quad \|\mathcal{P}_{\mathbb{T}}\mathcal{R}_{\Omega}\| \leq \frac{m}{L} + \frac{4m\sqrt{n}}{L}\sqrt{\frac{\beta\nu r n \log(n)}{3m}},$$

Then

$$\|\mathcal{P}_{\mathbb{T}_l}\mathcal{R}_{\Omega}\| \leq \left(\frac{m}{L} + 4\sqrt{\frac{8m\log(n)}{n}}\right)\frac{2\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\sigma_{\min}(\boldsymbol{X})} + \frac{m}{L} + \frac{4m\sqrt{n}}{L}\sqrt{\frac{\beta\nu r n \log(n)}{3m}}.$$

Proof of 2.B.2. This result follows from direct computation, as

$$\begin{aligned}
\|\mathcal{P}_{\mathbb{T}_l}\mathcal{R}_{\Omega}\| &= \|\mathcal{P}_{\mathbb{T}_l}\mathcal{R}_{\Omega} - \mathcal{P}_{\mathbb{T}}\mathcal{R}_{\Omega} + \mathcal{P}_{\mathbb{T}}\mathcal{R}_{\Omega}\| \\
&= \|(\mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_{\mathbb{T}})\mathcal{R}_{\Omega} + \mathcal{P}_{\mathbb{T}}\mathcal{R}_{\Omega}\| \\
&\leq \|\mathcal{R}_{\Omega}\|\|\mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_{\mathbb{T}}\| + \|\mathcal{P}_{\mathbb{T}}\mathcal{R}_{\Omega}\| \\
&\leq \left(\frac{m}{L} + 4\sqrt{\frac{8m\log(n)}{n}}\right)\frac{2\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\sigma_{\min}(\boldsymbol{X})} + \frac{m}{L} + \frac{4m\sqrt{n}}{L}\sqrt{\frac{\beta\nu r n \log(n)}{3m}},
\end{aligned}$$

where the second inequality follows from Lemma 1.8.3 and the assumptions of this lemma. This concludes the proof. □

Lemma 2.B.3 (RIP in a Local Neighborhood)**.** *Assume*

$$\left\|\frac{L}{m}\mathcal{P}_{\mathbb{T}}\mathcal{R}_{\Omega}\mathcal{P}_{\mathbb{T}} - \mathcal{P}_{\mathbb{T}}\right\| \leq \varepsilon_0 < 1, \tag{2.10}$$

$$\frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\sigma_{min}(\boldsymbol{X})} \leq \frac{\sqrt{m}\varepsilon_0}{16n^{5/4}\sqrt{\beta\nu r \log n}}, \quad \|\mathcal{R}_{\Omega}\| \leq \frac{m}{L} + 4\sqrt{\frac{8m\log(n)}{n}}, \tag{2.11}$$

$$\|\mathcal{R}_{\Omega}\mathcal{P}_{\mathbb{T}}\| \leq \frac{m}{L} + \frac{m\sqrt{n}}{L}\sqrt{\frac{8\beta\nu r n \log(n)}{3m}}, \quad \text{and} \quad \|\mathcal{P}_{\mathbb{T}}\mathcal{R}_{\Omega}\| \leq \frac{m}{L} + \frac{4m\sqrt{n}}{L}\sqrt{\frac{\beta\nu r n \log(n)}{3m}}. \tag{2.12}$$

Then

$$\left\| \mathcal{P}_{\mathbb{T}_l} - \frac{L}{m} \mathcal{P}_{\mathbb{T}_l} \mathcal{R}_\Omega \mathcal{P}_{\mathbb{T}_l} \right\| \leq 4\varepsilon_0.$$

Proof. First, notice that

$$\begin{aligned}
&\left\| \mathcal{P}_{\mathbb{T}_l} - \frac{L}{m} \mathcal{P}_{\mathbb{T}_l} \mathcal{R}_\Omega \mathcal{P}_{\mathbb{T}_l} \right\| \\
&\qquad \leq \|\mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_{\mathbb{T}}\| + \frac{L}{m} \|\mathcal{P}_{\mathbb{T}_l} \mathcal{R}_\Omega \mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_{\mathbb{T}_l} \mathcal{R}_\Omega \mathcal{P}_{\mathbb{T}}\| \\
&\qquad\quad + \frac{L}{m} \|\mathcal{P}_{\mathbb{T}_l} \mathcal{R}_\Omega \mathcal{P}_{\mathbb{T}} - \mathcal{P}_{\mathbb{T}} \mathcal{R}_\Omega \mathcal{P}_{\mathbb{T}}\| + \left\| \mathcal{P}_{\mathbb{T}} - \frac{L}{m} \mathcal{P}_{\mathbb{T}} \mathcal{R}_\Omega \mathcal{P}_{\mathbb{T}} \right\| \\
&\qquad \leq \|\mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_{\mathbb{T}}\| + \frac{L}{m} \|\mathcal{P}_{\mathbb{T}_l} \mathcal{R}_\Omega\| \|\mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_{\mathbb{T}}\| \\
&\qquad\quad + \frac{L}{m} \|\mathcal{R}_\Omega \mathcal{P}_{\mathbb{T}}\| \|\mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_{\mathbb{T}}\| + \left\| \mathcal{P}_{\mathbb{T}} - \frac{L}{m} \mathcal{P}_{\mathbb{T}} \mathcal{R}_\Omega \mathcal{P}_{\mathbb{T}} \right\| \\
&\qquad \leq \frac{2\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\sigma_{\min}(\boldsymbol{X})} \left(1 + \frac{L}{m} \|\mathcal{P}_{\mathbb{T}_l} \mathcal{R}_\Omega\| + \frac{L}{m} \|\mathcal{R}_\Omega \mathcal{P}_{\mathbb{T}}\| \right) + \left\| \mathcal{P}_{\mathbb{T}} - \frac{L}{m} \mathcal{P}_{\mathbb{T}} \mathcal{R}_\Omega \mathcal{P}_{\mathbb{T}} \right\|,
\end{aligned}$$

using the triangle inequality and the results gathered in Lemma 1.8.3.

We can now bound each of these terms using the assumptions and prior lemmas. First, notice that

$$\begin{aligned}
\frac{2L\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{m\sigma_{\min}(\boldsymbol{X})} \|\mathcal{R}_\Omega \mathcal{P}_{\mathbb{T}}\| &\leq \frac{2\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\sigma_{\min}(\boldsymbol{X})} + \sqrt{\frac{32\beta\nu r n^2 \log(n)}{3m}} \frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\sigma_{\min}(\boldsymbol{X})} \\
&\leq \frac{\sqrt{m}\varepsilon_0}{8n^{5/4}\sqrt{\beta\nu r \log n}} + \sqrt{\frac{\beta\nu r n^2 \log(n)}{24m}} \frac{\sqrt{m}\varepsilon_0}{n^{5/4}\sqrt{\beta\nu r \log n}} \\
&\leq \frac{\sqrt{m}}{n^{5/4}} \frac{\varepsilon_0}{8} + \frac{\varepsilon_0}{n^{1/4}\sqrt{24}} \\
&\leq \frac{\sqrt{L}}{n^{5/4}} \frac{\varepsilon_0}{8} + \frac{\varepsilon_0}{\sqrt{24}} \\
&\leq \frac{\varepsilon_0}{8} + \frac{\varepsilon_0}{\sqrt{24}} \\
&\leq \varepsilon_0,
\end{aligned}$$

where the first inequality comes from the assumption on $\|\mathcal{R}_\Omega \mathcal{P}_{\mathbb{T}}\|$ in (2.12), the second inequality comes from the local neighborhood assumption in (2.11), the third inequality comes from term cancellation and the fact that $\beta, \nu, r, \log(n) \geq 1$, the fourth inequality comes from the fact that $m \leq L$, the fifth inequality comes from the fact that $\frac{L}{n^2} < 1$, and the last inequality is a

numerical inequality on the fractions.

Next, notice that the conditions of Lemma 2.B.2 are satisfied, so

$$\begin{aligned}
&\frac{2L\|\boldsymbol{X}_l-\boldsymbol{X}\|_{\mathrm{F}}}{m\sigma_{\min}(\boldsymbol{X})}\|\mathcal{P}_{\mathbb{T}_l}\mathcal{R}_\Omega\| \\
&\qquad\leq \left(\frac{2L\|\boldsymbol{X}_l-\boldsymbol{X}\|_{\mathrm{F}}}{m\sigma_{\min}(\boldsymbol{X})}\right)\left(\left(\frac{m}{L}+4\sqrt{\frac{8m\log(n)}{n}}\right)\frac{2\|\boldsymbol{X}_l-\boldsymbol{X}\|_{\mathrm{F}}}{\sigma_{\min}(\boldsymbol{X})}\right. \\
&\qquad\qquad\left.+\frac{m}{L}+\frac{4m\sqrt{n}}{L}\sqrt{\frac{\beta\nu rn\log(n)}{3m}}\right) \\
&\qquad\leq \left(\frac{2\|\boldsymbol{X}_l-\boldsymbol{X}\|_{\mathrm{F}}}{\sigma_{\min}(\boldsymbol{X})}\right)^2\left(1+\frac{2n^2}{m}\sqrt{\frac{8m\log(n)}{n}}\right) \\
&\qquad\quad+\frac{2\|\boldsymbol{X}_l-\boldsymbol{X}\|_{\mathrm{F}}}{\sigma_{\min}(\boldsymbol{X})}\left(1+4\sqrt{\frac{\beta\nu rn^2\log(n)}{3m}}\right) \\
&\qquad\leq \frac{m\varepsilon_0^2}{64n^{5/2}\beta\nu r\log n}+\frac{\sqrt{m}\varepsilon_0}{8n^{5/4}\sqrt{\beta\nu r\log n}} \\
&\qquad\quad+\left(\frac{m\varepsilon_0^2}{64n^{5/2}\beta\nu r\log n}\right)\left(\frac{2n^2}{m}\sqrt{\frac{8m\log(n)}{n}}\right) \\
&\qquad\quad+\left(\frac{\sqrt{m}\varepsilon_0}{8n^{5/4}\sqrt{\beta\nu r\log n}}\right)\left(4\sqrt{\frac{\beta\nu rn^2\log(n)}{3m}}\right) \\
&\qquad\leq \frac{\varepsilon_0^2}{64}+\frac{\varepsilon_0}{8}+\left(\frac{m\varepsilon_0^2}{64n^{5/2}\beta\nu r\log n}\right)\left(\frac{2n^2}{m}\sqrt{\frac{8m\log(n)}{n}}\right) \\
&\qquad\qquad+\left(\frac{\sqrt{m}\varepsilon_0}{8n^{5/4}\sqrt{\beta\nu r\log n}}\right)\left(4\sqrt{\frac{\beta\nu rn^2\log(n)}{3m}}\right) \\
&\qquad= \frac{\varepsilon_0^2}{64}+\frac{\varepsilon_0}{8}+\frac{\sqrt{32m}\varepsilon_0^2}{64\beta\nu rn\log(n)}+\frac{\varepsilon_0}{n^{1/4}\sqrt{12}} \\
&\qquad\leq \frac{\varepsilon_0^2}{64}+\frac{\varepsilon_0}{8}+\frac{\varepsilon_0^2}{16}+\frac{\varepsilon_0}{\sqrt{12}} \\
&\qquad\leq \varepsilon_0,
\end{aligned}$$

where the first inequality follows from the assumptions on $\|\mathcal{P}_{\mathbb{T}}\mathcal{R}_\Omega\|$ in (2.12), the second inequality follows from rearrangement of terms and the fact that $\frac{L}{m}\leq\frac{n^2}{m}$, the third inequality comes from the local neighborhood assumption in (2.11), the fourth inequality comes from the fact that $\frac{m}{n^2}\leq\frac{L}{n^2}<1$ along with $\beta,\nu,r,\log(n)\geq 1$, the fifth line comes from multiplying out terms, the sixth inequality again comes from a bound on $\frac{m}{n^2}$ amongst other simplifications, and the last line comes from a numerical inequality about the fractions coupled with the fact that

$\varepsilon_0 < 1$ from (2.10), so $\varepsilon_0^2 \leq \varepsilon_0$. The desired statement follows from here, thus concluding the proof. □

For Algorithm 5, we will need what the authors of [WCCL20] call an asymmetric form of RIP for $\mathcal{R}_\Omega$. The statement and proof of this are below:

Lemma 2.B.4 (Asymmetric RIP of $\mathcal{R}_\Omega$)**.** *Let $\boldsymbol{X}_l = \boldsymbol{U}_l \boldsymbol{D}_l \boldsymbol{U}_l^\top$ and $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top$ be two fixed rank-r matrices. Assume*

$$\|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \leq \sqrt{\frac{\nu r}{2n}}, \qquad \|\mathcal{P}_U \boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \leq \sqrt{\frac{\nu r}{2n}}, \qquad \|\mathcal{P}_{U_l} \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \leq \sqrt{\frac{\nu r}{2n}}, \quad \text{and} \quad \|\mathcal{P}_{U_l} \boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \leq \sqrt{\frac{\nu r}{2n}},$$

for $\boldsymbol{\alpha} \in \mathbb{I}$. Let $|\Omega| = m$. For $m \geq \frac{4}{3}\beta\nu^2 r^2 n \log(n)$, with probability at least $1 - 2n^{1-\beta}$ for $\beta > 1$, the following estimate holds:

$$\left\| \frac{L}{m} \mathcal{P}_{\mathbb{T}_l} \mathcal{R}_\Omega (\mathcal{P}_U - \mathcal{P}_{U_l}) - \mathcal{P}_{\mathbb{T}_l} (\mathcal{P}_U - \mathcal{P}_{U_l}) \right\| \leq \sqrt{\frac{4\beta\nu^2 r^2 n \log(n)}{3m}}. \tag{2.13}$$

Proof of Lemma 2.B.4. First, note that for all $\boldsymbol{Z} \in \mathbb{R}^{n \times n}$,

$$\begin{aligned} (\mathcal{P}_U - \mathcal{P}_{U_l})\boldsymbol{Z} &= \sum_{\boldsymbol{\alpha} \in \mathbb{I}} \langle (\mathcal{P}_U - \mathcal{P}_{U_l})\boldsymbol{Z}, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \boldsymbol{v}_{\boldsymbol{\alpha}} \\ &= \sum_{\boldsymbol{\alpha} \in \mathbb{I}} \langle \boldsymbol{Z}, (\mathcal{P}_U - \mathcal{P}_{U_l})\boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \boldsymbol{v}_{\boldsymbol{\alpha}}, \end{aligned}$$

so it follows that

$$\mathcal{R}_\Omega (\mathcal{P}_U - \mathcal{P}_{U_l})\boldsymbol{Z} = \sum_{\boldsymbol{\alpha} \in \Omega} \langle \boldsymbol{Z}, (\mathcal{P}_U - \mathcal{P}_{U_l})\boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \boldsymbol{v}_{\boldsymbol{\alpha}},$$

and subsequently

$$\mathcal{P}_{\mathbb{T}_l} \mathcal{R}_\Omega (\mathcal{P}_U - \mathcal{P}_{U_l})\boldsymbol{Z} = \sum_{\boldsymbol{\alpha} \in \Omega} \langle \boldsymbol{Z}, (\mathcal{P}_U - \mathcal{P}_{U_l})\boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \mathcal{P}_{\mathbb{T}_l} \boldsymbol{v}_{\boldsymbol{\alpha}}.$$

We define $\mathcal{K}_{\boldsymbol{\alpha}}(\cdot) = \mathcal{P}_{\mathbb{T}_l}(\boldsymbol{v}_{\boldsymbol{\alpha}}) \otimes (\mathcal{P}_U - \mathcal{P}_{U_l})(\boldsymbol{w}_{\boldsymbol{\alpha}})(\cdot) = \langle \cdot, (\mathcal{P}_U - \mathcal{P}_{U_l})\boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \mathcal{P}_{\mathbb{T}_l} \boldsymbol{v}_{\boldsymbol{\alpha}}$. It follows that $\mathcal{P}_{\mathbb{T}_l} \mathcal{R}_\Omega (\mathcal{P}_U - \mathcal{P}_{U_l}) = \sum_{\boldsymbol{\alpha} \in \Omega} \mathcal{K}_{\boldsymbol{\alpha}}$, and that $\mathcal{K}_{\boldsymbol{\alpha}}^* = (\mathcal{P}_U - \mathcal{P}_{U_l})(\boldsymbol{w}_{\boldsymbol{\alpha}}) \otimes \mathcal{P}_{\mathbb{T}_l}(\boldsymbol{v}_{\boldsymbol{\alpha}})(\cdot) = \langle \cdot, \mathcal{P}_{\mathbb{T}_l} \boldsymbol{v}_{\boldsymbol{\alpha}} \rangle (\mathcal{P}_U - \mathcal{P}_{U_l})\boldsymbol{w}_{\boldsymbol{\alpha}}$. Additionally, note that $\mathbb{E}[\mathcal{K}_{\boldsymbol{\alpha}}] = \frac{1}{L}\mathcal{P}_{\mathbb{T}_l}(\mathcal{P}_U - \mathcal{P}_{U_l})$ and that $\mathbb{E}[\mathcal{K}_{\boldsymbol{\alpha}}^*] = \frac{1}{L}(\mathcal{P}_U - \mathcal{P}_{U_l})\mathcal{P}_{\mathbb{T}_l}$. We note that the variance term $\mathbb{E}\left[\left(\mathcal{K}_{\boldsymbol{\alpha}} - \frac{1}{L}\mathcal{P}_{\mathbb{T}_l}(\mathcal{P}_U - \mathcal{P}_{U_l})\right)\left(\mathcal{K}_{\boldsymbol{\alpha}} - \frac{1}{L}\mathcal{P}_{\mathbb{T}_l}(\mathcal{P}_U - \mathcal{P}_{U_l})\right)^*\right]$ can be

bounded as follows:

$$\begin{aligned}
&\mathbb{E}\left[\left(\mathcal{K}_{\boldsymbol{\alpha}} - \frac{1}{L}\mathcal{P}_{\mathbb{T}_l}(\mathcal{P}_U - \mathcal{P}_{U_l})\right)\left(\mathcal{K}_{\boldsymbol{\alpha}} - \frac{1}{L}\mathcal{P}_{\mathbb{T}_l}(\mathcal{P}_U - \mathcal{P}_{U_l})\right)^*\right] \\
&= \mathbb{E}\left[\mathcal{K}_{\boldsymbol{\alpha}}\mathcal{K}_{\boldsymbol{\alpha}}^* - \mathcal{K}_{\boldsymbol{\alpha}}\frac{1}{L}(\mathcal{P}_U - \mathcal{P}_{U_l})\mathcal{P}_{\mathbb{T}_l} - \frac{1}{L}\mathcal{P}_{\mathbb{T}_l}(\mathcal{P}_U - \mathcal{P}_{U_l})\mathcal{K}_{\boldsymbol{\alpha}}^* + \frac{1}{L^2}\mathcal{P}_{\mathbb{T}_l}(\mathcal{P}_U - \mathcal{P}_{U_l})\mathcal{P}_{\mathbb{T}_l}\right] \\
&= \mathbb{E}\left[\mathcal{K}_{\boldsymbol{\alpha}}\mathcal{K}_{\boldsymbol{\alpha}}^*\right] - \mathbb{E}\left[\mathcal{K}_{\boldsymbol{\alpha}}\right]\frac{1}{L}(\mathcal{P}_U - \mathcal{P}_{U_l})\mathcal{P}_{\mathbb{T}_l} - \frac{1}{L}\mathcal{P}_{\mathbb{T}_l}(\mathcal{P}_U - \mathcal{P}_{U_l})\mathbb{E}\left[\mathcal{K}_{\boldsymbol{\alpha}}^*\right] + \frac{1}{L^2}\mathcal{P}_{\mathbb{T}_l}(\mathcal{P}_U - \mathcal{P}_{U_l})\mathcal{P}_{\mathbb{T}_l} \\
&= \mathbb{E}\left[\mathcal{K}_{\boldsymbol{\alpha}}\mathcal{K}_{\boldsymbol{\alpha}}^*\right] - \frac{1}{L^2}\mathcal{P}_{\mathbb{T}_l}\left(\mathcal{P}_U - \mathcal{P}_{U_l}\right)^2\mathcal{P}_{\mathbb{T}_l} - \frac{1}{L^2}\mathcal{P}_{\mathbb{T}_l}\left(\mathcal{P}_U - \mathcal{P}_{U_l}\right)^2\mathcal{P}_{\mathbb{T}_l} + \frac{1}{L^2}\mathcal{P}_{\mathbb{T}_l}\left(\mathcal{P}_U - \mathcal{P}_{U_l}\right)^2\mathcal{P}_{\mathbb{T}_l} \\
&= \mathbb{E}\left[\mathcal{K}_{\boldsymbol{\alpha}}\mathcal{K}_{\boldsymbol{\alpha}}^*\right] - \frac{1}{L^2}\mathcal{P}_{\mathbb{T}_l}\left(\mathcal{P}_U - \mathcal{P}_{U_l}\right)^2\mathcal{P}_{\mathbb{T}_l}.
\end{aligned}$$

A similar computation holds for $\mathbb{E}\left[\left(\mathcal{K}_{\boldsymbol{\alpha}} - \frac{1}{L}\mathcal{P}_{\mathbb{T}_l}(\mathcal{P}_U - \mathcal{P}_{U_l})\right)^*\left(\mathcal{K}_{\boldsymbol{\alpha}} - \frac{1}{L}\mathcal{P}_{\mathbb{T}_l}(\mathcal{P}_U - \mathcal{P}_{U_l})\right)\right]$, indicating that it suffices to compute an upper bound on the following terms in order to leverage Theorem 1.7.2:

$$\left\|\mathcal{K}_{\boldsymbol{\alpha}} - \frac{1}{L}\mathcal{P}_{\mathbb{T}_l}(\mathcal{P}_U - \mathcal{P}_{U_l})\right\| \leq c,$$

$$\max\left\{\left\|\mathbb{E}[\mathcal{K}_\alpha\mathcal{K}_\alpha^*] - \frac{1}{L^2}\mathcal{P}_{\mathbb{T}_l}(\mathcal{P}_U - \mathcal{P}_{U_l})^2\mathcal{P}_{\mathbb{T}_l}\right\|, \left\|\mathbb{E}[\mathcal{K}_\alpha^*\mathcal{K}_\alpha] - \frac{1}{L^2}(\mathcal{P}_U - \mathcal{P}_{U_l})\mathcal{P}_{\mathbb{T}_l}(\mathcal{P}_U - \mathcal{P}_{U_l})\right\|\right\} \leq V_0.$$

For the first term, notice that

$$\begin{aligned}
\|\mathcal{K}_\alpha\| &\leq \|(\mathcal{P}_U - \mathcal{P}_{U_l})\boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}\|\mathcal{P}_{\mathbb{T}_l}\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \\
&\leq (\|\mathcal{P}_U\boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} + \|\mathcal{P}_{U_l}\boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}})\|\mathcal{P}_{\mathbb{T}_l}\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \\
&\leq \frac{\nu r}{n} \leq \frac{\nu^2 r^2}{n}.
\end{aligned}$$

So by the triangle inequality $\left\|\mathcal{K}_\alpha - \frac{1}{L}\mathcal{P}_{\mathbb{T}_l}(\mathcal{P}_U - \mathcal{P}_{U_l})\right\| \leq \frac{2\nu^2 r^2}{n} =: c$. For the second term, notice that

$$\begin{aligned}
\mathbb{E}[\mathcal{K}_\alpha\mathcal{K}_\alpha^*] &= \frac{1}{L}\sum_{\boldsymbol{\alpha}\in\mathbb{I}}\langle\cdot, (\mathcal{P}_U - \mathcal{P}_{U_l})\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\langle\mathcal{P}_{\mathbb{T}_l}\boldsymbol{v}_{\boldsymbol{\alpha}}, \mathcal{P}_{\mathbb{T}_l}\boldsymbol{v}_{\boldsymbol{\alpha}}\rangle(\mathcal{P}_U - \mathcal{P}_{U_l})\boldsymbol{w}_{\boldsymbol{\alpha}}, \\
\mathbb{E}[\mathcal{K}_\alpha^*\mathcal{K}_\alpha] &= \frac{1}{L}\sum_{\boldsymbol{\alpha}\in\mathbb{I}}\langle\cdot, \mathcal{P}_{\mathbb{T}_l}\boldsymbol{v}_{\boldsymbol{\alpha}}\rangle\langle(\mathcal{P}_U - \mathcal{P}_{U_l})\boldsymbol{w}_{\boldsymbol{\alpha}}, (\mathcal{P}_U - \mathcal{P}_{U_l})\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\mathcal{P}_{\mathbb{T}_l}\boldsymbol{v}_{\boldsymbol{\alpha}}.
\end{aligned}$$

As such,

$$\begin{aligned}\|\mathbb{E}[\mathcal{K}_\alpha \mathcal{K}_\alpha^*]\| &= \frac{1}{L} \max_{\boldsymbol{X}\in\mathbb{R}^{n\times n}, \|\boldsymbol{X}\|_{\mathrm{F}}=1} \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \boldsymbol{X}, (\mathcal{P}_U - \mathcal{P}_{U_l})\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \langle \mathcal{P}_{\mathbb{T}_l}\boldsymbol{v}_{\boldsymbol{\alpha}}, \mathcal{P}_{\mathbb{T}_l}\boldsymbol{v}_{\boldsymbol{\alpha}}\rangle \\ &\leq \frac{\nu r}{2nL} \max_{\boldsymbol{X}\in\mathbb{R}^{n\times n}, \|\boldsymbol{X}\|_{\mathrm{F}}=1} \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \boldsymbol{X}, \underbrace{(\mathcal{P}_U - \mathcal{P}_{U_l})\boldsymbol{w}_{\boldsymbol{\alpha}}}_{=:\tilde{\boldsymbol{w}}_{\boldsymbol{\alpha}}^l} \rangle^2 \\ &= \frac{\nu r}{2nL} \lambda_{\max}(\tilde{\boldsymbol{H}}^l),\end{aligned}$$

where $\tilde{\boldsymbol{H}}^l = [\langle \tilde{\boldsymbol{w}}_{\boldsymbol{\alpha}}^l, \tilde{\boldsymbol{w}}_{\boldsymbol{\beta}}^l\rangle] \in \mathbb{R}^{L\times L}$. To bound this, notice that if $\boldsymbol{\alpha}\cap\boldsymbol{\beta} = \emptyset$,

$$\begin{aligned}\langle \tilde{\boldsymbol{w}}_{\boldsymbol{\alpha}}^l, \tilde{\boldsymbol{w}}_{\boldsymbol{\beta}}^l\rangle &= \langle \boldsymbol{w}_{\boldsymbol{\alpha}}, (\mathcal{P}_U - \mathcal{P}_{U_l})^2 \boldsymbol{w}_{\boldsymbol{\beta}}\rangle \\ &= \mathrm{Trace}\left[\boldsymbol{w}_{\boldsymbol{\alpha}}(\mathcal{P}_U - \mathcal{P}_{U_l})^2 \boldsymbol{w}_{\boldsymbol{\beta}}\right] \\ &= \mathrm{Trace}\left[\boldsymbol{w}_{\boldsymbol{\beta}}\boldsymbol{w}_{\boldsymbol{\alpha}}(\mathcal{P}_U - \mathcal{P}_{U_l})^2\right] \\ &= 0,\end{aligned}$$

therefore preserving the same sparsity structure as $\tilde{\boldsymbol{H}}$. To bound the magnitude of the entries, we can see that

$$\begin{aligned}|\langle \tilde{\boldsymbol{w}}_{\boldsymbol{\alpha}}^l, \tilde{\boldsymbol{w}}_{\boldsymbol{\beta}}^l\rangle| &= |\langle \boldsymbol{w}_{\boldsymbol{\alpha}}, (\mathcal{P}_U - \mathcal{P}_{U_l})^2 \boldsymbol{w}_{\boldsymbol{\beta}}\rangle| = |\langle \boldsymbol{w}_{\boldsymbol{\alpha}}, (\mathcal{P}_U - \mathcal{P}_U\mathcal{P}_{U_l} - \mathcal{P}_{U_l}\mathcal{P}_U + \mathcal{P}_{U_l})\boldsymbol{w}_{\boldsymbol{\beta}}\rangle| \\ &\leq |\langle \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}, \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\beta}}\rangle| + |\langle \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}, \mathcal{P}_{U_l} \boldsymbol{w}_{\boldsymbol{\beta}}\rangle| + |\langle \mathcal{P}_{U_l} \boldsymbol{w}_{\boldsymbol{\alpha}}, \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\beta}}\rangle| + |\langle \mathcal{P}_{U_l} \boldsymbol{w}_{\boldsymbol{\alpha}}, \mathcal{P}_{U_l} \boldsymbol{w}_{\boldsymbol{\beta}}\rangle| \\ &\leq \frac{2\nu r}{n},\end{aligned}$$

giving us an upper bound on $\tilde{\boldsymbol{H}}^l$ of

$$\lambda_{\max}(\tilde{\boldsymbol{H}}^l) \leq \frac{2\nu r}{n}(2n-3) \leq 4\nu r.$$

This gives a bound of

$$\|\mathbb{E}[\mathcal{K}_\alpha \mathcal{K}_\alpha^*]\| \leq \frac{2\nu^2 r^2}{nL}.$$

Next, we can see that

$$\begin{aligned}\|\mathbb{E}[\mathcal{K}_\alpha^*\mathcal{K}_\alpha]\| &= \frac{1}{L}\max_{\boldsymbol{X}\in\mathbb{R}^{n\times n},\|\boldsymbol{X}\|_{\mathrm{F}}=1}\sum_{\boldsymbol{\alpha}\in\mathbb{I}}\langle(\mathcal{P}_U-\mathcal{P}_{U_l})\boldsymbol{w}_{\boldsymbol{\alpha}},(\mathcal{P}_U-\mathcal{P}_{U_l})\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\,\langle\boldsymbol{X},\mathcal{P}_{\mathbb{T}_l}\boldsymbol{v}_{\boldsymbol{\alpha}}\rangle^2\\ &\le \frac{2\nu r}{nL}\max_{\boldsymbol{X}\in\mathbb{T},\|\boldsymbol{X}\|_{\mathrm{F}}=1}\sum_{\boldsymbol{\alpha}\in\mathbb{I}}\langle\boldsymbol{X},\boldsymbol{v}_{\boldsymbol{\alpha}}\rangle^2\\ &\le \frac{2\nu r}{nL}\lambda_{\max}(\boldsymbol{H}^{-1})\\ &= \frac{\nu r}{nL}.\end{aligned}$$

Now, it follows that

$$\begin{aligned}\left\|\mathbb{E}[\mathcal{K}_\alpha\mathcal{K}_\alpha^*]-\frac{1}{L^2}\mathcal{P}_{\mathbb{T}_l}(\mathcal{P}_U-\mathcal{P}_{U_l})^2\mathcal{P}_{\mathbb{T}_l}\right\| &\le \frac{2\nu^2r^2}{nL}+\frac{4}{L^2}\le\frac{4\nu^2r^2}{nL}\\ \left\|\mathbb{E}[\mathcal{K}_\alpha^*\mathcal{K}_\alpha]-\frac{1}{L^2}(\mathcal{P}_U-\mathcal{P}_{U_l})\mathcal{P}_{\mathbb{T}_l}(\mathcal{P}_U-\mathcal{P}_{U_l})\right\| &\le \frac{\nu r}{nL}+\frac{4}{L^2}\le\frac{2\nu r}{nL},\end{aligned}$$

so $V_0 := \frac{4\nu^2r^2}{nL}$. Now, for $t < \frac{mV_0}{c} = \frac{2m}{L}$, the result follows from Theorem 1.7.2 with $t = \sqrt{\frac{4\beta\nu^2r^2n\log(n)}{3m}}$ with $m \ge \frac{4}{3}\nu^2r^2n\log(n)$. □

2.C Proof of Local Convergence (Theorem 2.2.2)

In this section, we will use the properties proven thus far to provide proof of local convergence of Riemannian Pseudogradient Descent. We begin with the following technical lemmas:

Lemma 2.C.1 (Stepsize Bounds)**.** *Assume that* $\|\mathcal{P}_{\mathbb{T}_l}-\frac{L}{m}\mathcal{P}_{\mathbb{T}_l}\mathcal{R}_\Omega\mathcal{P}_{\mathbb{T}_l}\|\le 4\varepsilon_0<1$. *Then the stepsize* α_l *in Riemannian Pseudogradient Descent can be bounded by*

$$\frac{L/m}{1+4\varepsilon_0}\le\alpha_l=\frac{\|\mathcal{P}_{\mathbb{T}}\boldsymbol{G}_l\|_{\mathrm{F}}^2}{\langle\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l,\mathcal{R}_\Omega\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l\rangle}\le\frac{L/m}{1-4\varepsilon_0}.$$

Proof of Lemma 2.C.1. We will prove this by leveraging the local RIP assumption. Notice the following:

$$\begin{aligned}\langle\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l,\mathcal{R}_\Omega\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l\rangle &= \langle\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l,\mathcal{P}_{\mathbb{T}_l}\mathcal{R}_\Omega\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l\rangle\\ &= \left\langle\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l,\mathcal{P}_{\mathbb{T}_l}\mathcal{R}_\Omega\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l-\frac{m}{L}\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l\right\rangle+\frac{m}{L}\langle\mathcal{P}_{\mathbb{T}}\boldsymbol{G}_l,\mathcal{P}_{\mathbb{T}}\boldsymbol{G}_l\rangle.\end{aligned}$$

We can now leverage the variational characterization of the spectral norm and local RIP, proven in Lemma 2.B.3, to bound the following:

$$-\frac{m}{L}(4\varepsilon_0)\|\mathcal{P}_{\mathbb{T}}\boldsymbol{G}_l\|_{\mathrm{F}}^2 \leq \left\langle \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l, \mathcal{P}_{\mathbb{T}_l}\mathcal{R}_\Omega\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l - \frac{m}{L}\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l \right\rangle \leq \frac{m}{L}(4\varepsilon_0)\|\mathcal{P}_{\mathbb{T}}\boldsymbol{G}_l\|_{\mathrm{F}}^2.$$

As such, we can now bound the denominator as

$$\frac{m}{L}(1-4\varepsilon_0)\|\mathcal{P}_{\mathbb{T}}\boldsymbol{G}_l\|_{\mathrm{F}}^2 \leq \langle \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l, \mathcal{R}_\Omega\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l\rangle \leq \frac{m}{L}(1+4\varepsilon_0)\|\mathcal{P}_{\mathbb{T}}\boldsymbol{G}_l\|_{\mathrm{F}}^2.$$

Rearrangement of this last expression yields the upper and lower bounds on the step size derived above. The condition that $4\varepsilon_0 < 1$ is required to enforce the positivity of the step size, as negative step sizes cause divergence in the contractive sequence. This is necessary as $\mathcal{R}_\Omega$ is not a self-adjoint positive semi-definite operator. This concludes the proof. □

Lemma 2.C.2 (I_1 Bound)**.** *Assume* $\left\|\mathcal{P}_{\mathbb{T}_l} - \frac{L}{m}\mathcal{P}_{\mathbb{T}_l}\mathcal{R}_\Omega\mathcal{P}_{\mathbb{T}_l}\right\| \leq 4\varepsilon_0$ *and* α_l *can be bounded as in Lemma 2.C.1. Then the spectral norm of* $\mathcal{P}_{\mathbb{T}_l} - \alpha_l\mathcal{P}_{\mathbb{T}_l}\mathcal{R}_\Omega\mathcal{P}_{\mathbb{T}_l}$ *can be bounded as*

$$\|\mathcal{P}_{\mathbb{T}_l} - \alpha_l\mathcal{P}_{\mathbb{T}_l}\mathcal{R}_\Omega\mathcal{P}_{\mathbb{T}_l}\| \leq \frac{8\varepsilon_0}{1-4\varepsilon_0}. \tag{2.14}$$

Proof of Lemma 2.C.2. From direct calculation, it follows that

$$\begin{aligned}
\|\mathcal{P}_{\mathbb{T}_l} - \alpha_l\mathcal{P}_{\mathbb{T}_l}\mathcal{R}_\Omega\mathcal{P}_{\mathbb{T}_l}\| &\leq \left\|\mathcal{P}_{\mathbb{T}_l} - \frac{L}{m}\mathcal{P}_{\mathbb{T}_l}\mathcal{R}_\Omega\mathcal{P}_{\mathbb{T}_l}\right\| + \left|\alpha_l - \frac{L}{m}\right| \|\mathcal{P}_{\mathbb{T}_l}\mathcal{R}_\Omega\mathcal{P}_{\mathbb{T}_l}\| \\
&\leq 4\varepsilon_0 + \left|\alpha_l - \frac{L}{m}\right| \left(\left\|\mathcal{P}_{\mathbb{T}_l}\mathcal{R}_\Omega\mathcal{P}_{\mathbb{T}_l} - \frac{m}{L}\mathcal{P}_{\mathbb{T}_l}\right\| + \frac{m}{L}\|\mathcal{P}_{\mathbb{T}_l}\|\right) \\
&\leq 4\varepsilon_0 + \left(\frac{L/m}{1-4\varepsilon_0} - \frac{L/m(1-4\varepsilon_0)}{1-4\varepsilon_0}\right)\left(4\varepsilon_0\frac{m}{L} + \frac{m}{L}\right) \\
&\leq 4\varepsilon_0 + \frac{4\varepsilon_0}{1-4\varepsilon_0}(1+4\varepsilon_0) \\
&= \frac{8\varepsilon_0}{1-4\varepsilon_0}.
\end{aligned}$$

This finishes the proof. □

We can now prove Theorem 2.2.2.

2.C.1 Proof of Theorem 2.2.2

Proof. First, it follows that

$$\|\boldsymbol{X}_{l+1} - \boldsymbol{X}\|_{\mathrm{F}} \leq \|\boldsymbol{X}_{l+1} - \boldsymbol{W}_l\|_{\mathrm{F}} + \|\boldsymbol{W}_l - \boldsymbol{X}\|_{\mathrm{F}} \leq 2\|\boldsymbol{W}_l - \boldsymbol{X}\|_{\mathrm{F}},$$

as $\boldsymbol{X}_{l+1}$ is the best rank-r approximation of $\boldsymbol{W}_l$. Plugging in $\boldsymbol{W}_l = \boldsymbol{X}_l + \alpha_l \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l$, we see that

$$\begin{aligned}
\|\boldsymbol{X}_{l+1} - \boldsymbol{X}\|_{\mathrm{F}} &\leq 2\left\|\boldsymbol{X}_l + \alpha_l \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l - \boldsymbol{X}\right\|_{\mathrm{F}} \\
&= 2\|\boldsymbol{X}_l - \boldsymbol{X} - \alpha_l \mathcal{P}_{\mathbb{T}_l}\mathcal{R}_\Omega(\boldsymbol{X}_l - \boldsymbol{X})\|_{\mathrm{F}} \\
&\leq \underbrace{2\|(\mathcal{P}_{\mathbb{T}_l} - \alpha_l \mathcal{P}_{\mathbb{T}_l}\mathcal{R}_\Omega\mathcal{P}_{\mathbb{T}_l})(\boldsymbol{X}_l - \boldsymbol{X})\|_{\mathrm{F}}}_{I_1} \\
&\quad + \underbrace{2\|(I - \mathcal{P}_{\mathbb{T}_l})(\boldsymbol{X}_l - \boldsymbol{X})\|_{\mathrm{F}}}_{I_2} \\
&\quad + \underbrace{2|\alpha_l|\|\mathcal{P}_{\mathbb{T}_l}\mathcal{R}_\Omega(I - \mathcal{P}_{\mathbb{T}_l})(\boldsymbol{X}_l - \boldsymbol{X})\|_{\mathrm{F}}}_{I_3}.
\end{aligned}$$

It remains to bound each term individually. Using Lemma 2.C.2, we see that

$$I_1 \leq \frac{16\varepsilon_0}{1 - 4\varepsilon_0}\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}.$$

Next, notice that from Lemma 1.8.3 and the fact that $\mathcal{P}_{\mathbb{T}_l}\boldsymbol{X}_l = \boldsymbol{X}_l$,

$$\begin{aligned}
I_2 &= 2\|(I - \mathcal{P}_{\mathbb{T}_l})\boldsymbol{X}\|_{\mathrm{F}} \\
&\leq \frac{2\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}^2}{\sigma_{\min}(\boldsymbol{X})} \\
&\leq \frac{\sqrt{m}\varepsilon_0}{8n^{5/4}\sqrt{\beta\nu r \log n}}\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}} \\
&\leq \frac{\varepsilon_0}{1 - 4\varepsilon_0}\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}},
\end{aligned}$$

using Lemma 1.8.3 and our initial assumption. Finally, we see that, following a similar argument as in the bound of I_2 and using Lemma 2.B.2,

$$I_3 \leq 2|\alpha_l|\|\mathcal{P}_{\mathbb{T}_l}\mathcal{R}_\Omega\|\|(I - \mathcal{P}_{\mathbb{T}_l})\boldsymbol{X}\|_{\mathrm{F}}$$

$$
\begin{aligned}
&\leq \frac{2L/m}{1-4\varepsilon_0}\left[\left(\frac{m}{L}+4\sqrt{\frac{8m\log(n)}{n}}\right)\frac{2\|\boldsymbol{X}_l-\boldsymbol{X}\|_{\mathrm{F}}}{\sigma_{\min}(\boldsymbol{X})}\right.\\
&\qquad\left.+\frac{m}{L}+\frac{4m\sqrt{n}}{L}\sqrt{\frac{\beta\nu rn\log(n)}{3m}}\right]\left(\frac{\|\boldsymbol{X}_l-\boldsymbol{X}\|_{\mathrm{F}}}{\sigma_{\min}(\boldsymbol{X})}\right)\|\boldsymbol{X}_l-\boldsymbol{X}\|_{\mathrm{F}}\\
&\leq \frac{1}{1-4\varepsilon_0}\left(\frac{\varepsilon_0^2}{128}+\frac{\varepsilon_0}{16}+\frac{\varepsilon_0^2}{32}+\frac{\varepsilon_0}{\sqrt{48}}\right)\|\boldsymbol{X}_l-\boldsymbol{X}\|_{\mathrm{F}}\\
&\leq \frac{\varepsilon_0}{1-4\varepsilon_0},
\end{aligned}
$$

where the second to last inequality follows from the same analysis conducted in Lemma 2.B.3, sclaed by a factor of $\frac{1}{2}$. Collecting these results, we get

$$
\|\boldsymbol{X}_{l+1}-\boldsymbol{X}\|_{\mathrm{F}} \leq \frac{18\varepsilon_0}{1-4\varepsilon_0}\|\boldsymbol{X}_l-\boldsymbol{X}\|_{\mathrm{F}}.
$$

By the assumption of the theorem, which holds for $l=0$, and as we have a contractive sequence, it inductively follows that the assumption holds for $l \geq 0$. This concludes the proof. □

2.D Initialization Results

Now that local convergence has been established, we can now prove quantitative guarantees for the initialization methods provided in the main text.

2.D.1 Proof of Lemma 2.2.3

We first start with a proof of the guarantee provided by one-step hard thresholding, detailed in Lemma 2.2.3:

Proof. First, notice that for $\boldsymbol{W}_0 = \frac{L}{m}\mathcal{R}_\Omega(\boldsymbol{X})$, we get

$$
\begin{aligned}
\|\boldsymbol{X}_0-\boldsymbol{X}\| &\leq \|\boldsymbol{W}_0-\boldsymbol{X}\|+\|\boldsymbol{W}_0-\boldsymbol{X}_0\|\\
&\leq 2\,\|\boldsymbol{W}_0-\boldsymbol{X}\|,
\end{aligned}
$$

where the first inequality follows from the triangle inequality and the second inequality follows from the fact that $\boldsymbol{W}_0$ is the best rank-r approximation of $\boldsymbol{X}_0$ by Eckart-Young-Mirsky[EY36]. We now need a bound for this last term. Notice that $\boldsymbol{W}_0-\boldsymbol{X} = \frac{L}{m}\sum_{\boldsymbol{\alpha}}\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \boldsymbol{v}_{\boldsymbol{\alpha}} - \boldsymbol{X}$ is a

sum of zero-mean i.i.d. random matrices, opening up use of Bernstein's inequality. In order to use this, define $\boldsymbol{Z}_{\boldsymbol{\alpha}} = \frac{L}{m}\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \boldsymbol{v}_{\boldsymbol{\alpha}} - \frac{1}{m}\boldsymbol{X}$. We need a bound on $\|\boldsymbol{Z}_{\boldsymbol{\alpha}}\|$ and $\left\|\mathbb{E}[\boldsymbol{Z}_{\boldsymbol{\alpha}}]^2\right\|$. First, notice that

$$\begin{aligned}
\|\boldsymbol{Z}_{\boldsymbol{\alpha}}\| &= \left\| \frac{L}{m}\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \boldsymbol{v}_{\boldsymbol{\alpha}} - \frac{1}{m}\boldsymbol{X} \right\| \\
&\leq \frac{L}{m} |\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle| \, \|\boldsymbol{v}_{\boldsymbol{\alpha}}\| + \frac{1}{m}\|\boldsymbol{X}\| \\
&\leq \frac{L}{m}\left(\max_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \right) + \frac{n}{m}\left(\max_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \right) \\
&\leq \frac{2L}{m}\left(\max_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \right) =: c,
\end{aligned}$$

where the third inequality comes from the fact that for centered Gram matrices $\boldsymbol{X}$ that $\|\boldsymbol{X}\| \leq \frac{n}{2}\left(\max_{\boldsymbol{\alpha}\in\mathbb{I}}\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\right)$, and as $\|\boldsymbol{v}_{\boldsymbol{\alpha}}\| < 1$ from Lemma 2.A.3. Next, notice that

$$\left\|\mathbb{E}[\boldsymbol{Z}_{\boldsymbol{\alpha}}^2]\right\| = \left\| \frac{L}{m^2}\sum_{\boldsymbol{\alpha}} \langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \boldsymbol{v}_{\boldsymbol{\alpha}}^2 - \frac{1}{m^2}\boldsymbol{X}^2 \right\|.$$

As both these matrices are positive semi-definite, it follows that

$$\left\|\mathbb{E}[\boldsymbol{Z}_{\boldsymbol{\alpha}}^2]\right\| \leq \max\left\{ \left\| \frac{L}{m^2}\sum_{\boldsymbol{\alpha}} \langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \boldsymbol{v}_{\boldsymbol{\alpha}}^2 \right\|, \left\| \frac{1}{m^2}\boldsymbol{X}^2 \right\| \right\}.$$

It follows that

$$\begin{aligned}
\left\| \frac{L}{m^2}\sum_{\boldsymbol{\alpha}} \langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \boldsymbol{v}_{\boldsymbol{\alpha}}^2 \right\| &= \max_{\boldsymbol{y}\in\mathbb{R}^n, \|\boldsymbol{y}\|_2 = 1} \frac{L}{m^2}\sum_{\boldsymbol{\alpha}} \langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \boldsymbol{y}^\top \boldsymbol{v}_{\boldsymbol{\alpha}}^2 \boldsymbol{y} \\
&\leq \frac{L}{m^2}\left(\max_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \right) \boldsymbol{y}^\top \left(\sum_{\boldsymbol{\alpha}} \boldsymbol{v}_{\boldsymbol{\alpha}}^2 \right) \boldsymbol{y} \\
&= \frac{L}{m^2}\left(\max_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \right) \lambda_{\max}\left(\sum_{\boldsymbol{\alpha}} \boldsymbol{v}_{\boldsymbol{\alpha}}^2 \right).
\end{aligned}$$

Now, from Lemma 2.A.4, $\sum_{\boldsymbol{\alpha}} \boldsymbol{v}_{\boldsymbol{\alpha}}^2 = \frac{n^2-2n+2}{4n}\boldsymbol{J}$. It follows that $\lambda_{\max}\left(\sum_{\boldsymbol{\alpha}} \boldsymbol{v}_{\boldsymbol{\alpha}}^2\right) = \frac{n^2-2n+2}{4n} \leq \frac{n}{4}$ as $\boldsymbol{J}$ is an orthogonal projection. Thus,

$$\left\|\mathbb{E}\left(\boldsymbol{Z}_{\boldsymbol{\alpha}}^2\right)\right\| \leq \frac{nL}{m^2}\left(\max_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \right) =: V_0.$$

Now to determine t, we note that

$$\begin{aligned}\frac{V}{c} &= \frac{nL}{m}\left(\max_{\boldsymbol{\alpha}\in\mathbb{I}}\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2\right)\frac{m}{2L\left(\max_{\boldsymbol{\alpha}\in\mathbb{I}}\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\right)} \\ &= \frac{n}{2}\left(\max_{\boldsymbol{\alpha}\in\mathbb{I}}\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\right) \\ &\geq \sqrt{\frac{8\beta n^3\log n}{3m}}\left(\max_{\boldsymbol{\alpha}\in\mathbb{I}}\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\right),\end{aligned}$$

for $m \geq \frac{32}{3}\beta n\log n$. It follows that

$$\begin{aligned}\mathbb{P}\left(\|\boldsymbol{X}_0 - \boldsymbol{X}\| > \sqrt{\frac{8\beta n^3\log n}{3m}}\left(\max_{\boldsymbol{\alpha}\in\mathbb{I}}\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\right)\right) &\leq 2n\exp\left(-\beta\log(n)\right) \\ &= 2n^{1-\beta},\end{aligned}$$

verifying the probabilistic bound. To complete the proof we use Theorem 1.A.3 and it follows that

$$\|\boldsymbol{X}_0 - \boldsymbol{X}\|_{\mathrm{F}} \leq \sqrt{2r}\|\boldsymbol{X}_0 - \boldsymbol{X}\| \leq \sqrt{\frac{16r\beta n^3\log(n)}{3m}}\left(\max_{\boldsymbol{\alpha}\in\mathbb{I}}\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\right) \leq \sqrt{\frac{\nu^2 r^3 n\log(n)}{3m}}\|\boldsymbol{X}\|,$$

thus concluding the proof. □

Now, we will prove a technical lemma about Algorithm 6:

Lemma 2.D.1 (Trimming Result)**.** *Let $\boldsymbol{Z}_l = \boldsymbol{U}_l\boldsymbol{D}_l\boldsymbol{U}_l^\top$ be a rank-r matrix such that*

$$\|\boldsymbol{Z}_l - \boldsymbol{X}\| \leq \frac{\sigma_{\min}(\boldsymbol{X})}{10\sqrt{2}}.$$

Then the matrix $\hat{\boldsymbol{Z}}_l$ returned by Algorithm 6 satisfies

$$\|\mathcal{P}_{\hat{\boldsymbol{U}}_l}\boldsymbol{e}_i\| \leq \frac{10}{9}\sqrt{\frac{\nu r}{128n}}, \qquad \|\mathcal{P}_{\hat{\boldsymbol{U}}_l}\boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \leq \frac{10}{9}\sqrt{\frac{\nu r}{8n}}, \qquad \text{and} \qquad \|\mathcal{P}_{\hat{\boldsymbol{U}}_l}\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \leq \frac{10}{9}\sqrt{\frac{\nu r}{2n}},$$

and furthermore

$$\|\hat{\boldsymbol{Z}}_l - \boldsymbol{X}\|_{\mathrm{F}} \leq 8\kappa\|\boldsymbol{Z}_l - \boldsymbol{X}\|_{\mathrm{F}}.$$

Proof of Lemma 2.D.1. The proof of the first and fourth statements can be found in [WCCL20].

To see the second and third statements, we can apply the same analysis as in Remark 2.2.1. This analysis is reproduced here for convenience. First, notice that if $\|\mathcal{P}_{\hat{\boldsymbol{U}}_l}\boldsymbol{e}_{ij}\|_2 \leq \frac{10}{9}\sqrt{\frac{\nu r}{128n}}$, then by the triangle inequality

$$\|\mathcal{P}_{\hat{\boldsymbol{U}}_l}\boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 \leq \frac{40}{9}\|\mathcal{P}_U\boldsymbol{e}_{ij}\|_{\mathrm{F}} \leq \frac{10}{9}\sqrt{\frac{\nu r}{8n}}.$$

This validates the second statement. To see the third result, notice that

$$\begin{aligned}\|\mathcal{P}_{\hat{\boldsymbol{U}}_l}\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} &= \left\|\mathcal{P}_{\hat{\boldsymbol{U}}_l}\left(\sum_{\boldsymbol{\beta}\in\mathbb{I}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}}\boldsymbol{w}_{\boldsymbol{\beta}}\right)\right\|_{\mathrm{F}} \\ &\leq \sum_{\boldsymbol{\beta}\in\mathbb{I}}\left|H^{\boldsymbol{\alpha}\boldsymbol{\beta}}\right|\|\mathcal{P}_{\hat{\boldsymbol{U}}_l}\boldsymbol{w}_{\boldsymbol{\beta}}\|_{\mathrm{F}} \\ &\leq \frac{10}{9}\sqrt{\frac{\nu r}{8n}}\sum_{\boldsymbol{\beta}\in\mathbb{I}}\left|H^{\boldsymbol{\alpha}\boldsymbol{\beta}}\right|,\end{aligned}$$

and from Lemma 2.A.3, it follows that $\sum_{\boldsymbol{\alpha}\in\mathbb{I}}|H^{\boldsymbol{\alpha}\boldsymbol{\beta}}| \leq 2$, so the last statement follows. □

2.D.2 Proof of Lemma 2.2.5

Proof. First, assume that at the l-th iteration of Algorithm 5,

$$\|\boldsymbol{Z}_l - \boldsymbol{X}\|_{\mathrm{F}} \leq \frac{\sigma_{\max}(\boldsymbol{X})}{256\kappa^2}.$$

This indicates that $\hat{\boldsymbol{Z}}_l$ is $\frac{100}{81}$-ν incoherent with respect to $\{\boldsymbol{w}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$ and that

$$\|\hat{\boldsymbol{Z}}_l - \boldsymbol{X}\|_{\mathrm{F}} \leq 8\kappa\|\boldsymbol{Z}_l - \boldsymbol{X}\|_{\mathrm{F}}.$$

Following a similar strategy as in the proof for Theorem 2.2.2, we can decompose the error at the $l+1$-th iteration as follows:

$$\begin{aligned}\|\boldsymbol{Z}_{l+1} - \boldsymbol{X}\|_{\mathrm{F}} \leq\ & \underbrace{2\left\|(\mathcal{P}_{\hat{\mathbb{T}}_l} - \frac{L}{m}\mathcal{P}_{\hat{\mathbb{T}}_l}\mathcal{R}_{\Omega_{l+1}}\mathcal{P}_{\hat{\mathbb{T}}_l})(\hat{\boldsymbol{Z}}_l - \boldsymbol{X})\right\|_{\mathrm{F}}}_{I_4} \\ & + \underbrace{2\left\|(\mathcal{I} - \mathcal{P}_{\hat{\mathbb{T}}_l})(\hat{\boldsymbol{Z}}_l - \boldsymbol{X})\right\|_{\mathrm{F}}}_{I_5}\end{aligned}$$

$$+ \underbrace{2 \left\| \frac{L}{m} \mathcal{P}_{\hat{\mathbb{T}}_l} \mathcal{R}_{\Omega_{l+1}} (\mathcal{I} - \mathcal{P}_{\hat{\mathbb{T}}_l})(\hat{\boldsymbol{Z}}_l - \boldsymbol{X}) \right\|_{\mathrm{F}}}_{I_6}.$$

Now, as $\hat{\boldsymbol{Z}}_l$ and Ω_{l+1} are independent, we can use Theorem 2.2.1 to bound I_4 as follows:

$$\begin{aligned} I_4 &\leq 2 \left\| \frac{L}{m} \mathcal{P}_{\hat{\mathbb{T}}_l} \mathcal{R}_{\Omega_{l+1}} \mathcal{P}_{\hat{\mathbb{T}}_l} - \mathcal{P}_{\hat{\mathbb{T}}_l} \right\| \|\hat{\boldsymbol{Z}}_l - \boldsymbol{X}\|_{\mathrm{F}} \\ &\leq 2\sqrt{\frac{8(100\nu/81)^2 r^2 \beta n \log(n)}{3\hat{m}}} \|\hat{\boldsymbol{Z}}_l - \boldsymbol{X}\|_{\mathrm{F}} \\ &\leq 16\kappa \sqrt{\frac{80000\nu^2 r^2 \beta n \log(n)}{19683\hat{m}}} \|\boldsymbol{Z}_l - \boldsymbol{X}\|_{\mathrm{F}} \\ &\leq 32\kappa \sqrt{\frac{\nu^2 r^2 \beta n \log(n)}{\hat{m}}} \|\boldsymbol{Z}_l - \boldsymbol{X}\|_{\mathrm{F}}, \end{aligned}$$

with probability at least $1 - 2n^{1-\beta}$ as long as $\hat{m} \geq 4\nu^2 r^2 \beta n \log(n)$.

Next, we can bound I_5 as follows:

$$\begin{aligned} I_5 &\leq \frac{2\|\hat{\boldsymbol{Z}}_l - \boldsymbol{X}\|_{\mathrm{F}}^2}{\sigma_{\min}(\boldsymbol{X})} \\ &\leq \frac{128\kappa^2 \|\boldsymbol{Z}_l - \boldsymbol{X}\|_{\mathrm{F}}^2}{\sigma_{\min}(\boldsymbol{X})} \\ &\leq \frac{1}{2} \|\boldsymbol{Z}_l - \boldsymbol{X}\|_{\mathrm{F}}, \end{aligned}$$

where the first inequality follows from Lemma 1.8.3, the second inequality comes from Lemma 2.D.1, and the last inequality comes from the starting assumption.

For the final result, again recall that $\hat{\boldsymbol{Z}}_l$ and Ω_{l+1} are independent, and $\hat{\boldsymbol{Z}}_l$ is $\frac{100}{81}$-ν incoherent with respect to $\{\boldsymbol{w}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathbb{I}}$. From Lemma 2.B.4 and the new incoherence parameter, we have that

$$\begin{aligned} \left\| \frac{L}{m} \mathcal{P}_{\hat{\mathbb{T}}_l} \mathcal{R}_{\Omega_{l+1}} \mathcal{P}_{\hat{\mathbb{T}}_l} (\mathcal{P}_U - \mathcal{P}_{\hat{U}_l}) - \mathcal{P}_{\hat{\mathbb{T}}_l} (\mathcal{P}_U - \mathcal{P}_{\hat{U}_l}) \right\| &\leq \sqrt{\frac{40000\beta\nu^2 r^2 n \log(n)}{19683\hat{m}}} \\ &\leq \sqrt{\frac{2\beta\nu^2 r^2 n \log(n)}{\hat{m}}}, \end{aligned}$$

with probability at least $1 - 2n^{1-\beta}$ given $\hat{m} \geq 2\beta\nu^2 r^2 n \log(n)$. Now, as

$$(\mathcal{I} - \mathcal{P}_{\hat{\mathbb{T}}_l})(\hat{\boldsymbol{Z}}_l - \boldsymbol{X}) = -(\mathcal{I} - \mathcal{P}_{\hat{\mathbb{T}}_l})(\boldsymbol{X})$$

$$
\begin{aligned}
&= -\boldsymbol{X} + \hat{\boldsymbol{U}}_l\hat{\boldsymbol{U}}_l^\top\boldsymbol{X} + \boldsymbol{X}\hat{\boldsymbol{U}}_l\hat{\boldsymbol{U}}_l^\top - \hat{\boldsymbol{U}}_l\hat{\boldsymbol{U}}_l^\top\boldsymbol{X}\hat{\boldsymbol{U}}_l\hat{\boldsymbol{U}}_l^\top \\
&= -\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{X} + \hat{\boldsymbol{U}}_l\hat{\boldsymbol{U}}_l^\top\boldsymbol{X} + \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{X}\hat{\boldsymbol{U}}_l\hat{\boldsymbol{U}}_l^\top - \hat{\boldsymbol{U}}_l\hat{\boldsymbol{U}}_l^\top\boldsymbol{X}\hat{\boldsymbol{U}}_l\hat{\boldsymbol{U}}_l^\top \\
&= -(\boldsymbol{U}\boldsymbol{U}^\top - \hat{\boldsymbol{U}}_l\hat{\boldsymbol{U}}_l^\top)\boldsymbol{X}(\boldsymbol{I} - \hat{\boldsymbol{U}}_l\hat{\boldsymbol{U}}_l^\top) \\
&= (\boldsymbol{U}\boldsymbol{U}^\top - \hat{\boldsymbol{U}}_l\hat{\boldsymbol{U}}_l^\top)(\hat{\boldsymbol{Z}}_l - \boldsymbol{X})(\boldsymbol{I} - \hat{\boldsymbol{U}}_l\hat{\boldsymbol{U}}_l^\top) \\
&= (\mathcal{P}_U - \mathcal{P}_{\hat{U}_l})(\hat{\boldsymbol{Z}}_l - \boldsymbol{X})(\mathcal{I} - \mathcal{P}_{\hat{U}_l}),
\end{aligned}
$$

where the first line follows from the fact that $\mathcal{P}_{\hat{\mathbb{T}}_l}\hat{\boldsymbol{Z}}_l = \hat{\boldsymbol{Z}}_l$, the second line follows from the definition of $\mathcal{P}_{\hat{\mathbb{T}}_l}$, the third line follows from the fact that $\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{X} = \boldsymbol{X}$, the fourth line is a rearrangement of terms, and the fifth line follows from the fact that $\hat{\boldsymbol{Z}}_l(\boldsymbol{I} - \hat{\boldsymbol{U}}_l\hat{\boldsymbol{U}}_l^\top) = \boldsymbol{0}$. It follows that

$$
\begin{aligned}
I_6 &= 2\left\|\frac{L}{m}\mathcal{P}_{\hat{\mathbb{T}}_l}\mathcal{R}_{\Omega_{l+1}}(\mathcal{P}_U - \mathcal{P}_{\hat{U}_l})(\hat{\boldsymbol{Z}}_l - \boldsymbol{X})(\mathcal{I} - \mathcal{P}_{\hat{U}_l})\right\|_{\mathrm{F}} \\
&= 2\left\|\frac{L}{m}\mathcal{P}_{\hat{\mathbb{T}}_l}\mathcal{R}_{\Omega_{l+1}}(\mathcal{P}_U - \mathcal{P}_{\hat{U}_l})(\hat{\boldsymbol{Z}}_l - \boldsymbol{X})(\mathcal{I} - \mathcal{P}_{\hat{U}_l}) - \underbrace{\mathcal{P}_{\hat{\mathbb{T}}_l}(\mathcal{P}_U - \mathcal{P}_{\hat{U}_l})(\hat{\boldsymbol{Z}}_l - \boldsymbol{X})(\mathcal{I} - \mathcal{P}_{\hat{U}_l})}_{=\boldsymbol{0}}\right\|_{\mathrm{F}} \\
&\leq 2\left\|\frac{L}{m}\mathcal{P}_{\hat{\mathbb{T}}_l}\mathcal{R}_{\Omega_{l+1}}(\mathcal{P}_U - \mathcal{P}_{\hat{U}_l}) - \mathcal{P}_{\hat{\mathbb{T}}_l}(\mathcal{P}_U - \mathcal{P}_{\hat{U}_l})\right\|\left\|(\hat{\boldsymbol{Z}}_l - \boldsymbol{X})(\mathcal{I} - \mathcal{P}_{\hat{U}_l})\right\|_{\mathrm{F}} \\
&\leq 2\left\|\frac{L}{m}\mathcal{P}_{\hat{\mathbb{T}}_l}\mathcal{R}_{\Omega_{l+1}}(\mathcal{P}_U - \mathcal{P}_{\hat{U}_l}) - \mathcal{P}_{\hat{\mathbb{T}}_l}(\mathcal{P}_U - \mathcal{P}_{\hat{U}_l})\right\|\left\|\hat{\boldsymbol{Z}}_l - \boldsymbol{X}\right\|_{\mathrm{F}} \\
&\leq 16\kappa\sqrt{\frac{2\beta\nu^2r^2n\log(n)}{\hat{m}}},
\end{aligned}
$$

where the first two lines follow from the computation above. Combining I_4, I_5, and I_6 gives

$$
\|\boldsymbol{Z}_{l+1} - \boldsymbol{X}\|_{\mathrm{F}} \leq \left(\frac{1}{2} + 64\kappa\sqrt{\frac{\beta\nu^2r^2n\log(n)}{\hat{m}}}\right)\|\boldsymbol{Z}_l - \boldsymbol{X}\|_{\mathrm{F}}, \tag{2.15}
$$

with probability at least $1-4n^{1-\beta}$. It follows that (2.15) is less than $\frac{5}{6}$ for $\hat{m} \geq (192\nu r\kappa)^2\beta n\log(n)$.

Now, as $\boldsymbol{Z}_0 = \mathcal{H}_r\left(\frac{L}{\hat{m}}\mathcal{R}_{\Omega_0}(\boldsymbol{X})\right)$, we can make $\|\boldsymbol{Z}_0 - \boldsymbol{X}\|_{\mathrm{F}} \leq \frac{\sigma_{\min}(\boldsymbol{X})}{256\kappa^2}$ using the one step hard thresholding result from Lemma 2.2.3 for $\hat{m} \geq (2.2\times10^4)\kappa^6\nu^2r^3\beta n\log(n)$, and the result follows from here. No attempts were made to optimize the constants. □

Chapter 3

An Improved Non-convex Riemannian Approach to the EDMC Problem

In this chapter, we define a new algorithm that fills in the performance gap between Riemannian Pseudogradient Descent and Restricted Frame Riemannian Gradient Descent for EDMC to answer Question 1.3.1. This is done by defining a novel operator to the EDMC literature, and this new algorithm requires a wholly novel set of techniques to the EDMC literature to prove local convergence. Much of this chapter has been adapted from [SCT25]. In this chapter, we define a new algorithm that fills in the performance gap between Riemannian Pseudogradient Descent and Restricted Frame Riemannian Gradient Descent for EDMC to answer Question 1.3.1. This is done by defining an operator novel to the EDMC literature, and this new algorithm requires a set of techniques wholly novel to the EDMC literature to prove local convergence. Much of this chapter has been adapted from [SCT25].

3.1 An Improved Operator

With the goal of translating the standard matrix completion problem to Gram matrix completion of a ground truth matrix $\boldsymbol{X} \in \mathbb{S}$, where $\mathbb{S} = \{\boldsymbol{Y} \in \mathbb{R}^{n\times n} | \boldsymbol{Y} = \boldsymbol{Y}^\top, \boldsymbol{Y}\mathbf{1} = \mathbf{0}\}$, the most direct adaptation of the work conducted in [WCCL20] would be defining an objective function

by analogy to (1.14) as follows:

$$\underset{\boldsymbol{Y}\in\mathbb{S}}{\text{minimize}}\ \langle \boldsymbol{Y}-\boldsymbol{X}, \mathcal{R}_\Omega(\boldsymbol{Y}-\boldsymbol{X})\rangle \ \text{ subject to } \ \text{rank}(\boldsymbol{Y}) = r.$$

However, a notable challenge arises: computing the Euclidean gradient of the objective function necessitates unavailable information in the form $\langle \boldsymbol{X}, \boldsymbol{v}_{\boldsymbol{\alpha}}\rangle$ from $\mathcal{R}^*_\Omega(\boldsymbol{X})$ as

$$\nabla_{\boldsymbol{Y}}\left(\langle \boldsymbol{Y}-\boldsymbol{X}, \mathcal{R}_\Omega(\boldsymbol{Y}-\boldsymbol{X})\rangle\right) = \mathcal{R}_\Omega(\boldsymbol{Y}-\boldsymbol{X}) + \mathcal{R}^*_\Omega(\boldsymbol{Y}-\boldsymbol{X}),$$

where $\nabla_{\boldsymbol{Y}}$ denotes the gradient with respect to $\boldsymbol{Y}$. This quantity is inaccessible under the given problem statement, since each $\boldsymbol{v}_{\boldsymbol{\alpha}}$ depends on all $\boldsymbol{w}_{\boldsymbol{\alpha}}$ through the relation $\boldsymbol{v}_{\boldsymbol{\alpha}} = \sum_{\boldsymbol{\alpha}\in\mathbb{I}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}}\boldsymbol{w}_{\boldsymbol{\beta}}$. Consequently, the computation $\langle \boldsymbol{X}, \boldsymbol{v}_{\boldsymbol{\alpha}}\rangle = \sum_{\boldsymbol{\alpha}\in\mathbb{I}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}}\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle$, requires complete knowledge of all pairwise distances. Therefore, evaluating $\mathcal{R}^*_\Omega(\boldsymbol{X})$ necessitates access to all information. To circumvent this difficulty, there has been exploration into self-adjoint alternatives to $\mathcal{R}_\Omega$ [TL21, TL18, SLCT23]. The novel surrogate introduced in this work, denoted $\mathcal{M}_\Omega$, allows for the definition of an objective function analogous to (1.14).

We now define $\mathcal{M}_\Omega : \mathbb{S} \to \mathbb{S}$. This operator samples indices from $\mathbb{I}$ with uniform Bernoulli probability p, and is defined as follows:

$$\mathcal{M}_\Omega(\cdot) = \sum_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\Omega} C_{\boldsymbol{\alpha}\boldsymbol{\beta}}\boldsymbol{w}_{\boldsymbol{\alpha}}\langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle \cdot, \boldsymbol{w}_{\boldsymbol{\beta}}\rangle, \tag{3.1}$$

where $C_{\boldsymbol{\alpha}\boldsymbol{\alpha}} = p$ for all $\boldsymbol{\alpha}$, and $C_{\boldsymbol{\alpha}\boldsymbol{\beta}} = 1$ for all $\boldsymbol{\alpha} \neq \boldsymbol{\beta}$. Previous literature introduced an unscaled form of this operator, i.e., $C_{\boldsymbol{\alpha}\boldsymbol{\alpha}} = 1$, which is equivalent to $\mathcal{R}^*_\Omega\mathcal{R}_\Omega$. This operator was discussed in Chapter 2 and was first introduced in [SLCT23]. However, $\mathcal{R}^*_\Omega\mathcal{R}_\Omega$ does not concentrate around $p^2\mathcal{I}$. The concentration around $p^2\mathcal{I}$ is a necessary condition for convergence of the algorithm presented in [SLCT23], and to rectify this discrepancy, a different operator must be considered. The diagonal re-scaling present in $\mathcal{M}_\Omega$ enforces that $\mathbb{E}[\mathcal{M}_\Omega] = p^2\mathcal{I}$, with proof provided in Theorem 3.B.1. Additionally, $\mathcal{M}_\Omega$ is self-adjoint, allowing us to define the following objective function for the EDMC problem using this operator:

$$\underset{\boldsymbol{Y}\in\mathbb{S}}{\text{minimize}}\ \frac{1}{2}\langle \boldsymbol{Y}-\boldsymbol{X}, \mathcal{M}_\Omega(\boldsymbol{Y}-\boldsymbol{X})\rangle \ \text{ subject to } \ \text{rank}(\boldsymbol{Y}) = r. \tag{3.2}$$

This object is a true quadratic form with a symmetric operator, and its Euclidean gradient is given solely by $\mathcal{M}_\Omega(\boldsymbol{Y} - \boldsymbol{X})$. As such, it can be approached identically to (1.14) following the principles outlined in Section 1.4.2. To perform this first-order retraction method, we need to define the tangent space at each iterate $\boldsymbol{X}_l \in \mathcal{N}_r$. Given $\boldsymbol{X}_l = \boldsymbol{U}_l \boldsymbol{\Lambda}_l \boldsymbol{U}_l^\top$ for $\boldsymbol{U}_l \in \mathbb{R}^{n\times r}$, we can define $\mathbb{T}_l$ as in (1.15). The projection onto $\mathbb{T}_l$, denoted $\mathcal{P}_{\mathbb{T}_l}$, is defined in (1.16). The retraction map is the hard thresholding operator $\mathcal{H}_r$, as defined in (1.17). We can now define Algorithm 7, the main object of study in this chapter:

Algorithm 7 Dual Basis Riemannian EDMC (DBRE)

Input: Ω, $\{\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\}_{\boldsymbol{\alpha}\in\Omega}$, $\operatorname{rank}(\boldsymbol{X}) = r$
Initialization: $\boldsymbol{X}_0 = \boldsymbol{U}_0 \boldsymbol{D}_0 \boldsymbol{U}_0^\top$
for $l = 0, 1, \cdots$ **do**
  $\boldsymbol{G}_l = \mathcal{M}_\Omega(\boldsymbol{X} - \boldsymbol{X}_l)$
  $\alpha_l = \frac{\|\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l\|_{\mathrm{F}}^2}{\langle \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l, \mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l\rangle}$
  $\boldsymbol{W}_l = \boldsymbol{X}_l + \alpha_l \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l$
  $\boldsymbol{X}_{l+1} = \mathcal{H}_r(\boldsymbol{W}_l)$
end for
Output: $\boldsymbol{X}_{\mathrm{rec}}$: Estimated Gram matrix.

We note that this algorithm is a standard projected gradient descent scheme on $\mathcal{N}_r$; the novelty presented in this work is in the construction of the measurement operator $\mathcal{M}_\Omega$ to build the proper objective function, which enables the convergence analysis of Algorithm 7, presented in Section 3.2.

In the approach seen in DBRE, the thin spectral decomposition in the gradient descent scheme is the most expensive, especially when n is large. As described previously, the authors in [WCCL20] found an efficient way to reduce the computational complexity of this decomposition from $\mathcal{O}(rn^2)$ to $\mathcal{O}(r^3 + r^2 n)$, substantially reducing the cost per iteration, which we implement as well. We note that in DBRE, the reconstruction of the ground truth Gram matrix $\boldsymbol{X}$ is equivalent to the reconstruction of $\boldsymbol{D}$, as there is a one-to-one correspondence between $\boldsymbol{X}$ and $\boldsymbol{D}$ through (1.2).

Remark 3.1.1: Operator Interpretation

We wish to provide an interpretation of the operators $\mathcal{M}_\Omega$ and $\mathcal{R}_\Omega^*\mathcal{R}_\Omega$, and why they are considered. Although $\mathcal{F}_\Omega$ concentrates around its expectation through a Bernstein inequality—see Theorem 3.B.2—its expectation is not the identity operator $\mathcal{I}$, and as such $\mathcal{F}_\mathbb{I} \neq \mathcal{I}$. In fact, the spectrum of $\mathcal{F}_\mathbb{I}$ is known to be equivalent to the spectrum of $\boldsymbol{H}$, and thus $\lambda_{\max}(\mathcal{F}_\mathbb{I}) = 2n$ [LT23]. As such, neither $\|\mathcal{F}_\Omega - p\mathcal{I}\|$ nor $\|\mathcal{P}_\mathbb{T}\mathcal{F}_\Omega\mathcal{P}_\mathbb{T} - p\mathcal{P}_\mathbb{T}\|$ are small. In particular, $\|\mathcal{P}_\mathbb{T}\mathcal{F}_\Omega\mathcal{P}_\mathbb{T} - p\mathcal{P}_\mathbb{T}\| = \mathcal{O}(1)$, but as the operator needs to exhibit RIP for the analysis of an algorithm like DBRE, this bound is not tight enough to guarantee convergence. We can instead consider re-scaling the geometry of the linear space $\mathbb{S}$ through some function f so that $\mathbb{E}[\mathcal{F}_\Omega(f(\cdot))] \approx p\mathcal{I}$. First, define $\mathcal{S}_\Omega : \mathbb{R}^{n\times n} \to \mathbb{R}^L$ as $(\mathcal{S}_\Omega(\boldsymbol{X}))_{\boldsymbol{\alpha}} = \langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle$ for $\boldsymbol{\alpha} \in \Omega$, and 0 otherwise. As such, one can show that $\mathcal{F}_\Omega = \mathcal{S}_\Omega^*\mathcal{S}_\Omega$, where $\mathcal{S}_\Omega^* : \mathbb{R}^L \to \mathbb{R}^{n\times n}$ is the adjoint of $\mathcal{S}_\Omega$. To re-scale $\mathcal{F}_\Omega$, one can instead consider $\mathcal{S}_\Omega^*\boldsymbol{H}^{-1}\mathcal{S}_\Omega$. This rescaling is done with $\boldsymbol{H}^{-1}$ to make it so that $\mathcal{S}_\Omega^*\boldsymbol{H}^{-1}\mathcal{S}_\Omega = \mathcal{I}$ when $\Omega = \mathbb{I}$. One can compute out $\mathcal{S}_\Omega^*\boldsymbol{H}^{-1}\mathcal{S}_\Omega$ and show that

$$\mathcal{S}_\Omega^*\boldsymbol{H}^{-1}\mathcal{S}_\Omega = \mathcal{R}_\Omega^*\mathcal{R}_\Omega.$$

Further investigation in Theorem 3.B.1 validates the necessity of considering a diagonally reweighted variant of $\mathcal{R}_\Omega^*\mathcal{R}_\Omega$ to ensure concentration around $\mathcal{I}$, resulting in $\mathcal{M}_\Omega$.

3.1.1 Implementation Efficiency

We use recent advances in Riemannian optimization from [WCCL20] and [SLCT23] to develop an efficient implementation of the proposed algorithm. Computation of $\mathcal{R}_\Omega(\boldsymbol{X})$ and $\mathcal{M}_\Omega(\boldsymbol{X})$ can be done efficiently, with a minimal complexity per iteration. For $\mathcal{R}_\Omega$, a given iterate $\boldsymbol{X}_l$ can be easily translated to its distance matrix $\boldsymbol{D}_l$ via (1.2), and through (1.10), $\mathcal{R}_\Omega(\boldsymbol{X}_l)$ can be computed in $\mathcal{O}(m)$ operations, for $|\Omega| = m$. First, we note that $\mathcal{M}_\Omega(\boldsymbol{X}_l) = \mathcal{R}_\Omega^*\mathcal{R}_\Omega(\boldsymbol{X}_l) - \|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_\mathrm{F}^2(1-p)\mathcal{F}_\Omega(\boldsymbol{X}_l)$ and $\|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_\mathrm{F}$ is constant for all $\boldsymbol{\alpha} \in \mathbb{I}$ (Theorem 2.A.3). This can be seen as

follows:

$$\begin{aligned}
\mathcal{M}_\Omega(\cdot) &= \sum_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\Omega} C_{\boldsymbol{\alpha}\boldsymbol{\beta}}\boldsymbol{w}_{\boldsymbol{\alpha}}\langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle \cdot, \boldsymbol{w}_{\boldsymbol{\beta}}\rangle \\
&= \sum_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta}\in\Omega \\ \boldsymbol{\alpha}=\boldsymbol{\beta}}} C_{\boldsymbol{\alpha}\boldsymbol{\beta}}\boldsymbol{w}_{\boldsymbol{\alpha}}\langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle \cdot, \boldsymbol{w}_{\boldsymbol{\beta}}\rangle + \sum_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta}\in\Omega \\ \boldsymbol{\alpha}\neq\boldsymbol{\beta}}} C_{\boldsymbol{\alpha}\boldsymbol{\beta}}\boldsymbol{w}_{\boldsymbol{\alpha}}\langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle \cdot, \boldsymbol{w}_{\boldsymbol{\beta}}\rangle \\
&= \sum_{\boldsymbol{\alpha}\in\Omega} C_{\boldsymbol{\alpha}\boldsymbol{\alpha}}\boldsymbol{w}_{\boldsymbol{\alpha}}\langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\alpha}}\rangle\langle \cdot, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle + \sum_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta}\in\Omega \\ \boldsymbol{\alpha}\neq\boldsymbol{\beta}}} C_{\boldsymbol{\alpha}\boldsymbol{\beta}}\boldsymbol{w}_{\boldsymbol{\alpha}}\langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle \cdot, \boldsymbol{w}_{\boldsymbol{\beta}}\rangle \\
&= p\|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2\mathcal{F}_\Omega(\cdot) + \sum_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta}\in\Omega \\ \boldsymbol{\alpha}\neq\boldsymbol{\beta}}} \boldsymbol{w}_{\boldsymbol{\alpha}}\langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle \cdot, \boldsymbol{w}_{\boldsymbol{\beta}}\rangle \\
&= p\|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2\mathcal{F}_\Omega(\cdot) + \sum_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\Omega} \boldsymbol{w}_{\boldsymbol{\alpha}}\langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle \cdot, \boldsymbol{w}_{\boldsymbol{\beta}}\rangle - \sum_{\boldsymbol{\alpha}\in\Omega}\langle \cdot, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\alpha}}\rangle\boldsymbol{w}_{\boldsymbol{\alpha}} \\
&= p\|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2\mathcal{F}_\Omega(\cdot) + \mathcal{R}_\Omega^*\mathcal{R}_\Omega(\cdot) - \|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2\mathcal{F}_\Omega(\cdot),
\end{aligned} \tag{3.3}$$

as expected. It is known that $\mathcal{R}_\Omega^*\mathcal{R}_\Omega(\boldsymbol{X}_l)$ is $\mathcal{O}(m)$-sparse and requires $\mathcal{O}(m)$ operations to compute [SLCT23]. The argument is outlined as follows. Let $\mathcal{T} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ denote the map defined by (1.2) and let $\mathcal{T}^*$ denote its adjoint. It was shown in [SLCT23] that, correcting for a previously missing minus sign, for a Gram matrix $\boldsymbol{X}$,

$$\mathcal{R}_\Omega^*\mathcal{R}_\Omega(\boldsymbol{X}) = -\frac{1}{4}\mathcal{T}^*\left(\mathcal{P}_\Omega\left(\boldsymbol{J}\mathcal{P}_\Omega(\mathcal{T}(\boldsymbol{X}))\boldsymbol{J}\right)\right).$$

For any matrix $\boldsymbol{Y}$, both $\mathcal{P}_\Omega(\boldsymbol{Y})$ and $\boldsymbol{J}\mathcal{P}_\Omega(\boldsymbol{Y})\boldsymbol{J}$ are computable in $\mathcal{O}(m)$ operations. The accessible information in the EDMC problem is of the form $\mathcal{P}_\Omega(\mathcal{T}(\boldsymbol{X}))$. Furthermore, $\mathcal{T}^*\left(\mathcal{P}_\Omega(\boldsymbol{Y})\right)$ for any $\boldsymbol{Y}$ is computable in $\mathcal{O}(m)$ operations as well.

Next, $\mathcal{F}_\Omega(\boldsymbol{X}_l)$ is efficiently computable in $\mathcal{O}(m)$ operations as each matrix $\boldsymbol{w}_{\boldsymbol{\alpha}}$ has 4 non-zero entries, allowing for easy computation given $\{\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\}_{\boldsymbol{\alpha}\in\Omega}$. As such, $\mathcal{F}_\Omega(\boldsymbol{X}_l)$ is $\mathcal{O}(m)$-sparse. Using the fact that $\mathcal{R}_\Omega^*\mathcal{R}_\Omega(\cdot)$ and $\mathcal{F}_\Omega(\cdot)$ are sparse, it can be easily argued that the sum of the three terms in (3.3) preserves a common sparsity pattern, and it can be computed in $\mathcal{O}(m)$ operations. Therefore $\mathcal{M}_\Omega(\boldsymbol{X}_l)$, and thus $\boldsymbol{G}_l$ in Step 4 of DBRE, is computable in $\mathcal{O}(m)$ operations.

Step 5 can be computed in $\mathcal{O}(n^2)$ operations, as $\mathcal{P}_{\mathbb{T}}\boldsymbol{G}_l$ is a dense matrix. Some calculation

yields that steps 6 and 7 can be computed with $n^2r+\mathcal{O}(nr^2+r^3)$ [WCCL20], giving a total cost per iteration of $n^2r+\mathcal{O}(m+nr^2+r^3)$. Note that the dominant cost is n^2r, which is less expensive than computing Step 7 using the truncated singular value decomposition directly. Although both approaches have the same asymptotic complexity, the latter incurs a significantly higher constant factor (e.g., a factor of 6 or 14 depending on the choice of algorithm; see, for example, Figure 8.6.1 in [GVL13]).

3.2 Theoretical Analysis

In this section, we will provide the main results of this work, which are the local convergence and recovery guarantees for DBRE, presented in Theorems 3.2.2 and 3.2.4. Prior to this, we formally state our incoherence assumptions, expanding upon the assumption first described in Section 1.6:

Assumption 3.2.1: Incoherence Assumption

Let $\boldsymbol{X} \in \mathbb{R}^{n\times n}$ be a rank-r matrix with eigenvalue decomposition $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top$. We assume that $\boldsymbol{X}$ is ν-incoherent to the basis $\{\boldsymbol{w}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$ and ν-incoherent to its dual basis $\{\boldsymbol{v}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$; that is, there exists a constant $\nu \geq 1$ such that for all $\boldsymbol{\alpha} = (i,j) \in \mathbb{I}$:

$$\|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \leq \sqrt{\frac{\nu r}{2n}}, \quad \text{and} \quad \|\mathcal{P}_U \boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \leq \sqrt{\frac{\nu r}{2n}}. \tag{3.4}$$

In addition to the above, we require that

$$\|\mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \leq \sqrt{\frac{\nu r}{2n}}, \quad \text{and} \quad \|\mathcal{P}_{\mathbb{T}} \boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \leq \sqrt{\frac{\nu r}{2n}}. \tag{3.5}$$

Notice that the two definitions in (3.4) and (3.5) are equivalent up to a small constant, as

$$\|\mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} = \|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}} + \boldsymbol{w}_{\boldsymbol{\alpha}}\mathcal{P}_U - \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\mathcal{P}_U\|_{\mathrm{F}} \leq 3\,\|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}\,,$$

where the first inequality follows from the triangle inequality and the self-adjointness of $\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}$,

and because

$$\|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} = \|\mathcal{P}_U \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \leq \|\mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}},$$

where the equality follows from the definition of $\mathcal{P}_U$ and $\mathcal{P}_{\mathbb{T}}$, and the inequality follows from Cauchy-Schwarz. As such, we pick a ν large enough such that the inequalities in (3.4) and (3.5) hold. We note that the constant difference in the condition stated above and in Section 1.6 is merely a matter of mathematical convenience. We also note that these incoherence conditions are similar to those seen in matrix completion with respect to the standard basis [Rec11a], as well as completion with respect to other bases [Gro11, TL18].

Remark 3.2.1: Incoherence Equivalence

We want to note that ν-incoherence with respect to $\boldsymbol{w}_{\boldsymbol{\alpha}}$ in both (3.4) and (3.5) implies, at worst, 4ν-incoherence with respect to $\boldsymbol{v}_{\boldsymbol{\alpha}}$. As such, we choose a ν large enough so that both $\boldsymbol{X}$ is ν-incoherent with respect to $\boldsymbol{w}_{\boldsymbol{\alpha}}$ and $\boldsymbol{v}_{\boldsymbol{\alpha}}$. See Theorem 1.A.2 for details.

We provide one further assumption for this work. As we are typically interested in large n, assuming that $n \geq 3$ produces uniform results for several numerical bounds in the appendix, and is formally stated as an assumption.

Assumption 3.2.2: Size Assumption

For the given ground truth rank-r matrix $\boldsymbol{X} \in \mathbb{R}^{n \times n}$, we assume that $n \geq 3$.

Throughout the remainder of this work, we will assume that our ground truth matrix $\boldsymbol{X} \in \mathbb{S}$ satisfies Assumption 3.2.1 with constant factor ν. As in [WCCL20], we identify a neighborhood in $\mathcal{N}_r$ around which any initial guess in this neighborhood converges linearly to the true solution with high probability using DBRE.

3.2.1 Local Convergence Analysis

The most critical property for a sampling operator to possess in matrix completion theory is the restricted isometry property, briefly discussed in Section 1.2.2. This property roughly states that, when restricted to the local structure (or tangent space) around the true low-rank matrix,

the partial observations preserve enough information to allow for faithful algorithmic recovery. We state this more formally with the following theorem:

Theorem 3.2.1 (RIP of $\mathcal{M}_\Omega$)**.** *Let $\boldsymbol{X} \in \mathbb{S}$ be the ground truth, rank-r, ν-incoherent Gram matrix with tangent space $\mathbb{T}$ in $\mathcal{N}_r$. Let Ω be sampled from $\mathbb{I}$ via a Bernoulli sampling process with parameter $p \geq \frac{128}{3}\beta\frac{\log n}{n}$ for any $\beta > 3$. With probability at least $1 - C'n^{3-\beta}$, where C' is an absolute constant, we have that for some absolute numerical constant $C > 0$ independent of n or $\boldsymbol{X}$ that*

$$p^{-2}\|\mathcal{P}_\mathbb{T}\mathcal{M}_\Omega\mathcal{P}_\mathbb{T} - p^2\mathcal{P}_\mathbb{T}\| \leq 7\sqrt{\frac{\nu^2 r^2 \beta \log n}{pn}} + C\sqrt{\beta}\nu r\frac{\log n}{pn}.$$

Furthermore, for any ε_0, if $p \geq \frac{C''\nu^2 r^2}{\varepsilon_0^2}\frac{\beta\log n}{n}$ for some sufficiently large numerical constant $C'' > 0$, then

$$p^{-2}\|\mathcal{P}_\mathbb{T}\mathcal{M}_\Omega\mathcal{P}_\mathbb{T} - p^2\mathcal{P}_\mathbb{T}\| \leq \varepsilon_0.$$

Proof sketch. This proof works by decomposing $\mathcal{M}_\Omega$ into diagonal and off-diagonal components. We recognize that estimating deviations from expectation in the off-diagonal terms in $\|\mathcal{P}_\mathbb{T}\mathcal{M}_\Omega\mathcal{P}_\mathbb{T} - p^2\mathcal{P}_\mathbb{T}\|$ is a second-order degenerate U-statistic—see [Lee19] for details on the theory of U-statistics—allowing the application of the decoupling inequality from [DlPG12], with the relevant theorem reproduced in Theorem 1.7.5. Once the random variables are decoupled, we can conditionally apply a non-commutative Bernstein inequality, reproduced in Theorem 1.7.2, and use the law of total probability to obtain a high-probability bound for the total deviation. The concentration of the diagonal terms in the full expression is equivalent to the concentration of $p\|\mathcal{P}_\mathbb{T}\mathcal{F}_\Omega\mathcal{P}_\mathbb{T} - p\mathcal{P}_\mathbb{T}\mathcal{F}_\mathbb{I}\mathcal{P}_\mathbb{T}\|$, which is more simply analyzed through Theorem 1.7.2 alone. See Appendix 3.B.2 for details. □

Remark 3.2.2: Sample Complexity

We note that this result is given in terms of the Bernoulli sampling probability, rather than the more traditional number of samples with replacement seen in the matrix completion literature. To provide a more direct comparison, and remarking that $\mathbb{E}[|\Omega|] = m$

and $p = \frac{m}{L}$, we have that for a sufficiently large constant $C > 0$ that

$$m \geq C \frac{\nu^2 r^2}{\varepsilon_0^2} \beta n \log n,$$

gives ε_0-RIP of $\mathcal{M}_\Omega$. We again note that, due to using the weaker Assumption 3.2.1 instead of the incoherence assumption in [GTK24], this is optimal up to constant factors and equivalent to the RIP established in [GTK24].

Now that RIP is established, we can prove local convergence of DBRE. This theorem describes a high-probability guarantee that DBRE exhibits linear convergence in an attractive basin near the solution, provided that $\mathcal{M}_\Omega$ exhibits RIP.

Theorem 3.2.2 (Local Convergence of Algorithm 7)**.** *Let $\boldsymbol{X} \in \mathbb{R}^{n \times n}$ be the ground truth rank-r, ν-incoherent matrix and let $\mathbb{T}$ be the tangent space of $\mathcal{N}_r$ at $\boldsymbol{X}$. Suppose that $p \geq C \frac{\beta \log n}{n}$ for some sufficiently large constant $C > 0$ and any $\beta > 3$. Further assume that*

$$p^{-2} \|\mathcal{P}_\mathbb{T} \mathcal{M}_\Omega \mathcal{P}_\mathbb{T} - p^2 \mathcal{P}_\mathbb{T}\| \leq \varepsilon_0, \tag{3.6}$$

$$\|\mathcal{M}_\Omega\| \leq 50 p^{3/2} \sqrt{\beta n \log n}, \tag{3.7}$$

$$\|\mathcal{M}_\Omega \mathcal{P}_\mathbb{T}\| \leq 25 p^{3/2} \sqrt{\nu r \beta \log n}, \tag{3.8}$$

$$\|\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l}\| \leq 100 p^{3/2} \sqrt{\beta n \log n} \frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_\mathrm{F}}{\lambda_r(\boldsymbol{X})} + 25 p^{3/2} \sqrt{\nu r \beta \log n}, \tag{3.9}$$

$$\frac{\|\boldsymbol{X}_0 - \boldsymbol{X}\|_\mathrm{F}}{\lambda_r(\boldsymbol{X})} \leq \frac{\varepsilon_0 p^{1/2}}{50 \left(\beta n \log n\right)^{1/4}}, \tag{3.10}$$

where ε_0 is a constant satisfying

$$\delta = \frac{18 \varepsilon_0}{1 - 4\varepsilon_0} < 1.$$

Then DBRE converges linearly as the iterates satisfy

$$\|\boldsymbol{X}_l - \boldsymbol{X}\|_\mathrm{F} \leq \delta^l \|\boldsymbol{X}_0 - \boldsymbol{X}\|_\mathrm{F}.$$

Proof Sketch. We first note that each of the above assumptions, save for (3.10), holds with high probability for $p \geq C \frac{\nu^2 r^2}{\varepsilon_0^2} \frac{\beta \log n}{n}$, where $C > 0$ is an absolute constant. See Appendix 3.C.1 for

details.

The proof begins first by simple linear algebra, as we have

$$\begin{aligned}\|\boldsymbol{X}_{l+1} - \boldsymbol{X}\|_{\mathrm{F}} &= \|\boldsymbol{X}_{l+1} - \boldsymbol{W}_l - \boldsymbol{X} + \boldsymbol{W}_l\|_{\mathrm{F}} \\ &\leq \|\boldsymbol{X}_{l+1} - \boldsymbol{W}_l\|_{\mathrm{F}} + \|\boldsymbol{X} - \boldsymbol{W}_l\|_{\mathrm{F}} \\ &\leq 2\|\boldsymbol{W}_l - \boldsymbol{X}\|_{\mathrm{F}},\end{aligned}$$

where the last inequality follows from $\boldsymbol{X}_{l+1}$ being the best rank-r approximation to $\boldsymbol{W}_l$ by Eckart-Young-Mirsky [EY36]. Next, plugging in $\boldsymbol{W}_l = \boldsymbol{X}_l + \alpha_l \mathcal{P}_{\mathbb{T}_l} \boldsymbol{G}_l$, we see that

$$\begin{aligned}\|\boldsymbol{X}_{l+1} - \boldsymbol{X}\|_{\mathrm{F}} &\leq 2\left\|\boldsymbol{X}_l + \alpha_l \mathcal{P}_{\mathbb{T}_l} \boldsymbol{G}_l - \boldsymbol{X}\right\|_{\mathrm{F}} \\ &= 2\|\boldsymbol{X}_l - \boldsymbol{X} - \alpha_l \mathcal{P}_{\mathbb{T}_l} \mathcal{M}_\Omega(\boldsymbol{X}_l - \boldsymbol{X})\|_{\mathrm{F}} \\ &\leq \underbrace{2\|(\mathcal{P}_{\mathbb{T}_l} - \alpha_l \mathcal{P}_{\mathbb{T}_l} \mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l})(\boldsymbol{X}_l - \boldsymbol{X})\|_{\mathrm{F}}}_{I_1} \\ &\quad + \underbrace{2\|(\mathcal{I} - \mathcal{P}_{\mathbb{T}_l})(\boldsymbol{X}_l - \boldsymbol{X})\|_{\mathrm{F}}}_{I_2} \\ &\quad + \underbrace{2|\alpha_l| \|\mathcal{P}_{\mathbb{T}_l} \mathcal{M}_\Omega(\mathcal{I} - \mathcal{P}_{\mathbb{T}_l})(\boldsymbol{X}_l - \boldsymbol{X})\|_{\mathrm{F}}}_{I_3}.\end{aligned}$$

The remainder of the proof is in the bounding of I_1, I_2, and I_3. I_1 is proven by showing that in a neighborhood of the solution, defined by (3.10), a local form of RIP for $\mathcal{M}_\Omega$ holds if (3.6) is true. This proof leverages the assumptions made in (3.8) and (3.9). I_2 follows from the neighborhood assumption of (3.10) in tandem with Theorem 1.8.3, and I_3 follows from bounds on the step size (seen in Theorem 3.C.1), the assumption in (3.8), and Theorem 1.8.3. The assumptions in (3.6), (3.7), and (3.8) are all proven via high probability guarantees using Theorems 1.7.2 and 1.7.5. The technical details are deferred to the appendix, see Appendix 3.C.1. See Figure 3.1 for a diagram of the main dependencies for the convergence proof. □

3.2.2 Initialization Results

In this section, we outline our initialization guarantees for DBRE. Given that the convergence of this algorithm is only local, initialization is important to consider in the context of sample

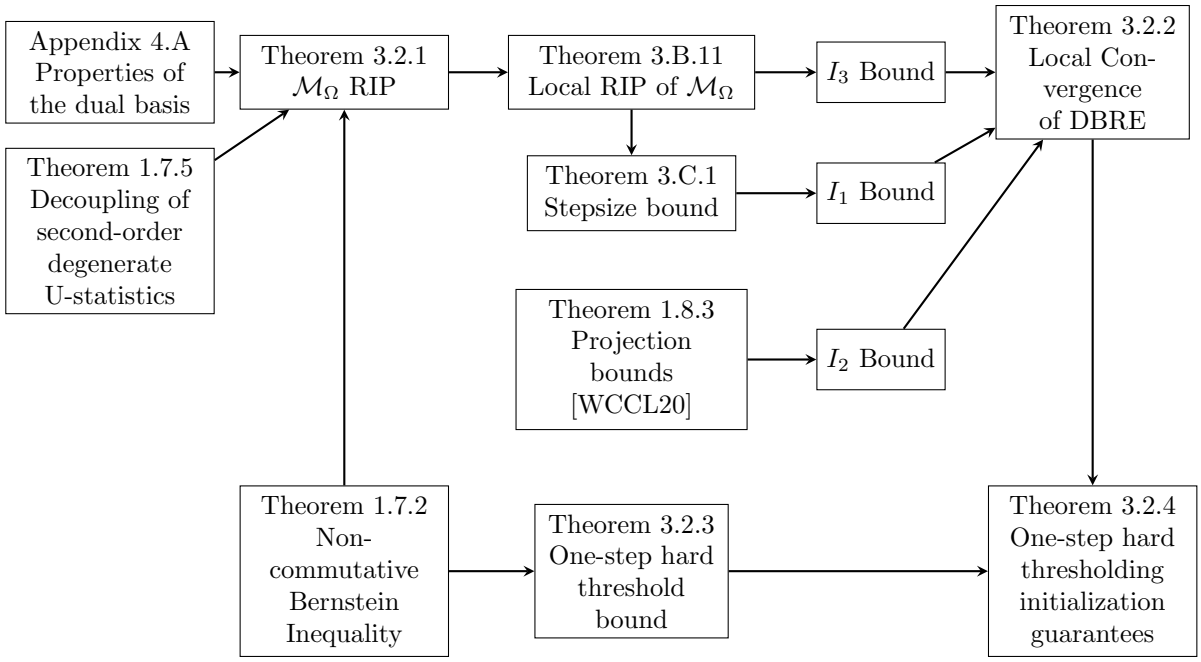


Figure 3.1: This diagram is a schematic of the overall proof of convergence for DBRE. Arrows indicate how results depend on one another, and how they link together to form the overall proof of convergence. Not every exact dependency is shown in this figure for legibility purposes, instead focusing on the key pieces of the overall flow of the argument.

complexity. The simplest initialization, a hard thresholding to $\mathcal{N}_r$ of the measured information, provides a reasonable starting point and is described in Algorithm 8, where $\mathcal{H}_r$ is as defined in (1.17). The following sections describe how close a one-step hard-thresholding initialization will be to the ground truth for DBRE. Following this, and in tandem with Theorem 3.2.2, we show recovery guarantees for DBRE.

Algorithm 8 1 Step Hard Thresholding Initialization

1: **Input:** Ω, $\{\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle\}_{\boldsymbol{\alpha} \in \Omega}$, $\operatorname{rank}(\boldsymbol{X}) = r$
2: $\mathcal{R}_\Omega(\boldsymbol{X}) = \sum_{\boldsymbol{\alpha} \in \Omega} \langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \boldsymbol{v}_{\boldsymbol{\alpha}}$
3: $\boldsymbol{X}_0 = \frac{1}{p} \mathcal{H}_r(\mathcal{R}_\Omega(\boldsymbol{X}))$
4: **Output:** $\boldsymbol{X}_0$: 1 Step Hard Thresholding Initialization.

Lemma 3.2.3. *Under a Bernoulli sampling parameter $p \geq \frac{128\beta \log n}{3n}$, then with probability at least $1 - 2n^{1-\beta}$ we have for $\boldsymbol{X}_0 = p^{-1}\mathcal{H}_r(\mathcal{R}_\Omega(\boldsymbol{X}))$ that*

$$\|\boldsymbol{X}_0 - \boldsymbol{X}\|_{\mathrm{F}} \leq \sqrt{2r}\|\boldsymbol{X}_0 - \boldsymbol{X}\| \leq \sqrt{\frac{2\beta n r \log n}{3p}} \max_{\boldsymbol{\alpha} \in \mathbb{I}} \langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \leq \sqrt{\frac{\beta \nu^2 r^3 \log(n)}{24pn}} \|\boldsymbol{X}\|.$$

Proof. See Appendix 3.D. □

Theorem 3.2.4 (Recovery Guarantee for Algorithm 7)**.** *For*

$$p \geq \max\left\{\frac{20\kappa\nu r^{3/2}}{\varepsilon_0}\frac{\beta^{3/4}\log^{3/4}(n)}{n^{1/4}}, C\frac{\nu^2 r^2}{\varepsilon_0^2}\frac{\beta\log n}{n}\right\},$$

where κ is the condition number of $\boldsymbol{X}$, $\beta > 3$, and $\varepsilon_0 < \frac{1}{22}$ for some sufficiently large constant $C > 0$, then with probability at least $1 - C'n^{3-\beta}$, where C' is an absolute constant, DBRE recovers the ground truth matrix $\boldsymbol{X}$ when initialized by Algorithm 8.

Proof. This result is a consequence of Theorem 3.2.3 and the local neighborhood assumption in (3.15). We can see this by increasing the sample complexity p to a sufficiently large value such that the initialization is smaller than the local neighborhood assumption. Furthermore, for a high enough sample complexity, all of the assumptions in Theorem 3.2.2 hold with high probability. The result then follows from the union bound. □

Remark 3.2.3: Resampling Initialization

We note here that a more delicate initialization through a resampling technique, such as the one in [WCCL20] or in Algorithm 5, could likely reduce the sample complexity from $p \gtrsim \frac{\log^{3/4} n}{n^{1/4}}$ to $p \gtrsim \frac{\log n}{n}$. Further investigation of initialization has been omitted from this work due to space constraints, but is an area of interest for future research.

Remark 3.2.4: Sample Complexity

For DBRE, we use a Bernoulli sampling model with parameter p, while other matrix completion methodologies use a uniform at random with replacement model. To provide a more direct sample complexity comparison, let $m = \mathbb{E}[|\Omega|]$ under a Bernoulli model. This implies that $p = \frac{m}{L}$. Theorem 3.2.4 therefore implies that, if

$$m \geq \max\left\{\frac{20\nu r^{3/2}\kappa\beta^{3/4}}{\varepsilon_0}n^{7/4}\log^{3/4}(n), C\frac{\nu^2 r^2\beta}{\varepsilon_0^2}n\log n\right\},$$

for some sufficiently large constant $C > 0$, DBRE recovers $\boldsymbol{X}$.

3.3 Robustness Guarantees

In many applications, the distance matrix may be corrupted, and understanding the sources of this corruption is central to designing robust recovery algorithms [KKM09, GQZ23, BLT+06, TD25, KTC25b, KTC25a]. Broadly, there are two main causes. First, even if distance measurements are perfectly accurate, the underlying point configuration may itself be perturbed due to physical factors. For instance, sensors placed in dynamic environments, such as the ocean, may drift over time. In such cases, the observed distances correspond to a perturbed version of the true point set. Second, the points themselves may be fixed, but the distance measurements are noisy. This can arise from various sources: sensor imprecision, environmental interference, or limited measurement resolution. In practice, both types of corruption may occur simultaneously. However, in this paper, we focus on the first scenario: perturbations in the point configuration. This assumption simplifies the analysis, since the resulting distance matrix remains a valid Euclidean distance matrix, and avoids challenges associated with arbitrary noise patterns that could violate geometric consistency. We believe that this setting is relevant to setting where environmental drift is more dominant than measurement noise. Moreover, the developed technical analysis for this setting could potentially serve as a foundation for future extensions to more general noise models.

In this section, we will provide robustness results for DBRE. To begin, we assume the following: for a given point matrix $\boldsymbol{P}$, we denote $\hat{\boldsymbol{P}} = \boldsymbol{P} + \boldsymbol{N}$, where $\boldsymbol{N} \in \mathbb{R}^{n\times r}$ is a random matrix. We denote $\hat{\boldsymbol{X}} = \hat{\boldsymbol{P}}\hat{\boldsymbol{P}}^\top$. To show robustness to noise, we first show that under noise bounded almost surely by $\|\boldsymbol{N}\|_\infty \leq C$ for some constant C that

$$\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_{\mathrm{F}} \leq C\mathcal{O}(\sqrt{n\|\boldsymbol{X}\|}) + C^2 nr,$$

in Theorem 3.E.1. This quantifies the amount that the drift noise can perturb the dataset by. Then, we show that for small enough noise the incoherence of the perturbed matrix is not drastically changed, and is bounded by an $\mathcal{O}(1)$ constant in Theorem 3.E.2. This result is important to demonstrate that the sample complexity required for recovery is not pushed out of a feasible range. With these results established, we can combine them together to show that

for bounded noise, given a sufficiently large Bernoulli sample complexity p, $\hat{\boldsymbol{X}}$ is recovered with DBRE with high probability and that the relative error between $\hat{\boldsymbol{X}}$ and $\boldsymbol{X}$ is small. This result is formally stated in the following theorem: To demonstrate robustness to noise, we first show that, under bounded noise satisfying $\|\boldsymbol{N}\|_\infty \leq C$ for some constant C, the following holds:

$$\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_{\mathrm{F}} \leq C\,\mathcal{O}\big(\sqrt{n\|\boldsymbol{X}\|}\big) + C^2 nr,$$

as established in Theorem 3.E.1. This quantifies the amount that the drift noise can perturb the dataset by. Then, we show that for small enough noise, the incoherence of the perturbed matrix is not drastically changed, and is bounded by an $\mathcal{O}(1)$ constant in Theorem 3.E.2. This result is important to demonstrate that the sample complexity required for recovery is not pushed out of a feasible range. With these results established, we can combine them together to show that for bounded noise, given a sufficiently large Bernoulli sample complexity p, $\hat{\boldsymbol{X}}$ is recovered with DBRE with high probability and that the relative error between $\hat{\boldsymbol{X}}$ and $\boldsymbol{X}$ is small. This result is formally stated in the following theorem:

Theorem 3.3.1 (Robustness Guarantee for Algorithm 7)**.** *Let $\boldsymbol{P} \in \mathbb{R}^{n\times r}$ and define $\hat{\boldsymbol{P}} = \boldsymbol{P} + \boldsymbol{N}$, where $\boldsymbol{N} \in \mathbb{R}^{n\times r}$ is a random matrix with $\mathbb{E}[N_{ij}] = 0$. Assume that the following quantity, which admits a signal-to-noise ratio (SNR) interpretation $\frac{\|\boldsymbol{X}\|^{1/2}}{\|\boldsymbol{N}\|_\infty}$, satisfies*

$$\frac{\|\boldsymbol{X}\|^{1/2}}{\|\boldsymbol{N}\|_\infty} \geq \frac{32\sqrt{2}\,\beta\kappa n \log n}{3\nu^{1/2}},$$

for some $\beta > \max\left\{3, \frac{3r}{8\log n}\right\}$, where κ denotes the condition number of $\boldsymbol{X}$.

Assume that the measured distances are of the form $\langle \hat{\boldsymbol{X}}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle$ and are sampled according to a Bernoulli model with parameter p satisfying

$$p \geq C\nu^2 r^2 \frac{\beta \log n}{n},$$

where $C > 3.92 \times 10^4$ is an absolute constant. Furthermore, suppose that DBRE is initialized at a point $\boldsymbol{X}_0$ satisfying the conditions of Theorem 3.2.2 with $\varepsilon_0 < \frac{1}{22}$.

Then, with probability at least $1 - C' n^{3-\beta}$, where C' is an absolute constant, DBRE recovers

$\hat{\boldsymbol{X}}$, *and*

$$\|\boldsymbol{X} - \hat{\boldsymbol{X}}\| \leq \left(\sqrt{\frac{\nu}{8n}} + \frac{9\nu r}{2048\kappa\beta^2 n \log^2 n}\right) \lambda_r(\boldsymbol{X}).$$

Proof of Theorem 3.3.1. This result follows first from Lemmas 3.E.1 and 3.E.2. From here, the sample complexity guarantee of Theorem 3.2.1, with sufficiently large constant C as stated above, coupled with the high probability guarantees of Assumptions (3.6)-(3.9) in Theorem 3.2.2, gives the desired result, provided an initialization satisfying (3.10). See Section 3.B for details on the guarantees of Assumptions (3.6)-(3.9). □

This result shows that recovery under noise depends intrinsically on the underlying geometry of the object. In particular, highly degenerate objects with large condition numbers can tolerate only low levels of noise, as captured by the ratio $\|\boldsymbol{X}\|^{1/2}/\|\boldsymbol{N}\|_\infty$ in Theorem 3.3.1. Moreover, as this ratio increases, the allowable perturbation in the incoherence parameter also increases, which in turn leads to a higher sample complexity required for recovery. While this dependence may initially seem counterintuitive, the interplay between this noise-to-signal ratio and the sample complexity shows that the most recoverable objects under noise are those that are both well-conditioned and highly incoherent, as a large $\lambda_r(\boldsymbol{X})$ facilitates stronger recovery. As a corollary to Theorem 3.3.1, we obtain the following relative error bound:

Corollary 3.3.2. *Under the assumptions of Theorem 3.3.1, the relative recovery error of DBRE satisfies, with probability at least* $1 - C'n^{3-\beta}$ *for some absolute constant* C'*,*

$$\frac{\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_{\mathrm{F}}}{\|\boldsymbol{X}\|_{\mathrm{F}}} \leq \sqrt{\frac{\nu}{4n}} + \frac{9\sqrt{2}\nu r}{2048\kappa\beta^2 n \log^2 n}.$$

Proof. This follows from Theorem 3.3.1 as

$$\frac{\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_{\mathrm{F}}}{\|\boldsymbol{X}\|_{\mathrm{F}}} \leq \frac{\sqrt{2r}\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|}{\sqrt{r}\lambda_r(\boldsymbol{X})} = \sqrt{2}\frac{\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|}{\lambda_r(\boldsymbol{X})}.$$

□

3.4 Numerical Results

All of the following experiments were conducted in MATLAB. The code used for the following experiments can be found in the GitHub repository at `https://github.com/chandlersmith2/NonConvexEDMC`.

3.4.1 Synthetic Data Experiment

In this section, we test the proposed algorithm on synthetic data. Various two and three dimensional datasets were used, and are referred to in Table 3.1 with their corresponding sizes. The goal of DBRE is to recover the full set of points $\boldsymbol{P}$ up to orthogonal transformation by sampling the entries above the diagonal of $\boldsymbol{D}$ uniformly with replacement, with a total of pL entries chosen for $p \in [0, 1]$. The algorithm reconstructs the Gram matrix $\boldsymbol{X}_{\text{rec}} = \boldsymbol{P}\boldsymbol{P}^\top$, from which $\boldsymbol{P}$ can be recovered using (1.3). The comparison referenced in Table 3.1 is the relative error between the recovered matrix $\boldsymbol{X}_{\text{rec}}$ and the ground truth matrix $\boldsymbol{X}$ in Frobenius norm. Each run was terminated at either 1000 iterations or when a relative Frobenius norm difference between iterates of 10^{-5} was achieved. This experiment was initialized using the one-step-hard-thresholding method described in Algorithm 8.

Table 3.1: Relative recovery error $\|\mathbf{X} - \mathbf{X}_{\text{rec}}\|_{\text{F}}/\|\mathbf{X}\|_{\text{F}}$ averaged over 25 trials.

Dataset \ p (%)	**10**	**7**	**5**	**3**	**2**	**1**
Sphere ($n = 1002$)	3.38×10^{-7}	4.61×10^{-7}	6.12×10^{-7}	1.48×10^{-6}	8.40×10^{-3}	6.81×10^{-1}
Cow ($n = 2601$)	4.41×10^{-7}	5.24×10^{-7}	6.04×10^{-7}	9.14×10^{-7}	2.47×10^{-4}	5.71×10^{-3}
Swiss Roll ($n = 2048$)	3.85×10^{-7}	4.70×10^{-7}	5.81×10^{-7}	9.47×10^{-7}	1.56×10^{-6}	6.40×10^{-2}

We note that the recovery completely fails for the sphere at 1% sampling, while recovery is partially successful for the other two datasets. This is because the other datasets are larger while maintaining the same rank, allowing for better scaling in the low sampling regime. In Figure 3.2, we show an image of the reconstruction of the figures described in Table 3.1.

3.4.2 Comparison to Existing Methods

We provide additional experiments to compare the efficacy of our algorithm DBRE to another provably convergent non-convex EDMC algorithm [GTK24], which we refer to as MatrixIRLS-

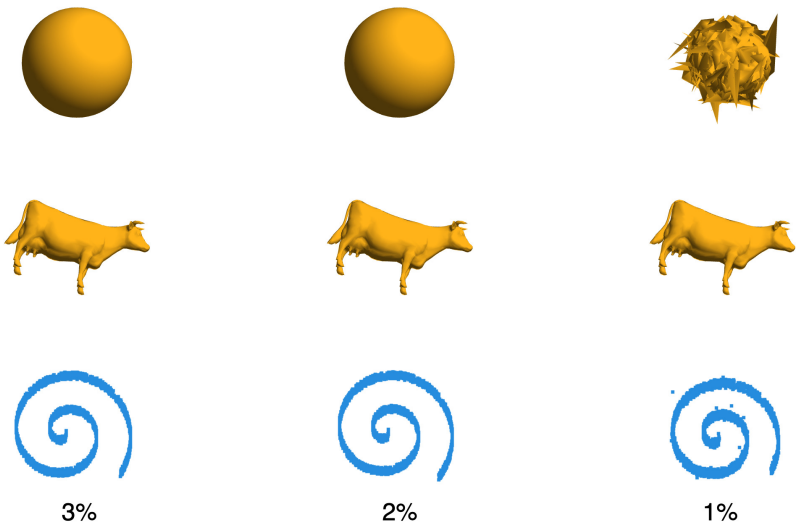


Figure 3.2: Reconstruction of the synthetic datasets referenced in Table 3.1. From left to right, the Bernoulli parameter is 0.03, 0.02, and 0.01.

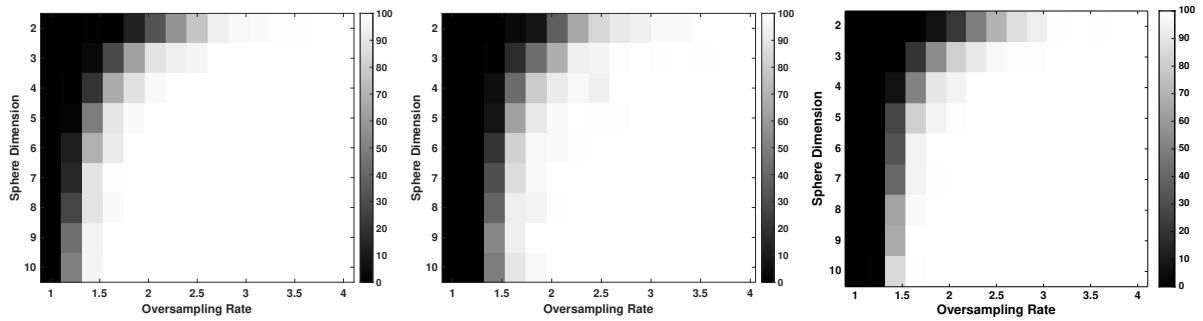
EDMC. For the first experiment, let $r \in [2, 10]$, and consider $n = 100$ points sampled from $\text{Unif}(S^{r-1})$, the uniform distribution on the sphere embedded in r dimensions. As the number of degrees of freedom in a rank-r $n \times n$ matrix is $nr - \frac{r(r-1)}{2}$, define the oversampling ratio ρ as

$$\rho = \frac{pL}{nr - \frac{r(r-1)}{2}},$$

as $\mathbb{E}[|\Omega|] = pL$ for Bernoulli random sampling with parameter p. In Figure 3.3, we compare the oversampling ratio versus the dimension of the sphere in a transition plot. Black indicates complete failure, classified as a relative Gram matrix error larger than 10^{-3}, and white indicates success. Each of these squares was run for 100 trials using DBRE, MatrixIRLS-EDMC, and the Augmented Lagrangian technique from [TL18]. DBRE was initialized with 10 iterations of the Augmented Lagrangian algorithm, and MatrixIRLS-EDMC was initialized using a least-squares methodology described in [GTK24].

As these experiments indicate, performance between the Augmented Lagrangian, MatrixIRLS for EDMC, and DBRE is comparable, with DBRE performing slightly better overall, but most noticeably in the higher rank regime.

Next, we compare the performance of DBRE, the Augmented Lagrangian, and the MatrixIRLS methods on the synthetic datasets seen in Figure 3.2. We initialize both DBRE and



(a) DBRE (b) MatrixIRLS-EDMC (c) Augmented Lagrangian

Figure 3.3: Oversampling ratio ρ versus dimension r for 100 points on the uniform distribution on S^{r-1}. Each parameter was tested 100 times.

the Augmented Lagrangian method with 20 iterations of the Augmented Lagrangian method, and we use the default settings for the MatrixIRLS algorithm. DBRE was run for 1000 iterations, or until a 10^{-8} relative difference between iterates was achieved.

Table 3.2: Relative recovery error $\|\boldsymbol{X} - \boldsymbol{X}_{\text{rec}}\|_{\text{F}} / \|\boldsymbol{X}\|_{\text{F}}$ between the recovered Gram matrix and the true Gram matrix averaged over 25 trials comparing DBRE, MatrixIRLS, and the Augmented Lagrangian approach.

Dataset \ p (%)	**10**	**7**	**5**	**3**	**2**	**1**
			DBRE			
Sphere	6.38×10^{-8}	5.72×10^{-8}	7.27×10^{-8}	1.58×10^{-7}	5.08×10^{-7}	6.47×10^{-1}
Cow	4.92×10^{-8}	5.67×10^{-8}	6.51×10^{-8}	9.93×10^{-8}	1.55×10^{-7}	6.28×10^{-7}
Swiss Roll	4.24×10^{-8}	4.88×10^{-8}	6.45×10^{-8}	1.01×10^{-7}	1.71×10^{-7}	2.29×10^{-6}
			Augmented Lagrangian			
Sphere	2.33×10^{-7}	2.13×10^{-7}	3.53×10^{-7}	9.29×10^{-7}	4.44×10^{-6}	1.05×10^{-1}
Cow	1.97×10^{-7}	3.03×10^{-7}	3.86×10^{-7}	5.69×10^{-7}	7.51×10^{-7}	2.30×10^{-7}
Swiss Roll	1.22×10^{-6}	3.27×10^{-6}	2.42×10^{-6}	4.16×10^{-6}	7.35×10^{-6}	2.45×10^{-6}
			MatrixIRLS			
Sphere	1.30×10^{-7}	1.33×10^{-7}	1.36×10^{-7}	3.81×10^{-6}	3.76×10^{-2}	6.09×10^{-1}
Cow	4.48×10^{-7}	5.33×10^{-7}	6.12×10^{-6}	3.75×10^{-3}	3.53×10^{-2}	1.12×10^{-1}
Swiss Roll	1.10×10^{-7}	1.23×10^{-7}	1.28×10^{-7}	5.14×10^{-7}	2.78×10^{-3}	4.78×10^{-1}

We note that for an initialization at the same point, both DBRE and the Augmented Lagrangian algorithm perform nearly identically in their recovery. MatrixIRLS is outperformed by both of the other algorithms in this experiment, but this is likely a result of comparing the default initialization to the Augmented Lagrangian initialization. Previous work [GTK24] shows similar performance between the Augmented Lagrangian and MatrixIRLS. Both Table 3.2 and

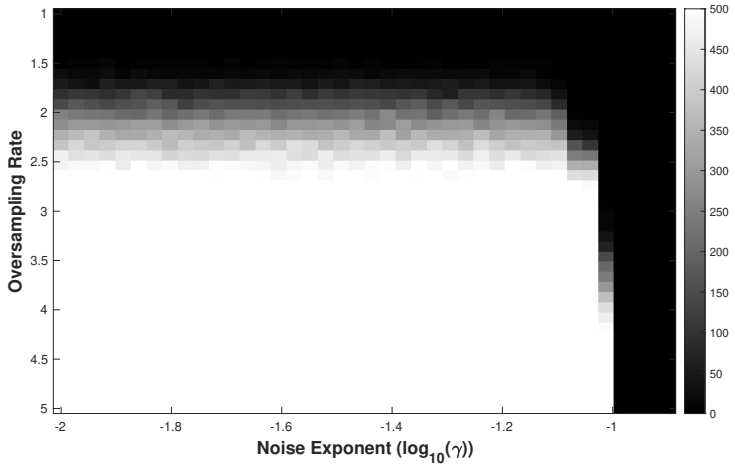


Figure 3.4: Oversampling ratio ρ versus noise level 10^{γ} for 100 points drawn i.i.d. from $\text{Unif}(S^2)$. Each parameter was tested 500 times.

Figure 3.3 both indicate the comparable to state-of-the-art performance of DBRE.

3.4.3 Experiments on Noisy Distance Measurements

Finally, we also ran an experiment with noise following the model in Section 3.3 using DBRE. Let $\{\boldsymbol{p}_i\}_{i=1}^{100} \sim \text{Unif}(S^2)$ be drawn i.i.d. and where $\boldsymbol{P} = [\boldsymbol{p}_1 \cdots \boldsymbol{p}_{100}]^\top \in \mathbb{R}^{100\times 3}$. We perturb $\boldsymbol{P}$ with a bounded, centered noise matrix $\boldsymbol{N} \in \mathbb{R}^{100\times 3}$ with $\|\boldsymbol{N}\|_\infty \leq 10^{\gamma}$ for $\gamma \in [-2, -1]$. Similar to the previous experiment, we set the oversampling ratio $\rho \in [1, 5]$. We set the success threshold at 10^{-2} relative difference, a relaxed value from previous experiments due to the addition of noise. Figure 3.4 shows the results over 500 trials.

Figure 3.4 indicates that recovery up to tolerance predominantly gets worse with added noise, although there is some clear dependence on the size of the noise impacting the reconstruction of the ground truth. This is most likely due to an increase in the incoherence of the dataset, requiring higher sample complexities to reconstruct. However, the noise level is still the dominant factor, and after a large enough noise value, reconstruction up to a certain tolerance is no longer viable.

We also present recovery results for the spherical point configuration shown in Figure 3.2 perturbed by Gaussian noise with varying standard deviations. For this experiment, the sampling rate p varies from 0.05 to 0.01, and the standard deviation of the Gaussian perturbations σ ranges from 10^{-6} to 1. The relative error between the Gram matrix recovered using DBRE

for 500 iterations, denoted $\hat{\boldsymbol{X}}_{\text{rec}}$, and the unperturbed Gram matrix $\boldsymbol{X}$ is presented in Table 3.3, averaged over 10 trials. For this experiment, DBRE was initialized with Algorithm 8.

Table 3.3: Relative recovery error $\left\|\boldsymbol{X}-\hat{\boldsymbol{X}}_{\text{rec}}\right\|_{\text{F}} / \|\boldsymbol{X}\|_{\text{F}}$ between the recovered perturbed Gram matrix and the true Gram matrix averaged over 10 trials using DBRE for the Sphere dataset referenced in Table 3.1 and Figure 3.2. σ is the standard deviation of the perturbing noise, and p is the sampling rate.

p \ σ	$\mathbf{10^0}$	$\mathbf{10^{-1}}$	$\mathbf{10^{-2}}$	$\mathbf{10^{-3}}$	$\mathbf{10^{-4}}$	$\mathbf{10^{-5}}$
0.05	9.68×10^{-1}	2.39×10^{-1}	2.45×10^{-3}	2.46×10^{-3}	2.44×10^{-4}	2.42×10^{-5}
0.04	9.67×10^{-1}	2.40×10^{-1}	2.45×10^{-2}	2.49×10^{-3}	2.43×10^{-4}	2.46×10^{-5}
0.03	9.68×10^{-1}	2.39×10^{-1}	2.44×10^{-2}	2.53×10^{-3}	1.03×10^{-2}	2.44×10^{-5}
0.02	9.70×10^{-1}	3.02×10^{-1}	6.21×10^{-2}	1.62×10^{-2}	4.02×10^{-2}	3.83×10^{-2}
0.01	1.08×10^{0}	1.35×10^{0}	1.39×10^{0}	1.34×10^{0}	1.36×10^{0}	1.38×10^{0}

The experimental results in Table 3.3 also indicate that the magnitude of the drift noise is the dominant factor in the reconstructability of the point cloud. While there are sample complexity effects, which are noticeable for $\gamma = 0.03$, the dominant effect is the noise level itself, further confirming the conclusion drawn from Figure 3.4.

3.5 Conclusion

In this chapter, we proposed a novel Riemannian gradient descent approach for solving the EDMC problem using a matrix completion approach on the manifold of rank-r matrices with the DBRE algorithm. In a local neighborhood, we proved that DBRE exhibits linear convergence with high probability. To the authors' knowledge, this is the first work to provide both initialization guarantees and robustness guarantees for a non-convex approach to the EDMC problem. The convergence analysis of DBRE was predicated on a statistical understanding of coupled terms in a random operator, and required novel analysis to the matrix completion literature. For our method, we provided numerical results to underline its efficacy, and DBRE performs comparably to other state-of-the-art non-convex methods. Additionally, we provided robustness analysis and corresponding convergence guarantees. Finally, we provided a novel interpretation of incoherence in the EDMC setting, highlighting potential areas for development of non-uniform sampling methods in this field. This is a primary avenue of future interest,

as improving the sample complexity through geometrically-optimal sampling schemes would represent a noteworthy development in the EDMC literature.

3.A Further Properties of the Dual Bases

As in Section 2.A, we will first provide some useful properties of the dual bases $\{\boldsymbol{w}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$ and $\{\boldsymbol{v}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$ and associated objects.

We begin by defining the following matrix $\boldsymbol{H}^{-1}_{\text{offdiag}} \in \mathbb{R}^{L\times L}$:

Definition 3.A.1: Definition of $\boldsymbol{H}^{-1}_{\text{offdiag}}$

Let $\boldsymbol{H}^{-1} \in \mathbb{R}^{L\times L}$ be as defined previously with $(\boldsymbol{H}^{-1})_{\boldsymbol{\alpha\beta}} = H^{\boldsymbol{\alpha\beta}} = \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle$. We define the following matrix $\boldsymbol{H}^{-1}_{\text{offdiag}}$ as follows:

$$(H^{-1}_{\text{offdiag}})_{\boldsymbol{\alpha\beta}} = \begin{cases} H^{\alpha\beta} & \boldsymbol{\alpha} \neq \boldsymbol{\beta}; \\ 0 & \boldsymbol{\alpha} = \boldsymbol{\beta}. \end{cases} \tag{3.11}$$

Lemma 3.A.1. *For $\boldsymbol{H}^{-1}_{\text{offdiag}}$ defined as in* (3.11)*, we have that*

$$\|\boldsymbol{H}^{-1}_{\text{offdiag}}\| \leq 2,$$

and that

$$\|\boldsymbol{H}^{-1}_{\text{offdiag}}\|^2_{2\to\infty} \leq \frac{2}{n},$$

where $\|\boldsymbol{A}\|^2_{2\to\infty}$ is the maximum Euclidean row norm of the matrix $\boldsymbol{A}$.

Proof. This statement follows directly from a Gershgorin circle argument, as

$$\begin{aligned} \|\boldsymbol{H}^{-1}_{\text{offdiag}}\| &\leq \max_{\boldsymbol{\alpha}} \sum_{\boldsymbol{\beta}\neq\boldsymbol{\alpha}} |H^{\boldsymbol{\alpha\beta}}| \\ &\leq \max_{\boldsymbol{\alpha}} \left((2n-3)\left(\frac{1}{2n} + \frac{1}{n^2}\right) + \frac{L}{n^2}\right) \\ &\leq 2, \end{aligned}$$

where the values for $H^{\boldsymbol{\alpha}\boldsymbol{\beta}}$ come from Theorem 2.A.3. The second statement follows from the fact that $\|\boldsymbol{H}^{-1}_{\text{offdiag}}\|_{2\to\infty}$ is equal to the maximum ℓ_2 norm of the rows of $\boldsymbol{H}^{-1}_{\text{offdiag}}$. As each row has the same entries, just in a different ordering, for any $\boldsymbol{\alpha} \in \mathbb{I}$ we have that

$$\begin{aligned}\|\boldsymbol{H}^{-1}_{\text{offdiag}}\|^2_{2\to\infty} &= \sum_{\boldsymbol{\beta}\neq\boldsymbol{\alpha}} (H^{\boldsymbol{\alpha}\boldsymbol{\beta}})^2 \\ &\leq (2n-3)\left(\frac{1}{2n}+\frac{1}{n^2}\right)^2 + \left(\frac{L}{n^2}\right)^2 \\ &\leq \frac{2}{n}.\end{aligned}$$

This concludes the proof. □

3.B Restricted Isometry Results

As RIP and its variants are critical to the analysis of DBRE in this paper, this section is dedicated to the proofs of RIP and similar results.

3.B.1 Supporting Concentration Results

The following results are various concentration identities used in the computation of $\|\mathcal{P}_{\mathbb{T}}\mathcal{M}_{\Omega}\mathcal{P}_{\mathbb{T}} - p^2\mathcal{P}_{\mathbb{T}}\|$, $\|\mathcal{P}_{\mathbb{T}}\mathcal{M}_{\Omega}\|$, and $\|\mathcal{M}_{\Omega}\|$.

Lemma 3.B.1 (Expectation of $\mathcal{M}_{\Omega}$)**.** *Let $\mathcal{M}_{\Omega}$ be as defined in* (3.1)*, and let $\mathcal{R}^*_{\Omega}\mathcal{R}_{\Omega}(\cdot) = \sum_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\Omega} \boldsymbol{w}_{\boldsymbol{\alpha}}\langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle \cdot, \boldsymbol{w}_{\boldsymbol{\beta}}\rangle$. Then $\mathbb{E}\left[\mathcal{R}^*_{\Omega}\mathcal{R}_{\Omega}\right] \neq p^2\mathcal{I}$, and $\mathbb{E}\left[\mathcal{M}_{\Omega}\right] = p^2\mathcal{I}$.*

Proof. First, notice that

$$\begin{aligned}\mathcal{R}^*_{\Omega}\mathcal{R}_{\Omega}(\cdot) &= \sum_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\Omega} \boldsymbol{w}_{\boldsymbol{\alpha}}\langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle \cdot, \boldsymbol{w}_{\boldsymbol{\beta}}\rangle \\ &= \sum_{\boldsymbol{\alpha}\in\Omega} \langle \cdot, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\alpha}}\rangle \boldsymbol{w}_{\boldsymbol{\alpha}} + \sum_{\boldsymbol{\alpha}\in\Omega} \boldsymbol{w}_{\boldsymbol{\alpha}} \sum_{\boldsymbol{\beta}\in\Omega,\boldsymbol{\beta}\neq\boldsymbol{\alpha}} \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle \cdot, \boldsymbol{w}_{\boldsymbol{\beta}}\rangle.\end{aligned}$$

This was decomposed suggestively into diagonal and off-diagonal elements of the matrix $\boldsymbol{H}^{-1}$ for the following reason. Previously, it was believed that $\mathbb{E}[\mathcal{R}^*_{\Omega}\mathcal{R}_{\Omega}] = p^2\mathcal{I}$, as it is true that $\mathcal{R}^*_{\mathbb{I}}\mathcal{R}_{\mathbb{I}} = \mathcal{I}^2 = \mathcal{I}$[SLCT23]. Let us consider the problem of computing the expectation of $\mathcal{R}^*_{\Omega}\mathcal{R}_{\Omega}$

in more detail now. We will assume that each entry of $\mathbb{I}$ is sampled with Bernoulli probability p, contrasted to the uniformly at random with replacement sampling strategy employed in [SLCT23]. Now we see that

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}_\Omega^*\mathcal{R}_\Omega(\cdot)\right] &= \mathbb{E}\left[\sum_{\boldsymbol{\alpha}\in\Omega}\boldsymbol{w}_{\boldsymbol{\alpha}}\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle\cdot,\boldsymbol{w}_{\boldsymbol{\beta}}\rangle\right] + \mathbb{E}\left[\sum_{\boldsymbol{\alpha}\in\Omega}\boldsymbol{w}_{\boldsymbol{\alpha}}\sum_{\boldsymbol{\beta}\in\Omega,\boldsymbol{\beta}\neq\boldsymbol{\alpha}}\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle\cdot,\boldsymbol{w}_{\boldsymbol{\beta}}\rangle\right] \\ &= p\sum_{\boldsymbol{\alpha}\in\mathbb{I}}\langle\cdot,\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\alpha}}\rangle\boldsymbol{w}_{\boldsymbol{\alpha}} + p^2\sum_{\boldsymbol{\alpha}\in\mathbb{I}}\boldsymbol{w}_{\boldsymbol{\alpha}}\sum_{\boldsymbol{\beta}\in\mathbb{I},\boldsymbol{\beta}\neq\boldsymbol{\alpha}}\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle\cdot,\boldsymbol{w}_{\boldsymbol{\beta}}\rangle.\end{aligned}$$

The reason for this difference is that the probability of each entry being selected in the diagonal terms is given by p, as per the Bernoulli definition. However, the probability of the off-diagonal elements requires looking at the probability of sampling two distinct entries at a time. This probability, for each of these options, is p^2. If we let $p = \frac{m}{L}$, where $\mathbb{E}|\Omega| = m$, p^2 is the exact scaling originally expected in [SLCT23] for the expectation of the operator to hold. This is incorrect however, and does not recognize that the diagonal entries are more likely to be sampled, as they only require the contributions of a single sample. As such, introducing a rescaling to de-bias the above operator leads to the definition of

$$\mathcal{M}_\Omega(\cdot) = \sum_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\Omega} C_{\boldsymbol{\alpha}\boldsymbol{\beta}}\boldsymbol{w}_{\boldsymbol{\alpha}}\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle\cdot,\boldsymbol{w}_{\boldsymbol{\beta}}\rangle,$$

where

$$C_{\boldsymbol{\alpha}\boldsymbol{\beta}} = \begin{cases} 1 & \boldsymbol{\alpha}\neq\boldsymbol{\beta}; \\ p & \boldsymbol{\alpha}=\boldsymbol{\beta}. \end{cases}$$

We can decompose $\mathcal{M}_\Omega$ into diagonal and off-diagonal elements again and see that

$$\mathcal{M}_\Omega(\cdot) = p\sum_{\boldsymbol{\alpha}\in\Omega}\langle\cdot,\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\alpha}}\rangle\boldsymbol{w}_{\boldsymbol{\alpha}} + \sum_{\boldsymbol{\alpha}\in\Omega}\boldsymbol{w}_{\boldsymbol{\alpha}}\sum_{\boldsymbol{\beta}\in\Omega,\boldsymbol{\beta}\neq\boldsymbol{\alpha}}\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle\cdot,\boldsymbol{w}_{\boldsymbol{\beta}}\rangle.$$

It now follows that

$$\mathbb{E}\left[\mathcal{M}_\Omega(\cdot)\right] = p\mathbb{E}\left[\sum_{\boldsymbol{\alpha}\in\Omega}\langle\cdot,\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\alpha}}\rangle\boldsymbol{w}_{\boldsymbol{\alpha}}\right] + \mathbb{E}\left[\sum_{\boldsymbol{\alpha}\in\Omega}\boldsymbol{w}_{\boldsymbol{\alpha}}\sum_{\boldsymbol{\beta}\in\Omega,\boldsymbol{\beta}\neq\boldsymbol{\alpha}}\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle\cdot,\boldsymbol{w}_{\boldsymbol{\beta}}\rangle\right]$$

$$
\begin{aligned}
&= p^2 \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \cdot, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\alpha}}\rangle \boldsymbol{w}_{\boldsymbol{\alpha}} + p^2 \sum_{\boldsymbol{\alpha}\in\Omega} \boldsymbol{w}_{\boldsymbol{\alpha}} \sum_{\boldsymbol{\beta}\in\Omega, \boldsymbol{\beta}\neq\boldsymbol{\alpha}} \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle \langle \cdot, \boldsymbol{w}_{\boldsymbol{\beta}}\rangle \\
&= p^2 \sum_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{I}} \boldsymbol{w}_{\boldsymbol{\alpha}} \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle \langle \cdot, \boldsymbol{w}_{\boldsymbol{\beta}}\rangle \\
&= p^2 \mathcal{R}_{\mathbb{I}}^* \mathcal{R}_{\mathbb{I}} \\
&= p^2 \mathcal{I},
\end{aligned}
$$

as $\mathcal{R}_{\mathbb{I}} = \mathcal{I}$, thus concluding the proof. □

Lemma 3.B.2. *Let $\hat{\mathcal{F}}_\Omega$ be either $\mathcal{P}_{\mathbb{T}}\mathcal{F}_\Omega\mathcal{P}_{\mathbb{T}}$ or $\mathcal{F}_\Omega$, and let $\hat{\mathcal{F}}_{\mathbb{I}}$ be $\mathcal{P}_{\mathbb{T}}\mathcal{F}_{\mathbb{I}}\mathcal{P}_{\mathbb{T}}$ or $\mathcal{F}_{\mathbb{I}}$, respectively. Let $\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}$ be either $\mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}$ or $\boldsymbol{w}_{\boldsymbol{\alpha}}$, respectively. Let $\hat{\boldsymbol{H}} = [\langle \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}, \hat{\boldsymbol{w}}_{\boldsymbol{\beta}}\rangle] \in \mathbb{R}^{L\times L}$ be the corresponding correlation matrix. For a ground truth rank r ν-incoherent matrix $\boldsymbol{X}$ with tangent space $\mathbb{T}$ on $\mathcal{N}_r$, we have that for any $\beta > 1$, and with probability at least $1 - 2n^{1-\beta}$, and for $p \geq \frac{8\beta \log n}{3n}$, that*

$$
\|\hat{\mathcal{F}}_\Omega - p\hat{\mathcal{F}}_{\mathbb{I}}\| \leq p\sqrt{\frac{8\beta \left(\max_{\boldsymbol{\alpha}} \|\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2\right) \lambda_{\max}(\hat{\boldsymbol{H}}) \log n}{3p}}.
$$

Proof. This proof will follow a standard Bernstein argument. In order to make this argument, we need to build out $\hat{\mathcal{F}}_\Omega$ as a sum of random operators, we need to bound each, and then to bound the variance term. First, let $\{\xi_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$ be i.i.d. Bernoulli random variables that are 1 with probability p and 0 with probability $1 - p$. It follows then that

$$
\hat{\mathcal{F}}_\Omega(\cdot) = \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \xi_{\boldsymbol{\alpha}} \langle \cdot, \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\rangle \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}} \qquad \text{and} \qquad \mathbb{E}[\hat{\mathcal{F}}_\Omega] = p \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \cdot, \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\rangle \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}} = p\hat{\mathcal{F}}_{\mathbb{I}}.
$$

Now, to prove concentration of the desired sum, let

$$
\boldsymbol{S}_{\boldsymbol{\alpha}} = (\xi_{\boldsymbol{\alpha}} - p) \langle \cdot, \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\rangle \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}},
$$

and notice that $\sum_{\boldsymbol{\alpha}\in\mathbb{I}} \boldsymbol{S}_{\boldsymbol{\alpha}} = \hat{\mathcal{F}}_\Omega - p\hat{\mathcal{F}}_{\mathbb{I}}$. We can now use a Bernstein inequality to bound the deviation of the spectral norm from 0. Now, first notice that

$$
\begin{aligned}
\|\boldsymbol{S}_{\boldsymbol{\alpha}}\| &= \|(\xi_{\boldsymbol{\alpha}} - p) \langle \cdot, \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\rangle \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\| \\
&\leq \|\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2
\end{aligned}
$$

$$\leq \max_{\boldsymbol{\alpha}} \|\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 =: c,$$

where the last inequality follows from Assumption 3.2.1. Next, we seek to bound the variance term, $\sigma^2 = \left\|\sum_{\boldsymbol{\alpha}\in\mathbb{I}} \mathbb{E}\left[\boldsymbol{S}_{\boldsymbol{\alpha}}^2\right]\right\|$. To see this, first notice that

$$\begin{aligned}
\left\|\mathbb{E}\left[\sum_{\boldsymbol{\alpha}\in\mathbb{I}} \boldsymbol{S}_{\boldsymbol{\alpha}}^2\right]\right\| &= \left\|\mathbb{E}\left[\sum_{\boldsymbol{\alpha}\in\mathbb{I}} (\xi_{\boldsymbol{\alpha}} - p)^2 \langle\cdot, \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\rangle\langle\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}, \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\rangle\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\right]\right\| \\
&= \left\|\mathbb{E}\left[\sum_{\boldsymbol{\alpha}\in\mathbb{I}} (\xi_{\boldsymbol{\alpha}} - 2\xi_{\boldsymbol{\alpha}} p + p^2)\langle\cdot, \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\rangle\langle\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}, \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\rangle\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\right]\right\| \\
&\leq p(1-p)\left(\max_{\boldsymbol{\alpha}} \|\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2\right)\left\|\sum_{\boldsymbol{\alpha}\in\mathbb{I}} \langle\cdot, \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\rangle\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\right\| \\
&\leq p\left(\max_{\boldsymbol{\alpha}} \|\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2\right)\lambda_{\max}(\hat{\boldsymbol{H}}),
\end{aligned}$$

where the final inequality follows from Theorem 1.8.2. Now, for $t = \sqrt{\frac{8}{3}\sigma^2\beta\log n}$, we have that $t \leq \frac{\sigma^2}{c} = \lambda_{\max}(\hat{\boldsymbol{H}})p$ if $p \geq \frac{8\beta\log n}{3n}$, allowing us to use the simplified form of Theorem 1.7.2. This yields

$$\mathbb{P}\left(\left\|\sum_{\boldsymbol{\alpha}\in\mathbb{I}} \boldsymbol{S}_{\boldsymbol{\alpha}}\right\| \geq p\sqrt{\frac{8\left(\max_{\boldsymbol{\alpha}} \|\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2\right)\lambda_{\max}(\hat{\boldsymbol{H}})\beta\log n}{3p}}\right) \leq 2n\exp\left(-\beta\log n\right) = 2n^{1-\beta},$$

as stated previously. □

Lemma 3.B.3. *For a ground truth rank-r, ν-incoherent matrix $\boldsymbol{X}$ with tangent space $\mathbb{T}$ on $\mathcal{N}_r$, we have that for any $\beta > 1$ that if $p \geq \frac{4}{3}\frac{\beta\log n}{n}$, with probability at least $1 - 2n^{1-\beta}$*

$$\|\mathcal{F}_\Omega\mathcal{P}_{\mathbb{T}} - p\mathcal{F}_{\mathbb{I}}\mathcal{P}_{\mathbb{T}}\| \leq \sqrt{\frac{128p\nu r\beta\log n}{3}}.$$

Proof. We will prove this result using Theorem 1.7.2. First, notice that

$$\mathcal{F}_\Omega\mathcal{P}_{\mathbb{T}}(\cdot) - p\mathcal{F}_{\mathbb{I}}\mathcal{P}_{\mathbb{T}}(\cdot) = \sum_{\boldsymbol{\alpha}\in\mathbb{I}} (\xi_{\boldsymbol{\alpha}} - p)\langle\cdot, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\boldsymbol{w}_{\boldsymbol{\alpha}},$$

so this is a sum of zero mean independent random variables and Bernstein's inequality holds.

We define $\boldsymbol{J}_{\boldsymbol{\alpha}} = (\xi_{\boldsymbol{\alpha}} - p)\langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \boldsymbol{w}_{\boldsymbol{\alpha}}$. Next, notice that

$$\begin{aligned} |\boldsymbol{J}_{\boldsymbol{\alpha}}| &= |(\xi_{\boldsymbol{\alpha}} - p)\langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \boldsymbol{w}_{\boldsymbol{\alpha}}| \\ &\leq |\xi_{\boldsymbol{\alpha}} \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \boldsymbol{w}_{\boldsymbol{\alpha}}| \\ &\leq \|\mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \|\boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \\ &\leq \sqrt{\frac{2\nu r}{n}}, \end{aligned}$$

where the first inequality follows from dropping the negative term, and the third inequality follows from Assumption 3.2.1 and $\|\boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} = 2$. Next, we note that

$$\begin{aligned} \mathbb{E}\left[\sum_{\boldsymbol{\alpha} \in \mathbb{I}} \boldsymbol{J}_{\boldsymbol{\alpha}} \boldsymbol{J}_{\boldsymbol{\alpha}}^*\right] &= \mathbb{E}\left[\sum_{\boldsymbol{\alpha} \in \mathbb{I}} (\xi_{\boldsymbol{\alpha}} - p)^2 \langle \cdot, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \langle \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}}, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \boldsymbol{w}_{\boldsymbol{\alpha}}\right] \\ &= \mathbb{E}\left[\sum_{\boldsymbol{\alpha} \in \mathbb{I}} (\xi_{\boldsymbol{\alpha}}(1-2p) + p^2) \langle \cdot, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \langle \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}}, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \boldsymbol{w}_{\boldsymbol{\alpha}}\right] \\ &= \sum_{\boldsymbol{\alpha} \in \mathbb{I}} p(1-p) \langle \cdot, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \langle \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}}, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \boldsymbol{w}_{\boldsymbol{\alpha}}, \end{aligned}$$

so

$$\begin{aligned} \left\| \mathbb{E}\left[\sum_{\boldsymbol{\alpha} \in \mathbb{I}} \boldsymbol{J}_{\boldsymbol{\alpha}} \boldsymbol{J}_{\boldsymbol{\alpha}}^*\right] \right\| &= \left\| \sum_{\boldsymbol{\alpha} \in \mathbb{I}} p(1-p) \langle \cdot, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \langle \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}}, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \boldsymbol{w}_{\boldsymbol{\alpha}} \right\| \\ &\leq p \frac{\nu r}{2n} \left\| \sum_{\boldsymbol{\alpha} \in \mathbb{I}} \langle \cdot, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \boldsymbol{w}_{\boldsymbol{\alpha}} \right\| \\ &\leq p \frac{\nu r}{2n} \lambda_{\max}(\boldsymbol{H}) \\ &= p\nu r =: \sigma_1^2, \end{aligned}$$

where the first inequality follows from Assumption 3.2.1, the second inequality comes from Theorem 1.8.2, and the final line comes from Theorem 2.A.3. Next, notice that

$$\begin{aligned} \mathbb{E}\left[\sum_{\boldsymbol{\alpha} \in \mathbb{I}} \boldsymbol{J}_{\boldsymbol{\alpha}}^* \boldsymbol{J}_{\boldsymbol{\alpha}}\right] &= \mathbb{E}\left[\sum_{\boldsymbol{\alpha} \in \mathbb{I}} (\xi_{\boldsymbol{\alpha}} - p)^2 \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \langle \boldsymbol{w}_{\boldsymbol{\alpha}}, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}}\right] \\ &= \mathbb{E}\left[\sum_{\boldsymbol{\alpha} \in \mathbb{I}} (\xi_{\boldsymbol{\alpha}}(1-2p) + p^2) \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \langle \boldsymbol{w}_{\boldsymbol{\alpha}}, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}}\right] \end{aligned}$$

$$= \sum_{\boldsymbol{\alpha} \in \mathbb{I}} p(1-p) \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \langle \boldsymbol{w}_{\boldsymbol{\alpha}}, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}},$$

so it follows that

$$\begin{aligned} \left\| \mathbb{E}\left[\sum_{\boldsymbol{\alpha} \in \mathbb{I}} \boldsymbol{J}_{\boldsymbol{\alpha}}^* \boldsymbol{J}_{\boldsymbol{\alpha}} \right] \right\| &= \left\| \sum_{\boldsymbol{\alpha} \in \mathbb{I}} p(1-p) \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \langle \boldsymbol{w}_{\boldsymbol{\alpha}}, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \right\| \\ &\leq 4p \left\| \sum_{\boldsymbol{\alpha} \in \mathbb{I}} \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \right\| \\ &\leq 16 p \nu r =: \sigma_2^2, \end{aligned}$$

where the first inequality comes from $\|\boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 = 4$, the second inequality comes from Theorem 1.8.2, and the final line comes from $\lambda_{\max}(\tilde{\boldsymbol{H}}) \leq \nu r$ in Theorem 2.A.1. Taking $\sigma^2 = \max\{\sigma_1^2, \sigma_2^2\} = 16 p \nu r$, we get that, if $p \geq \frac{4}{3} \frac{\beta \log n}{n}$, then for $t = \sqrt{\frac{128 p \nu r \beta \log n}{3}}$ we have that

$$\begin{aligned} \mathbb{P}\left[\|\mathcal{F}_{\Omega} \mathcal{P}_{\mathbb{T}} - p \mathcal{F}_{\mathbb{I}} \mathcal{P}_{\mathbb{T}}\| \geq \sqrt{\frac{128 p \nu r \beta \log n}{3}} \right] &\leq 2n \exp\left(\frac{-3}{8(16 p \nu r)} \frac{128 p \nu r \beta \log n}{3} \right) \\ &= 2n^{1-\beta}, \end{aligned}$$

following from the simplified form of Theorem 1.7.2. This concludes the proof. □

Lemma 3.B.4. *Let $\{\boldsymbol{A}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathbb{I}}, \{\boldsymbol{B}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathbb{I}} \subset \mathbb{R}^{n \times n}$ be any two collections of L matrices. Then*

$$\begin{aligned} \left\| \sum_{\boldsymbol{\alpha} \neq \boldsymbol{\beta}} (\xi_{\boldsymbol{\alpha}} \xi_{\boldsymbol{\beta}} - p^2) \boldsymbol{A}_{\boldsymbol{\alpha}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \boldsymbol{B}_{\boldsymbol{\beta}} \rangle \right\| \leq & \left\| \sum_{\boldsymbol{\alpha} \neq \boldsymbol{\beta}} (\xi_{\boldsymbol{\alpha}} - p)(\xi_{\boldsymbol{\beta}} - p) H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \boldsymbol{A}_{\boldsymbol{\alpha}} \langle \cdot, \boldsymbol{B}_{\boldsymbol{\beta}} \rangle \right\| \\ & + \left\| \sum_{\boldsymbol{\alpha} \neq \boldsymbol{\beta}} (\xi_{\boldsymbol{\alpha}} - p) p \boldsymbol{A}_{\boldsymbol{\alpha}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \boldsymbol{B}_{\boldsymbol{\beta}} \rangle \right\| \\ & + \left\| \sum_{\boldsymbol{\alpha} \neq \boldsymbol{\beta}} (\xi_{\boldsymbol{\beta}} - p) p \boldsymbol{A}_{\boldsymbol{\alpha}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \boldsymbol{B}_{\boldsymbol{\beta}} \rangle \right\|. \end{aligned}$$

Proof. This result follows from the fact that

$$\xi_{\boldsymbol{\alpha}} \xi_{\boldsymbol{\beta}} - p^2 = (\xi_{\boldsymbol{\alpha}} - p)(\xi_{\boldsymbol{\beta}} - p) + p(\xi_{\boldsymbol{\alpha}} - p) + p(\xi_{\boldsymbol{\beta}} - p).$$

As such, the expression becomes

$$\begin{aligned}
&\left\| \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} (\xi_{\boldsymbol{\alpha}}\xi_{\boldsymbol{\beta}} - p^2)\boldsymbol{A}_{\boldsymbol{\alpha}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \boldsymbol{B}_{\boldsymbol{\beta}} \rangle \right\| \\
&\qquad = \left\| \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} \left((\xi_{\boldsymbol{\alpha}} - p)(\xi_{\boldsymbol{\beta}} - p) + p(\xi_{\boldsymbol{\alpha}} - p) + p(\xi_{\boldsymbol{\beta}} - p)\right) \boldsymbol{A}_{\boldsymbol{\alpha}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \boldsymbol{B}_{\boldsymbol{\beta}} \rangle \right\| \\
&\qquad = \Bigg\| \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} (\xi_{\boldsymbol{\alpha}} - p)(\xi_{\boldsymbol{\beta}} - p)\boldsymbol{A}_{\boldsymbol{\alpha}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \boldsymbol{B}_{\boldsymbol{\beta}} \rangle + \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} p(\xi_{\boldsymbol{\alpha}} - p)\boldsymbol{A}_{\boldsymbol{\alpha}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \boldsymbol{B}_{\boldsymbol{\beta}} \rangle \\
&\qquad\qquad + \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} p(\xi_{\boldsymbol{\beta}} - p)\boldsymbol{A}_{\boldsymbol{\alpha}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \boldsymbol{B}_{\boldsymbol{\beta}} \rangle \Bigg\|.
\end{aligned}$$

The result then follows from the triangle inequality. □

Lemma 3.B.5. *Let $\boldsymbol{X}$ be a rank-r, ν-incoherent ground truth matrix with tangent space $\mathbb{T}$. Let $\boldsymbol{V}_{\boldsymbol{\alpha}} = vec(\mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}})$ or $vec(\boldsymbol{w}_{\boldsymbol{\alpha}})$ for all $\boldsymbol{\alpha} \in \mathbb{I}$, and let $\boldsymbol{G} = [\boldsymbol{V}_{\boldsymbol{\alpha}}^\top \boldsymbol{V}_{\boldsymbol{\beta}}] \in \mathbb{R}^{L\times L}$ be the Gram matrix of $\{\boldsymbol{V}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$. Defining $\boldsymbol{b}_{\boldsymbol{\alpha}} = \sum_{\boldsymbol{\beta}\neq\boldsymbol{\alpha}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}}(\xi'_{\boldsymbol{\beta}} - p)\boldsymbol{V}_{\boldsymbol{\beta}}$, and letting $\boldsymbol{B} = [\boldsymbol{b}_1 \cdots \boldsymbol{b}_L]$, we have that $\boldsymbol{B} = \sum_{\boldsymbol{\alpha}}(\xi'_{\boldsymbol{\alpha}} - p)\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{h}_{\boldsymbol{\alpha}}^\top$, where $\boldsymbol{h}_{\boldsymbol{\alpha}} \in \mathbb{R}^L$ is the $\boldsymbol{\alpha}$-th column of $\boldsymbol{H}_{\text{offdiag}}^{-1}$, and that with probability at least $1 - 2n^{2-\beta}$ that if $p \geq \frac{16\beta\log n}{3n}$, for $\boldsymbol{V}_{\boldsymbol{\alpha}} = vec(\mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}})$,*

$$\|\boldsymbol{B}\| \leq \sqrt{p\frac{32\nu r\beta\log n}{3n}},$$

and for $\boldsymbol{V}_{\boldsymbol{\alpha}} = vec(\boldsymbol{w}_{\boldsymbol{\alpha}})$,

$$\|\boldsymbol{B}\| \leq \sqrt{p\frac{128\beta\log n}{3}}.$$

Proof. The first statement follows from the rules of matrix multiplication, and can be seen explicitly by defining the matrix $\boldsymbol{W} = [(\xi'_1 - p)\boldsymbol{V}_1 \cdots (\xi'_L - p)\boldsymbol{V}_L] \in \mathbb{R}^{n^2\times L}$ and seeing that $\boldsymbol{B} = \boldsymbol{W}\boldsymbol{H}_{\text{offdiag}}^{-1}$. The first result then follows from the symmetry of $\boldsymbol{H}_{\text{offdiag}}^{-1}$. It follows that, as $\boldsymbol{B} = \sum_{\boldsymbol{\alpha}}(\xi'_{\boldsymbol{\alpha}} - p)\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{h}_{\boldsymbol{\alpha}}^\top$, that $\boldsymbol{B}$ is a sum of mean-zero, independent matrices. Therefore we can develop a high probability bound on $\|\boldsymbol{B}\|$ using Theorem 1.7.2. Defining each summand as $\boldsymbol{B}_{\boldsymbol{\alpha}}$ we first bound

$$\begin{aligned}
\|\boldsymbol{B}_{\boldsymbol{\alpha}}\| &= \|(\xi'_{\boldsymbol{\alpha}} - p)\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{h}_{\boldsymbol{\alpha}}^\top\| \\
&\leq \|\boldsymbol{V}_{\boldsymbol{\alpha}}\|_2\|\boldsymbol{h}_{\boldsymbol{\alpha}}\|_2
\end{aligned}$$

$$\leq \max_{\boldsymbol{\beta}} \|\boldsymbol{V}_{\boldsymbol{\beta}}\|_2 \|\boldsymbol{H}_{\text{offdiag}}^{-1}\|_{2\to\infty}$$
$$\leq \max_{\boldsymbol{\beta}} \|\boldsymbol{V}_{\boldsymbol{\beta}}\|_2 \sqrt{\frac{2}{n}} =: R,$$

where the second inequality follows from Theorem 3.A.1. Next, we bound the variance expressions. First, we have that

$$\begin{aligned}\left\|\sum_{\boldsymbol{\alpha}} \mathbb{E}[\boldsymbol{B}_{\boldsymbol{\alpha}}\boldsymbol{B}_{\boldsymbol{\alpha}}^*]\right\| &= \left\|\sum_{\boldsymbol{\alpha}} \mathbb{E}[(\xi'_{\boldsymbol{\alpha}} - p)^2]\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{h}_{\boldsymbol{\alpha}}^\top \boldsymbol{h}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}^\top\right\| \\ &= p(1-p)\left\|\sum_{\boldsymbol{\alpha}} \|\boldsymbol{h}_{\boldsymbol{\alpha}}\|_2^2 \boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}^\top\right\| \\ &\leq p\frac{2}{n}\left\|\sum_{\boldsymbol{\alpha}} \boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}^\top\right\| \\ &= \frac{2p}{n}\lambda_{\max}(\boldsymbol{G}),\end{aligned}$$

using Theorem 1.8.2 for the final line, followed by

$$\begin{aligned}\left\|\sum_{\boldsymbol{\alpha}} \mathbb{E}[\boldsymbol{B}_{\boldsymbol{\alpha}}^*\boldsymbol{B}_{\boldsymbol{\alpha}}]\right\| &= \left\|\sum_{\boldsymbol{\alpha}} \mathbb{E}[(\xi'_{\boldsymbol{\alpha}} - p)^2]\boldsymbol{h}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}^\top \boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{h}_{\boldsymbol{\alpha}}\right\| \\ &= p(1-p)\left\|\sum_{\boldsymbol{\alpha}} \|\boldsymbol{V}_{\boldsymbol{\alpha}}\|_2^2 \boldsymbol{h}_{\boldsymbol{\alpha}}\boldsymbol{h}_{\boldsymbol{\alpha}}^\top\right\| \\ &\leq p\max_{\boldsymbol{\beta}} \|\boldsymbol{V}_{\boldsymbol{\beta}}\|_2^2 \left\|\sum_{\boldsymbol{\alpha}} \boldsymbol{h}_{\boldsymbol{\alpha}}\boldsymbol{h}_{\boldsymbol{\alpha}}^\top\right\| \\ &= p\max_{\boldsymbol{\beta}} \|\boldsymbol{V}_{\boldsymbol{\beta}}\|_2^2 \left\|\boldsymbol{H}_{\text{offdiag}}^{-1}(\boldsymbol{H}_{\text{offdiag}}^{-1})^\top\right\| \\ &= p\max_{\boldsymbol{\beta}} \|\boldsymbol{V}_{\boldsymbol{\beta}}\|_2^2 \|\boldsymbol{H}_{\text{offdiag}}^{-1}\|^2 \\ &\leq 4p\max_{\boldsymbol{\beta}} \|\boldsymbol{V}_{\boldsymbol{\beta}}\|_2^2,\end{aligned}$$

where the last inequality follows from Theorem 3.A.1. Now, if $\boldsymbol{V}_{\boldsymbol{\alpha}} = \text{vec}(\mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}})$ it follows that taking the maximum of the variance terms, we have $\sigma^2 = p\frac{16\nu r}{n}$ and $R = \sqrt{\frac{\nu r}{n^2}}$. Now, as $\frac{\sigma^2}{R} = p\sqrt{16\nu r} \geq \sqrt{p\frac{128\nu r\beta\log n}{3n}}$ for $p \geq \frac{8\beta\log n}{3n}$, we have from (1.25) that

$$\mathbb{P}\left(\|\boldsymbol{B}\| \geq \sqrt{p\frac{128\nu r\beta\log n}{3n}}\right) \leq 2n^{2-\beta}.$$

If $\boldsymbol{V}_{\boldsymbol{\alpha}} = \text{vec}(\boldsymbol{w}_{\boldsymbol{\alpha}})$, then taking the maximum of the variance terms we have $\sigma^2 = 16p$ and $R = \frac{4}{\sqrt{n}}$. Now, as $\frac{\sigma^2}{R} = 4p\sqrt{n} \geq \sqrt{p\frac{128\beta\log n}{3}}$ for $p \geq \frac{4\beta\log n}{n}$, it follows again from (1.25) that

$$\mathbb{P}\left(\|\boldsymbol{B}\| \geq \sqrt{p\frac{128\beta\log n}{3}}\right) \leq 2n^{2-\beta}.$$

This concludes the proof. □

Lemma 3.B.6. *Let $\boldsymbol{X}$ be a rank-r, ν-incoherent ground truth matrix with tangent space $\mathbb{T}$. Let $\boldsymbol{V}_{\boldsymbol{\alpha}} = vec(\mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}})$ or $vec(\boldsymbol{w}_{\boldsymbol{\alpha}})$. Let $\boldsymbol{b}_{\boldsymbol{\alpha}} = \sum_{\boldsymbol{\beta}\neq\boldsymbol{\alpha}}(\xi'_{\boldsymbol{\beta}} - p)\boldsymbol{V}_{\boldsymbol{\beta}}$. For $p \geq \frac{2\beta\log n}{n}$, with probability at least $1 - n^{2-\beta}$ we have that for every $\boldsymbol{\alpha} \in \mathbb{I}$, if $\boldsymbol{V}_{\boldsymbol{\alpha}} = vec(\mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}})$,*

$$\|\boldsymbol{b}_{\boldsymbol{\alpha}}\|_2 \leq \sqrt{p\frac{4\beta\nu r\log n}{n^2}},$$

and if $\boldsymbol{V}_{\boldsymbol{\alpha}} = vec(\boldsymbol{w}_{\boldsymbol{\alpha}})$

$$\|\boldsymbol{b}_{\boldsymbol{\alpha}}\|_2 = \sqrt{p\frac{16\beta\log n}{n}}.$$

Proof. This result will follow from the Vector Bernstein inequality seen in Theorem 1.7.3. We first note that $\boldsymbol{b}_{\boldsymbol{\alpha}}$ is a sum of mean-zero independent random vectors. Defining the term $\boldsymbol{y}_{\boldsymbol{\beta}} = (\xi'_{\boldsymbol{\beta}} - p)H^{\boldsymbol{\alpha}\boldsymbol{\beta}}\boldsymbol{V}_{\boldsymbol{\beta}}$, we first need to bound $\|\boldsymbol{y}_{\boldsymbol{\beta}}\|_2$ via

$$\begin{aligned}\|\boldsymbol{y}_{\boldsymbol{\beta}}\|_2 &= |(\xi'_{\boldsymbol{\beta}} - p)|H^{\boldsymbol{\alpha}\boldsymbol{\beta}}\|\boldsymbol{V}_{\boldsymbol{\beta}}\|_2 \\ &\leq \frac{1}{n}\max_{\boldsymbol{\alpha}}\|\boldsymbol{V}_{\boldsymbol{\alpha}}\|_2 =: R.\end{aligned}$$

Next, to bound the variance term, we see that

$$\begin{aligned}\sum_{\boldsymbol{\beta}\neq\boldsymbol{\alpha}}\mathbb{E}[\|\boldsymbol{y}_{\boldsymbol{\beta}}\|_2^2] &= \sum_{\boldsymbol{\beta}\neq\boldsymbol{\alpha}}\mathbb{E}[(\xi'_{\boldsymbol{\beta}} - p)^2](H^{\boldsymbol{\alpha}\boldsymbol{\beta}})^2\|\boldsymbol{V}_{\boldsymbol{\beta}}\|_2^2 \\ &= p(1-p)\sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}}(H^{\boldsymbol{\alpha}\boldsymbol{\beta}})^2 \\ &\leq p\max_{\boldsymbol{\gamma}}\|\boldsymbol{V}_{\boldsymbol{\gamma}}\|_2^2\sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}}(H^{\boldsymbol{\alpha}\boldsymbol{\beta}})^2 \\ &\leq \frac{2p}{n}\max_{\boldsymbol{\gamma}}\|\boldsymbol{V}_{\boldsymbol{\gamma}}\|_2^2.\end{aligned}$$

When $\boldsymbol{V}_{\boldsymbol{\beta}} = \text{vec}(\mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\beta}})$, we have $\sigma^2 = p\frac{\nu r}{n^2}$ and $R = \sqrt{\frac{\nu r}{2n^3}}$. It follows then that $\frac{\sigma^2}{R} = p\sqrt{\frac{2\nu r}{n}} \geq \sqrt{p\frac{4\nu r\beta \log n}{n^2}}$ for $p \geq \frac{2\beta \log n}{n}$, we have from Theorem 1.7.3 that

$$\mathbb{P}\left(\|\boldsymbol{b}_{\boldsymbol{\alpha}}\|_2 \geq \sqrt{p\frac{8\nu r\beta \log n}{n^2}}\right) \leq 2n^{-\beta}.$$

Similarly, when $\boldsymbol{V}_{\boldsymbol{\beta}} = \text{vec}(\boldsymbol{w}_{\boldsymbol{\beta}})$, we have $\sigma^2 = \frac{4p}{n}$ and $R = \frac{2}{n}$. As such, $\frac{\sigma^2}{R} = 2p \geq \sqrt{p\frac{16\beta \log n}{3n}}$ for $p \geq \frac{2\beta \log n}{n}$, that

$$\mathbb{P}\left(\|\boldsymbol{b}_{\boldsymbol{\alpha}}\|_2 \geq \sqrt{p\frac{16\beta \log n}{n}}\right) \leq 2n^{-\beta}.$$

The stated results then follow from the union bound over every $\boldsymbol{\alpha} \in \mathbb{I}$. □

Lemma 3.B.7. *Let $\boldsymbol{X}$ be a rank-r, ν-incoherent ground truth matrix with tangent space $\mathbb{T}$. Then if $p \geq \frac{16\beta \log n}{3n}$ for any $\beta > 3$, with probability at least $1 - C'n^{3-\beta}$, where C' is an absolute constant, we have that, for some absolute numerical constants C_1, C_2, and C_3 independent of n or $\boldsymbol{X}$, the following three inequalities all hold:*

$$\left\|\sum_{\substack{\boldsymbol{\alpha}\in\mathbb{I}\\ \boldsymbol{\beta}\neq\boldsymbol{\alpha}}} (\xi_{\boldsymbol{\alpha}} - p)(\xi_{\boldsymbol{\beta}} - p)H^{\boldsymbol{\alpha}\boldsymbol{\beta}}\langle \cdot, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\beta}}\rangle \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}\right\| \leq p\frac{C_1\nu r\sqrt{\beta}\log n}{n},$$

$$\left\|\sum_{\substack{\boldsymbol{\alpha}\in\mathbb{I}\\ \boldsymbol{\beta}\neq\boldsymbol{\alpha}}} (\xi_{\boldsymbol{\alpha}} - p)(\xi_{\boldsymbol{\beta}} - p)H^{\boldsymbol{\alpha}\boldsymbol{\beta}}\langle \cdot, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\beta}}\rangle \boldsymbol{w}_{\boldsymbol{\alpha}}\right\| \leq pC_2\sqrt{\frac{\beta\nu r}{n}}\log n,$$

$$\left\|\sum_{\substack{\boldsymbol{\alpha}\in\mathbb{I}\\ \boldsymbol{\beta}\neq\boldsymbol{\alpha}}} (\xi_{\boldsymbol{\alpha}} - p)(\xi_{\boldsymbol{\beta}} - p)H^{\boldsymbol{\alpha}\boldsymbol{\beta}}\langle \cdot, \boldsymbol{w}_{\boldsymbol{\beta}}\rangle \boldsymbol{w}_{\boldsymbol{\alpha}}\right\| \leq pC_3\sqrt{\beta}\log n.$$

Proof. The goal is to develop a high probability bound for the operator norm of the previously mentioned operator. Prior to doing this, we define some simplifying notation. Let $\bar{\boldsymbol{w}}_{\boldsymbol{\alpha}} = \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}$ or $\boldsymbol{w}_{\boldsymbol{\alpha}}$, and let $\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}} = \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}$ or $\boldsymbol{w}_{\boldsymbol{\alpha}}$ with them not necessarily being the same. Let $\boldsymbol{G} = [\langle \bar{\boldsymbol{w}}_{\boldsymbol{\alpha}}, \bar{\boldsymbol{w}}_{\boldsymbol{\beta}}\rangle] \in \mathbb{R}^{L\times L}$ be the corresponding Gram matrix for $\{\bar{\boldsymbol{w}}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$.

We begin the proof by considering, for some $t > 0$, the expression

$$\mathbb{P}\left[\left\| \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} (\xi_{\boldsymbol{\alpha}} - p)(\xi_{\boldsymbol{\beta}} - p)\langle \cdot, \hat{\boldsymbol{w}}_{\boldsymbol{\beta}}\rangle H^{\boldsymbol{\alpha\beta}} \bar{\boldsymbol{w}}_{\boldsymbol{\alpha}} \right\| \geq t\right].$$

For notational simplicity, we let $\boldsymbol{U}_{\boldsymbol{\alpha}} = \text{vec}(\bar{\boldsymbol{w}}_{\boldsymbol{\alpha}}) \in \mathbb{R}^{n^2}$ denote the vectorized version of $\bar{\boldsymbol{w}}_{\boldsymbol{\alpha}}$. Similarly, we let $\boldsymbol{V}_{\boldsymbol{\beta}} = \text{vec}(\hat{\boldsymbol{w}}_{\boldsymbol{\beta}}) \in \mathbb{R}^{n^2}$ As such, we can see that this probabilistic expression can be equivalently represented as

$$\mathbb{P}\left[\left\| \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} (\xi_{\boldsymbol{\alpha}} - p)(\xi_{\boldsymbol{\beta}} - p) H^{\boldsymbol{\alpha\beta}} \boldsymbol{U}_{\boldsymbol{\alpha}} \boldsymbol{V}_{\boldsymbol{\beta}}^{\top} \right\| \geq t\right].$$

This term is a second order matrix chaos, and is a degenerate U-statistic—see [Lee19] for details on the theory of U-statistics, and see [DlPG12] for a detailed investigation of the decoupling of Banach-space-valued U-statistics. As seen in Theorem 1.7.5, we can study the distribution of the operator norm for this operator by considering a new set of i.i.d. random Bernoulli variables $\xi'_{\boldsymbol{\beta}}$. As such, Theorem 1.7.5 states that for some absolute constant C independent of n, $\boldsymbol{U}_{\boldsymbol{\alpha}}$, $\boldsymbol{V}_{\boldsymbol{\alpha}}$, or $\boldsymbol{H}^{-1}$, that

$$\mathbb{P}\left[\left\| \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} (\xi_{\boldsymbol{\alpha}} - p)(\xi_{\boldsymbol{\beta}} - p) H^{\boldsymbol{\alpha\beta}} \boldsymbol{U}_{\boldsymbol{\alpha}} \boldsymbol{V}_{\boldsymbol{\beta}}^{\top} \right\| \geq t\right] \leq C\mathbb{P}\left[\left\| \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} (\xi_{\boldsymbol{\alpha}} - p)(\xi'_{\boldsymbol{\beta}} - p) H^{\boldsymbol{\alpha\beta}} \boldsymbol{U}_{\boldsymbol{\alpha}} \boldsymbol{V}_{\boldsymbol{\beta}}^{\top} \right\| \geq t/C\right].$$

We note that optimal values of the constants C have not been explicitly computed for each chaos of order m, but that the constant itself is a function of m. For $m = 2$, which is the case we consider, this constant is usually rather small[DlPG12]. Using the law of total probability, conditioning on the random variables $\xi'_1, ..., \xi'_L$ and defining the vector $\boldsymbol{\xi'} = [\xi'_1 \ ... \ \xi'_L]^{\top}$ that

$$\begin{aligned}\mathbb{P}\left[\left\| \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} (\xi_{\boldsymbol{\alpha}} - p)(\xi'_{\boldsymbol{\beta}} - p) H^{\boldsymbol{\alpha\beta}} \boldsymbol{U}_{\boldsymbol{\alpha}} \boldsymbol{V}_{\boldsymbol{\beta}}^{\top} \right\| \geq t\right] \\ = \mathbb{E}\left[\mathbb{P}\left(\left\| \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} (\xi_{\boldsymbol{\alpha}} - p)(\xi'_{\boldsymbol{\beta}} - p) H^{\boldsymbol{\alpha\beta}} \boldsymbol{U}_{\boldsymbol{\alpha}} \boldsymbol{V}_{\boldsymbol{\beta}}^{\top} \right\| \geq t \;\middle|\; \boldsymbol{\xi'}\right)\right].\end{aligned}$$

We can study the conditional expression

$$\mathbb{P}\left(\left\| \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} (\xi_{\boldsymbol{\alpha}} - p)(\xi'_{\boldsymbol{\beta}} - p) H^{\boldsymbol{\alpha\beta}} \boldsymbol{U}_{\boldsymbol{\alpha}} \boldsymbol{V}_{\boldsymbol{\beta}}^{\top} \right\| \geq t \;\middle|\; \boldsymbol{\xi'}\right)$$

using Theorem 1.7.2. Defining the expression $\boldsymbol{b}_{\boldsymbol{\alpha}} = \sum_{\boldsymbol{\beta} \neq \boldsymbol{\alpha}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}}(\xi'_{\boldsymbol{\beta}} - p)\boldsymbol{V}_{\boldsymbol{\beta}}$, we notice that the sum can be re-written as

$$\mathbb{P}\left(\left\| \sum_{\boldsymbol{\alpha} \neq \boldsymbol{\beta}} (\xi_{\boldsymbol{\alpha}} - p)(\xi'_{\boldsymbol{\beta}} - p) H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \boldsymbol{U}_{\boldsymbol{\alpha}} \boldsymbol{V}_{\boldsymbol{\beta}}^\top \right\| \geq t \;\middle|\; \boldsymbol{\xi}' \right) = \mathbb{P}\left(\left\| \sum_{\boldsymbol{\alpha}} (\xi_{\boldsymbol{\alpha}} - p) \boldsymbol{U}_{\boldsymbol{\alpha}} \boldsymbol{b}_{\boldsymbol{\alpha}}^\top \right\| \geq t \;\middle|\; \boldsymbol{\xi}' \right).$$

In this form, conditioned on $\boldsymbol{\xi}'$ we see that this is a sum of mean-zero, independent random variables. To study its concentration with Theorem 1.7.2, we define $\boldsymbol{T}_{\boldsymbol{\alpha}} = (\xi_{\boldsymbol{\alpha}} - p)\boldsymbol{U}_{\boldsymbol{\alpha}} \boldsymbol{b}_{\boldsymbol{\alpha}}^\top$. The first task is bounding the operator norm

$$\|\boldsymbol{T}_{\boldsymbol{\alpha}}\| \leq \|\boldsymbol{U}_{\boldsymbol{\alpha}}\|_2 \|\boldsymbol{b}_{\boldsymbol{\alpha}}\|_2 \leq \max_{\boldsymbol{\beta}} \|\boldsymbol{U}_{\boldsymbol{\beta}}\|_2 \max_{\boldsymbol{\gamma}} \|\boldsymbol{b}_{\boldsymbol{\gamma}}\|_2 =: R.$$

Next, we need to estimate the variance terms $\|\sum_{\boldsymbol{\alpha}} \mathbb{E}[\boldsymbol{T}_{\boldsymbol{\alpha}} \boldsymbol{T}_{\boldsymbol{\alpha}}^* \mid \boldsymbol{\xi}']\|$ and $\|\sum_{\boldsymbol{\alpha}} \mathbb{E}[\boldsymbol{T}_{\boldsymbol{\alpha}}^* \boldsymbol{T}_{\boldsymbol{\alpha}} \mid \boldsymbol{\xi}']\|$. For the first term, notice that

$$\begin{aligned}
\left\| \sum_{\boldsymbol{\alpha}} \mathbb{E}[\boldsymbol{T}_{\boldsymbol{\alpha}} \boldsymbol{T}_{\boldsymbol{\alpha}}^* \mid \boldsymbol{\xi}'] \right\| &= \left\| \sum_{\boldsymbol{\alpha}} \mathbb{E}[(\xi_{\boldsymbol{\alpha}} - p)^2] \boldsymbol{U}_{\boldsymbol{\alpha}} \boldsymbol{b}_{\boldsymbol{\alpha}}^\top \boldsymbol{b}_{\boldsymbol{\alpha}} \boldsymbol{U}_{\boldsymbol{\alpha}}^\top \right\| \\
&= p(1-p) \left\| \sum_{\boldsymbol{\alpha}} \|\boldsymbol{b}_{\boldsymbol{\alpha}}\|_2^2 \boldsymbol{U}_{\boldsymbol{\alpha}} \boldsymbol{U}_{\boldsymbol{\alpha}}^\top \right\| \\
&\leq p \max_{\boldsymbol{\alpha}} \|\boldsymbol{b}_{\boldsymbol{\alpha}}\|_2^2 \left\| \sum_{\boldsymbol{\alpha}} \boldsymbol{U}_{\boldsymbol{\alpha}} \boldsymbol{U}_{\boldsymbol{\alpha}}^\top \right\| \\
&\leq p \max_{\boldsymbol{\alpha}} \|\boldsymbol{b}_{\boldsymbol{\alpha}}\|_2^2 \lambda_{\max}(\boldsymbol{G}).
\end{aligned}$$

The other variance term is given by

$$\begin{aligned}
\left\| \sum_{\boldsymbol{\alpha}} \mathbb{E}[\boldsymbol{T}_{\boldsymbol{\alpha}}^* \boldsymbol{T}_{\boldsymbol{\alpha}} \mid \boldsymbol{\xi}'] \right\| &= \left\| \sum_{\boldsymbol{\alpha}} \mathbb{E}[(\xi_{\boldsymbol{\alpha}} - p)^2] \boldsymbol{b}_{\boldsymbol{\alpha}} \boldsymbol{U}_{\boldsymbol{\alpha}}^\top \boldsymbol{U}_{\boldsymbol{\alpha}} \boldsymbol{b}_{\boldsymbol{\alpha}}^\top \right\| \\
&\leq p \max_{\boldsymbol{\gamma}} \|\boldsymbol{U}_{\boldsymbol{\gamma}}\|_2^2 \left\| \sum_{\boldsymbol{\alpha}} \boldsymbol{b}_{\boldsymbol{\alpha}} \boldsymbol{b}_{\boldsymbol{\alpha}}^\top \right\| \\
&= p \max_{\boldsymbol{\gamma}} \|\boldsymbol{U}_{\boldsymbol{\gamma}}\|_2^2 \|\boldsymbol{B}\boldsymbol{B}^\top\| \\
&= p \max_{\boldsymbol{\gamma}} \|\boldsymbol{U}_{\boldsymbol{\gamma}}\|_2^2 \|\boldsymbol{B}\|^2,
\end{aligned}$$

where the last matrix is defined as $\boldsymbol{B} = [\boldsymbol{b}_1 \cdots \boldsymbol{b}_L]$ and the spectral result follows from the vectorization lemma.

Defining the variance parameter $\sigma^2 = \max\{\|\sum_{\boldsymbol{\alpha}} \mathbb{E}[\boldsymbol{T}_{\boldsymbol{\alpha}}\boldsymbol{T}_{\boldsymbol{\alpha}}^* \mid \boldsymbol{\xi}']\|, \|\sum_{\boldsymbol{\alpha}} \mathbb{E}[\boldsymbol{T}_{\boldsymbol{\alpha}}^*\boldsymbol{T}_{\boldsymbol{\alpha}} \mid \boldsymbol{\xi}']\|\}$, we have from the monotonicity of expectation that

$$\mathbb{E}\left[\mathbb{P}\left(\left\|\sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}}(\xi_{\boldsymbol{\alpha}}-p)(\xi'_{\boldsymbol{\beta}}-p)H^{\boldsymbol{\alpha}\boldsymbol{\beta}}\boldsymbol{U}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\beta}}^\top\right\| \geq t \;\middle|\; \boldsymbol{\xi}'\right)\right] \leq \mathbb{E}\left[2n\exp\left(\frac{-t^2/2}{Rt/3+\sigma^2}\right)\right].$$

It follows that if high probability estimates exist for R and σ^2, referred heretofore as $R_{\max}$ and $\sigma^2_{\max}$ respectively, that we can effectively uncondition R and σ^2 as functions of $\boldsymbol{\xi}'$ as follows:

$$\begin{aligned}\mathbb{E}\left[2n\exp\left(\frac{-t^2/2}{Rt/3+\sigma^2}\right)\right] &\leq 2\mathbb{P}(\sigma^2\leq\sigma^2_{\max}, R\leq R_{\max})n\exp\left(\frac{-t^2/2}{R_{\max}t/3+\sigma^2_{\max}}\right)\\ &\quad + 2n\mathbb{P}(\sigma^2\geq\sigma^2_{\max}, R\geq R_{\max}).\end{aligned}$$

Plugging in the expressions in Theorem 3.B.5 and Theorem 3.B.6 and setting $t = \sqrt{\frac{8}{3}\sigma^2_{\max}\log n}$ in the expressions such that the simplification in Theorem 1.7.2 is possible, we have that, if $p \geq \frac{16\beta\log n}{n}$ for all three inequalities, the following inequality holds for each operator in the Lemma statement.

$$\begin{aligned}\mathbb{E}\left[2n\exp\left(\frac{-t^2/2}{Rt/3+\sigma^2}\right)\right] &\leq 2\mathbb{P}(\sigma^2\leq\sigma^2_{\max}, R\leq R_{\max})n\exp\left(\frac{-t^2/2}{R_{\max}t/3+\sigma^2_{\max}}\right)\\ &\quad + 2n\mathbb{P}(\sigma^2\geq\sigma^2_{\max} \text{ or } R\geq R_{\max})\\ &\leq 2\mathbb{P}(\sigma^2\leq\sigma^2_{\max}, R\leq R_{\max})n\exp\left(\frac{-3t^2}{8\sigma^2_{\max}}\right)\\ &\quad + 2n\mathbb{P}(\sigma^2\geq\sigma^2_{\max} \text{ or } R\geq R_{\max})\\ &\leq 2(1-4n^{2-\beta})n^{1-\beta} + 8n^{3-\beta}\\ &\leq 10n^{3-\beta}.\end{aligned}$$

Taking the union bound over the 3 inequalities and incorporating the constant C from Theorem 1.7.5 concludes the proof. □

Lemma 3.B.8. *Let $\boldsymbol{X}$ be a ground-truth, rank-r, ν-incoherent Gram matrix with tangent space $\mathbb{T}$. If the sampling probability $p \geq \frac{128\beta\log n}{3n}$ for any $\beta > 1$, then the following collection of*

inequalities hold with probability at least $1 - 8n^{1-\beta}$

$$\left\| \sum_{\boldsymbol{\alpha} \in \mathbb{I}} (\xi_{\boldsymbol{\alpha}} - p)\, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \sum_{\substack{\boldsymbol{\beta} \in \mathbb{I} \\ \boldsymbol{\alpha} \neq \boldsymbol{\beta}}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\beta}} \rangle \right\| \leq \sqrt{p \frac{64 \nu^2 r^2 \beta \log n}{3n}},$$

$$\left\| \sum_{\boldsymbol{\alpha} \in \mathbb{I}} (\xi_{\boldsymbol{\alpha}} - p)\, \boldsymbol{w}_{\boldsymbol{\alpha}} \sum_{\substack{\boldsymbol{\beta} \in \mathbb{I} \\ \boldsymbol{\alpha} \neq \boldsymbol{\beta}}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \boldsymbol{w}_{\boldsymbol{\beta}} \rangle \right\| \leq \sqrt{p \frac{256 n \beta \log n}{3}},$$

$$\left\| \sum_{\boldsymbol{\alpha} \in \mathbb{I}} (\xi_{\boldsymbol{\alpha}} - p)\, \boldsymbol{w}_{\boldsymbol{\alpha}} \sum_{\substack{\boldsymbol{\beta} \in \mathbb{I} \\ \boldsymbol{\alpha} \neq \boldsymbol{\beta}}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\beta}} \rangle \right\| \leq \sqrt{p \frac{128 \nu r \beta \log n}{3}},$$

$$\left\| \sum_{\boldsymbol{\alpha} \in \mathbb{I}} (\xi_{\boldsymbol{\alpha}} - p)\, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \sum_{\substack{\boldsymbol{\beta} \in \mathbb{I} \\ \boldsymbol{\alpha} \neq \boldsymbol{\beta}}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \boldsymbol{w}_{\boldsymbol{\beta}} \rangle \right\| \leq \sqrt{p \frac{128 \nu r \beta \log n}{3}}.$$

Proof. We first define $\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}} = \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}}$ or $\boldsymbol{w}_{\boldsymbol{\alpha}}$ and $\bar{\boldsymbol{w}}_{\boldsymbol{\alpha}} = \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}}$ or $\boldsymbol{w}_{\boldsymbol{\alpha}}$, where $\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}$ and $\bar{\boldsymbol{w}}_{\boldsymbol{\alpha}}$ are not necessarily equal. We also define the corresponding Gram matrices of $\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}$ and $\bar{\boldsymbol{w}}_{\boldsymbol{\alpha}}$ as $\hat{\boldsymbol{H}}$ and $\bar{\boldsymbol{H}}$, respectively. Similarly define the dual basis vectors $\bar{\boldsymbol{v}}_{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{v}}_{\boldsymbol{\alpha}}$. As such, all of these inequalities can be represented through studying the expression

$$\left\| \sum_{\boldsymbol{\alpha} \in \mathbb{I}} (\xi_{\boldsymbol{\alpha}} - p)\, \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}} \sum_{\substack{\boldsymbol{\beta} \in \mathbb{I} \\ \alpha \neq \beta}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \bar{\boldsymbol{w}}_{\boldsymbol{\beta}} \rangle \right\|.$$

Defining the vectorized version of our summands $\hat{\boldsymbol{U}}_{\boldsymbol{\alpha}} = \mathrm{vec}(\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}) \in \mathbb{R}^{n^2}$ and $\bar{\boldsymbol{V}}_{\boldsymbol{\alpha}} = \mathrm{vec}(\bar{\boldsymbol{w}}_{\boldsymbol{\alpha}}) \in \mathbb{R}^{n^2}$, we can see that the above expression is equivalent to

$$\left\| \sum_{\boldsymbol{\alpha} \in \mathbb{I}} (\xi_{\boldsymbol{\alpha}} - p)\, \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}} \sum_{\substack{\boldsymbol{\beta} \in \mathbb{I} \\ \alpha \neq \beta}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \bar{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\top} \right\|.$$

This expression is the operator norm bound on a sum of mean-zero, independent random matrices. We define each summand as $\boldsymbol{G}_{\boldsymbol{\alpha}} = (\xi_{\boldsymbol{\alpha}} - p) \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}} \sum_{\boldsymbol{\alpha} \neq \boldsymbol{\beta}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}}^{\top}$.To proceed, we use Theorem 1.7.2.

First, we bound $\|\boldsymbol{G}_{\boldsymbol{\alpha}}\|$:

$$\begin{aligned}\|\boldsymbol{G}_{\boldsymbol{\alpha}}\| &\leq \|\hat{\boldsymbol{U}}_{\boldsymbol{\alpha}}\|_2 \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} |H^{\boldsymbol{\alpha}\boldsymbol{\beta}}| \|\bar{\boldsymbol{V}}_{\boldsymbol{\beta}}^\top\|_2 \\ &\leq 2 \max_{\boldsymbol{\beta}} \|\hat{\boldsymbol{U}}_{\boldsymbol{\alpha}}\|_2 \max_{\boldsymbol{\gamma}} \|\bar{\boldsymbol{V}}_{\boldsymbol{\gamma}}\|_2,\end{aligned}$$

following from Theorem 2.A.3. Next, we bound the variance terms $\|\sum_{\boldsymbol{\alpha}} \mathbb{E}[\boldsymbol{G}_{\boldsymbol{\alpha}}\boldsymbol{G}_{\boldsymbol{\alpha}}^*]\|$ and $\|\sum_{\boldsymbol{\alpha}} \mathbb{E}[\boldsymbol{G}_{\boldsymbol{\alpha}}^*\boldsymbol{G}_{\boldsymbol{\alpha}}]\|$. It follows that, defining $\boldsymbol{d}_{\boldsymbol{\alpha}} = \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \boldsymbol{V}_{\boldsymbol{\beta}}$

$$\begin{aligned}\left\|\sum_{\boldsymbol{\alpha}} \mathbb{E}[\boldsymbol{G}_{\boldsymbol{\alpha}}\boldsymbol{G}_{\boldsymbol{\alpha}}^*]\right\| &= \left\|\sum_{\boldsymbol{\alpha}} \mathbb{E}[(\xi_{\boldsymbol{\alpha}} - p)^2] \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}} \boldsymbol{d}_{\boldsymbol{\alpha}}^\top \boldsymbol{d}_{\boldsymbol{\alpha}} \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}}^\top\right\| \\ &\leq p \left\|\sum_{\boldsymbol{\alpha}} \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}} \boldsymbol{d}_{\boldsymbol{\alpha}}^\top \boldsymbol{d}_{\boldsymbol{\alpha}} \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}}^\top\right\|.\end{aligned}$$

Now, $\boldsymbol{d}_{\boldsymbol{\alpha}} + H^{\boldsymbol{\alpha}\boldsymbol{\alpha}} \bar{\boldsymbol{V}}_{\boldsymbol{\alpha}} = \mathrm{vec}(\bar{\boldsymbol{v}}_{\boldsymbol{\alpha}})$ by duality and the linearity of both the vec operator and the projection operator $\mathcal{P}_{\mathbb{T}}$, if present in the relevant definition of $\bar{\boldsymbol{v}}_{\boldsymbol{\alpha}}$. Defining $\bar{\boldsymbol{Q}}_{\boldsymbol{\alpha}} = \mathrm{vec}(\bar{\boldsymbol{v}}_{\boldsymbol{\alpha}})$, we see that

$$\begin{aligned}\left\|\sum_{\boldsymbol{\alpha}} \mathbb{E}[\boldsymbol{G}_{\boldsymbol{\alpha}}\boldsymbol{G}_{\boldsymbol{\alpha}}^*]\right\| &\leq p \left\|\sum_{\boldsymbol{\alpha}} \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}} \left(\bar{\boldsymbol{Q}}_{\boldsymbol{\alpha}} - H^{\boldsymbol{\alpha}\boldsymbol{\alpha}} \bar{\boldsymbol{V}}_{\boldsymbol{\alpha}}\right)^\top \left(\bar{\boldsymbol{Q}}_{\boldsymbol{\alpha}} - H^{\boldsymbol{\alpha}\boldsymbol{\alpha}} \bar{\boldsymbol{V}}_{\boldsymbol{\alpha}}\right) \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}}^\top\right\| \\ &= p \Bigg\| \sum_{\boldsymbol{\alpha}} \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}} \bar{\boldsymbol{Q}}_{\boldsymbol{\alpha}}^\top \bar{\boldsymbol{Q}}_{\boldsymbol{\alpha}} \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}}^\top - H^{\boldsymbol{\alpha}\boldsymbol{\alpha}} \sum_{\boldsymbol{\alpha}} \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}} \bar{\boldsymbol{Q}}_{\boldsymbol{\alpha}}^\top \bar{\boldsymbol{V}}_{\boldsymbol{\alpha}} \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}}^\top \\ &\qquad - H^{\boldsymbol{\alpha}\boldsymbol{\alpha}} \sum_{\boldsymbol{\alpha}} \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}} \bar{\boldsymbol{V}}_{\boldsymbol{\alpha}}^\top \bar{\boldsymbol{Q}}_{\boldsymbol{\alpha}} \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}}^\top + (H^{\boldsymbol{\alpha}\boldsymbol{\alpha}})^2 \sum_{\boldsymbol{\alpha}} \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}} \bar{\boldsymbol{V}}_{\boldsymbol{\alpha}}^\top \bar{\boldsymbol{V}}_{\boldsymbol{\alpha}} \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}}^\top \Bigg\| \\ &\leq p \left\|\sum_{\boldsymbol{\alpha}} \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}} \bar{\boldsymbol{Q}}_{\boldsymbol{\alpha}}^\top \bar{\boldsymbol{Q}}_{\boldsymbol{\alpha}} \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}}^\top\right\| + 2p \left\|H^{\boldsymbol{\alpha}\boldsymbol{\alpha}} \sum_{\boldsymbol{\alpha}} \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}} \bar{\boldsymbol{Q}}_{\boldsymbol{\alpha}}^\top \bar{\boldsymbol{V}}_{\boldsymbol{\alpha}} \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}}^\top\right\| \\ &\quad + p \left\|(H^{\boldsymbol{\alpha}\boldsymbol{\alpha}})^2 \sum_{\boldsymbol{\alpha}} \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}} \bar{\boldsymbol{V}}_{\boldsymbol{\alpha}}^\top \bar{\boldsymbol{V}}_{\boldsymbol{\alpha}} \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}}^\top\right\| \\ &\leq p \max_{\boldsymbol{\beta}} \|\bar{\boldsymbol{Q}}_{\boldsymbol{\beta}}\|_2^2 \left\|\sum_{\boldsymbol{\alpha}} \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}} \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}}^\top\right\| + 2p \left\|H^{\boldsymbol{\alpha}\boldsymbol{\alpha}} \sum_{\boldsymbol{\alpha}} |\bar{\boldsymbol{Q}}_{\boldsymbol{\alpha}}^\top \bar{\boldsymbol{V}}_{\boldsymbol{\alpha}}| \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}} \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}}^\top\right\| \\ &\quad + p \max_{\boldsymbol{\beta}} \|\bar{\boldsymbol{V}}_{\boldsymbol{\beta}}\|_2^2 \left\|(H^{\boldsymbol{\alpha}\boldsymbol{\alpha}})^2 \sum_{\boldsymbol{\alpha}} \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}} \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}}^\top\right\| \\ &\leq p \left(\max_{\boldsymbol{\beta}} \|\bar{\boldsymbol{Q}}_{\boldsymbol{\beta}}\|_2^2 + \max_{\boldsymbol{\beta}} \|\bar{\boldsymbol{V}}_{\boldsymbol{\beta}}\|_2^2 + 2 \max_{\boldsymbol{\beta}} \|\bar{\boldsymbol{Q}}_{\boldsymbol{\beta}}\|_2 \max_{\boldsymbol{\gamma}} \|\bar{\boldsymbol{V}}_{\boldsymbol{\gamma}}\|_2\right) \left\|\sum_{\boldsymbol{\alpha}} \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}} \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}}^\top\right\| \\ &\leq p \left(\max_{\boldsymbol{\beta}} \|\bar{\boldsymbol{Q}}_{\boldsymbol{\beta}}\|_2^2 + \max_{\boldsymbol{\beta}} \|\bar{\boldsymbol{V}}_{\boldsymbol{\beta}}\|_2^2 + 2 \max_{\boldsymbol{\beta}} \|\bar{\boldsymbol{Q}}_{\boldsymbol{\beta}}\|_2 \max_{\boldsymbol{\gamma}} \|\bar{\boldsymbol{V}}_{\boldsymbol{\gamma}}\|_2\right) \lambda_{\max}(\hat{\boldsymbol{H}}).\end{aligned}$$

Where the first inequality follows from the triangle inequality, the second inequality follows from Cauchy-Schwarz, the third inequality follows from the fact that $|H^{\boldsymbol{\alpha}\boldsymbol{\alpha}}| < 1$ for every $\boldsymbol{\alpha} \in \mathbb{I}$, and the last inequality follows from Theorem 1.8.2 and Theorem 2.A.3.

To analyze the other variance term, we begin by defining the matrix $\boldsymbol{E} = [\boldsymbol{d}_1 \cdots \boldsymbol{d}_L] \in \mathbb{R}^{n^2 \times L}$. We now see that

$$\begin{aligned}
\left\| \sum_{\boldsymbol{\alpha}} \mathbb{E}[\boldsymbol{G}_{\boldsymbol{\alpha}}^* \boldsymbol{G}_{\boldsymbol{\alpha}}] \right\| &= \left\| \sum_{\boldsymbol{\alpha}} \mathbb{E}[(\xi_{\boldsymbol{\alpha}} - p)^2] \boldsymbol{d}_{\boldsymbol{\alpha}} \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}}^\top \hat{\boldsymbol{U}}_{\boldsymbol{\alpha}} \boldsymbol{d}_{\boldsymbol{\alpha}}^\top \right\| \\
&\leq p \max_{\boldsymbol{\beta}} \|\hat{\boldsymbol{U}}_{\boldsymbol{\alpha}}\|_2^2 \left\| \sum_{\boldsymbol{\alpha}} \boldsymbol{d}_{\boldsymbol{\alpha}} \boldsymbol{d}_{\boldsymbol{\alpha}}^\top \right\| \\
&= p \max_{\boldsymbol{\beta}} \|\hat{\boldsymbol{U}}_{\boldsymbol{\alpha}}\|_2^2 \|\boldsymbol{E}\boldsymbol{E}^\top\|.
\end{aligned}$$

Similarly to Theorem 3.B.5, we can see that $\boldsymbol{E} = \boldsymbol{W}\boldsymbol{H}_{\text{offdiag}}^{-1}$. As such, notice that

$$\begin{aligned}
\|\boldsymbol{E}\boldsymbol{E}^\top\| &= \max_{\|\boldsymbol{y}\|_2=1} \boldsymbol{y}^\top \boldsymbol{E}\boldsymbol{E}^\top \boldsymbol{y} \\
&= \max_{\|\boldsymbol{y}\|_2=1} \boldsymbol{y}^\top \boldsymbol{W}\boldsymbol{H}_{\text{offdiag}}^{-1}\boldsymbol{H}_{\text{offdiag}}^{-1}\boldsymbol{W}^\top \boldsymbol{y} \\
&= \max_{\|\boldsymbol{y}\|_2=1} \|\boldsymbol{H}_{\text{offdiag}}^{-1}\boldsymbol{W}^\top \boldsymbol{y}\|_2^2 \\
&\leq \|\boldsymbol{H}_{\text{offdiag}}^{-1}\|^2 \max_{\|\boldsymbol{y}\|_2=1} \|\boldsymbol{W}^\top \boldsymbol{y}\|_2^2 \\
&\leq 4 \max_{\|\boldsymbol{y}\|_2=1} \boldsymbol{y}^\top \boldsymbol{W}\boldsymbol{W}^\top \boldsymbol{y} \\
&= 4\|\boldsymbol{W}\boldsymbol{W}^\top\| \\
&= 4\|\boldsymbol{W}^\top\boldsymbol{W}\| \\
&= 4\lambda_{\max}(\bar{\boldsymbol{H}}),
\end{aligned}$$

where the first inequality follows from Cauchy-Schwarz, and the second inequality follows from Theorem 3.A.1. This yields a variance bound as follows:

$$\begin{aligned}
\sigma^2 &= \max\left\{ \left\| \sum_{\boldsymbol{\alpha}} \mathbb{E}[\boldsymbol{G}_{\boldsymbol{\alpha}}^* \boldsymbol{G}_{\boldsymbol{\alpha}}] \right\|, \left\| \sum_{\boldsymbol{\alpha}} \mathbb{E}[\boldsymbol{G}_{\boldsymbol{\alpha}} \boldsymbol{G}_{\boldsymbol{\alpha}}^*] \right\| \right\} \\
&\leq p \max\left\{ 4 \max_{\boldsymbol{\beta}} \|\hat{\boldsymbol{U}}_{\boldsymbol{\alpha}}\|_2^2 \lambda_{\max}(\bar{\boldsymbol{H}}),
\end{aligned}$$

$$\left(\max_{\boldsymbol{\beta}} \|\bar{\boldsymbol{Q}}_{\boldsymbol{\beta}}\|_2^2 + \max_{\boldsymbol{\beta}} \|\bar{\boldsymbol{V}}_{\boldsymbol{\beta}}\|_2^2 + 2 \max_{\boldsymbol{\beta}} \|\bar{\boldsymbol{Q}}_{\boldsymbol{\beta}}\|_2 \max_{\boldsymbol{\gamma}} \|\bar{\boldsymbol{V}}_{\boldsymbol{\gamma}}\|_2 \right) \lambda_{\max}(\hat{\boldsymbol{H}}) \Big\}.$$

This leads to the following variance bounds and norm bounds for the corresponding combinations of $\bar{\boldsymbol{w}}_{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}$:

$$\begin{aligned}
\bar{\boldsymbol{w}}_{\boldsymbol{\alpha}} = \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}},\ \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}} = \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} &\Rightarrow \sigma^2 = 8p\frac{\nu^2 r^2}{n},\ R = \frac{\nu r}{n}, \\
\bar{\boldsymbol{w}}_{\boldsymbol{\alpha}} = \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}},\ \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}} = \boldsymbol{w}_{\boldsymbol{\alpha}} &\Rightarrow \sigma^2 = 64p\nu r,\ R = \sqrt{\frac{8\nu r}{n}}, \\
\bar{\boldsymbol{w}}_{\boldsymbol{\alpha}} = \boldsymbol{w}_{\boldsymbol{\alpha}},\ \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}} = \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} &\Rightarrow \sigma^2 = 16p\nu r,\ R = \sqrt{\frac{8\nu r}{n}}, \\
\bar{\boldsymbol{w}}_{\boldsymbol{\alpha}} = \boldsymbol{w}_{\boldsymbol{\alpha}},\ \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}} = \boldsymbol{w}_{\boldsymbol{\alpha}} &\Rightarrow \sigma^2 = 32pn,\ R = 4.
\end{aligned}$$

Setting $t = \sqrt{\frac{8}{3}\sigma^2 \beta \log n}$ for each expression, which uniformly satisfies $t \leq \frac{\sigma^2}{R}$ for each bound as long as $p \geq \frac{128\beta \log n}{3n}$, yields each of the desired inequalities with failure probabilities of $1 - 2n^{1-\beta}$ for each expression using Theorem 1.7.2. This concludes the proof. □

3.B.2 Proof of Theorem 3.2.1

We are now ready to prove Theorem 3.2.1.

Proof. First, notice that since $\mathcal{M}_\Omega = \mathcal{M}_\Omega^*$,

$$\begin{aligned}
\|\mathcal{P}_{\mathbb{T}} \mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}} - p^2 \mathcal{P}_{\mathbb{T}}\| &= \left\| \sum_{\boldsymbol{\alpha},\boldsymbol{\beta} \in \Omega} \xi_{\boldsymbol{\alpha}} \xi_{\boldsymbol{\beta}} C_{\boldsymbol{\alpha}\boldsymbol{\beta}} \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}} \rangle \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\beta}} \rangle - p^2 \mathcal{P}_{\mathbb{T}} \right\| \\
&= \left\| \sum_{\boldsymbol{\alpha},\boldsymbol{\beta} \in \Omega} \xi_{\boldsymbol{\alpha}} \xi_{\boldsymbol{\beta}} C_{\boldsymbol{\alpha}\boldsymbol{\beta}} \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}} \rangle \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\beta}} \rangle - p^2 \sum_{\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathbb{I}} \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}} \rangle \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\beta}} \rangle \right\| \\
&= \Bigg\| p \sum_{\boldsymbol{\alpha} \in \Omega} \xi_{\boldsymbol{\alpha}} \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\alpha}} \rangle + \sum_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta} \in \Omega \\ \boldsymbol{\alpha} \neq \boldsymbol{\beta}}} \xi_{\boldsymbol{\alpha}} \xi_{\boldsymbol{\beta}} \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}} \rangle \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\beta}} \rangle \\
&\qquad - p^2 \sum_{\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathbb{I}} \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}} \rangle \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\beta}} \rangle \Bigg\| \\
&= \left\| p \|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 \sum_{\boldsymbol{\alpha} \in \mathbb{I}} (\xi_{\boldsymbol{\alpha}} - p) \langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} + \sum_{\boldsymbol{\alpha} \neq \boldsymbol{\beta}} (\xi_{\boldsymbol{\alpha}} \xi_{\boldsymbol{\beta}} - p^2) \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\beta}} \rangle \right\|
\end{aligned}$$

$$
\begin{aligned}
&\leq \left\| p\|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 \sum_{\boldsymbol{\alpha}\in\mathbb{I}} (\xi_{\boldsymbol{\alpha}} - p)\langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}} \right\| + \left\| \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} (\xi_{\boldsymbol{\alpha}}\xi_{\boldsymbol{\beta}} - p^2)\mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\beta}}\rangle \right\| \\
&= p\|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 \, \|\mathcal{P}_{\mathbb{T}}\mathcal{F}_{\Omega}\mathcal{P}_{\mathbb{T}} - p\mathcal{P}_{\mathbb{T}}\mathcal{F}_{\mathbb{I}}\mathcal{P}_{\mathbb{T}}\| + \left\| \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} (\xi_{\boldsymbol{\alpha}}\xi_{\boldsymbol{\beta}} - p^2)\mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\beta}}\rangle \right\| \\
&\leq \underbrace{p\|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 \, \|\mathcal{P}_{\mathbb{T}}\mathcal{F}_{\Omega}\mathcal{P}_{\mathbb{T}} - p\mathcal{P}_{\mathbb{T}}\mathcal{F}_{\mathbb{I}}\mathcal{P}_{\mathbb{T}}\|}_{B_1} + \underbrace{\left\| \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} (\xi_{\boldsymbol{\alpha}} - p)(\xi_{\boldsymbol{\beta}} - p)\mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\beta}}\rangle \right\|}_{B_2} \\
&\quad + \underbrace{2p \left\| \sum_{\boldsymbol{\alpha}\in\mathbb{I}} (\xi_{\boldsymbol{\alpha}} - p)\, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}} \sum_{\substack{\boldsymbol{\beta}\in\mathbb{I} \\ \boldsymbol{\alpha}\neq\boldsymbol{\beta}}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\beta}}\rangle \right\|}_{B_3},
\end{aligned}
$$

where the first inequality follows from the triangle inequality and the final inequality follows from Theorem 3.B.4. We note that term B_1 is bounded by Theorem 3.B.2, and by the fact that $\|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}$ is constant for all $\boldsymbol{\alpha} \in \mathbb{I}$ and that $\|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 \leq \frac{1}{2}$ from Theorem 2.A.3. We bound B_2 using Theorem 3.B.7 and we bound B_3 using Theorem 3.B.8. This concludes the proof. □

Lemma 3.B.9. *Let $\Omega \subset \mathbb{I}$ be sampled with uniform Bernoulli probability p. If $p \geq \frac{128}{3}\frac{\beta \log n}{n}$ for any $\beta > 3$, then with probability at least $1 - C'n^{3-\beta}$, where C' is an absolute constant, we have that*

$$
\|\mathcal{M}_{\Omega}\| \leq p^2 \left(1 + 20\sqrt{\frac{\beta n \log n}{3p}}\right) + Cp\beta^{1/2} \log n,
$$

for some absolute constant $C > 0$. Furthermore, for $p \geq \frac{C''\beta \log n}{n}$ for some sufficiently large $C'' > 0$, this simplifies to

$$
\|\mathcal{M}_{\Omega}\| \leq 50p^{3/2}\sqrt{\beta n \log n}.
$$

Proof. This proof follows directly from Lemmas 3.B.2 and 3.B.4. First, notice that

$$
\|\mathcal{M}_{\Omega}\| = \|\mathcal{M}_{\Omega} - p^2\mathcal{I} + p^2\mathcal{I}\| \leq p^2 + \|\mathcal{M}_{\Omega} - p^2\mathcal{I}\|.
$$

This second term can be analyzed in the same way as in the proof of Theorem 3.2.1, seen in

Section 3.B.2:

$$\|\mathcal{M}_\Omega - p^2\mathcal{I}\| \leq p\|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2\|\mathcal{F}_\Omega - p\mathcal{F}_{\mathbb{I}}\| + \left\|\sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}}(\xi_{\boldsymbol{\alpha}}\xi_{\boldsymbol{\beta}} - p^2)\boldsymbol{w}_{\boldsymbol{\alpha}}H^{\boldsymbol{\alpha}\boldsymbol{\beta}}\langle\cdot,\boldsymbol{w}_{\boldsymbol{\beta}}\rangle\right\|.$$

Applying the results of Theorem 3.B.2, Theorem 3.B.4 , Theorem 3.B.7, and Theorem 3.B.8 gives the desired result. For the simplification, we notice that

$$p^2 \leq 5p^{3/2}\sqrt{\beta n\log n} \Leftrightarrow p \leq 25\beta n\log n,$$

and that

$$25p^{3/2}\sqrt{\beta n\log n} \geq Cp\beta^{1/2}\log n \Leftrightarrow p \geq \frac{C^2}{625}\frac{\log n}{n},$$

so for some sufficiently large $C' > 0$ the stated simplification holds. □

Lemma 3.B.10. *Let $\Omega \subset \mathbb{I}$ be sampled with uniform Bernoulli probability p, and let $\mathbb{T}$ be the tangent space on $\mathcal{N}_r$ for a rank-r, ν-incoherent ground truth matrix $\boldsymbol{X}$. If $p \geq \frac{128}{3}\frac{\beta\log n}{n}$ for any $\beta > 3$, then with probability at least $1 - C'n^{3-\beta}$, where C' is an absolute constant, we have that, for some absolute constant $C > 0$,*

$$\|\mathcal{M}_\Omega\mathcal{P}_{\mathbb{T}}\| \leq p^2 + Cp\sqrt{\frac{\nu r\beta}{n}}\log n + 20p^{3/2}\sqrt{\nu r\beta\log n}.$$

If $p \geq \frac{C''\log n}{n}$ for some sufficiently large constant $C'' > 0$ independent of n, ν and r, this simplifies to

$$\|\mathcal{M}_\Omega\mathcal{P}_{\mathbb{T}}\| \leq 50p^{3/2}\sqrt{\nu r\beta\log n}.$$

Furthermore, we have that

$$\|\mathcal{M}_\Omega\mathcal{P}_{\mathbb{T}_l}\| \leq 2\|\mathcal{M}_\Omega\|\frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\lambda_r(\boldsymbol{X})} + \|\mathcal{M}_\Omega\mathcal{P}_{\mathbb{T}}\|,$$

and with probability at least $1 - 14n^{3-\beta}$, for some sufficiently large constant $C''' > 0$ independent

of n, ν and r, if $p \geq \frac{C''' \log n}{n}$, then

$$\|\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l}\| \leq 100p^{3/2}\sqrt{\beta n \log n}\frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\lambda_r(\boldsymbol{X})} + 50p^{3/2}\sqrt{\beta \log n}.$$

Proof. We first notice that

$$\begin{aligned}\|\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}}\| &\leq \|\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}} - \mathbb{E}[\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}}]\| + \|\mathbb{E}[\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}}]\| \\ &= \|\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}} - \mathbb{E}[\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}}]\| + p^2.\end{aligned}$$

Similarly to the proofs of Theorem 3.2.1 and Theorem 3.B.9, we can decompose the difference between $\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}}$ and $\mathbb{E}[\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}}]$ as concentration of $\mathcal{F}_\Omega \mathcal{P}_{\mathbb{T}} - \mathbb{E}[\mathcal{F}_\Omega]\mathcal{P}_{\mathbb{T}}$ and the off-diagonal term. As such, we can see that

$$\|\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}}\| \leq p^2 + p\|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 \|\mathcal{F}_\Omega \mathcal{P}_{\mathbb{T}} - p\mathcal{F}_{\mathbb{I}}\mathcal{P}_{\mathbb{T}}\| + \left\| \sum_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathbb{I} \\ \boldsymbol{\alpha} \neq \boldsymbol{\beta}}} (\xi_{\boldsymbol{\alpha}}\xi_{\boldsymbol{\beta}} - p^2)\boldsymbol{w}_{\boldsymbol{\alpha}} \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}} \rangle \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\beta}} \rangle \right\|.$$

From Theorem 3.B.3, Theorem 3.B.7, and Theorem 3.B.8, the first result follows. The simplification follows a similar strategy as in Theorem 3.B.9. For the last result, notice that

$$\begin{aligned}\|\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l}\| &= \|\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l} - \mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}} + \mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}}\| \\ &\leq \|\mathcal{M}_\Omega (\mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_{\mathbb{T}})\| + \|\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}}\| \\ &\leq \|\mathcal{M}_\Omega\| \frac{2\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\lambda_r(\boldsymbol{X})} + \|\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}}\|.\end{aligned}$$

For the final result, the result follows from using the simplified expressions for both $\|\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}}\|$ and $\|\mathcal{M}_\Omega\|$, picking a sufficiently large constant such that both simplifications hold. □

Lemma 3.B.11 (Local RIP of $\mathcal{M}_\Omega$)**.** *Assume that*

$$p^{-2}\|\mathcal{P}_{\mathbb{T}}\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}} - p^2 \mathcal{P}_{\mathbb{T}}\| \leq \varepsilon_0, \tag{3.12}$$

$$\|\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}}\| \leq 50p^{3/2}\sqrt{\beta \nu r \log n}, \tag{3.13}$$

$$\|\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l}\| \leq 100p^{3/2}\sqrt{\beta n \log n}\frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\lambda_r(\boldsymbol{X})} + 50p^{3/2}\sqrt{\beta \nu r \log n}, \tag{3.14}$$

$$\frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\lambda_r(\boldsymbol{X})} \leq \frac{\varepsilon_0 p^{1/2}}{100\,(\beta n \log n)^{1/4}}. \tag{3.15}$$

Then

$$p^{-2}\|\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega\mathcal{P}_{\mathbb{T}_l} - p^2\mathcal{P}_{\mathbb{T}_l}\| \leq 4\varepsilon_0.$$

Proof.

$$\begin{aligned}
\|\mathcal{P}_{\mathbb{T}_l} - p^{-2}\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega\mathcal{P}_{\mathbb{T}_l}\| &= \|\mathcal{P}_{\mathbb{T}_l} - p^{-2}\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega\mathcal{P}_{\mathbb{T}_l} \\
&\quad + \mathcal{P}_{\mathbb{T}} - \mathcal{P}_{\mathbb{T}} + p^{-2}\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega\mathcal{P}_{\mathbb{T}} - p^{-2}\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega\mathcal{P}_{\mathbb{T}} + p^{-2}\mathcal{P}_{\mathbb{T}}\mathcal{M}_\Omega\mathcal{P}_{\mathbb{T}} - p^{-2}\mathcal{P}_{\mathbb{T}}\mathcal{M}_\Omega\mathcal{P}_{\mathbb{T}}\| \\
&\leq \|\mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_{\mathbb{T}}\| + p^{-2}\|\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega\mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega\mathcal{P}_{\mathbb{T}}\| + p^{-2}\|\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega\mathcal{P}_{\mathbb{T}} - \mathcal{P}_{\mathbb{T}}\mathcal{M}_\Omega\mathcal{P}_{\mathbb{T}}\| \\
&\quad + \left\|\mathcal{P}_{\mathbb{T}} - p^{-2}\mathcal{P}_{\mathbb{T}}\mathcal{M}_\Omega\mathcal{P}_{\mathbb{T}}\right\| \\
&\leq \|\mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_{\mathbb{T}}\| + p^{-2}\|\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega\|\|\mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_{\mathbb{T}}\| \\
&\quad + p^{-2}\|\mathcal{M}_\Omega\mathcal{P}_{\mathbb{T}}\|\|\mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_{\mathbb{T}}\| + \left\|\mathcal{P}_{\mathbb{T}} - p^{-2}\mathcal{P}_{\mathbb{T}}\mathcal{M}_\Omega\mathcal{P}_{\mathbb{T}}\right\| \\
&\leq \frac{2\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\lambda_r(\boldsymbol{X})}\left(1 + p^{-2}\|\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega\| + p^{-2}\|\mathcal{M}_\Omega\mathcal{P}_{\mathbb{T}}\|\right) + \left\|\mathcal{P}_{\mathbb{T}} - p^{-2}\mathcal{P}_{\mathbb{T}}\mathcal{M}_\Omega\mathcal{P}_{\mathbb{T}}\right\| \\
&\leq 2\varepsilon_0 + p^{-2}\frac{2\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\lambda_r(\boldsymbol{X})}\left(100p^{3/2}\sqrt{\beta n \log n}\frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\lambda_r(\boldsymbol{X})} + 100p^{3/2}\sqrt{\beta\nu r \log n}\right) \\
&= 2\varepsilon_0 + 200p^{-1/2}\sqrt{\beta n \log n}\left(\frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\lambda_r(\boldsymbol{X})}\right)^2 + 100p^{-1/2}\sqrt{\beta\nu r \log n}\frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\lambda_r(\boldsymbol{X})} \\
&\leq 4\varepsilon_0,
\end{aligned}$$

where the first inequality is the triangle inequality, the second inequality is Cauchy-Schwarz, the third inequality is due to Theorem 1.8.3 and (3.12), the fourth inequality is a result of (3.13) and (3.14), and the final inequality is due to (3.15). □

3.C Local Convergence Results

We begin with the following technical lemmas used in the proof of local convergence.

Lemma 3.C.1 (Algorithm 7 Stepsize Bounds)**.** *Assume that* $\|\mathcal{P}_{\mathbb{T}_l} - p^{-2}\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega\mathcal{P}_{\mathbb{T}_l}\| \leq 4\varepsilon_0 < 1$. *Then the stepsize* α_l *in DBRE can be bounded by*

$$\frac{p^{-2}}{1 + 4\varepsilon_0} \leq \alpha_l = \frac{\|\mathcal{P}_{\mathbb{T}}\boldsymbol{G}_l\|_{\mathrm{F}}^2}{\langle \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l, \mathcal{M}_\Omega\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l\rangle} \leq \frac{p^{-2}}{1 - 4\varepsilon_0}.$$

Proof. We will prove this by leveraging the local RIP assumption. Notice the following:

$$\begin{aligned}\langle \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l, \mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l\rangle &= \langle \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l, \mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l\rangle \\ &= \left\langle \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l, \mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l - p^2\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l\right\rangle + p^2\langle \mathcal{P}_{\mathbb{T}}\boldsymbol{G}_l, \mathcal{P}_{\mathbb{T}}\boldsymbol{G}_l\rangle.\end{aligned}$$

We can now leverage the variational characterization of the spectral norm and local RIP, proven in Theorem 3.B.11, to bound the following:

$$-p^2(4\varepsilon_0)\|\mathcal{P}_{\mathbb{T}}\boldsymbol{G}_l\|_{\mathrm{F}}^2 \le \left\langle \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l, \mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l - p^2\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l\right\rangle \le p^2(4\varepsilon_0)\|\mathcal{P}_{\mathbb{T}}\boldsymbol{G}_l\|_{\mathrm{F}}^2.$$

As such, we can now bound the denominator as

$$p^2(1-4\varepsilon_0)\|\mathcal{P}_{\mathbb{T}}\boldsymbol{G}_l\|_{\mathrm{F}}^2 \le \langle \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l, \mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l\rangle \le p^2(1+4\varepsilon_0)\|\mathcal{P}_{\mathbb{T}}\boldsymbol{G}_l\|_{\mathrm{F}}^2.$$

Rearrangement of this last expression yields the upper and lower bounds on the step size derived above. □

Lemma 3.C.2 (I_1 Bound)**.** *Assume* $\left\|\mathcal{P}_{\mathbb{T}_l} - p^{-2}\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l}\right\| \le 4\varepsilon_0$ *and* α_l *can be bounded as in Theorem 3.C.1. Then the spectral norm of* $\mathcal{P}_{\mathbb{T}_l} - \alpha_l \mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l}$ *can be bounded as*

$$\|\mathcal{P}_{\mathbb{T}_l} - \alpha_l \mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l}\| \le \frac{8\varepsilon_0}{1-4\varepsilon_0}. \tag{3.16}$$

Proof. From direct calculation, it follows that

$$\begin{aligned}\|\mathcal{P}_{\mathbb{T}_l} - \alpha_l \mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l}\| &\le \left\|\mathcal{P}_{\mathbb{T}_l} - p^{-2}\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l}\right\| + \left|\alpha_l - p^{-2}\right| \|\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l}\| \\ &\le 4\varepsilon_0 + \left|\alpha_l - p^{-2}\right| \left(\left\|\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l} - p^2\mathcal{P}_{\mathbb{T}_l}\right\| + p^2\|\mathcal{P}_{\mathbb{T}_l}\|\right) \\ &\le 4\varepsilon_0 + \left(\frac{p^{-2}}{1-4\varepsilon_0} - \frac{p^{-2}(1-4\varepsilon_0)}{1-4\varepsilon_0}\right)\left(\left\|\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l} - p^2\mathcal{P}_{\mathbb{T}_l}\right\| + p^2\|\mathcal{P}_{\mathbb{T}_l}\|\right) \\ &\le 4\varepsilon_0 + \left(\frac{p^{-2}}{1-4\varepsilon_0} - \frac{p^{-2}(1-4\varepsilon_0)}{1-4\varepsilon_0}\right)\left(4\varepsilon_0 p^2 + p^2\right) \\ &= 4\varepsilon_0 + \frac{4\varepsilon_0}{1-4\varepsilon_0}(1+4\varepsilon_0) \\ &= \frac{8\varepsilon_0}{1-4\varepsilon_0},\end{aligned}$$

where the first inequality comes from the triangle inequality, the second inequality comes from Local RIP in Theorem 3.B.11, the third inequality comes from the stepsize bound in Theorem 3.C.1, the fourth inequality again comes from Theorem 3.B.11, and the remainder comes from algebraic simplification of terms. This finishes the proof. □

3.C.1 Proof of Theorem 3.2.2

We can now prove Theorem 3.2.2.

Proof. First, it follows that

$$\|\boldsymbol{X}_{l+1} - \boldsymbol{X}\|_{\mathrm{F}} \leq \|\boldsymbol{X}_{l+1} - \boldsymbol{W}_l\|_{\mathrm{F}} + \|\boldsymbol{W}_l - \boldsymbol{X}\|_{\mathrm{F}} \leq 2\|\boldsymbol{W}_l - \boldsymbol{X}\|_{\mathrm{F}},$$

as $\boldsymbol{X}_{l+1}$ is the best rank-r approximation of $\boldsymbol{W}_l$. Plugging in $\boldsymbol{W}_l = \boldsymbol{X}_l + \alpha_l \mathcal{P}_{\mathbb{T}_l} \boldsymbol{G}_l$, we see that

$$\begin{aligned}
\|\boldsymbol{X}_{l+1} - \boldsymbol{X}\|_{\mathrm{F}} &\leq 2\left\|\boldsymbol{X}_l + \alpha_l \mathcal{P}_{\mathbb{T}_l} \boldsymbol{G}_l - \boldsymbol{X}\right\|_{\mathrm{F}} \\
&= 2\|\boldsymbol{X}_l - \boldsymbol{X} - \alpha_l \mathcal{P}_{\mathbb{T}_l} \mathcal{M}_\Omega(\boldsymbol{X}_l - \boldsymbol{X})\|_{\mathrm{F}} \\
&\leq \underbrace{2\|(\mathcal{P}_{\mathbb{T}_l} - \alpha_l \mathcal{P}_{\mathbb{T}_l} \mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l})(\boldsymbol{X}_l - \boldsymbol{X})\|_{\mathrm{F}}}_{I_1} \\
&\quad + \underbrace{2\|(\mathcal{I} - \mathcal{P}_{\mathbb{T}_l})(\boldsymbol{X}_l - \boldsymbol{X})\|_{\mathrm{F}}}_{I_2} \\
&\quad + \underbrace{2|\alpha_l| \|\mathcal{P}_{\mathbb{T}_l} \mathcal{M}_\Omega (\mathcal{I} - \mathcal{P}_{\mathbb{T}_l})(\boldsymbol{X}_l - \boldsymbol{X})\|_{\mathrm{F}}}_{I_3}.
\end{aligned}$$

It remains to bound each term individually. Using Theorem 3.C.2, we see that

$$I_1 \leq \frac{16\varepsilon_0}{1 - 4\varepsilon_0} \|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}.$$

Next, notice that from Theorem 1.8.3 and the fact that $\mathcal{P}_{\mathbb{T}_l} \boldsymbol{X}_l = \boldsymbol{X}_l$,

$$\begin{aligned}
I_2 &= 2\|(\mathcal{I} - \mathcal{P}_{\mathbb{T}_l})\boldsymbol{X}_l - (\mathcal{I} - \mathcal{P}_{\mathbb{T}_l})\boldsymbol{X}\|_{\mathrm{F}} \\
&= 2\|(\mathcal{I} - \mathcal{P}_{\mathbb{T}_l})\boldsymbol{X}\|_{\mathrm{F}} \\
&\leq \frac{2\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}^2}{\lambda_r(\boldsymbol{X})}
\end{aligned}$$

$$\leq \frac{\varepsilon_0 p^{1/2}}{32\left(\beta n \log n\right)^{1/4}} \|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}$$
$$\leq \varepsilon_0 \|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}$$
$$\leq \frac{\varepsilon_0}{1-4\varepsilon_0} \|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}},$$

using Theorem 1.8.3 and our initial local neighborhood assumption. Finally, we see that, following a similar argument as in the bound of I_2,

$$\begin{aligned}
I_3 &\leq 2|\alpha_l| \|\mathcal{P}_{\mathbb{T}_l} \mathcal{M}_\Omega\| \|(\mathcal{I} - \mathcal{P}_{\mathbb{T}_l})\boldsymbol{X}\|_{\mathrm{F}} \\
&\leq \frac{2p^{-2}}{1-4\varepsilon_0} \Bigg[100 p^{3/2} \sqrt{\beta n \log n} \frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\lambda_r(\boldsymbol{X})} \\
&\qquad + 25 p^{3/2} \sqrt{\beta \nu r \log n} \Bigg] \left(\frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\lambda_r(\boldsymbol{X})} \right) \|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}} \\
&\leq \frac{2}{1-4\varepsilon_0} \Bigg[100 p^{-1/2} \sqrt{\beta n \log n} \left(\frac{\varepsilon_0 p^{1/2}}{100\left(\beta n \log n\right)^{1/4}} \right)^2 \\
&\qquad + 50 p^{-1/2} \sqrt{\beta \nu r \log n} \frac{\varepsilon_0 p^{1/2}}{100\left(\beta n \log n\right)^{1/4}} \Bigg] \|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}} \\
&\leq \frac{\varepsilon_0}{1-4\varepsilon_0} \|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}},
\end{aligned}$$

where the second to last inequality follows from the same analysis conducted in Theorem 3.B.11, just divided by 2. Collecting these results, we get

$$\|\boldsymbol{X}_{l+1} - \boldsymbol{X}\|_{\mathrm{F}} \leq \frac{18\varepsilon_0}{1-4\varepsilon_0} \|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}.$$

By the assumption of the theorem, which holds for $l = 0$, and as we have a contractive sequence, it inductively follows that the assumption holds for $l \geq 0$. This concludes the proof. □

3.D Initialization Results (Proof of Lemma 3.2.3)

Proof. First, notice that for $\boldsymbol{W}_0 = p^{-1}\mathcal{R}_\Omega(\boldsymbol{X})$, we get

$$\begin{aligned}
\|\boldsymbol{X}_0 - \boldsymbol{X}\| &\leq \|\boldsymbol{W}_0 - \boldsymbol{X}\| + \|\boldsymbol{W}_0 - \boldsymbol{X}_0\| \\
&\leq 2\|\boldsymbol{W}_0 - \boldsymbol{X}\|,
\end{aligned}$$

where the first inequality follows from the triangle inequality and the second inequality follows from the fact that $\boldsymbol{W}_0$ is the best rank-r approximation of $\boldsymbol{X}_0$ by Eckart-Young-Mirsky[EY36]. We now need a bound for this last term. Notice that $\boldsymbol{W}_0 - \boldsymbol{X} = \sum_{\boldsymbol{\alpha}\in\mathbb{I}}(p^{-1}\xi_{\boldsymbol{\alpha}} - 1)\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \boldsymbol{v}_{\boldsymbol{\alpha}}$ is a sum of zero-mean i.i.d. random matrices, opening up use of Bernstein's inequality. In order to use this, define $\boldsymbol{Z}_{\boldsymbol{\alpha}} = (p^{-1}\xi_{\boldsymbol{\alpha}} - 1)\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \boldsymbol{v}_{\boldsymbol{\alpha}}$. We need a bound on $\|\boldsymbol{Z}_{\boldsymbol{\alpha}}\|$ and $\left\|\mathbb{E}\left[\sum_{\boldsymbol{\alpha}\in\mathbb{I}} \boldsymbol{Z}_{\boldsymbol{\alpha}}\right]^2\right\|$. First, notice that

$$\begin{aligned}
\|\boldsymbol{Z}_{\boldsymbol{\alpha}}\| &= \left\|(p^{-1}\xi_{\boldsymbol{\alpha}} - 1)\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \boldsymbol{v}_{\boldsymbol{\alpha}}\right\| \\
&\leq (p^{-1} + 1)|\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle| \, \|\boldsymbol{v}_{\boldsymbol{\alpha}}\| \\
&\leq 2p^{-1}\left(\max_{\boldsymbol{\alpha}\in\mathbb{I}} |\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle|\right) =: c,
\end{aligned}$$

where the second inequality comes from the fact that $p \leq 1$ and $\|\boldsymbol{v}_{\boldsymbol{\alpha}}\| < 1$ from Theorem 2.A.3. Next, notice that

$$\begin{aligned}
\left\|\mathbb{E}\left[\sum_{\boldsymbol{\alpha}\in\mathbb{I}} \boldsymbol{Z}_{\boldsymbol{\alpha}}^2\right]\right\| &= \left\|\sum_{\boldsymbol{\alpha}\in\mathbb{I}} \mathbb{E}\left[p^{-2}\xi_{\boldsymbol{\alpha}} - 2\xi_{\boldsymbol{\alpha}}p^{-1} + 1\right] \langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \boldsymbol{v}_{\boldsymbol{\alpha}}^2\right\| \\
&= \left\|\sum_{\boldsymbol{\alpha}\in\mathbb{I}} \left(p^{-1} - 1\right) \langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \boldsymbol{v}_{\boldsymbol{\alpha}}^2\right\| \\
&\leq p^{-1}\left(\max_{\boldsymbol{\alpha}\in\mathbb{I}} |\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle|\right)^2 \lambda_{\max}\left(\sum_{\boldsymbol{\alpha}} \boldsymbol{v}_{\boldsymbol{\alpha}}^2\right).
\end{aligned}$$

Now, as Theorem 2.A.4, $\sum_{\boldsymbol{\alpha}} \boldsymbol{v}_{\boldsymbol{\alpha}}^2 = \frac{n^2-2n+2}{4n}\boldsymbol{J}$. It follows that $\lambda_{\max}\left(\sum_{\boldsymbol{\alpha}} \boldsymbol{v}_{\boldsymbol{\alpha}}^2\right) = \frac{n^2-2n+2}{4n} \leq \frac{n}{4}$ as $\boldsymbol{J}$ is an orthogonal projection matrix. Thus,

$$\left\|\mathbb{E}\left[\sum_{\boldsymbol{\alpha}\in\mathbb{I}} \boldsymbol{Z}_{\boldsymbol{\alpha}}^2\right]\right\| \leq \frac{np^{-1}}{4}\left(\max_{\boldsymbol{\alpha}\in\mathbb{I}} |\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle|\right) =: \sigma^2.$$

Now to determine t, we note that

$$\begin{aligned}
\frac{\sigma^2}{c} &= \frac{np^{-1}\left(\max_{\boldsymbol{\alpha}\in\mathbb{I}} |\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle|\right)^2}{8p^{-1}\left(\max_{\boldsymbol{\alpha}\in\mathbb{I}} |\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle|\right)} \\
&= \frac{n}{8}\left(\max_{\boldsymbol{\alpha}\in\mathbb{I}} |\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle|\right) \\
&\geq \sqrt{\frac{2\beta n \log n}{3p}}\left(\max_{\boldsymbol{\alpha}\in\mathbb{I}} |\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle|\right),
\end{aligned}$$

for $p \geq \frac{128\beta \log n}{3n}$. It follows that

$$\mathbb{P}\left(\|\boldsymbol{X}_0 - \boldsymbol{X}\| > \sqrt{\frac{2\beta n \log n}{3p}}\left(\max_{\boldsymbol{\alpha}\in\mathbb{I}} |\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle|\right)\right) \leq 2n\exp\left(-\beta\log(n)\right)$$
$$= 2n^{1-\beta},$$

verifying the probabilistic bound. To complete the proof, we use Theorem 1.A.3, from which it follows that

$$\|\boldsymbol{X}_0 - \boldsymbol{X}\|_{\mathrm{F}} \leq \sqrt{2r}\|\boldsymbol{X}_0 - \boldsymbol{X}\| \leq \sqrt{\frac{2\beta n r \log n}{3p}}\left(\max_{\boldsymbol{\alpha}\in\mathbb{I}} |\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle|\right) \leq \sqrt{\frac{\beta\nu^2 r^3 \log(n)}{24pn}}\|\boldsymbol{X}\|.$$

This concludes the proof. □

3.E Robustness Guarantees

In this section, we will prove Theorem 3.3.1. We first control the perturbation induced by additive noise.

Lemma 3.E.1. *Let $\hat{\boldsymbol{P}} = \boldsymbol{P} + \boldsymbol{N}$ where $\boldsymbol{N} \in \mathbb{R}^{n\times r}$ is a matrix with independent mean-zero entries, and $\hat{\boldsymbol{X}} = \hat{\boldsymbol{P}}\hat{\boldsymbol{P}}^\top$. Let $\lambda_1 \geq \cdots \geq \lambda_r > 0$ be the non-zero eigenvalues of $\boldsymbol{X}$ with corresponding eigenvectors $\boldsymbol{u}^i$. Assume that $\|\boldsymbol{N}\|_\infty \leq C$ for $C > 0$. Additionally, let $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{n}^i \boldsymbol{n}^{i\top}]$ be the covariance matrix of the columns of $\boldsymbol{N}$. Then*

$$\|\boldsymbol{X} - \hat{\boldsymbol{X}}\| \leq C\left(\frac{16}{3}n^{1/2}\beta\lambda_1^{1/2}\log n\right) + nrC^2,$$

with probability at least $1 - 2n^{1-\beta}$.

Proof. First, notice that

$$\hat{\boldsymbol{P}}\hat{\boldsymbol{P}}^\top = \boldsymbol{P}\boldsymbol{P}^\top + \boldsymbol{N}\boldsymbol{P}^\top + \boldsymbol{P}\boldsymbol{N}^\top + \boldsymbol{N}\boldsymbol{N}^\top,$$

so

$$\|\boldsymbol{X} - \hat{\boldsymbol{X}}\| = \|\boldsymbol{N}\boldsymbol{P}^\top + \boldsymbol{P}\boldsymbol{N}^\top + \boldsymbol{N}\boldsymbol{N}^\top\| \leq 2\|\boldsymbol{N}\boldsymbol{P}^\top\| + \|\boldsymbol{N}\boldsymbol{N}^\top\|.$$

We will first bound the term $\|\boldsymbol{N}\boldsymbol{N}^\top\| \le \|\boldsymbol{N}\boldsymbol{N}^\top\|_{\mathrm{F}}$. Notice that

$$\|\boldsymbol{N}\boldsymbol{N}^\top\|_{\mathrm{F}}^2 = \sum_{i,j=1}^{n} \left(\sum_{k=1}^{r} N_{ik} N_{jk} \right)^2 \le \sum_{i,j=1}^{n} \left(\sum_{k=1}^{r} C^2 \right)^2 = n^2 r^2 C^4,$$

where the inequality follows from the bound on $\|\boldsymbol{N}\|_\infty$. We now bound $\|\mathbf{N}\mathbf{P}^\top\|$. Since $\mathbf{P} = \mathbf{U}\boldsymbol{\Lambda}^{1/2}$, where $\mathbf{U}$ has orthonormal columns and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$, we have $\|\mathbf{p}^{(k)}\|_2^2 = \lambda_k$. We write $\mathbf{N}\mathbf{P}^\top$ as follows:

$$\mathbf{N}\mathbf{P}^\top = \sum_{k=1}^{r} \mathbf{n}^{(k)} \mathbf{p}^{(k)\top} = \sum_{k=1}^{r} \mathbf{Z}_k.$$

The matrices $\mathbf{Z}_k$ are independent, mean-zero, and rank one. Moreover,

$$\|\mathbf{Z}_k\| = \|\mathbf{n}^{(k)}\|_2 \, \|\mathbf{p}^{(k)}\|_2 \le C\sqrt{n\lambda_1} =: L.$$

We next compute the Bernstein variance parameter. Let $\boldsymbol{\Sigma}_k = \mathbb{E}\big[\mathbf{n}^{(k)}\mathbf{n}^{(k)\top}\big]$. As the entries of $\boldsymbol{N}$ are independent, $\boldsymbol{\Sigma} = \sum_k \boldsymbol{\Sigma}_k$ is diagonal, and as we have a bound on $\|\boldsymbol{N}\|_\infty$ it follows that

$$\|\boldsymbol{\Sigma}\| \le C^2.$$

and $\mathrm{Trace}(\boldsymbol{\Sigma}) = nC^2$. Therefore,

$$\left\| \sum_{k=1}^{r} \mathbb{E}\big[\mathbf{Z}_k \mathbf{Z}_k^\top\big] \right\| = \left\| \sum_{k=1}^{r} \lambda_k \boldsymbol{\Sigma}_k \right\| \le C^2 \sum_{k=1}^{r} \lambda_k \le rC^2\lambda_1,$$

while

$$\left\| \sum_{k=1}^{r} \mathbb{E}\big[\mathbf{Z}_k^\top \mathbf{Z}_k\big] \right\| = \left\| \sum_{k=1}^{r} \mathbf{p}^{(k)} \mathbf{p}^{(k)\top} \mathbb{E}\|\mathbf{n}^{(k)}\|_2^2 \right\| \le nC^2 \left\| \sum_{k=1}^{r} \mathbf{p}^{(k)} \mathbf{p}^{(k)\top} \right\| = nC^2\lambda_1.$$

Since $n \ge r$, we set $\sigma^2 = nC^2\lambda_1$. As

$$\frac{\sigma^2}{c} = \frac{nC^2\lambda_1}{n^{1/2}C\lambda_1^{1/2}} = n^{1/2}C\lambda_1^{1/2},$$

it follows that for $t = \frac{8}{3}n^{1/2}C\beta\lambda_1^{1/2}\log n$ we have that

$$\begin{aligned}\mathbb{P}\left[\left\|\boldsymbol{N}\boldsymbol{P}^\top\right\| \geq t\right] &\leq 2n\exp\left(\frac{-3t}{8c}\right)\\ &= 2n\exp\left(\frac{n^{1/2}C\beta\lambda_1^{1/2}\log n}{n^{1/2}C\lambda_1^{1/2}}\right)\\ &= 2n\exp\left(-\beta\log n\right).\end{aligned}$$

This finishes the proof. □

Next we will prove the following Lemma showing that bounded noise on the points does not change the incoherence of a Gram matrix substantially.

Lemma 3.E.2. *For $\hat{\boldsymbol{P}} = \boldsymbol{P} + \boldsymbol{N}$, where $\boldsymbol{N}$ is a mean-zero random matrix. Let $\boldsymbol{X} = \boldsymbol{P}\boldsymbol{P}^\top$ and $\hat{\boldsymbol{X}} = \hat{\boldsymbol{P}}\hat{\boldsymbol{P}}^\top$, and let $\lambda_1 \geq \cdots \geq \lambda_r > 0$ be the eigenvalues of $\boldsymbol{X}$. If $\|\boldsymbol{N}\|_\infty \leq C$ for some $C > 0$, where κ is the condition number of $\boldsymbol{X}$, then*

$$\left\|\mathcal{P}_{\hat{U}}\boldsymbol{w}_{\boldsymbol{\alpha}}\right\|_{\mathrm{F}} \leq \sqrt{\frac{\nu r}{2n}} + \frac{32}{3}C\frac{n^{1/2}}{\lambda_r^{1/2}(\boldsymbol{X})}\kappa^{1/2}\log n + \frac{2nr}{\lambda_r(\boldsymbol{X})}C^2,$$

with probability at least $1 - 2n^{1-\beta}$. Furthermore, if $C \leq \frac{3}{32\sqrt{2}}\frac{\nu^{1/2}}{n}\frac{\lambda_r(\boldsymbol{X})^{1/2}}{\beta\kappa^{1/2}\log n}$, then

$$\left\|\mathcal{P}_{\hat{U}}\boldsymbol{w}_{\boldsymbol{\alpha}}\right\|_{\mathrm{F}} \leq \frac{9\nu r}{2n}.$$

Proof. This result will follow from the classic Davis-Kahan $\sin\Theta$ Theorem, seen in Theorem 1.8.1. Let $\mathbb{U}$, $\hat{\mathbb{U}}$ be the subspace spanned by the columns of $\boldsymbol{U}, \hat{\boldsymbol{U}}$ respectively. First, as $\left\|\mathcal{P}_U - \mathcal{P}_{\hat{U}}\right\|_{\mathrm{F}} = \|\sin\Theta(\mathbb{U}, \hat{\mathbb{U}})\|_{\mathrm{F}}$ from [Bha13], we can see that

$$\left\|\mathcal{P}_U - \mathcal{P}_{\hat{U}}\right\|_{\mathrm{F}} = \|\sin\Theta(\mathbb{U}, \hat{\mathbb{U}})\|_{\mathrm{F}} \leq \frac{\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_{\mathrm{F}}}{\lambda_r}.$$

Next, notice that

$$\begin{aligned}\left\|\mathcal{P}_{\hat{U}}\boldsymbol{w}_{\boldsymbol{\alpha}}\right\|_{\mathrm{F}} &\leq \left\|\left(\mathcal{P}_{\hat{U}} - \mathcal{P}_U\right)\right\|\|\boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} + \|\mathcal{P}_U\boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}\\ &\leq \frac{2\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|}{\lambda_r(\boldsymbol{X})} + \sqrt{\frac{\nu r}{2n}}.\end{aligned}$$

The Lemma statement therefore comes from the bound in Theorem 3.E.1. The second expression comes as follows:

$$\begin{aligned}\sqrt{\frac{\nu r}{2n}} + \frac{32}{3} C \frac{n^{1/2}}{\lambda_r^{1/2}(\boldsymbol{X})} \kappa^{1/2} \beta \log n + \frac{2nr}{\lambda_r(\boldsymbol{X})} C^2 &\leq 2\sqrt{\frac{\nu r}{2n}} + \frac{\nu r}{n} \frac{9}{2^{11} \beta^2 \kappa \log^2 n} \\ &\leq 2\sqrt{\frac{\nu r}{2n}} + \frac{\nu r}{n} \\ &= 2\sqrt{\frac{\nu r}{2n}} + \nu^{1/2} \frac{\nu^{1/2} r}{n} \\ &\leq 2\sqrt{\frac{\nu r}{2n}} + \frac{n^{1/2}}{r^{1/2}} \frac{\nu^{1/2} r}{n} \\ &= 3\sqrt{\frac{\nu r}{2n}},\end{aligned}$$

where the first inequality comes from substituting in the stated bound on C, the second inequality follows from the fact that $\beta^2 \kappa \log^2 n \geq 1$, and the third inequality follows from the fact that $\nu \leq \frac{n}{r}$, thus concluding the proof. □

Chapter 4

Non-Uniform Sampling Approaches to EDMC

This chapter is dedicated to loosening the uniform sampling assumption of Question 1.3.1. In Chapter 2 and Chapter 3, the theoretical guarantees for recovery are highly sensitive to the coherence of the underlying point cloud. In this chapter, we will consider a new Bernoulli sampling scheme that will allow for the recovery of arbitrarily coherent datasets, and show that recovery is algorithmically possible under this sampling scheme. Much of this chapter is taken from a current work in progress, and will be available in full as a pre-print in the coming months.

4.1 A Leverage-Score based approach to LRMC

A natural observation one might have when investigating the EDMC problem in an LRMC framework is that the uniform sampling assumption, while theoretically nice and a decent starting place, is quite poor as a general mechanism for guaranteeing successful recovery. We have already seen that, presented in Section 1.6, the incoherence constant ν in the EDMC problem has a simple geometric interpretation as

$$\max_{\boldsymbol{\alpha}\in\mathbb{I}} (\boldsymbol{p}_i - \boldsymbol{p}_j)^\top (\boldsymbol{P}^\top \boldsymbol{P})^{-1} (\boldsymbol{p}_i - \boldsymbol{p}_j) \leq \frac{2\nu r}{n}.$$

If we consider the following datasets in Section 4.1, we note that the image on the left is

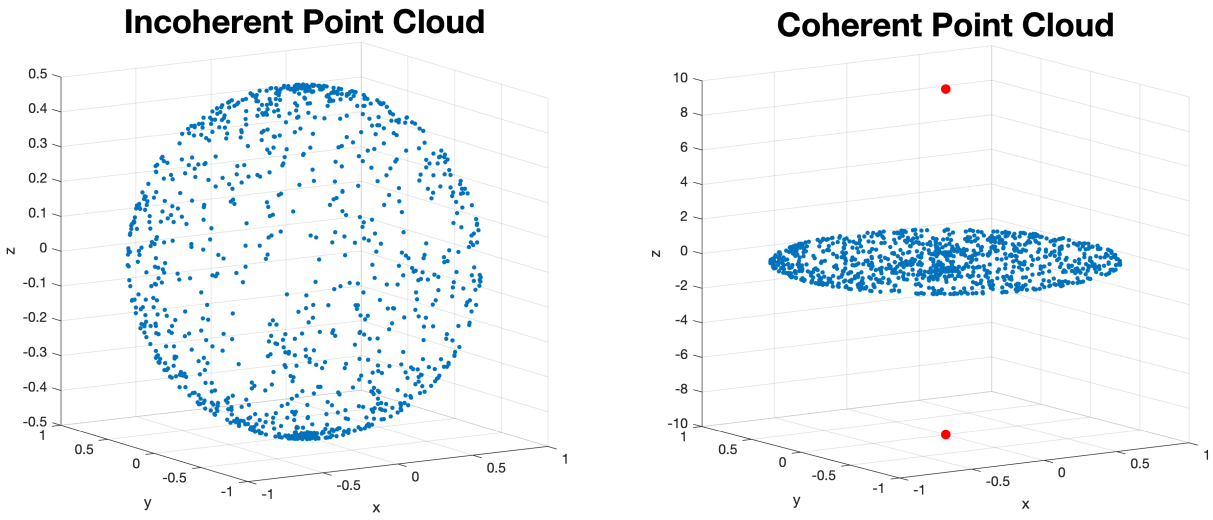


Figure 4.1: Incoherent and Coherent Datasets

highly incoherent, and the image on the right is highly coherent due to the outlier points in a low-variance direction. We note that, outside of the outlier points, the rest of the dataset is identical to the incoherent point cloud on the left.

This begs the following question:

Question 4.1.1: Leverage Score Motivating Question

If we prioritize sampling the distances that contribute the most to the incoherence constant ν, can we improve the sample complexity required to recover $\boldsymbol{P}$?

Although the incoherence constant is a maximum over all pairwise distances, for a coherent dataset there might be many distances that are close to the maximum value. In previous work such as [SCT25, GTK24, TL18], recovery guarantees for the EDMC problem are of the form $|\Omega| \geq Cn\nu^2 r^2 \log n$. This works well for data that has low incoherence numbers, and in [SCT25] it was shown that subgaussian distributions satisfy $\nu \leq \mathcal{O}(\kappa + \frac{K^2 \log n}{r^{1/2}})$, where κ is the condition number of the covariance matrix of the distribution and K is the subgaussian norm. However, this bound indicates that recovery of data drawn from either heavy-tailed distributions or ill-conditioned distributions, regardless of tail behavior, is an algorithmically challenging task. It is also difficult to recover datasets with outliers such as Section 4.1. The work done in this Chapter is dedicated to answer Question 4.1.1 in the affirmative: yes, sampling the "bad" distances more frequently will enable recovery with fewer pairwise distances needed.

4.2 Motivating Work in LRMC

In [CBSW14], the authors propose a sampling scheme that more heavily weights the sampling probabilities of entries that contribute more to a higher incoherence of the matrix. For simplicity, let $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ be a rank-r matrix with reduced SVD $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, and recall that the standard incoherence μ_0 is defined as the constant such that the following inequalities hold:

$$\max_i \|\boldsymbol{U}^\top \boldsymbol{e}_i\|_2^2 \leq \frac{\mu_0 r}{n}, \quad \max_j \|\boldsymbol{V}^\top \boldsymbol{e}_j\|_2^2 \leq \frac{\mu_0 r}{n}.$$

In the same vein as Question 4.1.1, we notice that certain matrices might have quite poor *global* incoherence constants, but only concentrated in a small handful of entries. To address this, the authors in [CBSW14] associate to each entry M_{ij} a *leverage score* μ_{ij} defined as

$$\mu_{ij} = \|\boldsymbol{U}^\top \boldsymbol{e}_i\|_2^2 + \|\boldsymbol{V}^\top \boldsymbol{e}_j\|_2^2.$$

We note that this notation is slightly different than what is provided in [CBSW14], but this is in order to highlight the parallels between our methodology and theirs. Specifically, framing the leverage score as a direct sum of the row and column subspace projection magnitudes clarifies how both frameworks independently isolate and quantify localized structural information to define the underlying sampling distribution. This then allowed the authors to consider a convex recovery approach with respect to this leverage-score based sampling strategy, which provided the following recovery guarantees: if

$$p_{ij} \geq \min\left\{c\frac{\mu_{ij}}{n}\log^2 n, 1\right\}, \quad p_{ij} \geq \frac{1}{n^{10}},$$

then with probability at least $1 - Cn^{-10}$, the nuclear norm minimization

$$\min_{\boldsymbol{Y}\in\mathbb{R}^{n\times n}} \|\boldsymbol{Y}\|_* \quad \text{subject to}\, \mathcal{P}_\Omega(\boldsymbol{Y}) = \mathcal{P}_\Omega(\boldsymbol{M}).$$

recovers $\boldsymbol{M}$. This paper captures the idea that if one can enforce a higher likelihood of sampling the "bad" entries, which are characterized by highly localized mass in the singular vectors, then the sample complexity required can be made substantially smaller than with uniform

sampling alone. Crucially, whereas standard uniform sampling approaches require a sample size bounded by the maximum global coherence parameter μ_0, which severely penalizes highly coherent matrices, this leverage-score strategy achieves exact recovery with an overall sample complexity on the order of $\mathcal{O}(rn \log^2 n)$, effectively decoupling the theoretical guarantees from the strict incoherence assumption.

4.3 Geometric Leverage Scores and Sampling Scheme

Motivated by [CBSW14] and the geometric interpretation of incoherence outlined in Section 1.6, we can define *geometric leverage scores*, denoted $\nu_{\boldsymbol{\alpha}}$ for each multi-index $\boldsymbol{\alpha}$, as follows:

$$\nu_{\boldsymbol{\alpha}} = \max\left\{\frac{15}{16}\|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 + \frac{1}{4}\|\mathcal{P}_U \boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2, \frac{4r}{n}\right\}. \tag{4.1}$$

This definition captures the geometric information encoded in $\|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2$ and $\|\mathcal{P}_U \boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2$ but still encourages sampling of points close together due to the maximum with $\frac{4r}{n}$ preventing degeneracy in the definition of $\nu_{\boldsymbol{\alpha}}$. This allows for arbitrarily close points to still have their pairwise distances sampled and recovered, which will be important for the technical analysis.

As such, we can define the following Bernoulli sampling scheme:

$$p_{\boldsymbol{\alpha}} = \min\left\{\frac{m\nu_{\boldsymbol{\alpha}}}{2nr}, 1\right\}, \tag{4.2}$$

where m is a surrogate for the total number of distances sampled to allow for more direct comparison to previous works in EDMC.

Now, as the number of samples $|\Omega| = \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \xi_{\boldsymbol{\alpha}}$, it follows that $\mathbb{E}\left[|\Omega|\right] = \sum_{\boldsymbol{\alpha}\in\mathbb{I}} p_{\boldsymbol{\alpha}}$. Most of the probabilities will be less than 1, but for very poor geometries it follows that some points are guaranteed to be sampled in this scheme. As such, we partition the index set into the disjoint union $\mathbb{I} = \mathbb{I}_p \sqcup \mathbb{I}_d$, where $\mathbb{I}_p = \{\boldsymbol{\alpha} \mid p_{\boldsymbol{\alpha}} < 1\}$ and $\mathbb{I}_d = \{\boldsymbol{\alpha} \mid p_{\boldsymbol{\alpha}} = 1\}$. We first prove that the size of $\mathbb{I}_d$ does not exceed m.

Lemma 4.3.1. *Let* $\mathbb{I}_d = \{\boldsymbol{\alpha} \mid p_{\boldsymbol{\alpha}} = 1\}$. *Then*

$$|\mathbb{I}_d| \leq m.$$

Proof. We first notice that from the proof of Theorem 1.6.1, $\sum_{\boldsymbol{\alpha}\in\mathbb{I}} \|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 = 2nr$. Furthermore, from the computation of $\|\mathcal{P}_U \boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2$ seen in Theorem 4.3.3, we have that

$$\|\mathcal{P}_U \boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 = \frac{1}{4}\left(\|\boldsymbol{u}_i\|_2^2 + \|\boldsymbol{u}_j\|_2^2\right) - \frac{1}{4n}\|\boldsymbol{u}_i + \boldsymbol{u}_j\|_2^2.$$

As such, we can compute that

$$\sum_{\boldsymbol{\alpha}\in\mathbb{I}} \left(\frac{15}{16}\|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 + \frac{1}{4}\|\mathcal{P}_U \boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2\right) \leq 2nr.$$

Now, by definition $\boldsymbol{\alpha} \in \mathbb{I}_d$ if and only if $p_{\boldsymbol{\alpha}} = 1$. From the definition of $p_{\boldsymbol{\alpha}}$, this implies:

$$\frac{m\nu_{\boldsymbol{\alpha}}}{2nr} \geq 1 \implies \nu_{\boldsymbol{\alpha}} \geq \frac{2nr}{m}.$$

Given that $m \leq L = \frac{n(n-1)}{2}$, we have that for all $\boldsymbol{\alpha} \in \mathbb{I}_d$, the geometric leverage score $\nu_{\boldsymbol{\alpha}}$ equals the sum of the projection norms:

$$\nu_{\boldsymbol{\alpha}} = \frac{15}{16}\|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 + \frac{1}{4}\|\mathcal{P}_U \boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2.$$

Summing $\nu_{\boldsymbol{\alpha}}$ strictly over the subset $\mathbb{I}_d$ yields the lower bound:

$$|\mathbb{I}_d|\frac{nr}{m} \leq \sum_{\boldsymbol{\alpha}\in\mathbb{I}_d} \nu_{\boldsymbol{\alpha}} = \sum_{\boldsymbol{\alpha}\in\mathbb{I}_d} \left(\frac{15}{16}\|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 + \frac{1}{4}\|\mathcal{P}_U \boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2\right).$$

Because $\mathbb{I}_d \subseteq \mathbb{I}$, the sum over $\mathbb{I}_d$ is upper-bounded by the sum over the entire index set $\mathbb{I}$:

$$|\mathbb{I}_d|\frac{2nr}{m} \leq \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \left(\frac{15}{16}\|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 + \frac{1}{4}\|\mathcal{P}_U \boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2\right).$$

Applying the fact that the sum over $\mathbb{I}$ is bounded by $2nr$, the inequality becomes:

$$|\mathbb{I}_d|\frac{2nr}{m} \leq 2nr.$$

Isolating $|\mathbb{I}_d|$ yields the required bound:

$$|\mathbb{I}_d| \leq m.$$

□

The next thing we must understand is how the cardinality of the probabilistic sampling set $\mathbb{I}_p$ concentrates around its expectation. We investigate this and state our result on the cardinality of $|\Omega|$ in the following Lemma.

Lemma 4.3.2. *Let $\mathbb{I}_p = \{\boldsymbol{\alpha} \mid p_{\boldsymbol{\alpha}} < 1\}$. Then with probability at least $1 - 2n^{-\beta}$, the following pair of inequalities holds if $m \geq \frac{16}{3}\beta \log n$:*

$$|\mathbb{I}_p| \leq 3m,$$
$$|\Omega| \leq 4m.$$

Proof. This is a straightforward application of Bernstein's inequality. As before, $|\mathbb{I}_p| = \sum_{\alpha \in \mathbb{I}_p} \xi_{\boldsymbol{\alpha}}$ and $\mathbb{E}\left[|\mathbb{I}_p|\right] = \sum_{\alpha \in \mathbb{I}_p} p_{\boldsymbol{\alpha}}$. Notice that

$$\begin{aligned}
\sum_{\alpha \in \mathbb{I}_p} \frac{m}{2nr} \max\left\{ \frac{15}{16} \|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 + \frac{1}{4} \|\mathcal{P}_U \boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2, \frac{4r}{n} \right\} & \\
&\leq \frac{m}{2nr} \sum_{\alpha \in \mathbb{I}_p} \left(\frac{15}{16} \|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 + \frac{1}{4} \|\mathcal{P}_U \boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 \right) + \frac{m}{2nr} \sum_{\alpha \in \mathbb{I}_p} \frac{4r}{n} \\
&\leq m + \frac{2m}{n^2} L \\
&\leq 2m.
\end{aligned}$$

Next, notice that $|\xi_{\boldsymbol{\alpha}} - p_{\boldsymbol{\alpha}}| < 1$ for all $\alpha \in \mathbb{I}_p$, and that

$$\begin{aligned}
\mathbb{E}\left[\sum_{\alpha \in \mathbb{I}_p} (\xi_{\boldsymbol{\alpha}} - p_{\boldsymbol{\alpha}})^2 \right] &= \sum_{\alpha \in \mathbb{I}_p} p_{\boldsymbol{\alpha}}(1 - p_{\boldsymbol{\alpha}}) \\
&\leq \sum_{\alpha \in \mathbb{I}_p} p_{\boldsymbol{\alpha}} \\
&= 2m.
\end{aligned}$$

The result follows from an application of Theorem 1.7.1 and the triangle inequality. □

As noted previously, the m in the definition of the sampling scheme is a surrogate for the total number of observed distances, as they differ by a small constant. To address this discrepancy, we report our sample complexity results in both m and in $|\Omega|$.

At this point, the astute reader might question the novelty and necessity of this methodology, especially when compared to the work done in [CBSW14]. We note that standard incoherence applied directly to the EDM alone does contain geometric information, outlined in Section 4.3.2, with this section also detailing that the global incoherence constants are tied to one another. As we will see in Section 4.6, the direct completion on $\boldsymbol{D}$ does work but this methodology still outperforms this method in this specific problem.

By defining a set of weighted probabilities, we can define a sequence of independent Bernoulli random variables as $\{\xi_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$ where

$$\xi_{\boldsymbol{\alpha}} = \begin{cases} 1 & p_{\boldsymbol{\alpha}}; \\ 0 & 1 - p_{\boldsymbol{\alpha}}. \end{cases}$$

This allows us to define a new sampling operator $\mathcal{M}_{p,\Omega} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ as

$$\mathcal{M}_{p,\Omega}(\cdot) = \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} \langle \cdot, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 \boldsymbol{w}_{\boldsymbol{\alpha}} + \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} \frac{\xi_{\boldsymbol{\alpha}}\xi_{\boldsymbol{\beta}}}{p_{\boldsymbol{\alpha}}p_{\boldsymbol{\beta}}} \boldsymbol{w}_{\boldsymbol{\alpha}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \boldsymbol{w}_{\boldsymbol{\beta}} \rangle. \tag{4.3}$$

We note that if $p_{\boldsymbol{\alpha}} = p$ for all $\boldsymbol{\alpha}$, this operator is identical to $p^{-2}\mathcal{M}_{\Omega}$.

4.3.1 Geometric Interpretation of Dual-Basis Geometric Leverage Scores

In Section 1.6, we provided a geometric interpretation for $\|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2$ that guided the discussion and interpretation of this technique. In this section, we will now do the same for $\|\mathcal{P}_U \boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2$.

Lemma 4.3.3. *Let $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ be the reduced eigenvalue decomposition of a Gram matrix $\boldsymbol{X}$, and let $\boldsymbol{\alpha} = (i, j)$. Then*

$$\|\mathcal{P}_U \boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 = \frac{1}{4}\left(\|\boldsymbol{u}_i\|_2^2 + \|\boldsymbol{u}_j\|_2^2\right) - \frac{1}{4n}\|\boldsymbol{u}_i + \boldsymbol{u}_j\|_2^2.$$

Proof. Consider the pair $\boldsymbol{\alpha} = (i, j)$ and define the vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$ as:

$$\begin{aligned}\boldsymbol{a} &= \boldsymbol{e}_i - \frac{1}{n}\mathbf{1},\\ \boldsymbol{b} &= \boldsymbol{e}_j - \frac{1}{n}\mathbf{1},\end{aligned}$$

where $\mathbf{1}$ is the all-ones vector. Let the symmetric matrix $\boldsymbol{v}_{\boldsymbol{\alpha}}$ be defined as:

$$\boldsymbol{v}_{\boldsymbol{\alpha}} = -\frac{1}{2}(\boldsymbol{a}\boldsymbol{b}^\top + \boldsymbol{b}\boldsymbol{a}^\top).$$

We wish to compute the squared Frobenius norm of the projected matrix $\mathcal{P}_U \boldsymbol{v}_{\boldsymbol{\alpha}}$:

$$\|\mathcal{P}_U \boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 = \mathrm{Trace}((\mathcal{P}_U \boldsymbol{v}_{\boldsymbol{\alpha}})^\top (\mathcal{P}_U \boldsymbol{v}_{\boldsymbol{\alpha}})) = \mathrm{Trace}(\boldsymbol{v}_{\boldsymbol{\alpha}} \mathcal{P}_U \boldsymbol{v}_{\boldsymbol{\alpha}}),$$

where we used the symmetry of $\boldsymbol{v}_{\boldsymbol{\alpha}}$ and the idempotence of $\mathcal{P}_U$. First, we determine the action of $\mathcal{P}_U$ on $\boldsymbol{a}$ and $\boldsymbol{b}$. Since $\boldsymbol{U}^\top \mathbf{1} = \mathbf{0}$, we have:

$$\begin{aligned}\mathcal{P}_U \boldsymbol{a} &= \boldsymbol{U}\boldsymbol{U}^\top \left(\boldsymbol{e}_i - \frac{1}{n}\mathbf{1}\right) = \boldsymbol{U}\boldsymbol{u}_i,\\ \mathcal{P}_U \boldsymbol{b} &= \boldsymbol{U}\boldsymbol{U}^\top \left(\boldsymbol{e}_j - \frac{1}{n}\mathbf{1}\right) = \boldsymbol{U}\boldsymbol{u}_j.\end{aligned}$$

This yields the following quadratic forms, as $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}$:

$$\begin{aligned}\boldsymbol{a}^\top \mathcal{P}_U \boldsymbol{a} &= \boldsymbol{u}_i^\top \boldsymbol{U}^\top \boldsymbol{U} \boldsymbol{u}_i = \|\boldsymbol{u}_i\|^2,\\ \boldsymbol{b}^\top \mathcal{P}_U \boldsymbol{b} &= \boldsymbol{u}_j^\top \boldsymbol{U}^\top \boldsymbol{U} \boldsymbol{u}_j = \|\boldsymbol{u}_j\|^2,\\ \boldsymbol{a}^\top \mathcal{P}_U \boldsymbol{b} &= \boldsymbol{u}_i^\top \boldsymbol{U}^\top \boldsymbol{U} \boldsymbol{u}_j = \boldsymbol{u}_i^\top \boldsymbol{u}_j.\end{aligned}$$

Next, we compute the standard scalar products of a and b (assuming $i \neq j$):

$$\begin{aligned}\boldsymbol{a}^\top \boldsymbol{a} &= \left(\boldsymbol{e}_i - \frac{1}{n}\mathbf{1}\right)^\top \left(\boldsymbol{e}_i - \frac{1}{n}\mathbf{1}\right) = 1 - \frac{1}{n},\\ \boldsymbol{b}^\top \boldsymbol{b} &= 1 - \frac{1}{n},\\ \boldsymbol{a}^\top \boldsymbol{b} &= \left(\boldsymbol{e}_i - \frac{1}{n}\mathbf{1}\right)^\top \left(\boldsymbol{e}_j - \frac{1}{n}\mathbf{1}\right) = -\frac{1}{n}.\end{aligned}$$

Substituting $\boldsymbol{v_\alpha}$ into the trace expression gives:

$$\begin{aligned}\operatorname{tr}(\boldsymbol{v_\alpha}\mathcal{P}_U\boldsymbol{v_\alpha}) &= \frac{1}{4}\operatorname{tr}\left((\boldsymbol{a}\boldsymbol{b}^\top + \boldsymbol{b}\boldsymbol{a}^\top)\mathcal{P}_U(\boldsymbol{a}\boldsymbol{b}^\top + \boldsymbol{b}\boldsymbol{a}^\top)\right) \\ &= \frac{1}{4}\Big[(\boldsymbol{a}^\top P\boldsymbol{a})(\boldsymbol{b}^\top\boldsymbol{b}) + (\boldsymbol{b}^\top P\boldsymbol{b})(\boldsymbol{a}^\top\boldsymbol{a}) + 2(\boldsymbol{a}^\top P\boldsymbol{b})(\boldsymbol{a}^\top\boldsymbol{b})\Big].\end{aligned}$$

Substituting the geometric and scalar terms derived above:

$$\begin{aligned}\|\mathcal{P}_U\boldsymbol{v_\alpha}\|_{\mathrm{F}}^2 &= \frac{1}{4}\left[\|\boldsymbol{u}_i\|^2\left(1-\frac{1}{n}\right) + \|\boldsymbol{u}_j\|^2\left(1-\frac{1}{n}\right) + 2(\boldsymbol{u}_i^\top\boldsymbol{u}_j)\left(-\frac{1}{n}\right)\right] \\ &= \frac{1}{4}\left[\left(\|\boldsymbol{u}_i\|^2 + \|\boldsymbol{u}_j\|^2\right) - \frac{1}{n}\left(\|\boldsymbol{u}_i\|^2 + \|\boldsymbol{u}_j\|^2 + 2\boldsymbol{u}_i^\top\boldsymbol{u}_j\right)\right].\end{aligned}$$

Recognizing the perfect square $\|\boldsymbol{u}_i + \boldsymbol{u}_j\|^2 = \|\boldsymbol{u}_i\|^2 + \|\boldsymbol{u}_j\|^2 + 2\boldsymbol{u}_i^\top\boldsymbol{u}_j$, we arrive at the final simplified result:

$$\|\mathcal{P}_U\boldsymbol{v_\alpha}\|_{\mathrm{F}}^2 = \frac{1}{4}\left(\|\boldsymbol{u}_i\|^2 + \|\boldsymbol{u}_j\|^2\right) - \frac{1}{4n}\|\boldsymbol{u}_i + \boldsymbol{u}_j\|^2.$$

□

4.3.2 Geometric Interpretation of Standard Leverage Scores

If we were to directly apply the method of [CBSW14] to the EDMC problem directly on the pairwise distance matrix $\boldsymbol{D}$, it would be beneficial to understand the geometric information encoded in the leverage scores. We do so in the following Lemma.

Lemma 4.3.4. *Let $\boldsymbol{P} \in \mathbb{R}^{n\times r}$ be the coordinate matrix of n points spanning $\mathbb{R}^r$, centered at the origin such that $\mathbf{1}^\top\boldsymbol{P} = \mathbf{0}$. Let $\boldsymbol{D} \in \mathbb{R}^{n\times n}$ be the corresponding Euclidean Distance Matrix, and let $\boldsymbol{D} = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ be its compact eigendecomposition. The statistical leverage score of the i-th row of $\boldsymbol{D}$, defined as $\ell_i = \|\boldsymbol{V}^\top\boldsymbol{e}_i\|_2^2$, admits the exact decomposition:*

$$\ell_i = \frac{1}{n} + \boldsymbol{p}_i^\top(\boldsymbol{P}^\top\boldsymbol{P})^{-1}\boldsymbol{p}_i + \frac{(v_i^\perp)^2}{\|\boldsymbol{v}^\perp\|_2^2}.$$

where $\boldsymbol{p}_i^\top$ is the i-th row of $\boldsymbol{P}$, $\boldsymbol{v} \in \mathbb{R}^n$ is the vector of squared Euclidean norms with $v_i = \|\boldsymbol{p}_i\|_2^2$, and $\boldsymbol{v}^\perp$ is the projection of $\boldsymbol{v}$ onto the orthogonal complement of the subspace spanned by $\mathbf{1}$ and the columns of $\boldsymbol{P}$.

Proof. The Euclidean Distance Matrix $\boldsymbol{D}$ can be factored algebraically into the sum of three rank-1 or rank-r matrices, first seen in (1.1):

$$\boldsymbol{D} = \boldsymbol{v}\mathbf{1}^\top + \mathbf{1}\boldsymbol{v}^\top - 2\boldsymbol{P}\boldsymbol{P}^\top.$$

It follows immediately that the column space of $\boldsymbol{D}$, denoted $\mathcal{R}(\boldsymbol{D})$, is a subspace of the span of the vectors generating these terms:

$$\mathcal{R}(\boldsymbol{D}) \subseteq \text{span}\{\mathbf{1}, \boldsymbol{P}, \boldsymbol{v}\}.$$

Assuming the points do not lie on a hypersphere centered at the origin (which would cause $\boldsymbol{v}$ to be linearly dependent on $\mathbf{1}$), $\boldsymbol{D}$ has rank $r+2$, and $\mathcal{R}(\boldsymbol{D}) = \text{span}\{\mathbf{1}, \boldsymbol{P}, \boldsymbol{v}\}$.

The matrix $\boldsymbol{V}\boldsymbol{V}^\top$ is the orthogonal projection operator onto $\mathcal{R}(\boldsymbol{D})$. The leverage score of the i-th row is the i-th diagonal element of this projection matrix:

$$\ell_i = \boldsymbol{e}_i^\top \boldsymbol{V}\boldsymbol{V}^\top \boldsymbol{e}_i.$$

To compute this diagonal element explicitly, we construct an orthogonal basis for $\mathcal{R}(\boldsymbol{D})$ by orthogonalizing the spanning set $\{\mathbf{1}, \boldsymbol{P}, \boldsymbol{v}\}$:

1. **The translation-invariant component:** The normalized all-ones vector provides the projection matrix $\frac{1}{n}\mathbf{1}\mathbf{1}^\top$.

2. **The spatial component:** By hypothesis, the point configuration is centered, meaning $\boldsymbol{P}^\top\mathbf{1} = \mathbf{0}$. Therefore, the column space of $\boldsymbol{P}$ is already strictly orthogonal to $\mathbf{1}$. The orthogonal projection matrix onto $\mathcal{R}(\boldsymbol{P})$ is given by $\boldsymbol{P}(\boldsymbol{P}^\top\boldsymbol{P})^{-1}\boldsymbol{P}^\top$. We note that this is equivalent to the projection operator onto the column space of $\boldsymbol{X}$.

3. **The radial component:** We extract the component of $\boldsymbol{v}$ that resides in the orthogonal complement of the subspace spanned by $\mathbf{1}$ and $\boldsymbol{P}$. Using the orthogonal projectors derived above, we define this residual vector $\boldsymbol{v}^\perp$:

$$\boldsymbol{v}^\perp = \left(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top - \boldsymbol{P}(\boldsymbol{P}^\top\boldsymbol{P})^{-1}\boldsymbol{P}^\top\right)\boldsymbol{v}.$$

The orthogonal projection matrix onto the one-dimensional subspace spanned by $\boldsymbol{v}^\perp$ is $\frac{\boldsymbol{v}^\perp(\boldsymbol{v}^\perp)^\top}{\|\boldsymbol{v}^\perp\|_2^2}$.

Because the subspaces spanned by $\mathbf{1}$, $\boldsymbol{P}$, and $\boldsymbol{v}^\perp$ are mutually orthogonal and their direct sum equals $\mathcal{R}(\boldsymbol{D})$, the total projection operator $\boldsymbol{V}\boldsymbol{V}^\top$ is the strict sum of the individual orthogonal projectors:

$$\boldsymbol{V}\boldsymbol{V}^\top = \frac{1}{n}\mathbf{1}\mathbf{1}^\top + \boldsymbol{P}(\boldsymbol{P}^\top\boldsymbol{P})^{-1}\boldsymbol{P}^\top + \frac{\boldsymbol{v}^\perp(\boldsymbol{v}^\perp)^\top}{\|\boldsymbol{v}^\perp\|_2^2}.$$

Evaluating the quadratic form $\boldsymbol{e}_i^\top(\boldsymbol{V}\boldsymbol{V}^\top)\boldsymbol{e}_i$ extracts the i-th diagonal element:

$$\ell_i = \frac{1}{n}\boldsymbol{e}_i^\top\mathbf{1}\mathbf{1}^\top\boldsymbol{e}_i + \boldsymbol{e}_i^\top\boldsymbol{P}(\boldsymbol{P}^\top\boldsymbol{P})^{-1}\boldsymbol{P}^\top\boldsymbol{e}_i + \frac{\boldsymbol{e}_i^\top\boldsymbol{v}^\perp(\boldsymbol{v}^\perp)^\top\boldsymbol{e}_i}{\|\boldsymbol{v}^\perp\|_2^2}.$$

Substituting the scalar identities $\mathbf{1}^\top\boldsymbol{e}_i = 1$, $\boldsymbol{e}_i^\top\boldsymbol{P} = \boldsymbol{p}_i^\top$, and $\boldsymbol{e}_i^\top\boldsymbol{v}^\perp = v_i^\perp$ yields the final identity:

$$\ell_i = \frac{1}{n} + \boldsymbol{p}_i^\top(\boldsymbol{P}^\top\boldsymbol{P})^{-1}\boldsymbol{p}_i + \frac{(v_i^\perp)^2}{\|\boldsymbol{v}^\perp\|_2^2}.$$

This completes the proof. □

We offer a geometric interpretation for each constituent term of the leverage score. The constant term, $\frac{1}{n}$, represents the uniform baseline influence intrinsic to any point within the configuration. The second term provides information on the displacement of the point $\boldsymbol{p}_i$ relative to the variance of the original point cloud $\boldsymbol{P}$, in terms of the principal components. This is the Mahalanobis distance, first discussed in Section 1.6, and it captures the spatial displacement relative to the empirical covariance structure of the point cloud $\boldsymbol{P}$.

The radial component evaluates the structural deviation of the squared Euclidean norms. Notice that if all points $\boldsymbol{p}_i$ reside on an origin-centered hypersphere in $\mathbb{R}^r$, the vector of squared norms is uniform, such that $\boldsymbol{v} = c\mathbf{1}$ for some $c > 0$. The Euclidean distance matrix then simplifies to $\boldsymbol{D} = 2c\mathbf{1}\mathbf{1}^\top - 2\boldsymbol{P}\boldsymbol{P}^\top$. This linear dependence restricts the matrix to $\operatorname{rank}(\boldsymbol{D}) = r+1$, causing the orthogonal projection of the radial variance to vanish identically. Consequently, the radial component of ℓ_i measures the magnitude by which the squared norm of $\boldsymbol{p}_i$ deviates from the idealized geometry of a concentric spherical configuration.

Lemma 4.3.5. *Let $\boldsymbol{P} \in \mathbb{R}^{n\times r}$ be a centered coordinate matrix such that $\mathbf{1}^\top \boldsymbol{P} = \mathbf{0}$, and let $M_{ij}^2 = (\boldsymbol{p}_i - \boldsymbol{p}_j)^\top (\boldsymbol{P}^\top \boldsymbol{P})^{-1}(\boldsymbol{p}_i - \boldsymbol{p}_j)$ denote the squared pairwise Mahalanobis distance. The distance M_{ij}^2 is linked to the Euclidean Distance Matrix leverage scores ℓ_i and ℓ_j via the exact decomposition:*

$$M_{ij}^2 = \ell_i + \ell_j - \frac{2}{n} - (R_i + R_j) - 2Q_{ij},$$

where $R_i \geq 0$ is the radial variance penalty and $Q_{ij} = \boldsymbol{p}_i^\top (\boldsymbol{P}^\top \boldsymbol{P})^{-1}\boldsymbol{p}_j$ is the spatial correlation between the points.

Proof. Let M_{ij}^2 denote the squared pairwise Mahalanobis distance between points $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$, which from Section 1.6 is equivalent to $\frac{1}{2}\|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2$ for $\boldsymbol{\alpha} = (i, j)$. Expanding the quadratic form yields:

$$M_{ij}^2 = (\boldsymbol{p}_i - \boldsymbol{p}_j)^\top (\boldsymbol{P}^\top \boldsymbol{P})^{-1}(\boldsymbol{p}_i - \boldsymbol{p}_j),$$

$$M_{ij}^2 = \boldsymbol{p}_i^\top (\boldsymbol{P}^\top \boldsymbol{P})^{-1}\boldsymbol{p}_i + \boldsymbol{p}_j^\top (\boldsymbol{P}^\top \boldsymbol{P})^{-1}\boldsymbol{p}_j - 2\boldsymbol{p}_i^\top (\boldsymbol{P}^\top \boldsymbol{P})^{-1}\boldsymbol{p}_j.$$

Recognizing the constituent terms as the elements of the projection matrix $\boldsymbol{Q} = \boldsymbol{P}(\boldsymbol{P}^\top \boldsymbol{P})^{-1}\boldsymbol{P}^\top$, we substitute Q_{ii}, Q_{jj}, and Q_{ij}:

$$M_{ij}^2 = Q_{ii} + Q_{jj} - 2Q_{ij}.$$

By Theorem 4.3.4, the leverage score for the i-th row is $\ell_i = \frac{1}{n} + Q_{ii} + R_i$. Isolating the spatial projection term provides $Q_{ii} = \ell_i - \frac{1}{n} - R_i$. Substituting this identity for both Q_{ii} and Q_{jj} gives:

$$M_{ij}^2 = \left(\ell_i - \frac{1}{n} - R_i\right) + \left(\ell_j - \frac{1}{n} - R_j\right) - 2Q_{ij}.$$

Simplifying the expression produces the exact algebraic link:

$$M_{ij}^2 = \ell_i + \ell_j - \frac{2}{n} - (R_i + R_j) - 2Q_{ij}.$$

□

Now that we have established a link between standard leverage scores on an EDM, and the geometric leverage scores outlined in this Chapter, we can notice the following relationships

between standard leverage scores and geometric leverage scores.

Lemma 4.3.6. *The relationship between the pairwise Mahalanobis distance M_{ij}^2 and the leverage scores ℓ_i, ℓ_j enforces the following strict asymmetric constraints:*

1. ***Necessity of large leverage scores for large distances:*** *By the Cauchy-Schwarz inequality, $|Q_{ij}| \leq \sqrt{Q_{ii}Q_{jj}}$. Substituting this into the distance decomposition yields the strict upper bound:*
$$M_{ij}^2 \leq \left(\sqrt{\ell_i - \frac{1}{n}} + \sqrt{\ell_j - \frac{1}{n}}\right)^2.$$
A large pairwise Mahalanobis distance mathematically requires at least one of the participating points to possess a large leverage score.

2. ***Insufficiency of large leverage scores for large distances:*** *The converse is false. It is possible to construct configurations where $M_{ij}^2 = 0$ while $\ell_i, \ell_j \geq 0.5$, indicating non-equivalence between the geometric leverage scores presented in this work and standard incoherence applied to an EDM. This arises when $\boldsymbol{p}_i = \boldsymbol{p}_j$ and both points represent severe outliers that dominate the principal axis variance of the empirical covariance structure.*

3. ***Global suppression:*** *A globally small maximum pairwise distance restricts the maximum leverage score. Because the mean squared distance from $\boldsymbol{p}_i$ is $Q_{ii} + \frac{k}{n}$, a small $\max_{i,j} M_{ij}^2$ forces $Q_{ii} \to 0$ for all i. This universally bounds the individual squared Euclidean norms $\|\boldsymbol{p}_i\|_2^2$, prohibiting radial outliers and forcing $R_i \to 0$, thereby globally suppressing all ℓ_i.*

Proof. **Part 1: Necessity.** The term $Q_{ij} = \boldsymbol{p}_i^\top (\boldsymbol{P}^\top \boldsymbol{P})^{-1} \boldsymbol{p}_j$ defines a valid inner product. Applying the Cauchy-Schwarz inequality yields the lower bound $Q_{ij} \geq -\sqrt{Q_{ii}Q_{jj}}$. Substituting this into the expanded distance $M_{ij}^2 = Q_{ii} + Q_{jj} - 2Q_{ij}$ provides the upper bound:

$$M_{ij}^2 \leq Q_{ii} + Q_{jj} + 2\sqrt{Q_{ii}Q_{jj}} = \left(\sqrt{Q_{ii}} + \sqrt{Q_{jj}}\right)^2.$$

Since the radial variance penalty is non-negative ($R_i \geq 0$), we have $Q_{ii} \leq \ell_i - \frac{1}{n}$. Substituting this inequality yields the final bound:

$$M_{ij}^2 \leq \left(\sqrt{\ell_i - \frac{1}{n}} + \sqrt{\ell_j - \frac{1}{n}}\right)^2.$$

Part 2: Insufficiency. We construct a counterexample to demonstrate that $M_{ij}^2 = 0$ does not imply small leverage scores ℓ_i, ℓ_j. Let the number of points be $n \geq 3$, and let the dimension $r = 1$. Let $\boldsymbol{P} \in \mathbb{R}^{n \times 1}$ be a mean-centered configuration matrix ($\mathbf{1}^\top \boldsymbol{P} = 0$). Assume two points are identical, setting $\boldsymbol{p}_1 = \boldsymbol{p}_2 = K$ for some non-zero constant K. To maintain the centering constraint $\sum_{k=1}^{n} \boldsymbol{p}_k = 0$, the remaining $n - 2$ points must sum to $-2K$. We distribute this remaining sum equally among the $n - 2$ points to minimize the variance, defining:

$$\boldsymbol{p}_k = -\frac{2K}{n-2} \quad \text{for } k = 3, \ldots, n.$$

Because $\boldsymbol{p}_1 = \boldsymbol{p}_2$, their vector difference is exactly zero. Substituting this into the Mahalanobis distance yields:

$$M_{12}^2 = (\boldsymbol{p}_1 - \boldsymbol{p}_2)^\top (\boldsymbol{P}^\top \boldsymbol{P})^{-1} (\boldsymbol{p}_1 - \boldsymbol{p}_2) = 0.$$

Next, we compute the empirical variance $\boldsymbol{P}^\top \boldsymbol{P}$ as follows:

$$\begin{aligned}
\boldsymbol{P}^\top \boldsymbol{P} &= \boldsymbol{p}_1^2 + \boldsymbol{p}_2^2 + \sum_{k=3}^{n} \boldsymbol{p}_k^2 \\
&= 2K^2 + \sum_{k=3}^{n} \left(-\frac{2K}{n-2} \right)^2 \\
&= 2K^2 + (n-2) \frac{4K^2}{(n-2)^2} \\
&= 2K^2 + \frac{4K^2}{n-2} \\
&= 2K^2 \left(1 + \frac{2}{n-2} \right) \\
&= 2K^2 \left(\frac{n}{n-2} \right).
\end{aligned}$$

The spatial projection term Q_{11} for the first point evaluates to:

$$Q_{11} = \boldsymbol{p}_1^\top (\boldsymbol{P}^\top \boldsymbol{P})^{-1} \boldsymbol{p}_1 = \frac{K^2}{2K^2 \left(\frac{n}{n-2} \right)} = \frac{n-2}{2n} = \frac{1}{2} - \frac{1}{n}.$$

By definition, the full leverage score for the i-th row is $\ell_i = \frac{1}{n} + Q_{ii} + R_i$. Substituting the exact

value of Q_{11} we can see that

$$\ell_1 = \frac{1}{n} + \left(\frac{1}{2} - \frac{1}{n}\right) + R_1 = \frac{1}{2} + R_1,$$

and as $R_1 \geq 0$, it follows that:

$$\ell_1 \geq \frac{1}{2}.$$

By symmetry, $\ell_2 \geq \frac{1}{2}$, indicating that large Mahalanobis distances corresponding to geometric leverage scores do not necessarily imply large standard leverage scores.

Part 3: Global suppression. Let $M^2_{\max} = \max_{i,j} M^2_{ij}$. We fix i and compute the mean squared pairwise distance to all n points:

$$\frac{1}{n}\sum_{j=1}^{n} M^2_{ij} = \frac{1}{n}\sum_{j=1}^{n}(Q_{ii} + Q_{jj} - 2Q_{ij})$$

$$= Q_{ii} + \frac{1}{n}\operatorname{Tr}(\boldsymbol{Q}) - \frac{2}{n}\boldsymbol{p}_i^\top(\boldsymbol{P}^\top\boldsymbol{P})^{-1}\boldsymbol{P}^\top\mathbf{1}.$$

Because the configuration is centered, $\boldsymbol{P}^\top\mathbf{1} = \mathbf{0}$. As $\boldsymbol{Q}$ is an orthogonal projector onto a r-dimensional subspace, $\operatorname{Tr}(\boldsymbol{Q}) = r$. Thus, the mean distance is strictly $Q_{ii} + \frac{r}{n}$. The maximum must upper bound the mean:

$$M^2_{\max} \geq Q_{ii} + \frac{r}{n} \implies Q_{ii} \leq M^2_{\max} - \frac{r}{n}.$$

Let $\boldsymbol{P} = \boldsymbol{U}_P\boldsymbol{\Sigma}_P\boldsymbol{V}_P^\top$. The squared Euclidean norm of $\boldsymbol{p}_i$ is bounded by the principal singular value σ_1:

$$\|\boldsymbol{p}_i\|_2^2 \leq \sigma_1^2\|\boldsymbol{U}_P^\top\boldsymbol{e}_i\|_2^2 = \sigma_1^2 Q_{ii}.$$

If $M^2_{\max} \to \frac{r}{n}$, then $Q_{ii} \to 0$ for all i. Consequently, $\|\boldsymbol{p}_i\|_2^2 \to 0$ globally. Lacking any variance in the squared norms, the orthogonal radial vector $\boldsymbol{v}^\perp$ vanishes, forcing $R_i \to 0$. Therefore, $\ell_i \to \frac{1}{n}$ globally. □

4.4 Algorithm

As in [SCT25], we consider a first-order retraction method on the manifold of rank-r matrices, denoted $\mathcal{N}_r$. In that work, the relevant objective function was defined as

$$\underset{\boldsymbol{Y}\in\mathbb{S}}{\text{minimize}} \quad \frac{1}{2}\langle \boldsymbol{Y}-\boldsymbol{X}, \mathcal{M}_\Omega(\boldsymbol{Y}-\boldsymbol{X})\rangle \text{ subject to } \operatorname{rank}(\boldsymbol{Y}) = r. \tag{4.4}$$

As $\mathcal{M}_\Omega$ is self-adjoint with respect to the trace inner product, we can see that the Euclidean gradient of (4.4) is $\mathcal{M}_\Omega(\boldsymbol{X}-\boldsymbol{Y})$, and the Riemannian gradient at a point $\boldsymbol{X}_l \in \mathcal{N}_r$ with tangent space $\mathbb{T}_l$ is $\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega(\boldsymbol{X}-\boldsymbol{Y})$, where $\mathcal{P}_{\mathbb{T}_l}$ is the projection operator onto $\mathbb{T}_l$. We can translate this to the non-uniform setting by defining the new objective function with the sampling operator from (4.3) as

$$\underset{\boldsymbol{Y}\in\mathbb{S}}{\text{minimize}} \quad \frac{1}{2}\langle \boldsymbol{X}-\boldsymbol{Y}, \mathcal{M}_{p,\Omega}(\boldsymbol{X}-\boldsymbol{Y})\rangle \text{ subject to } \operatorname{rank}(\boldsymbol{Y}) = r. \tag{4.5}$$

Here, our Euclidean gradient is $\mathcal{M}_{p,\Omega}(\boldsymbol{X}-\boldsymbol{Y})$, with a Riemannian gradient given by $\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_{p,\Omega}(\boldsymbol{X}-\boldsymbol{Y})$. This structural similarity to [SCT25] allows for a similar algorithmic definition. The retraction map is the hard thresholding operator as before, denoted $\mathcal{H}_r : \mathbb{T}_l \to \mathcal{N}_r$, and re-defined as follows:

$$\mathcal{H}_r(\boldsymbol{Y}) = \sum_{i=1}^{r} \lambda_i(\boldsymbol{Y})\boldsymbol{U}_i\boldsymbol{U}_i^\top,$$

where $\boldsymbol{U}_i$ is the i-th eigenvector of $\boldsymbol{Y}$ corresponding to the eigenvalue with the i-th largest magnitude $\lambda_i(\boldsymbol{Y})$. We note that for matrices $\boldsymbol{Y}$ with $\operatorname{rank}(\boldsymbol{Y}) \geq r$, $\operatorname{rank}(\mathcal{H}_r(\boldsymbol{Y})) = r$. We can now define Algorithm 9, the main object of study in this Chapter:

Algorithm 9 Weighted Dual Basis Riemannian EDMC (W-DBRE)

Input: Ω, $\{\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\}_{\boldsymbol{\alpha}\in\Omega}$, $\operatorname{rank}(\boldsymbol{X}) = r$, $\{p_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$
Initialization: $\boldsymbol{X}_0 = \boldsymbol{U}_0\boldsymbol{D}_0\boldsymbol{U}_0^\top$
for $l = 0, 1, \cdots$ **do**
  $\boldsymbol{G}_l = \mathcal{M}_{p,\Omega}(\boldsymbol{X}-\boldsymbol{X}_l)$
  $\alpha_l = \frac{\|\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l\|_{\mathrm{F}}^2}{\langle \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l, \mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l\rangle}$
  $\boldsymbol{W}_l = \boldsymbol{X}_l + \alpha_l\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l$
  $\boldsymbol{X}_{l+1} = \mathcal{H}_r(\boldsymbol{W}_l)$
end for
Output: $\boldsymbol{X}_{\mathrm{rec}}$: Estimated Gram matrix.

This algorithm is a standard first-order retraction scheme on $\mathcal{N}_r$, and obeys many of the standard convergence guarantees presented for these algorithms. The novelty introduced in this paper is the leverage score sampling scheme and in the construction of the corresponding operator $\mathcal{M}_{p,\Omega}$. We note that this sampling scheme would provide algorithmic convergence in both the convex and non-convex frameworks presented in [TL18, GTK24, SCT24, LS24] as well. The convergence analysis of these frameworks all rely on the preservation of restricted isometry of a corresponding sampling operator, which we prove in the following section for $\mathcal{M}_{p,\Omega}$. Concentration of any other EDMC operator will behave similarly.

4.5 Theoretical Guarantees

4.5.1 Restricted Isometry Property

One of the key properties required for the convergence of W-DBRE is the restricted isometry property for $\mathcal{M}_{p,\Omega}$. It is known that $\mathcal{M}_{\Omega}$ exhibits RIP with high probability if $p \gtrsim \frac{\nu^2 r^2 \log n}{n}$. We now present one of the key contributions of this work, which is the proof of RIP for $\mathcal{M}_{p,\Omega}$:

Theorem 4.5.1. *Let $\boldsymbol{X} \in \mathbb{S}$ be the ground truth, rank-r, ν-incoherent Gram matrix with tangent space $\mathbb{T}$ in $\mathcal{N}_r$. Let Ω be sampled from $\mathbb{I}$ via a Bernoulli sampling process defined in* (4.2)*, and let $m \geq \frac{1152}{3}\beta n \log n$ for any $\beta > 3$ and for some sufficiently large constant C. With probability at least $1 - C' n^{3-\beta}$, where C' is an absolute constant, we have that for some absolute numerical constant $C'' > 0$ independent of n or $\boldsymbol{X}$ that*

$$\|\mathcal{P}_{\mathbb{T}}\mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}} - \mathcal{P}_{\mathbb{T}}\| \leq \min\left\{\sqrt{\frac{144\nu r^2 n \log n}{m}}, \frac{40 n^{3/2} r \log n}{m}\right\} + \frac{C'' n r \beta^{1/2} \log n}{m}.$$

Proof Sketch. This proof will work by decomposing $\mathcal{M}_{p,\Omega}$ into $\mathcal{F}_{p,\Omega}$ and an off-diagonal expression. The off-diagonal expression will be convertible into an degenerate, second order matrix chaos, allowing us to use the decoupling inequalities of de la Peña and Giné. $\mathcal{F}_{p,\Omega}$ exhibits standard concentration properties using Theorem 1.7.2. □

We note that our result for RIP is better than the current state-of-the-art for uniform sampling in the EDMC literature. Given the definition of ν as previously written, the current

state of the art is a dependence on ν^2 as the definition in [GTK24] is stronger than our current definition. So, even in the low-ν regime the above result indicates that this geometry-aware sampling method provides better concentration properties. On the other hand, we note that even arbitrarily poor geometries are recoverable, something that was not viable in the uniform sampling framework investigated in the prior EDMC literature. This is because the upper bound of $n^{3/2}\log n$, while suboptimal, is still sharper than $\nu^2 r^2 n \log n$ for $\nu = \mathcal{O}(n)$.

Now that RIP has been established, we state the local convergence result for W-DBRE.

Theorem 4.5.2 (Local Convergence of Algorithm 9)**.** *Let $\boldsymbol{X} \in \mathbb{R}^{n\times n}$ be the ground truth rank-r, ν-incoherent matrix and let $\mathbb{T}$ be the tangent space of $\mathcal{N}_r$ at $\boldsymbol{X}$. Suppose that $m \geq \max\{n\beta\nu r\log n, n^{3/2}r\beta\log n\}$ for some sufficiently large constant $C > 0$ and any $\beta > 3$. Further assume that*

$$\|\boldsymbol{X}_0 - \boldsymbol{X}\|_{\mathrm{F}} \leq \frac{\varepsilon_0}{Cn^3},$$
$$\|\mathcal{P}_{\mathbb{T}}\mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}} - \mathcal{P}_{\mathbb{T}}\| \leq \varepsilon_0,$$

where ε_0 is a constant satisfying

$$\delta = \frac{18\varepsilon_0}{1-4\varepsilon_0} < 1.$$

Then W-DBRE converges linearly as the iterates satisfy

$$\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}} \leq \delta^l \|\boldsymbol{X}_0 - \boldsymbol{X}\|_{\mathrm{F}}.$$

Again, we note that the convergence of this operator is to be expected once the isometry of $\mathcal{M}_{p,\Omega}$ (or the concentration of $\mathcal{F}_{p,\Omega}$, or the isometry of $\mathcal{R}_{p,\Omega}$) is established. The algorithmic approach is straightforward once the theoretical backbone provided by the sampling scheme and the operator is established. Additionally, there is substantial room for this result to be sharpened, particularly by the addition of an initialization guarantee along with increasing a rather small radius of convergence around $\boldsymbol{X}$. The main source of looseness is in the deterministic bounds on $\|\mathcal{M}_{p,\Omega}\|$ and $\|\mathcal{P}_{\mathbb{T}}\mathcal{M}_{p,\Omega}\|$, seen in Theorem 4.B.7. This could be tightened by using a similar strategy to Theorem 3.B.9 and Theorem 3.B.10.

Remark 4.5.1: Sample Complexity

We note that Theorem 4.5.1 and Theorem 4.5.2, coupled with Theorem 4.3.2, provide a recovery guarantee that can be interpreted in terms of the total number of observed distances as follows:

With probability at least $1 - C'n^{3-\beta}$ for some $\beta > 3$ and some sufficiently large C', we have that if

$$|\Omega| \geq c \min\left\{n^{3/2} r\beta \log n, n\nu r^2 \log n\right\},$$

for some sufficiently large constant $c > 0$, W-DBRE recovers the ground truth matrix $\boldsymbol{X}$ provided a good initialization.

4.6 Numerical Results

All of the following experiments were run on a 2021 MacBook Pro with 16 GB integrated memory and an M1 processor in MATLAB. All of the convex programs run in this section were done using the SDPT3 solver for CVX in MATLAB [CR12, GB08].

In this section, we present on the efficacy of the recovery using W-DBRE. We compare this algorithm to DBRE under a uniform sampling regime with the same expected number of measurements, as well as the scheme outlined in [CBSW14]. We also compare this to uniform sampling for completion on the distance matrix alone using the methodology of [Rec11a], described in Section 1.2.2. The experimental setup is $n = 125$ uniformly distributed points on the sphere for the ellipse, and $n = 123$ points distributed on the sphere with 2 outlier points that lie a far distance away from the sphere; see Section 4.1. The relative Gram matrix error was measured following each trial per algorithm, and the averaged results are presented in Table 4.1. We define the sampling rate throughout this section as $p = \frac{|\Omega|}{L}$, but we note that this is not a uniform sampling probability unless explicitly used for uniform methodologies, such as DBRE.

From the results in Table 4.1, we see that the methodology outlined in [CBSW14] is relatively competitive with W-DBRE for both incoherent and coherent datasets across sample complexity regimes. Thus, we see that the leveraged methods like W-DBRE and [CBSW14] significantly outperform the unleveraged methods, such as DBRE and the convex method in [Rec11a].

Table 4.1: Comparative mean over 10 trials for relative Gram matrix error results by algorithm, dataset, and percentage of samples. The incoherence constant for the ellipse with outliers is $\nu \approx 38 \approx \frac{n}{3}$, and the incoherence constant for the standard ellipse is $\nu \approx 2$.

Dataset	**25%**	**20%**	**15%**	**10%**	**5%**
			W-DBRE		
Ellipse with Outliers	1.72×10^{-7}	2.68×10^{-7}	3.76×10^{-7}	9.83×10^{-2}	4.26×10^{-1}
Ellipse	3.79×10^{-8}	5.87×10^{-8}	1.71×10^{-7}	9.60×10^{-3}	1.37×10^{0}
			DBRE		
Ellipse with Outliers	2.13×10^{-1}	1.13×10^{-1}	3.62×10^{-1}	7.12×10^{-1}	1.35×10^{0}
Ellipse	1.04×10^{-7}	1.44×10^{-7}	4.64×10^{-7}	3.37×10^{-1}	4.70×10^{0}
			[**Rec11a**]		
Ellipse with Outliers	2.46×10^{0}	2.29×10^{0}	2.24×10^{0}	1.99×10^{0}	1.54×10^{0}
Ellipse	3.84×10^{-9}	1.11×10^{-3}	9.61×10^{-2}	4.43×10^{-1}	9.16×10^{-1}
			[**CBSW14**]		
Ellipse with Outliers	1.18×10^{-7}	1.05×10^{-6}	9.85×10^{-7}	1.02×10^{-2}	1.12×10^{0}
Ellipse	1.31×10^{-9}	6.07×10^{-9}	2.33×10^{-8}	1.49×10^{-1}	7.21×10^{-1}

The primary advantage of W-DBRE for the EDMC problem comes from its relative recovery speed as compared to convex nuclear norm minimization, We provide a timing comparison table in Table 4.2. We note that the substantial gap in the timing is a critical motivator for the investigation of these non-convex algorithmic approaches for matrix completion.

Table 4.2: Mean computational time across sampling rates for W-DBRE and [CBSW14] using the SDPT3 solver, sorted by sampling rate.

Method	**25%**	**20%**	**15%**	**10%**	**5%**
W-DBRE (s)	0.23	0.34	0.30	3.2	5.7
[CBSW14] (s)	53	51	47	90	48

4.6.1 Approximate Leverage Scores

We note that full access to all geometric leverage scores in practice is unrealistic. Instead, it is better to take a small uniform sample and to estimate the leverage scores from this technique. To do this, we sample uniformly with parameter $p = 0.05$, estimating $\boldsymbol{X}_0 = \frac{1}{p^2}\mathcal{M}_\Omega(\boldsymbol{X})$. Using $\boldsymbol{X}_0$, we can compute estimated leverage scores $\{\hat{\nu}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$ to approximate the sampling probabilities $\{p_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$, with the approximations denoted as $\{\hat{p}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$.

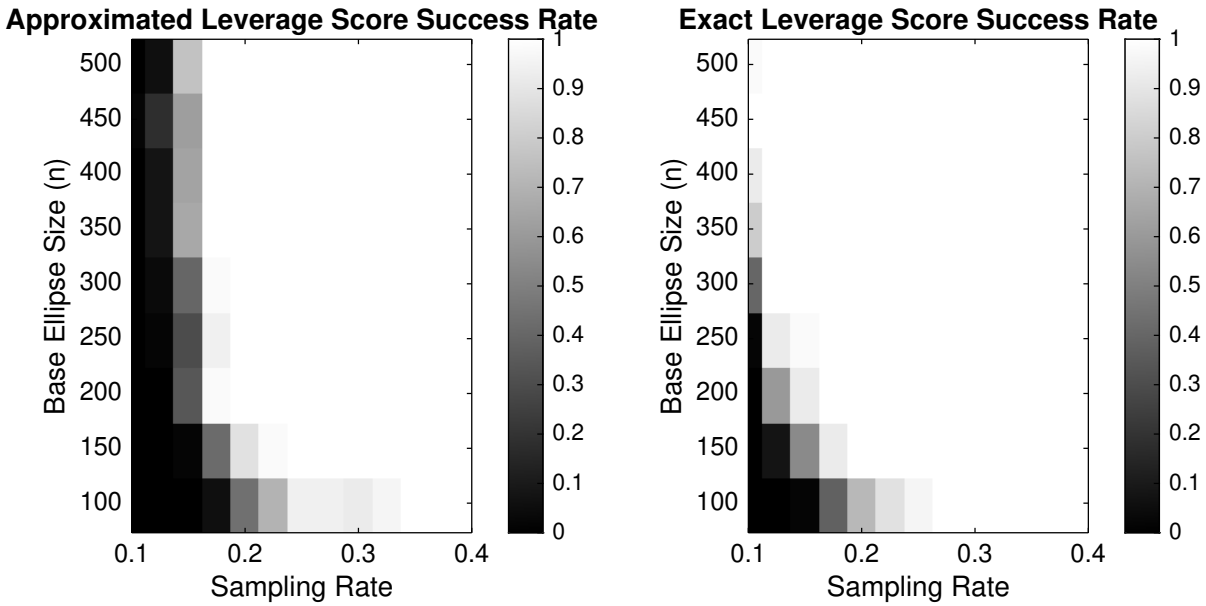


Figure 4.2: Phase Transition plot on the ellipse with outlier datapoints as a function of the size of the dataset and the total percentage of revealed samples. All-white indicates 100% success rate (classified as a relative error $< 10^{-3}$, and all-black indicates 0% success. Each square was averaged over 25 trials.

We present the following results on a coherent dataset using a full set of uniform samples, the estimated leverage score sampling techniques, and the true sampling in Table 4.3. This experiment was again run on the ellipse with 2 outlier points, and the total size of the dataset was $n = 50$. The relative Gram matrix error was averaged over 10 trials.

Table 4.3: Mean relative Gram matrix recovery error using W-DBRE with varying leverage scores. For this experiment, $\nu \approx 15 \approx \frac{n}{3}$.

Method	**50%**	**40%**	**30%**	**20%**	**10%**
Approximate Scores	2.26×10^{-7}	4.83×10^{-7}	6.20×10^{-7}	7.14×10^{-2}	5.33×10^{-1}
Exact Scores	5.16×10^{-8}	7.50×10^{-8}	1.45×10^{-7}	3.03×10^{-2}	1.09
Uniform Sampling	3.83×10^{-3}	1.20×10^{-1}	1.15×10^{-1}	4.89×10^{-1}	7.56×10^{-1}

To study the performance of the approximate leverage scores in greater detail, we study the same ellipse of varying sizes (from $n = 100$ to $n = 500$) with the sampling rate seen on the x-axis. This image can be seen in Figure 4.2.

As we can see in Figure 4.2, the exact leverage scores give better numerical performance. This is to be expected, as these leverage scores should contain the requisite geometric information to reconstruct coherent geometries. However, it is clear from Figure 4.2 that there is

a rather large regime where merely approximating the leverage scores is sufficient for a high likelihood of recovery, indicating the practical utility of sampling a small subset of the overall distances first to develop an estimate for the leverage scores. Further work will be done to investigate what sorts of theoretical guarantees we may be able to build out of this technique.

4.6.2 Approximating W-DBRE and Other Leverage Scores

As noted in Table 4.1, for small datasets [CBSW14] outperforms W-DBRE given exact leverage scores for highly coherent data. In this section, we show that approximating the leverage scores following the method of the previous section and using W-DBRE outperforms the leverage score approximation in [CBSW14], and the provable leverage scores of [EWW18] for nuclear norm minimization in the matrix completion problem as well. We provide the following experiment indicating this result on the outlier ellipse figure with $n = 35$, averaged over 25 trials.

Table 4.4: Mean relative Gram matrix error across varying sampling rates via approximate leverage scores. The incoherence constant is $\nu \approx 11 \approx \frac{n}{3}$.

Method	**50%**	**40%**	**30%**	**20%**	**10%**
W-DBRE	0.030	0.0010	0.031	0.33	0.73
[CBSW14]	0.30	0.17	0.15	0.37	1.13
[EWW18]	0.30	0.18	0.15	0.36	0.32

Table 4.4 indicates that the leverage scores used for W-DBRE are in some sense more easily estimated than the leverage scores in either [CBSW14] or [EWW18]. This is an important direction to explore more thoroughly both theoretically and with other datasets.

4.7 Conclusion

In this section, we introduced the novel operator $\mathcal{M}_{p,\Omega}$ and the corresponding W-DBRE algorithm, alongside a novel incoherence-based sampling scheme. This algorithm was shown to exhibit state-of-the-art numerical performance among non-convex methodologies, and also illustrated stronger numerical performance when the geometric leverage scores were approximated, as compared to convex methods. In addition, local convergence guarantees were provided and the restricted isometry property was proven for $\mathcal{M}_{p,\Omega}$. Future work will seek to provide provable

guarantees on the approximated leverage scores, as well as investigate the complex relationship between incoherence of a point cloud, its condition number, and the recoverability of said point cloud.

4.A Useful Lemmas

In this section, we will provide two Lemmas that will prove essential to the remainder of the theory developed in the Appendix.

Lemma 4.A.1 (Bounds on $p_{\boldsymbol{\alpha}}$)**.** *Let $p_{\boldsymbol{\alpha}}$ be as defined in* (4.2)*, with $\nu_{\boldsymbol{\alpha}}$ defined as in* (4.1)*. Then the following bounds for $p_{\boldsymbol{\alpha}}$ hold for $\alpha \in \mathbb{I}_p$:*

$$\begin{aligned}\frac{1}{p_{\boldsymbol{\alpha}}} &\leq \frac{3nr}{m}\frac{1}{\|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2} \\ \frac{1}{p_{\boldsymbol{\alpha}}} &\leq \frac{8nr}{m}\frac{1}{\|\mathcal{P}_U \boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2} \\ \frac{1}{p_{\boldsymbol{\alpha}}} &\leq \frac{n^2}{8m}.\end{aligned}$$

Proof. We note that this follows directly from the definitions in (4.2) and in (4.1). For the first, as we are considering all $\boldsymbol{\alpha}$ such that $\alpha \in \mathbb{I}_p$

$$\begin{aligned}\frac{1}{p_{\boldsymbol{\alpha}}} &= \max\left\{\frac{2nr}{m}\frac{1}{\nu_{\boldsymbol{\alpha}}}, 1\right\} \\ &\leq \frac{2nr}{m}\frac{1}{\nu_{\boldsymbol{\alpha}}}.\end{aligned}$$

The three bounds in the theorem statement then follow from (4.1) and from the fact that $\|\mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \leq 3\|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}$. □

Lemma 4.A.2. *Let $\tilde{\boldsymbol{H}}_p \in \mathbb{R}^{|\mathbb{I}_p|\times|\mathbb{I}_p|}$ be a matrix defined entrywise by, for $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{I}_p$ as $\left(\tilde{H}_p\right)_{\boldsymbol{\alpha}\boldsymbol{\beta}} = \frac{1}{p_{\boldsymbol{\alpha}}^{1/2} p_{\boldsymbol{\beta}}^{1/2}}\langle \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}, \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\beta}}\rangle$. Then*

$$\lambda_{\max}\left(\tilde{\boldsymbol{H}}_p\right) \leq \frac{3n^2 r}{m}.$$

Proof. We will use a Gershgorin argument to prove this upper bound. First, notice that if

$\boldsymbol{\alpha} \cap \boldsymbol{\beta} = \emptyset$,

$$\begin{aligned}\langle \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}, \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\beta}} \rangle &= \text{Trace}(\boldsymbol{w}_{\boldsymbol{\alpha}} \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\beta}}) \\ &= \text{Trace}(\boldsymbol{w}_{\boldsymbol{\beta}} \boldsymbol{w}_{\boldsymbol{\alpha}} \mathcal{P}_U) \\ &= 0,\end{aligned}$$

following from the definition of $\boldsymbol{w}_{\boldsymbol{\alpha}}$ and $\boldsymbol{w}_{\boldsymbol{\beta}}$. As such, we can bound the maximum eigenvalue of $\tilde{\boldsymbol{H}}_p$ as

$$\begin{aligned}\lambda_{\max}\left(\tilde{\boldsymbol{H}}_p\right) &\leq \max_{\boldsymbol{\beta} \in \mathbb{I}_p} \sum_{\substack{\boldsymbol{\alpha} \in \mathbb{I}_p \\ \boldsymbol{\alpha} \cap \boldsymbol{\beta} \neq \emptyset}} \frac{1}{p_{\boldsymbol{\alpha}}^{1/2} p_{\boldsymbol{\beta}}^{1/2}} \left| \langle \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}, \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\beta}} \rangle \right| \\ &\leq \max_{\boldsymbol{\beta} \in \mathbb{I}_p} \sum_{\substack{\boldsymbol{\alpha} \in \mathbb{I}_p \\ \boldsymbol{\alpha} \cap \boldsymbol{\beta} \neq \emptyset}} \frac{\|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\text{F}}}{p_{\boldsymbol{\alpha}}^{1/2}} \frac{\|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\beta}}\|_{\text{F}}}{p_{\boldsymbol{\beta}}^{1/2}} \\ &\leq \max_{\boldsymbol{\beta} \in \mathbb{I}_p} \sum_{\substack{\boldsymbol{\alpha} \in \mathbb{I}_p \\ \boldsymbol{\alpha} \cap \boldsymbol{\beta} \neq \emptyset}} \frac{3nr}{m} \\ &\leq \frac{3n^2 r}{m},\end{aligned}$$

thus concluding the proof. □

4.B Restricted Isometry Results

As RIP and its variants are critical to the analysis of W-DBRE in this paper, this section is dedicated to the proofs of RIP and similar results.

4.B.1 Supporting Concentration Results

The following results are various concentration identities used in the computation of $\|\mathcal{P}_{\mathbb{T}} \mathcal{M}_{p,\Omega} \mathcal{P}_{\mathbb{T}} - \mathcal{P}_{\mathbb{T}}\|$.

Lemma 4.B.1. *For a ground truth rank r ν-incoherent matrix $\boldsymbol{X}$ with tangent space $\mathbb{T}$ on $\mathcal{N}_r$,*

we have that for any $\beta > 1$, and with probability at least $1 - 2n^{1-\beta}$, and for $m \geq \frac{72}{3} n \log n$, that

$$\|\mathcal{P}_{\mathbb{T}} \mathcal{F}_{p,\Omega} \mathcal{P}_{\mathbb{T}} - \mathcal{P}_{\mathbb{T}} \mathcal{F}_{\mathbb{I}} \mathcal{P}_{\mathbb{T}}\| \leq \min \left\{ \sqrt{\frac{72 n^3 r^2 \beta \log n}{3m^2}}, \sqrt{\frac{24 n \nu r \beta \log n}{3m}} \right\}.$$

Proof. This argument will follow a standard Matrix Bernstein argument, seen in Theorem 1.7.2. As

$$\hat{\mathcal{F}}_{p,\Omega} - \mathcal{F}_{\mathbb{I}} = \sum_{\alpha \in \mathbb{I}_p} \left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1 \right) \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}},$$

we can decompose this expression as a sum of independent random variables. Let $\mathcal{T}_\alpha = \left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1 \right) \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}}$. First, we have that

$$\|\mathcal{T}_\alpha\| \leq \max_{\alpha \in \mathbb{I}_p} \frac{\|\mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2}{p_{\boldsymbol{\alpha}}} =: R.$$

Next, to bound the variance we have that, as for $\boldsymbol{\alpha} \in \mathbb{I}_p$ it follows that $\|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 \leq \frac{3nr}{m}$,

$$\begin{aligned}
\left\| \sum_{\boldsymbol{\alpha} \in \mathbb{I}} \mathbb{E}\left[\mathcal{T}_\alpha^2 \right] \right\| &= \left\| \sum_{\boldsymbol{\alpha} \in \mathbb{I}} \left(\frac{1}{p_{\boldsymbol{\alpha}}} - 1 \right) \|\mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \right\| \\
&\leq \frac{3nr}{m} \left\| \sum_{\alpha \in \mathbb{I}_p} \frac{1}{p_{\boldsymbol{\alpha}}} \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \right\| \\
&\leq \frac{12nr}{m} \lambda_{\max} \left(\tilde{\boldsymbol{H}}_p \right).
\end{aligned}$$

From Theorem 4.A.2, we note that our variance bound is now $\sigma^2 = \min \left\{ \frac{36 n^3 r^2}{m^2}, \frac{12 \nu r^2 n}{m} \right\}$. Now, for $t \leq \frac{\sigma^2}{R} = \min \left\{ \nu, n^{1/2} \right\}$, our result follows from Theorem 1.7.2 as long as $m \geq n\beta \log n$. □

Lemma 4.B.2. *Let $\{\boldsymbol{A}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathbb{I}}, \{\boldsymbol{B}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathbb{I}} \subset \mathbb{R}^{n \times n}$ be any two collections of L matrices. Then*

$$\begin{aligned}
\left\| \sum_{\boldsymbol{\alpha} \neq \boldsymbol{\beta}} \left(\frac{\xi_{\boldsymbol{\alpha}} \xi_{\boldsymbol{\beta}}}{p_{\boldsymbol{\alpha}} p_{\boldsymbol{\beta}}} - 1 \right) \boldsymbol{A}_{\boldsymbol{\alpha}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \boldsymbol{B}_{\boldsymbol{\beta}} \rangle \right\| &\leq \left\| \sum_{\boldsymbol{\alpha} \neq \boldsymbol{\beta}} \left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1 \right) \left(\frac{\xi_{\boldsymbol{\beta}}}{p_{\boldsymbol{\beta}}} - 1 \right) H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \boldsymbol{A}_{\boldsymbol{\alpha}} \langle \cdot, \boldsymbol{B}_{\boldsymbol{\beta}} \rangle \right\| \\
&\quad + \left\| \sum_{\boldsymbol{\alpha} \neq \boldsymbol{\beta}} \left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1 \right) \boldsymbol{A}_{\boldsymbol{\alpha}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \boldsymbol{B}_{\boldsymbol{\beta}} \rangle \right\| \\
&\quad + \left\| \sum_{\boldsymbol{\alpha} \neq \boldsymbol{\beta}} \left(\frac{\xi_{\boldsymbol{\beta}}}{p_{\boldsymbol{\beta}}} - 1 \right) \boldsymbol{A}_{\boldsymbol{\alpha}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \boldsymbol{B}_{\boldsymbol{\beta}} \rangle \right\|.
\end{aligned}$$

Proof. This result follows from the fact that

$$\frac{\xi_{\boldsymbol{\alpha}}\xi_{\boldsymbol{\beta}}}{p_{\boldsymbol{\alpha}}p_{\boldsymbol{\beta}}} - 1 = \left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1\right)\left(\frac{\xi_{\boldsymbol{\beta}}}{p_{\boldsymbol{\beta}}} - 1\right) + \left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1\right) + \left(\frac{\xi_{\boldsymbol{\beta}}}{p_{\boldsymbol{\beta}}} - 1\right).$$

As such, the expression becomes

$$\begin{aligned}
&\left\| \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} \left(\frac{\xi_{\boldsymbol{\alpha}}\xi_{\boldsymbol{\beta}}}{p_{\boldsymbol{\alpha}}p_{\boldsymbol{\beta}}} - 1\right) \boldsymbol{A}_{\boldsymbol{\alpha}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \boldsymbol{B}_{\boldsymbol{\beta}} \rangle \right\| \\
&\quad = \left\| \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} \left[\left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1\right)\left(\frac{\xi_{\boldsymbol{\beta}}}{p_{\boldsymbol{\beta}}} - 1\right) + \left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1\right) + \left(\frac{\xi_{\boldsymbol{\beta}}}{p_{\boldsymbol{\beta}}} - 1\right) \right] \boldsymbol{A}_{\boldsymbol{\alpha}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \boldsymbol{B}_{\boldsymbol{\beta}} \rangle \right\| \\
&\quad = \Bigg\| \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} \left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1\right)\left(\frac{\xi_{\boldsymbol{\beta}}}{p_{\boldsymbol{\beta}}} - 1\right) \boldsymbol{A}_{\boldsymbol{\alpha}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \boldsymbol{B}_{\boldsymbol{\beta}} \rangle + \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} \left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1\right) \boldsymbol{A}_{\boldsymbol{\alpha}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \boldsymbol{B}_{\boldsymbol{\beta}} \rangle \\
&\qquad + \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} \left(\frac{\xi_{\boldsymbol{\beta}}}{p_{\boldsymbol{\beta}}} - 1\right) \boldsymbol{A}_{\boldsymbol{\alpha}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \boldsymbol{B}_{\boldsymbol{\beta}} \rangle \Bigg\|.
\end{aligned}$$

The result then follows from the triangle inequality. □

Lemma 4.B.3. *Let $\boldsymbol{X}$ be a rank-r, ν-incoherent ground truth matrix with tangent space $\mathbb{T}$. Let $\boldsymbol{V}_{\boldsymbol{\alpha}} = \mathrm{vec}(\mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}})$ for all $\boldsymbol{\alpha} \in \mathbb{I}$, and let $\boldsymbol{G} = \left[\frac{\boldsymbol{V}_{\boldsymbol{\alpha}}^\top \boldsymbol{V}_{\boldsymbol{\beta}}}{p_{\boldsymbol{\alpha}}^{1/2} p_{\boldsymbol{\beta}}^{1/2}}\right] \in \mathbb{R}^{|\mathbb{I}_p|\times|\mathbb{I}_p|}$ be a diagonally re-scaled Gram matrix of $\{\boldsymbol{V}_{\boldsymbol{\alpha}}\}_{\alpha\in\mathbb{I}_p}$. Defining $\boldsymbol{b}_{\boldsymbol{\alpha}} = \sum_{\boldsymbol{\beta}\neq\boldsymbol{\alpha}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \left(\frac{\xi'_{\boldsymbol{\beta}}}{p_{\boldsymbol{\beta}}} - 1\right) \boldsymbol{V}_{\boldsymbol{\beta}}$, and letting $\boldsymbol{B} = [\boldsymbol{b}_1 \cdots \boldsymbol{b}_L]$, we have that $\boldsymbol{B} = \sum_{\boldsymbol{\alpha}} \left(\frac{\xi'_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1\right) \boldsymbol{V}_{\boldsymbol{\alpha}} \boldsymbol{h}_{\boldsymbol{\alpha}}^\top$, where $\boldsymbol{h}_{\boldsymbol{\alpha}} \in \mathbb{R}^L$ is the $\boldsymbol{\alpha}$-th column of $\boldsymbol{H}_{\mathrm{offdiag}}^{-1}$, and that with probability at least $1 - 2n^{2-\beta}$ that if $m \geq \frac{96}{3}\beta n \log n$,*

$$\|\boldsymbol{B}\| \leq \sqrt{36 \frac{nr\beta \log n}{m}}.$$

Proof. The first statement follows from the rules of matrix multiplication, and can be seen explicitly by defining the matrix $\boldsymbol{W} = \left[\left(\frac{\xi'_1}{p_{\boldsymbol{\alpha}}} - 1\right) \boldsymbol{V}_1 \cdots \left(\frac{\xi'_L}{p_{\boldsymbol{\alpha}}} - 1\right) \boldsymbol{V}_L\right] \in \mathbb{R}^{n^2 \times L}$ and seeing that $\boldsymbol{B} = \boldsymbol{W}\boldsymbol{H}_{\mathrm{offdiag}}^{-1}$. The first result then follows from the symmetry of $\boldsymbol{H}_{\mathrm{offdiag}}^{-1}$. It follows that, as $\boldsymbol{B} = \sum_{\boldsymbol{\alpha}} \left(\frac{\xi'_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1\right) \boldsymbol{V}_{\boldsymbol{\alpha}} \boldsymbol{h}_{\boldsymbol{\alpha}}^\top$, that $\boldsymbol{B}$ is a sum of mean-zero, independent matrices. Therefore we can develop a high probability bound on $\|\boldsymbol{B}\|$ using Theorem 1.7.2. Defining each summand as $\boldsymbol{B}_{\boldsymbol{\alpha}}$ we first bound

$$\|\boldsymbol{B}_{\boldsymbol{\alpha}}\| = \left\| \left(\frac{\xi'_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1\right) \boldsymbol{V}_{\boldsymbol{\alpha}} \boldsymbol{h}_{\boldsymbol{\alpha}}^\top \right\|$$

$$
\begin{aligned}
&\leq \frac{\|\boldsymbol{V}_{\boldsymbol{\alpha}}\|_2}{p_{\boldsymbol{\alpha}}}\|\boldsymbol{h}_{\boldsymbol{\alpha}}\|_2 \\
&\leq \left(\max_{\boldsymbol{\beta}\in\mathbb{I}_p}\frac{\|\boldsymbol{V}_{\boldsymbol{\beta}}\|_2}{p_{\boldsymbol{\beta}}^{1/2}}\right)\max_{\boldsymbol{\alpha}\in\mathbb{I}_p}\frac{1}{p_{\boldsymbol{\alpha}}^{1/2}}\|\boldsymbol{H}_{\text{offdiag}}^{-1}\|_{2\to\infty} \\
&\leq \left(\max_{\boldsymbol{\beta}\in\mathbb{I}_p}\frac{\|\boldsymbol{V}_{\boldsymbol{\beta}}\|_2}{p_{\boldsymbol{\beta}}^{1/2}}\right)\max_{\boldsymbol{\alpha}\in\mathbb{I}_p}\frac{1}{p_{\boldsymbol{\alpha}}^{1/2}}\sqrt{\frac{2}{n}} \\
&\leq \left(\max_{\boldsymbol{\beta}\in\mathbb{I}_p}\frac{\|\boldsymbol{V}_{\boldsymbol{\beta}}\|_2}{p_{\boldsymbol{\beta}}^{1/2}}\right)\frac{n^{1/2}}{m^{1/2}} \\
&\leq \frac{2nr^{1/2}}{m} =: R,
\end{aligned}
$$

where the second inequality follows from Theorem 3.A.1 and the last inequality follows from Theorem 4.A.1. Next, we bound the variance expressions. First, we have that

$$
\begin{aligned}
\left\|\sum_{\boldsymbol{\alpha}}\mathbb{E}[\boldsymbol{B}_{\boldsymbol{\alpha}}\boldsymbol{B}_{\boldsymbol{\alpha}}^*]\right\| &= \left\|\sum_{\boldsymbol{\alpha}}\mathbb{E}\left[\left(\frac{\xi'_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}}-1\right)^2\right]\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{h}_{\boldsymbol{\alpha}}^\top\boldsymbol{h}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}^\top\right\| \\
&= \left\|\sum_{\boldsymbol{\alpha}\in\mathbb{I}_p}\left(\frac{1}{p_{\boldsymbol{\alpha}}}-1\right)\|\boldsymbol{h}_{\boldsymbol{\alpha}}\|_2^2\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}^\top\right\| \\
&\leq \frac{2}{n}\left\|\sum_{\boldsymbol{\alpha}\in\mathbb{I}_p}\frac{1}{p_{\boldsymbol{\alpha}}}\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}^\top\right\| \\
&= \frac{2}{n}\lambda_{\max}(\boldsymbol{G}) \\
&\leq \frac{6nr}{m},
\end{aligned}
$$

using Theorem 1.8.2 for the final line, followed by Theorem 4.A.2. The other variance expression is then

$$
\begin{aligned}
\left\|\sum_{\boldsymbol{\alpha}}\mathbb{E}[\boldsymbol{B}_{\boldsymbol{\alpha}}^*\boldsymbol{B}_{\boldsymbol{\alpha}}]\right\| &= \left\|\sum_{\boldsymbol{\alpha}}\mathbb{E}\left[\left(\frac{\xi'_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}}-1\right)^2\right]\boldsymbol{h}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}^\top\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{h}_{\boldsymbol{\alpha}}\right\| \\
&\leq \left\|\sum_{\boldsymbol{\alpha}}\frac{\|\boldsymbol{V}_{\boldsymbol{\alpha}}\|_2^2}{p_{\boldsymbol{\alpha}}}\boldsymbol{h}_{\boldsymbol{\alpha}}\boldsymbol{h}_{\boldsymbol{\alpha}}^\top\right\| \\
&\leq \left(\max_{\boldsymbol{\beta}\in\mathbb{I}_p}\frac{\|\boldsymbol{V}_{\boldsymbol{\beta}}\|_2^2}{p_{\boldsymbol{\beta}}}\right)\left\|\sum_{\boldsymbol{\alpha}}\boldsymbol{h}_{\boldsymbol{\alpha}}\boldsymbol{h}_{\boldsymbol{\alpha}}^\top\right\| \\
&= \left(\max_{\boldsymbol{\beta}\in\mathbb{I}_p}\frac{\|\boldsymbol{V}_{\boldsymbol{\beta}}\|_2^2}{p_{\boldsymbol{\beta}}}\right)\left\|\boldsymbol{H}_{\text{offdiag}}^{-1}(\boldsymbol{H}_{\text{offdiag}}^{-1})^\top\right\|
\end{aligned}
$$

$$
\begin{aligned}
&= \left(\max_{\boldsymbol{\beta} \in \mathbb{I}_p} \frac{\|\boldsymbol{V}_{\boldsymbol{\beta}}\|_2^2}{p_{\boldsymbol{\beta}}} \right) \|\boldsymbol{H}_{\text{offdiag}}^{-1}\|^2 \\
&\leq 4 \left(\max_{\boldsymbol{\beta} \in \mathbb{I}_p} \frac{\|\boldsymbol{V}_{\boldsymbol{\beta}}\|_2^2}{p_{\boldsymbol{\beta}}} \right),
\end{aligned}
$$

where the last inequality follows from Theorem 3.A.1. Now, as $\boldsymbol{V}_{\boldsymbol{\alpha}} = \text{vec}(\mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}})$ it follows that taking the maximum of the variance terms, we have $\sigma^2 = 12\frac{nr}{m}$ and $R = 2\frac{nr^{1/2}}{m}$. Now, as $\frac{\sigma^2}{R} = 6r^{1/2} \geq \sqrt{36\frac{nr\beta \log n}{m}}$ for $m \geq \frac{96}{3}\beta n \log n$, we have from (1.25) that

$$
\mathbb{P}\left(\|\boldsymbol{B}\| \geq \sqrt{36\frac{nr\beta \log n}{m}} \right) \leq 2n^{2-\beta}.
$$

This concludes the proof. □

Lemma 4.B.4. *Let $\boldsymbol{X}$ be a rank-r, ν-incoherent ground truth matrix with tangent space $\mathbb{T}$. Let $\boldsymbol{V}_{\boldsymbol{\alpha}} = \text{vec}(\mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}})$. Let $\boldsymbol{b}_{\boldsymbol{\alpha}} = \sum_{\boldsymbol{\beta} \neq \boldsymbol{\alpha}} \left(\frac{\xi_{\boldsymbol{\beta}}}{p_{\boldsymbol{\beta}}} - 1 \right) \boldsymbol{V}_{\boldsymbol{\beta}}$. For $m \geq \frac{48}{3} n\beta \log n$, with probability at least $1 - n^{2-\beta}$ we have that for every $\boldsymbol{\alpha} \in \mathbb{I}$,*

$$
\|\boldsymbol{b}_{\boldsymbol{\alpha}}\|_2 \leq \sqrt{\frac{9 \log n}{m}},
$$

Proof. This result will follow from the Vector Bernstein inequality seen in Theorem 1.7.3. We first note that $\boldsymbol{b}_{\boldsymbol{\alpha}}$ is a sum of mean-zero independent random vectors. Defining the term $\boldsymbol{y}_{\boldsymbol{\beta}} = \left(\frac{\xi'_{\boldsymbol{\beta}}}{p_{\boldsymbol{\beta}}} - 1 \right) H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \boldsymbol{V}_{\boldsymbol{\beta}}$, we first need to bound $\|\boldsymbol{y}_{\boldsymbol{\beta}}\|_2$ via

$$
\begin{aligned}
\|\boldsymbol{y}_{\boldsymbol{\beta}}\|_2 &= \left| \left(\frac{\xi'_{\boldsymbol{\beta}}}{p_{\boldsymbol{\beta}}} - 1 \right) \right| H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \|\boldsymbol{V}_{\boldsymbol{\beta}}\|_2 \\
&\leq \frac{1}{n} \max_{\boldsymbol{\alpha}} \frac{\|\boldsymbol{V}_{\boldsymbol{\alpha}}\|_2}{p_{\boldsymbol{\alpha}}} \\
&\leq \frac{2n^{1/2} r^{1/2}}{m} =: R.
\end{aligned}
$$

Next, to bound the variance term, we see that

$$
\sum_{\boldsymbol{\beta} \neq \boldsymbol{\alpha}} \mathbb{E}[\|\boldsymbol{y}_{\boldsymbol{\beta}}\|_2^2] = \sum_{\boldsymbol{\beta} \neq \boldsymbol{\alpha}} \mathbb{E}\left[\left(\frac{\xi'_{\boldsymbol{\beta}}}{p_{\boldsymbol{\alpha}}} - 1 \right)^2 \right] (H^{\boldsymbol{\alpha}\boldsymbol{\beta}})^2 \|\boldsymbol{V}_{\boldsymbol{\beta}}\|_2^2
$$

$$
\begin{aligned}
&= \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} \left(\frac{\xi'_{\boldsymbol{\beta}}}{p_{\boldsymbol{\beta}}} - 1\right) (H^{\boldsymbol{\alpha}\boldsymbol{\beta}})^2 \|\boldsymbol{V}_{\boldsymbol{\beta}}\|_2^2 \\
&\leq \max_{\boldsymbol{\gamma}} \left(\frac{\|\boldsymbol{V}_{\boldsymbol{\gamma}}\|_2^2}{p_{\boldsymbol{\gamma}}}\right) \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} (H^{\boldsymbol{\alpha}\boldsymbol{\beta}})^2 \\
&\leq \frac{2}{n}\frac{3nr}{m} \\
&= \frac{6r}{m} =: \sigma^2.
\end{aligned}
$$

It follows then that $\frac{\sigma^2}{R} = \frac{3}{n^{1/2}} \geq \sqrt{9\frac{\log n}{m}}$ for $m \geq \frac{48}{3}\beta n \log n$, we have from Theorem 1.7.3 that

$$
\mathbb{P}\left(\|\boldsymbol{b}_{\boldsymbol{\alpha}}\|_2 \geq \sqrt{\frac{9\log n}{m}}\right) \leq 2n^{-\beta}.
$$

□

Lemma 4.B.5. *Let $\boldsymbol{X}$ be a rank-r, ν-incoherent ground truth matrix with tangent space $\mathbb{T}$. Then if $m \geq \frac{1152\beta n \log n}{3n}$ for any $\beta > 3$, with probability at least $1 - C'n^{3-\beta}$, where C' is an absolute constant, we have that, for some absolute numerical constants C_1 independent of n or $\boldsymbol{X}$, the following inequality holds:*

$$
\left\| \sum_{\substack{\boldsymbol{\alpha}\in\mathbb{I} \\ \boldsymbol{\beta}\neq\boldsymbol{\alpha}}} \left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1\right)\left(\frac{\xi_{\boldsymbol{\beta}}}{p_{\boldsymbol{\beta}}} - 1\right) H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\beta}}\rangle \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}} \right\| \leq \frac{C_1 nr\beta^{1/2}\log n}{m},
$$

Proof. The goal is to develop a high probability bound for the operator norm of the previously mentioned operator. We begin the proof by considering, for some $t > 0$, the expression

$$
\mathbb{P}\left[\left\| \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} \left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1\right)\left(\frac{\xi_{\boldsymbol{\beta}}}{p_{\boldsymbol{\beta}}} - 1\right) \langle \cdot, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\beta}}\rangle H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}} \right\| \geq t\right].
$$

For notational simplicity, we let $\boldsymbol{U}_{\boldsymbol{\alpha}} = \mathrm{vec}(\mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}) \in \mathbb{R}^{n^2}$ denote the vectorized version of $\mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}$. As such, we can see that this probabilistic expression can be equivalently represented as

$$
\mathbb{P}\left[\left\| \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} \left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1\right)\left(\frac{\xi_{\boldsymbol{\beta}}}{p_{\boldsymbol{\beta}}} - 1\right) H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \boldsymbol{U}_{\boldsymbol{\alpha}}\boldsymbol{U}_{\boldsymbol{\beta}}^\top \right\| \geq t\right].
$$

This term is a second order matrix chaos, and is a degenerate U-statistic—see [Lee19] for details on the theory of U-statistics, and see [DlPG12] for a detailed investigation of the decoupling of Banach-space-valued U-statistics. As seen in Theorem 1.7.5, we can study the distribution of the operator norm for this operator by considering a new set of i.i.d. random Bernoulli variables $\xi'_{\boldsymbol{\beta}}$. As such, Theorem 1.7.5 states that for some absolute constant C independent of n, $\boldsymbol{U}_{\boldsymbol{\alpha}}$, or $\boldsymbol{H}^{-1}$, that

$$\mathbb{P}\left[\left\| \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} \left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1\right)\left(\frac{\xi_{\boldsymbol{\beta}}}{p_{\boldsymbol{\beta}}} - 1\right) H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \boldsymbol{U}_{\boldsymbol{\alpha}} \boldsymbol{U}_{\boldsymbol{\beta}}^\top \right\| \geq t\right] \\ \leq C\mathbb{P}\left[\left\| \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} \left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1\right)\left(\frac{\xi'_{\boldsymbol{\beta}}}{p_{\boldsymbol{\beta}}} - 1\right) H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \boldsymbol{U}_{\boldsymbol{\alpha}} \boldsymbol{U}_{\boldsymbol{\beta}}^\top \right\| \geq t/C\right].$$

We note that optimal values of the constants C have not been explicitly computed for each chaos of order p, but that the constant itself is a function of p. For $p = 2$, which is the case we consider, this constant is usually rather small, with the expression in [DlPG12] using a constant of 6 for the second-order chaos. Using the law of total probability, conditioning on the random variables $\xi'_1, ..., \xi'_L$ and defining the vector $\boldsymbol{\xi'} = [\xi'_1 \ ... \ \xi'_L]^\top$, we have that

$$\mathbb{P}\left[\left\| \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} \left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1\right)\left(\frac{\xi_{\boldsymbol{\beta}}}{p_{\boldsymbol{\beta}}} - 1\right) H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \boldsymbol{U}_{\boldsymbol{\alpha}} \boldsymbol{U}_{\boldsymbol{\beta}}^\top \right\| \geq t\right] \\ = \mathbb{E}\left[\mathbb{P}\left(\left\| \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} \left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1\right)\left(\frac{\xi'_{\boldsymbol{\beta}}}{p_{\boldsymbol{\beta}}} - 1\right) H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \boldsymbol{U}_{\boldsymbol{\alpha}} \boldsymbol{U}_{\boldsymbol{\beta}}^\top \right\| \geq t \;\middle|\; \boldsymbol{\xi'}\right)\right].$$

We can study the conditional expression

$$\mathbb{P}\left(\left\| \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} \left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1\right)\left(\frac{\xi'_{\boldsymbol{\beta}}}{p_{\boldsymbol{\beta}}} - 1\right) H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \boldsymbol{U}_{\boldsymbol{\alpha}} \boldsymbol{U}_{\boldsymbol{\beta}}^\top \right\| \geq t \;\middle|\; \boldsymbol{\xi'}\right)$$

using Theorem 1.7.2. Defining the expression $\boldsymbol{b}_{\boldsymbol{\alpha}} = \sum_{\boldsymbol{\beta}\neq\boldsymbol{\alpha}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \left(\frac{\xi'_{\boldsymbol{\beta}}}{p_{\boldsymbol{\beta}}} - 1\right) \boldsymbol{U}_{\boldsymbol{\beta}}$, we notice that the sum can be re-written as

$$\mathbb{P}\left(\left\| \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} \left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1\right)\left(\frac{\xi'_{\boldsymbol{\beta}}}{p_{\boldsymbol{\beta}}} - 1\right) H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \boldsymbol{U}_{\boldsymbol{\alpha}} \boldsymbol{U}_{\boldsymbol{\beta}}^\top \right\| \geq t \;\middle|\; \boldsymbol{\xi'}\right) = \mathbb{P}\left(\left\| \sum_{\boldsymbol{\alpha}} (\xi_{\boldsymbol{\alpha}} - p) \boldsymbol{U}_{\boldsymbol{\alpha}} \boldsymbol{b}_{\boldsymbol{\alpha}}^\top \right\| \geq t \;\middle|\; \boldsymbol{\xi'}\right).$$

In this form, conditioned on $\boldsymbol{\xi'}$ we see that this is a sum of mean-zero, independent random

variables. To study its concentration with Theorem 1.7.2, we define $\boldsymbol{T}_{\boldsymbol{\alpha}} = \left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1\right) \boldsymbol{U}_{\boldsymbol{\alpha}} \boldsymbol{b}_{\boldsymbol{\alpha}}^{\top}$. The first task is bounding the operator norm

$$\|\boldsymbol{T}_{\boldsymbol{\alpha}}\| \leq \frac{\|\boldsymbol{U}_{\boldsymbol{\alpha}}\|_2}{p_{\boldsymbol{\alpha}}^{1/2}} \frac{1}{p_{\boldsymbol{\alpha}}^{1/2}} \|\boldsymbol{b}_{\boldsymbol{\alpha}}\|_2 \leq \frac{n^{3/2} r}{m} \|\boldsymbol{b}_{\boldsymbol{\alpha}}\|_2.$$

Next, we need to estimate the variance terms $\|\sum_{\boldsymbol{\alpha}} \mathbb{E}[\boldsymbol{T}_{\boldsymbol{\alpha}} \boldsymbol{T}_{\boldsymbol{\alpha}}^* \mid \boldsymbol{\xi}']\|$ and $\|\sum_{\boldsymbol{\alpha}} \mathbb{E}[\boldsymbol{T}_{\boldsymbol{\alpha}}^* \boldsymbol{T}_{\boldsymbol{\alpha}} \mid \boldsymbol{\xi}']\|$. For the first term, notice that

$$\begin{aligned}
\left\| \sum_{\boldsymbol{\alpha}} \mathbb{E}[\boldsymbol{T}_{\boldsymbol{\alpha}} \boldsymbol{T}_{\boldsymbol{\alpha}}^* \mid \boldsymbol{\xi}'] \right\| &= \left\| \sum_{\boldsymbol{\alpha}} \mathbb{E}\left[\left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1 \right)^2 \right] \boldsymbol{U}_{\boldsymbol{\alpha}} \boldsymbol{b}_{\boldsymbol{\alpha}}^{\top} \boldsymbol{b}_{\boldsymbol{\alpha}} \boldsymbol{U}_{\boldsymbol{\alpha}}^{\top} \right\| \\
&\leq \left\| \sum_{\boldsymbol{\alpha}} \|\boldsymbol{b}_{\boldsymbol{\alpha}}\|_2^2 \frac{1}{p_{\boldsymbol{\alpha}}} \boldsymbol{U}_{\boldsymbol{\alpha}} \boldsymbol{U}_{\boldsymbol{\alpha}}^{\top} \right\| \\
&\leq \max_{\boldsymbol{\alpha}} \|\boldsymbol{b}_{\boldsymbol{\alpha}}\|_2^2 \left\| \frac{1}{p_{\boldsymbol{\alpha}}} \sum_{\boldsymbol{\alpha}} \boldsymbol{U}_{\boldsymbol{\alpha}} \boldsymbol{U}_{\boldsymbol{\alpha}}^{\top} \right\| \\
&\leq \max_{\boldsymbol{\alpha}} \|\boldsymbol{b}_{\boldsymbol{\alpha}}\|_2^2 \lambda_{\max}(\tilde{\boldsymbol{H}}_p) \\
&\leq \frac{12 n^2 r^2}{m} \max_{\boldsymbol{\alpha}} \|\boldsymbol{b}_{\boldsymbol{\alpha}}\|_2^2.
\end{aligned}$$

The other variance term is given by

$$\begin{aligned}
\left\| \sum_{\boldsymbol{\alpha}} \mathbb{E}[\boldsymbol{T}_{\boldsymbol{\alpha}}^* \boldsymbol{T}_{\boldsymbol{\alpha}} \mid \boldsymbol{\xi}'] \right\| &= \left\| \sum_{\boldsymbol{\alpha}} \mathbb{E}\left[\left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1 \right)^2 \right] \boldsymbol{b}_{\boldsymbol{\alpha}} \boldsymbol{U}_{\boldsymbol{\alpha}}^{\top} \boldsymbol{U}_{\boldsymbol{\alpha}} \boldsymbol{b}_{\boldsymbol{\alpha}}^{\top} \right\| \\
&\leq \max_{\boldsymbol{\gamma} \in \mathbb{I}_p} \frac{\|\boldsymbol{U}_{\boldsymbol{\gamma}}\|_2^2}{p_{\boldsymbol{\gamma}}} \left\| \sum_{\boldsymbol{\alpha}} \boldsymbol{b}_{\boldsymbol{\alpha}} \boldsymbol{b}_{\boldsymbol{\alpha}}^{\top} \right\| \\
&\leq \frac{12 n r}{m} \|\boldsymbol{B} \boldsymbol{B}^{\top}\| \\
&= \frac{12 n r}{m} \|\boldsymbol{B}\|^2,
\end{aligned}$$

where the last matrix is defined as $\boldsymbol{B} = [\boldsymbol{b}_1 \cdots \boldsymbol{b}_L]$ and the spectral result follows from the vectorization lemma. Defining the variance parameter
$\sigma^2 = \max\left\{ \|\sum_{\boldsymbol{\alpha}} \mathbb{E}[\boldsymbol{T}_{\boldsymbol{\alpha}} \boldsymbol{T}_{\boldsymbol{\alpha}}^* \mid \boldsymbol{\xi}']\|, \|\sum_{\boldsymbol{\alpha}} \mathbb{E}[\boldsymbol{T}_{\boldsymbol{\alpha}}^* \boldsymbol{T}_{\boldsymbol{\alpha}} \mid \boldsymbol{\xi}']\| \right\}$, we have from the monotonicity of expectation that

$$\mathbb{E}\left[\mathbb{P}\left(\left\| \sum_{\boldsymbol{\alpha} \neq \boldsymbol{\beta}} (\xi_{\boldsymbol{\alpha}} - p)(\xi'_{\boldsymbol{\beta}} - p) H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \boldsymbol{U}_{\boldsymbol{\alpha}} \boldsymbol{U}_{\boldsymbol{\beta}}^{\top} \right\| \geq t \;\middle|\; \boldsymbol{\xi}' \right) \right] \leq \mathbb{E}\left[2n \exp\left(\frac{-t^2/2}{Rt/3 + \sigma^2} \right) \right],$$

where σ^2 and R are functions of $\boldsymbol{\xi}'$. It follows that if high probability estimates exist for R and σ^2, referred to hereafter as $R_{\max}$ and $\sigma^2_{\max}$ respectively, we can effectively uncondition R and σ^2 as follows:

$$\mathbb{E}\left[2n\exp\left(\frac{-t^2/2}{Rt/3+\sigma^2}\right)\right] \leq 2\mathbb{P}(\sigma^2 \leq \sigma^2_{\max}, R \leq R_{\max})n\exp\left(\frac{-t^2/2}{R_{\max}t/3+\sigma^2_{\max}}\right) + 2n\mathbb{P}(\sigma^2 \geq \sigma^2_{\max} \text{ or } R \geq R_{\max}).$$

Plugging in the expressions in Theorem 4.B.3 and Theorem 4.B.4 and setting

$$t = \sqrt{\frac{8}{3}\sigma^2_{\max}\beta\log n} = \frac{1152nr\beta^{1/2}\log n}{m},$$

using the simplification in Theorem 1.7.2 we have that, if $m \geq \frac{1152}{3}\beta n\log n$ for all three inequalities, the following inequality holds for each operator in the Lemma statement.

$$\begin{aligned}
&\mathbb{E}\left[2n\exp\left(\frac{-t^2/2}{Rt/3+\sigma^2}\right)\right] \\
&\quad\leq 2\mathbb{P}(\sigma^2 \leq \sigma^2_{\max}, R \leq R_{\max})n\exp\left(\frac{-t^2/2}{R_{\max}t/3+\sigma^2_{\max}}\right) + 2n\mathbb{P}(\sigma^2 \geq \sigma^2_{\max} \text{ or } R \geq R_{\max}) \\
&\quad\leq 2\mathbb{P}(\sigma^2 \leq \sigma^2_{\max}, R \leq R_{\max})n\exp\left(\frac{-3t^2}{8\sigma^2_{\max}}\right) + 2n\mathbb{P}(\sigma^2 \geq \sigma^2_{\max} \text{ or } R \geq R_{\max}) \\
&\quad\leq 2(1-4n^{2-\beta})n^{1-\beta} + 8n^{3-\beta} \\
&\quad\leq 10n^{3-\beta}.
\end{aligned}$$

Incorporating the constant C from Theorem 1.7.5 concludes the proof. □

Lemma 4.B.6. *Let $\boldsymbol{X}$ be a ground-truth, rank-r, ν-incoherent Gram matrix with tangent space $\mathbb{T}$. If the sampling probability $p \geq \frac{128\beta\log n}{3n}$ for any $\beta > 1$, then the following collection of inequalities hold with probability at least $1 - 8n^{1-\beta}$*

$$\left\|\sum_{\boldsymbol{\alpha}\in\mathbb{I}}\left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}}-1\right)\mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}\sum_{\substack{\boldsymbol{\beta}\in\mathbb{I}\\ \alpha\neq\beta}}H^{\boldsymbol{\alpha}\boldsymbol{\beta}}\langle\cdot,\mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\beta}}\rangle\right\| \leq \min\left\{\frac{20n^{3/2}r\beta^{1/2}\log n}{m}, \sqrt{\frac{96n\nu r^2\beta\log n}{3m}}\right\}.$$

Proof. We first notice that, for a fixed $\boldsymbol{\alpha}$, the sum $\sum_{\boldsymbol{\beta}\neq\boldsymbol{\alpha}}H^{\boldsymbol{\alpha}\boldsymbol{\beta}}\mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\beta}} = \mathcal{P}_{\mathbb{T}}\boldsymbol{v}_{\boldsymbol{\alpha}} - H^{\boldsymbol{\alpha}\boldsymbol{\alpha}}\mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}$ from

the definition of $\boldsymbol{v}_{\boldsymbol{\alpha}}$ in (1.9). As such, the above expression can be bounded as

$$\begin{aligned}
&\left\| \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1 \right) \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \sum_{\substack{\boldsymbol{\beta}\in\mathbb{I} \\ \alpha\neq\beta}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\beta}} \rangle \right\| \\
&\qquad = \left\| \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1 \right) \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \left(\langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{v}_{\boldsymbol{\alpha}} \rangle - H^{\boldsymbol{\alpha}\boldsymbol{\alpha}} \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \right) \right\| \\
&\qquad \leq \left\| \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1 \right) \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{v}_{\boldsymbol{\alpha}} \rangle \right\| \\
&\qquad\quad + \left\| \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1 \right) \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} H^{\boldsymbol{\alpha}\boldsymbol{\alpha}} \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \right\| \\
&\qquad = \| \mathcal{P}_{\mathbb{T}} \mathcal{R}^*_{p,\Omega} \mathcal{P}_{\mathbb{T}} - \mathcal{P}_{\mathbb{T}} \| + \| \mathcal{P}_{\mathbb{T}} \mathcal{F}_{p,\Omega} \mathcal{P}_{\mathbb{T}} - \mathcal{P}_{\mathbb{T}} \mathcal{F}_{\mathbb{I}} \mathcal{P}_{\mathbb{T}} \|.
\end{aligned}$$

We note that the latter expression has already been bounded in Theorem 4.B.1. We now seek to prove a concentration result for the first expression.

We note that this is the sum of independent random variables $\mathcal{T}_\alpha = \left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1 \right) \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{v}_{\boldsymbol{\alpha}} \rangle \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}}$. We can first see that

$$\| \mathcal{T}_\alpha \| \leq \frac{\| \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \|_{\mathrm{F}}}{p_{\boldsymbol{\alpha}}^{1/2}} \frac{\| \mathcal{P}_{\mathbb{T}} \boldsymbol{v}_{\boldsymbol{\alpha}} \|_{\mathrm{F}}}{p_{\boldsymbol{\alpha}}^{1/2}} \leq \frac{6nr}{m}.$$

Next, we notice that

$$\begin{aligned}
\left\| \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \mathbb{E} \left[\mathcal{T}_\alpha \mathcal{T}_\alpha^* \right] \right\| &\leq \left\| \sum_{\alpha\in\mathbb{I}_p} \frac{1}{p_{\boldsymbol{\alpha}}} \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \langle \mathcal{P}_{\mathbb{T}} \boldsymbol{v}_{\boldsymbol{\alpha}}, \mathcal{P}_{\mathbb{T}} \boldsymbol{v}_{\boldsymbol{\alpha}} \rangle \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \right\| \\
&\leq \frac{48nr}{m} \left\| \sum_{\alpha\in\mathbb{I}_p} \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \right\| \\
&\leq \frac{48nr}{m} \lambda_{\max} \left(\tilde{\boldsymbol{H}}_p \right) \\
&\leq \max \left\{ \frac{144n^3 r^2}{m^2}, \frac{12n\nu r^2}{m} \right\}.
\end{aligned}$$

The other variance term can be analyzed as

$$\left\| \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \mathbb{E} \left[\mathcal{T}_\alpha^* \mathcal{T}_\alpha \right] \right\| \leq \left\| \sum_{\alpha\in\mathbb{I}_p} \frac{1}{p_{\boldsymbol{\alpha}}} \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{v}_{\boldsymbol{\alpha}} \rangle \langle \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}}, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \mathcal{P}_{\mathbb{T}} \boldsymbol{v}_{\boldsymbol{\alpha}} \right\|$$

$$\begin{aligned}
&\leq \frac{12nr}{m} \left\| \sum_{\alpha \in \mathbb{I}_p} \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{v}_{\boldsymbol{\alpha}} \rangle \mathcal{P}_{\mathbb{T}} \boldsymbol{v}_{\boldsymbol{\alpha}} \right\| \\
&\leq \frac{12nr}{m} \lambda_{\max}(\boldsymbol{H}^{-1}) \\
&\leq \frac{6nr}{m}.
\end{aligned}$$

We have now from Theorem 1.7.2 that

$$\mathbb{P}\left(\|\mathcal{P}_{\mathbb{T}} \mathcal{R}_{p,\Omega} \mathcal{P}_{\mathbb{T}} - \mathcal{P}_{\mathbb{T}}\| \geq \max\left\{ \sqrt{\frac{96n\nu r^2 \beta \log n}{3m}}, \frac{20n^{3/2} r \beta^{1/2} \log n}{m} \right\} \right) \leq 2n^{1-\beta},$$

following the simplification seen in Theorem 1.7.2, thus yielding the stated result when combined with Theorem 4.B.1. □

4.B.2 Proof of Theorem 4.5.1

We are now ready to prove Theorem 4.5.1.

Proof. First, notice that since $\mathcal{M}_{p,\Omega} = \mathcal{M}^*_{p,\Omega}$,

$$\begin{aligned}
\|\mathcal{P}_{\mathbb{T}} \mathcal{M}_{p,\Omega} \mathcal{P}_{\mathbb{T}} - p^2 \mathcal{P}_{\mathbb{T}}\| &= \left\| \sum_{\boldsymbol{\alpha},\boldsymbol{\beta} \in \Omega} \frac{\xi_{\boldsymbol{\alpha}} \xi_{\boldsymbol{\beta}}}{p_{\boldsymbol{\alpha}} p_{\boldsymbol{\beta}}} C_{\boldsymbol{\alpha}\boldsymbol{\beta}} \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}} \rangle \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\beta}} \rangle - \mathcal{P}_{\mathbb{T}} \right\| \\
&= \left\| \sum_{\boldsymbol{\alpha},\boldsymbol{\beta} \in \Omega} \frac{\xi_{\boldsymbol{\alpha}} \xi_{\boldsymbol{\beta}}}{p_{\boldsymbol{\alpha}} p_{\boldsymbol{\beta}}} C_{\boldsymbol{\alpha}\boldsymbol{\beta}} \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}} \rangle \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\beta}} \rangle - \sum_{\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathbb{I}} \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}} \rangle \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\beta}} \rangle \right\| \\
&= \left\| \sum_{\boldsymbol{\alpha} \in \Omega} \frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\alpha}} \rangle + \sum_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta} \in \Omega \\ \boldsymbol{\alpha} \neq \boldsymbol{\beta}}} \frac{\xi_{\boldsymbol{\alpha}} \xi_{\boldsymbol{\beta}}}{p_{\boldsymbol{\alpha}} p_{\boldsymbol{\beta}}} \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}} \rangle \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\beta}} \rangle \right. \\
&\qquad \left. \sum_{\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathbb{I}} \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}} \rangle \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\beta}} \rangle \right\| \\
&= \left\| \|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 \sum_{\boldsymbol{\alpha} \in \mathbb{I}} \left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1 \right) \langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} + \sum_{\boldsymbol{\alpha} \neq \boldsymbol{\beta}} \left(\frac{\xi_{\boldsymbol{\alpha}} \xi_{\boldsymbol{\beta}}}{p_{\boldsymbol{\alpha}} p_{\boldsymbol{\beta}}} - 1 \right) \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\beta}} \rangle \right\| \\
&\leq \left\| \|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 \sum_{\boldsymbol{\alpha} \in \mathbb{I}} \left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1 \right) \langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \right\| + \left\| \sum_{\boldsymbol{\alpha} \neq \boldsymbol{\beta}} \left(\frac{\xi_{\boldsymbol{\alpha}} \xi_{\boldsymbol{\beta}}}{p_{\boldsymbol{\alpha}} p_{\boldsymbol{\beta}}} - 1 \right) \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\beta}} \rangle \right\| \\
&= \|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 \, \|\mathcal{P}_{\mathbb{T}} \mathcal{F}_{p,\Omega} \mathcal{P}_{\mathbb{T}} - \mathcal{P}_{\mathbb{T}} \mathcal{F}_{\mathbb{I}} \mathcal{P}_{\mathbb{T}}\| + \left\| \sum_{\boldsymbol{\alpha} \neq \boldsymbol{\beta}} \left(\frac{\xi_{\boldsymbol{\alpha}} \xi_{\boldsymbol{\beta}}}{p_{\boldsymbol{\alpha}} p_{\boldsymbol{\beta}}} - 1 \right) \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\beta}} \rangle \right\|
\end{aligned}$$

$$\leq \underbrace{\|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 \, \|\mathcal{P}_{\mathbb{T}}\mathcal{F}_{p,\Omega}\mathcal{P}_{\mathbb{T}} - \mathcal{P}_{\mathbb{T}}\mathcal{F}_{\mathbb{I}}\mathcal{P}_{\mathbb{T}}\|}_{B_1} + \underbrace{\left\| \sum_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}} \left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1\right)\left(\frac{\xi_{\boldsymbol{\beta}}}{p_{\boldsymbol{\beta}}} - 1\right) \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\beta}}\rangle \right\|}_{B_2}$$

$$+ 2\underbrace{\left\| \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1\right) \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}} \sum_{\substack{\boldsymbol{\beta}\in\mathbb{I} \\ \boldsymbol{\alpha}\neq\boldsymbol{\beta}}} H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\beta}}\rangle \right\|}_{B_3},$$

where the first inequality follows from the triangle inequality and the final inequality follows from Theorem 4.B.2. We note that term B_1 is bounded by Theorem 4.B.1, and by the fact that $\|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}$ is constant for all $\boldsymbol{\alpha} \in \mathbb{I}$ and that $\|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 \leq \frac{1}{2}$ from Theorem 2.A.3. We bound B_2 using Theorem 4.B.5 and we bound B_3 using Theorem 4.B.6. This concludes the proof. □

Lemma 4.B.7. *Let $\mathcal{M}_{p,\Omega}$ defined as above. Then for $m \geq \sqrt{5}n \log n$ we have that*

$$\|\mathcal{M}_{p,\Omega}\| \leq n^3.$$

Proof. We notice that

$$\begin{aligned}
\|\mathcal{M}_{p,\Omega}\| &= \left\| \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1\right) \langle \cdot, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 \boldsymbol{w}_{\boldsymbol{\alpha}} + \sum_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{I} \\ \boldsymbol{\alpha}\neq\boldsymbol{\beta}}} \left(\frac{\xi_{\boldsymbol{\alpha}}\xi_{\boldsymbol{\beta}}}{p_{\boldsymbol{\alpha}}p_{\boldsymbol{\beta}}} - 1\right) \langle \cdot, \boldsymbol{w}_{\boldsymbol{\beta}}\rangle H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \boldsymbol{w}_{\boldsymbol{\alpha}} \right\| \\
&\leq \left\| \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \left(\frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} - 1\right) \langle \cdot, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 \boldsymbol{w}_{\boldsymbol{\alpha}} \right\| + \left\| \sum_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{I} \\ \boldsymbol{\alpha}\neq\boldsymbol{\beta}}} \left(\frac{\xi_{\boldsymbol{\alpha}}\xi_{\boldsymbol{\beta}}}{p_{\boldsymbol{\alpha}}p_{\boldsymbol{\beta}}} - 1\right) \langle \cdot, \boldsymbol{w}_{\boldsymbol{\beta}}\rangle H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \boldsymbol{w}_{\boldsymbol{\alpha}} \right\| \\
&\leq \left\| \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \frac{\xi_{\boldsymbol{\alpha}}}{p_{\boldsymbol{\alpha}}} \langle \cdot, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 \boldsymbol{w}_{\boldsymbol{\alpha}} \right\| + \left\| \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \cdot, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 \boldsymbol{w}_{\boldsymbol{\alpha}} \right\| \\
&\quad + \left\| \sum_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{I} \\ \boldsymbol{\alpha}\neq\boldsymbol{\beta}}} \frac{\xi_{\boldsymbol{\alpha}}\xi_{\boldsymbol{\beta}}}{p_{\boldsymbol{\alpha}}p_{\boldsymbol{\beta}}} \langle \cdot, \boldsymbol{w}_{\boldsymbol{\beta}}\rangle H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \boldsymbol{w}_{\boldsymbol{\alpha}} \right\| + \left\| \sum_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{I} \\ \boldsymbol{\alpha}\neq\boldsymbol{\beta}}} \langle \cdot, \boldsymbol{w}_{\boldsymbol{\beta}}\rangle H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \boldsymbol{w}_{\boldsymbol{\alpha}} \right\| \\
&\leq \left(\frac{n^2}{m} + 1\right) \left\| \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \cdot, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 \boldsymbol{w}_{\boldsymbol{\alpha}} \right\| + \left(\frac{n^4}{m^2} + 1\right) \left\| \sum_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{I} \\ \boldsymbol{\alpha}\neq\boldsymbol{\beta}}} \langle \cdot, \boldsymbol{w}_{\boldsymbol{\beta}}\rangle H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \boldsymbol{w}_{\boldsymbol{\alpha}} \right\|
\end{aligned}$$

$$
\begin{aligned}
&\leq \left(\frac{n^2}{m}+1\right)\lambda_{\max}(\boldsymbol{H}) + \left(\frac{n^4}{m^2}+1\right)\left\|\sum_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{I}}\langle\cdot,\boldsymbol{w}_{\boldsymbol{\beta}}\rangle H^{\boldsymbol{\alpha}\boldsymbol{\beta}}\boldsymbol{w}_{\boldsymbol{\alpha}} - \sum_{\boldsymbol{\alpha}\in\mathbb{I}}\langle\cdot,\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle H^{\boldsymbol{\alpha}\boldsymbol{\alpha}}\boldsymbol{w}_{\boldsymbol{\alpha}}\right\| \\
&\leq 2n\left(\frac{n^2}{m}+1\right) + \left(\frac{n^4}{m^2}+1\right)\left(\left\|\sum_{\boldsymbol{\alpha}\in\mathbb{I}}\langle\cdot,\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\boldsymbol{v}_{\boldsymbol{\alpha}}\right\| + H^{\boldsymbol{\alpha}\boldsymbol{\alpha}}\left\|\sum_{\boldsymbol{\alpha}\in\mathbb{I}}\langle\cdot,\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\boldsymbol{w}_{\boldsymbol{\alpha}}\right\|\right) \\
&\leq 2n\left(\frac{n^2}{m}+1\right) + \left(\frac{n^4}{m^2}+1\right)(1+2n) \\
&\leq \frac{5n^5}{m^2} \\
&\leq \frac{n^3}{\log n} \\
&\leq n^3,
\end{aligned}
$$

where the first two inequalities follow from the triangle inequality, the third inequality follows from Theorem 4.A.1, the fourth inequality follows from the definition of $\boldsymbol{v}_{\boldsymbol{\alpha}}$, the sixth inequality follows from the duality of $\boldsymbol{w}_{\boldsymbol{\alpha}}$ and $\boldsymbol{v}_{\boldsymbol{\alpha}}$, along with the triangle inequality, and the second to last inequality follows from the stipulation that $m \geq \sqrt{5}n\log n$. □

Lemma 4.B.8 (Local RIP of $\mathcal{M}_{p,\Omega}$)**.** *Assume that*

$$\|\mathcal{P}_{\mathbb{T}}\mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}} - \mathcal{P}_{\mathbb{T}}\| \leq \varepsilon_0, \tag{4.6}$$

$$\frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\lambda_r(\boldsymbol{X})} \leq \frac{\varepsilon_0}{2n^3}. \tag{4.7}$$

Then

$$\|\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_{\mathbb{T}_l}\| \leq 4\varepsilon_0.$$

Proof. Notice that

$$
\begin{aligned}
\|\mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_{\mathbb{T}_l}\mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}_l}\| &= \|\mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_{\mathbb{T}_l}\mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}_l} \\
&\qquad + \mathcal{P}_{\mathbb{T}} - \mathcal{P}_{\mathbb{T}} + \mathcal{P}_{\mathbb{T}_l}\mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}} - \mathcal{P}_{\mathbb{T}_l}\mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}} + \mathcal{P}_{\mathbb{T}}\mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}} - \mathcal{P}_{\mathbb{T}}\mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}}\| \\
&\leq \|\mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_{\mathbb{T}}\| + \|\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_{\mathbb{T}_l}\mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}}\| \\
&\quad + \|\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}} - \mathcal{P}_{\mathbb{T}}\mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}}\| + \|\mathcal{P}_{\mathbb{T}} - \mathcal{P}_{\mathbb{T}}\mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}}\| \\
&\leq \|\mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_{\mathbb{T}}\| + \|\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_{p,\Omega}\|\|\mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_{\mathbb{T}}\| \\
&\quad + \|\mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}}\|\|\mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_{\mathbb{T}}\| + \|\mathcal{P}_{\mathbb{T}} - \mathcal{P}_{\mathbb{T}}\mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}}\|
\end{aligned}
$$

$$
\begin{aligned}
&\leq \frac{2\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\lambda_r(\boldsymbol{X})}\left(1 + \|\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_{p,\Omega}\| + \|\mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}}\|\right) + \|\mathcal{P}_{\mathbb{T}} - \mathcal{P}_{\mathbb{T}}\mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}}\| \\
&\leq \frac{2\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\lambda_r(\boldsymbol{X})}\left(1 + 2\|\mathcal{M}_{p,\Omega}\|\right) + \|\mathcal{P}_{\mathbb{T}} - \mathcal{P}_{\mathbb{T}}\mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}}\| \\
&\leq 2\varepsilon_0 + \frac{2\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\lambda_r(\boldsymbol{X})}\left(2n^3\right) \\
&= 2\varepsilon_0 + 2\frac{\varepsilon_0}{2n^3}2n^3 \\
&\leq 4\varepsilon_0,
\end{aligned}
$$

where the first inequality is the triangle inequality, the second inequality is Cauchy-Schwarz, the third inequality is due to Theorem 1.8.3 and (4.6), the fourth inequality is a result of Theorem 4.B.7, and the final inequality is due to (4.7). □

4.C Local Convergence Results

We begin with the following technical lemmas used in the proof of local convergence.

Lemma 4.C.1 (Algorithm 9 Stepsize Bounds)**.** *Assume that* $\|\mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_{\mathbb{T}_l}\mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}_l}\| \leq 4\varepsilon_0 < 1$. *Then the stepsize* α_l *in W-DBRE can be bounded by*

$$
\frac{1}{1+4\varepsilon_0} \leq \alpha_l = \frac{\|\mathcal{P}_{\mathbb{T}}\boldsymbol{G}_l\|_{\mathrm{F}}^2}{\langle \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l, \mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l\rangle} \leq \frac{1}{1-4\varepsilon_0}.
$$

Proof. We will prove this by leveraging the local RIP assumption. Notice the following:

$$
\begin{aligned}
\langle \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l, \mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l\rangle &= \langle \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l, \mathcal{P}_{\mathbb{T}_l}\mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l\rangle \\
&= \langle \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l, \mathcal{P}_{\mathbb{T}_l}\mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l - \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l\rangle + \langle \mathcal{P}_{\mathbb{T}}\boldsymbol{G}_l, \mathcal{P}_{\mathbb{T}}\boldsymbol{G}_l\rangle.
\end{aligned}
$$

We can now leverage the variational characterization of the spectral norm and local RIP, proven in Theorem 4.B.8, to bound the following:

$$
-(4\varepsilon_0)\|\mathcal{P}_{\mathbb{T}}\boldsymbol{G}_l\|_{\mathrm{F}}^2 \leq \langle \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l, \mathcal{P}_{\mathbb{T}_l}\mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l - \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l\rangle \leq (4\varepsilon_0)\|\mathcal{P}_{\mathbb{T}}\boldsymbol{G}_l\|_{\mathrm{F}}^2.
$$

As such, we can now bound the denominator as

$$(1-4\varepsilon_0)\|\mathcal{P}_{\mathbb{T}}\boldsymbol{G}_l\|_{\mathrm{F}}^2 \leq \langle \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l, \mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l\rangle \leq (1+4\varepsilon_0)\|\mathcal{P}_{\mathbb{T}}\boldsymbol{G}_l\|_{\mathrm{F}}^2.$$

Rearrangement of this last expression yields the upper and lower bounds on the step size derived above. □

Lemma 4.C.2 (I_1 Bound)**.** *Assume $\|\mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_{\mathbb{T}_l}\mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}_l}\| \leq 4\varepsilon_0$ and α_l can be bounded as in Theorem 4.C.1. Then the spectral norm of $\mathcal{P}_{\mathbb{T}_l} - \alpha_l\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}_l}$ can be bounded as*

$$\|\mathcal{P}_{\mathbb{T}_l} - \alpha_l\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}_l}\| \leq \frac{8\varepsilon_0}{1-4\varepsilon_0}. \tag{4.8}$$

Proof. From direct calculation, it follows that

$$\begin{aligned}
\|\mathcal{P}_{\mathbb{T}_l} - \alpha_l\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}_l}\| &\leq \|\mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_{\mathbb{T}_l}\mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}_l}\| + |\alpha_l - 1|\,\|\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}_l}\| \\
&\leq 4\varepsilon_0 + |\alpha_l - 1|\,(\|\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_{\mathbb{T}_l}\| + \|\mathcal{P}_{\mathbb{T}_l}\|) \\
&\leq 4\varepsilon_0 + \left(\frac{1}{1-4\varepsilon_0} - \frac{(1-4\varepsilon_0)}{1-4\varepsilon_0}\right)(\|\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_{p,\Omega}\mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_{\mathbb{T}_l}\| + \|\mathcal{P}_{\mathbb{T}_l}\|) \\
&\leq 4\varepsilon_0 + \left(\frac{1}{1-4\varepsilon_0} - \frac{(1-4\varepsilon_0)}{1-4\varepsilon_0}\right)(4\varepsilon_0 + 1) \\
&= 4\varepsilon_0 + \frac{4\varepsilon_0}{1-4\varepsilon_0}(1+4\varepsilon_0) \\
&= \frac{8\varepsilon_0}{1-4\varepsilon_0},
\end{aligned}$$

where the first inequality comes from the triangle inequality, the second inequality comes from Local RIP in Theorem 4.B.8, the third inequality comes from the stepsize bound in Theorem 4.C.1, the fourth inequality again comes from Theorem 4.B.8, and the remainder comes from algebraic simplification of terms. This finishes the proof. □

4.C.1 Proof of Theorem 4.5.2

We can now prove Theorem 4.5.2.

Proof. First, it follows that

$$\|\boldsymbol{X}_{l+1} - \boldsymbol{X}\|_{\mathrm{F}} \leq \|\boldsymbol{X}_{l+1} - \boldsymbol{W}_l\|_{\mathrm{F}} + \|\boldsymbol{W}_l - \boldsymbol{X}\|_{\mathrm{F}} \leq 2\|\boldsymbol{W}_l - \boldsymbol{X}\|_{\mathrm{F}},$$

as $\boldsymbol{X}_{l+1}$ is the best rank-r approximation of $\boldsymbol{W}_l$. Plugging in $\boldsymbol{W}_l = \boldsymbol{X}_l + \alpha_l \mathcal{P}_{\mathbb{T}_l} \boldsymbol{G}_l$, we see that

$$\begin{aligned}
\|\boldsymbol{X}_{l+1} - \boldsymbol{X}\|_{\mathrm{F}} &\leq 2\left\|\boldsymbol{X}_l + \alpha_l \mathcal{P}_{\mathbb{T}_l} \boldsymbol{G}_l - \boldsymbol{X}\right\|_{\mathrm{F}} \\
&= 2\|\boldsymbol{X}_l - \boldsymbol{X} - \alpha_l \mathcal{P}_{\mathbb{T}_l} \mathcal{M}_{p,\Omega}(\boldsymbol{X}_l - \boldsymbol{X})\|_{\mathrm{F}} \\
&\leq \underbrace{2\|(\mathcal{P}_{\mathbb{T}_l} - \alpha_l \mathcal{P}_{\mathbb{T}_l} \mathcal{M}_{p,\Omega} \mathcal{P}_{\mathbb{T}_l})(\boldsymbol{X}_l - \boldsymbol{X})\|_{\mathrm{F}}}_{I_1} \\
&\quad + \underbrace{2\|(I - \mathcal{P}_{\mathbb{T}_l})(\boldsymbol{X}_l - \boldsymbol{X})\|_{\mathrm{F}}}_{I_2} \\
&\quad + \underbrace{2|\alpha_l| \|\mathcal{P}_{\mathbb{T}_l} \mathcal{M}_{p,\Omega}(I - \mathcal{P}_{\mathbb{T}_l})(\boldsymbol{X}_l - \boldsymbol{X})\|_{\mathrm{F}}}_{I_3}.
\end{aligned}$$

It remains to bound each term individually. Using Theorem 4.C.2, we see that

$$I_1 \leq \frac{16\varepsilon_0}{1 - 4\varepsilon_0} \|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}.$$

Next, notice that from Theorem 1.8.3 and the fact that $\mathcal{P}_{\mathbb{T}_l} \boldsymbol{X}_l = \boldsymbol{X}_l$,

$$\begin{aligned}
I_2 &= 2\|(I - \mathcal{P}_{\mathbb{T}_l})\boldsymbol{X}_l - (I - \mathcal{P}_{\mathbb{T}_l})\boldsymbol{X})\|_{\mathrm{F}} \\
&= 2\|(I - \mathcal{P}_{\mathbb{T}_l})\boldsymbol{X}\|_{\mathrm{F}} \\
&\leq \frac{2\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}^2}{\lambda_r(\boldsymbol{X})} \\
&\leq \frac{\varepsilon_0}{Cn^3} \|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}} \\
&\leq \varepsilon_0 \|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}} \\
&\leq \frac{\varepsilon_0}{1 - 4\varepsilon_0} \|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}},
\end{aligned}$$

using Theorem 1.8.3 and our initial local neighborhood assumption. Finally, we see that,

following a similar argument as in the bound of I_2,

$$\begin{aligned} I_3 &\leq 2|\alpha_l| \|\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_{p,\Omega}\| \|(I - \mathcal{P}_{\mathbb{T}_l})\boldsymbol{X}\|_{\mathrm{F}} \\ &\leq \frac{2}{1-4\varepsilon_0}\|\mathcal{M}_{p,\Omega}\| \left(\frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\lambda_r(\boldsymbol{X})}\right) \|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}} \\ &\leq \frac{2}{1-4\varepsilon_0} 2n^3 \frac{\varepsilon_0}{2n^3} \|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}} \\ &\leq \frac{\varepsilon_0}{1-4\varepsilon_0} \|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}, \end{aligned}$$

where the second to last inequality follows from the same analysis conducted in Theorem 4.B.8, scaled by a factor of $\frac{1}{2}$. Collecting these results, we get

$$\|\boldsymbol{X}_{l+1} - \boldsymbol{X}\|_{\mathrm{F}} \leq \frac{18\varepsilon_0}{1-4\varepsilon_0} \|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}.$$

By the assumption of the theorem, which holds for $l = 0$, and as we have a contractive sequence, it inductively follows that the assumption holds for $l \geq 0$. This concludes the proof. □

Chapter 5

Conclusion and Future Outlook

5.1 Summary of Results

Throughout the course of this thesis, we have investigated various first-order Riemannian techniques to solve the Euclidean Distance Matrix Completion (EDMC) problem. By reformulating the completion task from the distance matrix directly onto the manifold of low-rank Gram matrices using a dual-basis approach, we developed and analyzed several novel algorithms.

In Chapter 2, we introduced a Riemannian pseudo-gradient descent method based on the asymmetric sampling operator $\mathcal{R}_\Omega$ in Algorithm 4 (Riemannian Pseudogradient Descent). We established the Restricted Isometry Property (RIP) for this operator under uniform sampling and provided rigorous local linear convergence and initialization guarantees. We also introduced $\mathcal{F}_\Omega$ and the corresponding Algorithm 3 (RF-RGD), but we do not provide a proof of convergence in this Chapter. We noted that, although Riemannian Pseudogradient Descent is provably convergent, it is numerically outperformed by Algorithm 3 and the non-convex algorithm in [TL18].

In Chapter 3, we improved upon this foundation by constructing $\mathcal{M}_\Omega$, a symmetric and diagonally re-weighted operator that allows for true Riemannian gradient descent (DBRE), seen in Algorithm 7. Through the application of advanced decoupling inequalities and non-commutative concentration bounds, we proved RIP, local convergence, and critically, robustness guarantees against geometric perturbations. This algorithm bridged the theoretical gap between Riemannian Pseudogradient Descent and RF-RGD, exhibiting provable guarantees and state-

of-the-art numerical performance.

Chapter 4 marked a significant departure from standard non-convex EDMC literature by dropping the uniform sampling assumption. To recover highly coherent or pathological point configurations, we designed a non-uniform sampling scheme based on geometric leverage scores. We proposed the weighted operator $\mathcal{M}_{p,\Omega}$ and corresponding algorithm in Algorithm 9 (W-DBRE), and successfully established its theoretical convergence. This approach bridged the gap for the recoverability of coherent datapoints, and even for approximated geometric leverage scores Algorithm 9 exhibited state-of-the-art recovery. This concluded the main body of the thesis.

For the curious reader, we finish the story by providing a sketch of the convergence analysis of RF-RGD in Section A.1. Although $\mathcal{F}_\Omega$ lacks standard RIP concentration around the identity, the provided proof sketch that demonstrates that local convergence can still be achieved by leveraging a second-order approximation of the retraction map, offering a non-RIP analytical avenue for algorithmic recovery. This convergence result is weaker than those presented in Chapters 2-4 due to the technique used, as the second-order approximation of the retraction map requires a very small local neighborhood to control perturbations in the approximation, and a more rigorous proof is a possible future direction to investigate.

5.2 Future Work

We will now outline some immediate future directions of inquiry in this domain.

5.2.1 Sampling Models

The most interesting line of inquiry in the EDMC problem, by my estimation, is the extension of rigorous convergence analysis to various sampling models. There are a handful of sampling models that are well-studied: examples include

- Uniform sampling—discussed in this thesis as well as, non-exhaustively, in [LL17, TL18, GTK24, LS25].

- Leverage score sampling—see Chapter 4.

- Uniform random geometric graphs—see [JM12]. This paper specifically shows that for random geometric graphs uniformly distributed in $[-0.5, 0.5]^d$ that recoverability is possible for an edge connection radius $r \gtrsim (\log(n)/n)^{1/d}$. See Definition 5.2.1 for a definition of random geometric graphs.

- Anchor-based strategies using trilateration [LLM08, LLMM14].

This is a rich body of theory in the EDG/EDMC problem, but this omits a wide class of interesting models to sample. Of particular interest are the following models:

- **Random d-regular graphs**—One direction of study we have attempted during the course of my thesis work is sampling using random d-regular graphs. For clarity, this could be thought of as where each point $\boldsymbol{p}_i$ has d observed distances connected to it; mathematically, for each fixed i, $|\{j : \|\boldsymbol{p}_i - \boldsymbol{p}_j\|_2^2 \text{ is known}\}| = d$. This is inspired from a line of work in the matrix completion literature, where reconstruction guarantees were derived for the standard matrix completion problem in [BJ14, BV20, XLL24]. The cited approaches use deterministic, linear-algebraic properties of a certain class of d-regular graphs that heavily leverage structural sparsity of the operator $\mathcal{P}_\Omega$, along with a nuclear norm minimization algorithm styled after [Rec11a]. It remains to be seen if this is possible for any of the studied operators, such as $\mathcal{R}_\Omega$, $\mathcal{F}_\Omega$, or $\mathcal{M}_\Omega$, although the dual-basis structure would likely make direct computation infeasible. Existing techniques in this paper, leveraging tools like Theorem 1.7.2, are inapplicable due to the highly dependent structure of the graph—each edge is dependent on the total number of edges connected to a given node, which eliminates the ability to decompose any measurement operator as the sum of independent random operators.

- **Other Random Geometric Graphs**—Although we stated that there is a paper that discusses the recoverability under a random geometric graph model in [JM12], to my knowledge there is still more that can be done in this area. For clarity, we define a random geometric graph as follows.

Definition 5.2.1: Random Geometric Graph

Let $(\mathcal{X}, d)$ be a metric space, and let $V = \{X_1, X_2, \ldots, X_n\}$ be a set of n random variables representing points distributed independently and identically over $\mathcal{X}$ according to a specified probability measure (typically the uniform distribution). For a given threshold radius $r > 0$, the **random geometric graph**, denoted $G(n, r)$, is defined as the undirected graph with vertex set V, where an edge e_{ij} exists between distinct vertices X_i and X_j if and only if:

$$d(X_i, X_j) \leq r.$$

In [JM12], an SDP solver was shown to recover random geometric graphs that generated from a uniform distribution on $[-0.5, 0.5]^d$. It remains to be seen if non-convex methodologies can exhibit provable recovery in this setting. Additionally, it would be interesting to see if other random geometric graphs on $[-0.5, 0.5]^d$ that are not uniformly distributed would be recoverable.

5.2.2 Other Algorithms

Semidefinite Approximation

Much of the work in this paper was dedicated to the convergence of Algorithm 3, Riemannian Pseudogradient Descent, and Algorithm 7 under the uniform sampling assumption, and subsequent convergence guarantees were provided in each of Chapters 2, 3, and 5. However, it is possible that other, faster non-convex algorithms exist with similar convergence guarantees. Take for example the Augmented Lagrangian algorithm outlined in [TL18]. This non-convex algorithm does not currently have any known theoretical convergence guarantees. However, in Table 2.1 and Figure 3.3, [TL18] performs either comparably-to our outperforms the non-convex algorithms studied in this work and in [GTK24]. To add onto this, the algorithm in [TL18] is simpler than the construction for Algorithm 7, and is computationally faster as well.

Before discussing a possible approach to the convergence theory of [TL18], we briefly introduce the Burer-Monteiro factorization of a semidefinite program (SDP). The Burer-Monteiro factorization is a non-convex formulation used to efficiently solve large-scale SDPs by replacing

the PSD matrix variable with a low-rank factorization. The standard SDP, for some linear operator $\mathcal{A} : \mathbb{R}^{n\times n} \to \mathbb{R}^m$, can be expressed as follows:

$$\min_{\boldsymbol{X}\succeq 0} \ \langle \boldsymbol{C}, \boldsymbol{X}\rangle \quad \text{subject to} \quad \mathcal{A}(\boldsymbol{X}) = \boldsymbol{b}. \tag{5.1}$$

The Burer-Monteiro approach factorizes the matrix variable $\boldsymbol{X} \in \mathbb{R}^{n\times n}$ as $\boldsymbol{X} = \boldsymbol{Y}\boldsymbol{Y}^\top$, where $\boldsymbol{Y} \in \mathbb{R}^{n\times p}$ and the rank p is chosen such that $p \ll n$. The problem is thus reformulated as:

$$\min_{\boldsymbol{Y}\in\mathbb{R}^{n\times p}} \ \langle \boldsymbol{C}, \boldsymbol{Y}\boldsymbol{Y}^\top\rangle \quad \text{subject to} \quad \mathcal{A}(\boldsymbol{Y}\boldsymbol{Y}^\top) = \boldsymbol{b}.$$

This approximation is non-convex, but is computationally more tractable for large-scale optimization tasks [BM03, BVB16]. In fact, the approximation in [TL18] came about from the difficulties of performing trace minimization directly on Gram matrices in the EDMC problem.

Recently in [DLY25], the authors proved the local convergence of the sub-problems of an Augmented Lagrangian under a Burer-Monteiro approach to SDPs. This convergence analysis represents an important step forward in understanding non-convex approximations to SDPs, and when convergence theory can be rigorously established. The convergence analysis relies on a handful of "well-posedness" properties of the underlying SDP. As the EDMC theory in [TL18] is itself a special case of an SDP (setting $\boldsymbol{C} = \boldsymbol{I}$ in (5.1)), it is possible that convergence analysis for the non-convex algorithm can be completed. This would be a meaningful addition to the EDMC literature, as much of the investigation into the complicated surrogates in this work is a result of a lack of existing convergence analysis of [TL18].

A Better Convergence Proof for Algorithm 3

In Section A.1, we provide a proof sketch of convergence of RF-RGD in a very small radius around a ground truth solution matrix $\boldsymbol{X}$. In order to make this proof rigorous, a tighter understanding of the second-order approximation of the retraction map on $\mathcal{N}_r$ is needed. Another approach, however, could prove more beneficial. If the eigenmatrices of $\mathcal{F}_{\mathbb{I}}$ were better characterized, and the majority of a given residual $\boldsymbol{X}_l - \boldsymbol{X}$ lies in a better-conditioned part of the spectrum of $\mathcal{F}_{\mathbb{I}}$, we could adapt the proof in Section A.1 to use the concentration result to

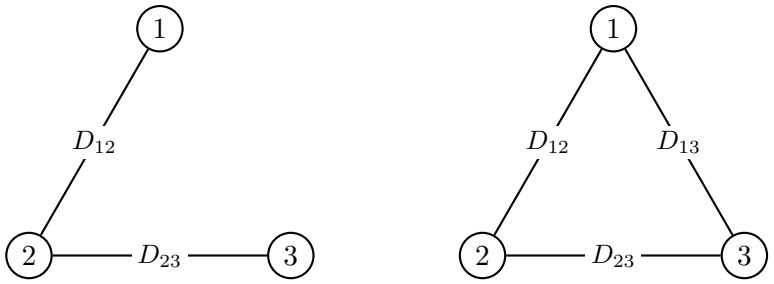


Figure 5.1: Comparison between a rigid and non-rigid graph

bound $\left\|\mathcal{P}_{\mathbb{T}}\mathcal{F}_{\Omega}\mathcal{P}_{\mathbb{T}} - \frac{m}{L}\mathcal{P}_{\mathbb{T}}\mathcal{F}_{\mathbb{I}}\mathcal{P}_{\mathbb{T}}\right\|$ and then use spectral information on $\mathcal{F}_{\mathbb{I}}$ to build a contractive sequence. It remains to be seen if such a characterization is feasible, but this approach would allow for more careful convergence theory that does not rely on the complicated second-order approximation of $\mathcal{N}_r$.

5.2.3 Connections to Rigidity Theory

One final direction of future interest is the connection between rigidity theory and the optimization-based approaches studied in this thesis. Rigidity theory investigates the uniqueness of graph realizations embedded in $\mathbb{R}^d$, with edge lengths constrained by the graph weights. For example, in Figure 5.1, the left-most figure depicts a non-rigid graph, meaning its realization remains valid even if Node 1 rotates around Node 2; the physical embedding is not a rigid body in $\mathbb{R}^2$. Contrast this with the right-most figure, where any continuous motion of the vertices must either result in a trivial isometry (translation or rotation) or require a deformation that changes a given edge length; this figure is a rigid body in $\mathbb{R}^2$. It is known that a globally rigid graph corresponds to a unique point configuration up to isometry [LLMM14]. In particular, it is also known that Erdös-Rényi graphs $G(n,p)$ are globally rigid with high probability if $p \gtrsim \frac{\log n}{n}$, for a generic set of edge lengths [LNPR23]. Compare this with the theory in Chapter 3, where algorithmic recovery is only possible for $p \gtrsim \frac{\log n}{n}$ for ν-incoherent, low-rank matrices with $\nu = \mathcal{O}(1)$. A current open question centers around the connection between the global rigidity of a graph and the incoherence of the corresponding point cloud; specifically, are highly coherent structures less likely to be globally rigid when connected via random edges? A connection between these two disparate areas appears natural, but as of yet remains unstudied in the literature.

5.3 Conclusion

We now return to Question 1.1.2 and Question 1.1.3 at the end of this thesis. In many cases, we can answer both Question 1.1.2 and Question 1.1.3 in the affirmative; yes, oftentimes you can reconstruct points in space from partial observations alone. It is not always easy, however, and the simplicity of the questions belies the mathematical richness and beauty of Question 1.1.3. There are likely endless variations on the theme of Question 1.1.3, with two seen in Question 1.3.1 and Question 4.1.1, and this problem will hopefully inspire mathematicians for years to come. Thank you for reading.

Appendix A

Appendix

A.1 A Proof Sketch for the Convergence of Algorithm 3 (RF-RGD)

In this section of the appendix, we provide a heuristic sketch of a convergence proof of Algorithm 3 (RF-RGD), the Restricted Frame Riemannian Gradient Descent for EDMC algorithm presented in Chapter 2. This proof does not use an RIP mechanism. This chapter has not been published previously, and is more of a formal proof of convergence rather than a rigorous one. In particular, the characterization of the retraction map on $\mathcal{N}_r$ is treated in a loose way. Future work could be done to make this proof of convergence rigorous, and could also improve local neighborhood assumptions for this idea, which are currently quite stringent.

A.1.1 Theoretical Analysis

We first reproduce an important result from [Van12] on a Taylor series expansion of the retraction map.

Lemma A.1.1 (Second-Order Retraction on $\mathcal{N}_r$)**.** *Let $\mathcal{N}_r$ be the Riemannian manifold of matrices of fixed rank r embedded in $\mathbb{R}^{m\times n}$. For any $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \in \mathcal{N}_r$ and any tangent vector $\boldsymbol{\xi} \in T_{\boldsymbol{X}}\mathcal{N}_r$ parameterized by the orthogonal decomposition $\boldsymbol{\xi} = \boldsymbol{U}\boldsymbol{M}\boldsymbol{V}^\top + \boldsymbol{U}_p\boldsymbol{V}^\top + \boldsymbol{U}\boldsymbol{V}_p^\top$, the*

mapping $R_{\boldsymbol{X}}^{(2)} : T_{\boldsymbol{X}}\mathcal{N}_r \to \mathcal{N}_r$ *defined by:*

$$R_{\boldsymbol{X}}^{(2)}(\boldsymbol{\xi}) = Z_U Z_V^\top,$$

where the factors are explicitly constructed as

$$\begin{aligned} Z_U &:= \boldsymbol{U}\left(\boldsymbol{\Sigma} + \frac{1}{2}\boldsymbol{M} - \frac{1}{8}\boldsymbol{M}\boldsymbol{\Sigma}^{-1}\boldsymbol{M}\right) + \boldsymbol{U}_p\left(\boldsymbol{I}_r - \frac{1}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{M}\right), \\ Z_V &:= \boldsymbol{V}\left(\boldsymbol{I}_r + \frac{1}{2}\boldsymbol{M}^\top\boldsymbol{\Sigma}^{-1} - \frac{1}{8}\boldsymbol{M}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{M}^\top\boldsymbol{\Sigma}^{-1}\right) + \boldsymbol{V}_p\left(\boldsymbol{\Sigma}^{-1} - \frac{1}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{M}^\top\boldsymbol{\Sigma}^{-1}\right), \end{aligned}$$

is a second-order approximation of the exponential map on $(g_X, \mathcal{N}_r)$*. Furthermore, the expansion of this mapping satisfies:*

$$R_{\boldsymbol{X}}^{(2)}(\boldsymbol{\xi}) = X + \boldsymbol{\xi} + \boldsymbol{U}_p\boldsymbol{\Sigma}^{-1}\boldsymbol{V}_p^\top + \mathcal{O}(\|\boldsymbol{\xi}\|_F^3) \quad as \quad \|\boldsymbol{\xi}\|_F \to 0.$$

Additionally, the deviation from the exact metric projection $R_{\boldsymbol{X}}(\boldsymbol{\xi}) = \mathcal{P}_{\mathcal{N}_r}(\boldsymbol{X} + \boldsymbol{\xi})$ *satisfies:*

$$R_{\boldsymbol{X}}^{(2)}(\boldsymbol{\xi}) - R_{\boldsymbol{X}}(\boldsymbol{\xi}) = \mathcal{O}(\|\boldsymbol{\xi}\|_F^3) \quad as \quad \|\boldsymbol{\xi}\|_F \to 0.$$

We will use this technique purely as a theoretical tool. First, we note that in RF-RGD $\mathcal{H}_r(\boldsymbol{X}_l + \alpha_l^\star \mathcal{P}_{\mathbb{T}_l}\mathcal{F}_\Omega(\boldsymbol{X}_l - \boldsymbol{X})) = R_{\boldsymbol{X}_l}(-\alpha_l^\star \mathcal{P}_{\mathbb{T}_l}\mathcal{F}_\Omega(\boldsymbol{X}_l - \boldsymbol{X}))$. We will use the deviations of the map $R_{\boldsymbol{X}_l}$ and $R_{\boldsymbol{X}}$ to prove convergence.

Again, we remark that this result is more formal than rigorous. In order to prove convergence of RF-RGD, one would need to more carefully characterize $R_{\boldsymbol{X}}$ beyond just an asymptotic expansion. This next theorem is more to illustrate how a convergence mechanism might work due to the second-order approximation around $\mathcal{N}_r$, rather than to be a convergence proof on the level with the preceding arguments seen in this document.

Theorem A.1.2. *Let* $\boldsymbol{X} \in \mathcal{N}_r$ *be a symmetric,* ν*-incoherent, rank-*r *ground truth Gram matrix accessed through partial Bernoulli random sampling of* $\{\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\}_{\boldsymbol{\alpha}\in\mathbb{I}}$*. Assume that the following*

initialization holds for some sufficiently small constant c:

$$\frac{\|\boldsymbol{X}_0 - \boldsymbol{X}\|_{\mathrm{F}}}{\lambda_r(\boldsymbol{X})} \leq \min\Bigg\{\sqrt{\frac{\nu r}{n}}, \frac{1}{3(36\nu r+2)}, \frac{m}{12nL(36\nu r+2)}, \tag{A.1}$$

$$c\frac{m^{3/2}(36\nu r+2)}{\lambda_r(\boldsymbol{X})\sqrt{172}n^{3/2}L^{3/2}}, \frac{m^2}{48nL^2(36\nu r+2)}\Bigg\}. \tag{A.2}$$

If $m \geq 3072\nu^2 r^2 n \log n$*, then RF-RGD recovers* $\boldsymbol{X}$ *with probability at least* $1 - 2n^{1-\beta}$ *with a per-iterate contraction factor* ρ_l *bounded by*

$$\rho_l < 1 - \frac{1}{3(36\nu r+2)}.$$

Proof. We first begin this theorem by noticing that, for $\boldsymbol{X}_{l+1} = \mathcal{H}_r(\boldsymbol{W}_l)$ and $\boldsymbol{W}_l = \boldsymbol{X}_l - \alpha_l \mathcal{P}_{\mathbb{T}_l}\mathcal{F}_\Omega(\boldsymbol{X}_l - \boldsymbol{X}) = \boldsymbol{X}_l + \boldsymbol{G}_l$ and for some sufficiently large $c > 0$

$$\begin{aligned}
\|\boldsymbol{X}_{l+1} - \boldsymbol{X}\|_{\mathrm{F}} &= \|\mathcal{H}_r(\boldsymbol{W}_l) - \boldsymbol{X}\|_{\mathrm{F}} \\
&= \|R_{\boldsymbol{X}_l}(\boldsymbol{G}_l) - \boldsymbol{X}\|_{\mathrm{F}} \\
&\leq \|R^{(2)}_{\boldsymbol{X}_l}(\boldsymbol{G}_l)\| + \|R_{\boldsymbol{X}_l}(\boldsymbol{G}_l) - R_{\boldsymbol{X}_l}(\boldsymbol{G}_l)\| \\
&= \|\boldsymbol{X}_l - \boldsymbol{X} + \boldsymbol{G}_l + \boldsymbol{U}_p\boldsymbol{\Sigma}^{-1}\boldsymbol{U}_p^\top\|_{\mathrm{F}} + c\|\boldsymbol{G}_l\|_{\mathrm{F}}^3 \\
&\leq \|\boldsymbol{X}_l - \boldsymbol{X} + \alpha_l\mathcal{P}_{\mathbb{T}_l}\mathcal{F}_\Omega(\boldsymbol{X} - \boldsymbol{X}_l)\|_{\mathrm{F}} + \frac{1}{\lambda_{\min}(\boldsymbol{X}_l)}\|\boldsymbol{U}_p\|_{\mathrm{F}}^2 + c\|\boldsymbol{G}_l\|_{\mathrm{F}}^3 \\
&\leq \|(\mathcal{I} - \alpha_l\mathcal{P}_{\mathbb{T}_l}\mathcal{F}_\Omega)(\boldsymbol{X}_l - \boldsymbol{X})\|_{\mathrm{F}} + \underbrace{\frac{1}{\lambda_{\min}(\boldsymbol{X}_l)}\|\boldsymbol{G}_l\|_{\mathrm{F}}^2 + c\|\boldsymbol{G}_l\|_{\mathrm{F}}^3}_{\text{Approximation Terms}},
\end{aligned}$$

where the first inequality follows from Theorem A.1.1 and the triangle inequality, the second inequality follows from Cauchy-Schwarz, and the last inequality follows from the orthogonal decomposition of the tangent space elements described in Theorem A.1.1. Examining the first term in the final inequality in a way analogous to previous proofs of convergence, such as in Theorem 2.2.2 or in Theorem 3.2.2, we can see that

$$\begin{aligned}
\|(\mathcal{I} - \alpha_l\mathcal{P}_{\mathbb{T}_l}\mathcal{F}_\Omega)(\boldsymbol{X}_l - \boldsymbol{X})\|_{\mathrm{F}} \leq \|(\mathcal{P}_{\mathbb{T}_l} - \alpha_l\mathcal{P}_{\mathbb{T}_l}\mathcal{F}_\Omega\mathcal{P}_{\mathbb{T}_l})(\boldsymbol{X}_l - \boldsymbol{X})\|_{\mathrm{F}} \\
+ \underbrace{\|(\mathcal{I} - \mathcal{P}_{\mathbb{T}_l})(\boldsymbol{X}_l - \boldsymbol{X})\|_{\mathrm{F}}}_{I_1}
\end{aligned}$$

$$
\begin{aligned}
&+ \underbrace{\|\alpha_l \mathcal{P}_{\mathbb{T}_l} \mathcal{F}_\Omega (\mathcal{I} - \mathcal{P}_{\mathbb{T}_l})(\boldsymbol{X}_l - \boldsymbol{X})\|_{\mathrm{F}}}_{I_2} \\
&\leq I_1 + I_2 + \underbrace{\alpha_l \left\| (\mathcal{P}_{\mathbb{T}_l} \mathcal{F}_\Omega \mathcal{P}_{\mathbb{T}_l} - \frac{m}{L} \mathcal{P}_{\mathbb{T}_l} \mathcal{F}_{\mathbb{I}} \mathcal{P}_{\mathbb{T}_l})(\boldsymbol{X}_l - \boldsymbol{X}) \right\|_{\mathrm{F}}}_{I_3} \\
&+ \underbrace{\left\| (\mathcal{P}_{\mathbb{T}_l} - \alpha_l \frac{m}{L} \mathcal{P}_{\mathbb{T}_l} \mathcal{F}_{\mathbb{I}} \mathcal{P}_{\mathbb{T}_l})(\boldsymbol{X}_l - \boldsymbol{X}) \right\|_{\mathrm{F}}}_{I_4}.
\end{aligned}
$$

The bounds for I_1, I_2, and I_3 follow a similar approach as in Chapters 2-4. The novel piece is now the analysis of I_4. We will now derive an optimal per-iteration step size and show that this algorithm is locally a contraction.

In order to do this, we first must bound $\lambda_{\max}(\mathcal{P}_{\mathbb{T}_l} \mathcal{F}_{\mathbb{I}} \mathcal{P}_{\mathbb{T}_l})$ and $\lambda_{\min}(\mathcal{P}_{\mathbb{T}} \mathcal{F}_{\mathbb{I}} \mathcal{P}_{\mathbb{T}})$. For the upper bound, notice that this is equivalent to the bounds in Theorem 3.B.2. As such, we have that

$$
\lambda_{\max}(\mathcal{P}_{\mathbb{T}_l} \mathcal{F}_{\mathbb{I}} \mathcal{P}_{\mathbb{T}_l}) \leq 4\nu_l r, \tag{A.3}
$$

where ν_l is the incoherence of the l-th iterate $\boldsymbol{X}_l$. From Theorem A.1.3, we know that $\nu_l \leq 36\nu$ if $\frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\lambda_{\min}(\boldsymbol{X})} \leq \sqrt{\frac{\nu r}{n}}$.

We now need to consider the lower bound $\lambda_{\min}(\mathcal{P}_{\mathbb{T}_l} \mathcal{F}_{\mathbb{I}} \mathcal{P}_{\mathbb{T}_l})$ when restricted to the subspace $\mathbb{T}_l$. It follows that

$$
\begin{aligned}
\lambda_{\min}(\mathcal{P}_{\mathbb{T}_l} \mathcal{F}_{\mathbb{I}} \mathcal{P}_{\mathbb{T}_l}) &= \min_{\boldsymbol{Y} \in \mathbb{T}_l, \|\boldsymbol{Y}\|_{\mathrm{F}} = 1} \sum_{\boldsymbol{\alpha} \in \mathbb{I}} \langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}_l} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle^2 \\
&= \min_{\boldsymbol{Y} \in \mathbb{T}_l, \|\boldsymbol{Y}\|_{\mathrm{F}} = 1} \sum_{\boldsymbol{\alpha} \in \mathbb{I}} \langle \boldsymbol{Y}, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle^2 \\
&\geq \min_{\|\boldsymbol{Y}\|_{\mathrm{F}} = 1} \sum_{\boldsymbol{\alpha} \in \mathbb{I}} \langle \boldsymbol{Y}, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle^2 \\
&\geq 2,
\end{aligned}
$$

as $\lambda_{\min}(\boldsymbol{H}) = 2$ from [LT23].

So now we can bound the expression I_4. Notice that for $\|\mathcal{P}_{\mathbb{T}_l} - \alpha_l \frac{m}{L} \mathcal{P}_{\mathbb{T}_l} \mathcal{F}_{\mathbb{I}} \mathcal{P}_{\mathbb{T}_l}\| < 1$, we need

the following pair of equations to hold:

$$-1 < 1 - \alpha_l \frac{36m\nu r}{L} < 1$$
$$-1 < 1 - \alpha_l \frac{2m}{L} < 1.$$

Rearranging these two equations, we can see that

$$0 < \alpha_l < \frac{L}{18m\nu r}$$
$$0 < \alpha_l < \frac{L}{m}.$$

As $\nu r \geq 1$, we have that contraction occurs if $\alpha_l \in \left(0, \frac{L}{18m\nu r}\right)$. The optimal contraction rate occurs when, for optimal step size $\alpha_l^\star$,

$$\left|1 - \alpha_l^\star \frac{36m\nu r}{L}\right| = \left|1 - \alpha_l^\star \frac{2m}{L}\right|.$$

Rearranging this expression, we see that the optimal step size $\alpha_l^\star$ is

$$\alpha_l^\star = \frac{2L/m}{36\nu r + 2}.$$

Plugging this into the contraction expression above, we see that the optimal theoretical contraction rate is

$$\left\|\mathcal{P}_{\mathbb{T}} - \alpha_l^\star \frac{m}{L}\mathcal{P}_{\mathbb{T}}\mathcal{F}_{\mathbb{I}}\mathcal{P}_{\mathbb{T}}\right\| = \rho^\star = \frac{36\nu r - 2}{36\nu r + 2} < 1.$$

As such, $I_4 < \frac{36\nu r - 2}{36\nu r + 2}\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}$. To prove convergence of RF-RGD, we need to show that $I_1 + I_2 + I_3 + \frac{1}{\lambda_{\min}(\boldsymbol{X}_l)}\|\boldsymbol{G}_l\|_{\mathrm{F}}^2 + c\|\boldsymbol{G}_l\|_{\mathrm{F}}^3 < \frac{2}{36\nu r + 2}$. To do this, we will leverage the local neighborhood assumption made in (A.1).

We now bound each of these expressions individually. We will first bound I_1. Notice that

$$\begin{aligned}\|(\mathcal{I} - \mathcal{P}_{\mathbb{T}_l})(\boldsymbol{X}_l - \boldsymbol{X})\|_{\mathrm{F}} &\leq \frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\lambda_{\min}(\boldsymbol{X})}\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}} \\ &\leq \frac{1}{3(36\nu r + 2)}\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}},\end{aligned}$$

by assumption. Next, we notice that

$$\begin{aligned} I_2 &= \|\alpha_l^\star \mathcal{P}_{\mathbb{T}_l} \mathcal{F}_\Omega (\mathcal{I} - \mathcal{P}_{\mathbb{T}_l})(\boldsymbol{X}_l - \boldsymbol{X})\|_{\mathrm{F}} \\ &\leq \frac{2L/m}{36\nu r + 2} \|\mathcal{F}_\Omega\| \frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\lambda_{\min}(\boldsymbol{X})} \|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}} \\ &\leq \frac{2L/m}{36\nu r + 2} \|\mathcal{F}_{\mathbb{I}}\| \frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\lambda_{\min}(\boldsymbol{X})} \|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}} \\ &= \frac{4nL/m}{36\nu r + 2} \frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\lambda_{\min}(\boldsymbol{X})} \|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}} \\ &\leq \frac{1}{3(36\nu r + 2)} \|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}, \end{aligned}$$

where the first inequality follows from the sub-multiplicativity of the operator norm, the second inequality follows from the fact that $\mathcal{F}_\Omega$ is a positive semi-definite sum of rank-one outer products, and the final inequality follows from (A.1). Now, we can bound I_3 using Theorem 3.B.2 and the assumption on $m \geq 3072\nu^2 r^2 n \log n$, noticing that

$$\begin{aligned} I_3 &= \frac{2L/m}{36\nu r + 2} \left\| \mathcal{P}_{\mathbb{T}_l} \mathcal{F}_\Omega \mathcal{P}_{\mathbb{T}_l} - \frac{m}{L} \mathcal{P}_{\mathbb{T}_l} \mathcal{F}_{\mathbb{I}} \mathcal{P}_{\mathbb{T}_l} \right\| \\ &\leq \frac{2}{36\nu r + 2} \frac{1}{6} \\ &= \frac{1}{3(36\nu r + 2)}, \end{aligned}$$

where the inequality holds with probability at least $1 - 2n^{1-\beta}$ from Theorem 3.B.2 and Theorem A.1.4. Next, we must bound the quadratic and cubic terms in $\|\boldsymbol{G}_l\|_{\mathrm{F}}$. First, we consider the curvature bound $\frac{1}{\lambda_{\min}(\boldsymbol{X}_l)} \|\boldsymbol{G}_l\|_{\mathrm{F}}^2$. One thing that needs to be taken into consideration for this expression is the factor of $\frac{1}{\lambda_{\min}(\boldsymbol{X}_l)}$, as this expression can become unbounded if the iterates approach lower-rank values, i.e. $\operatorname{rank}(\boldsymbol{X}_l) \to r - 1$. In a local neighborhood, however, we can use Weyl's inequality [Bha13] to show that

$$|\lambda_r(\boldsymbol{X}_l) - \lambda_r(\boldsymbol{X})| \leq \|\boldsymbol{X}_l - \boldsymbol{X}\| \leq \|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}.$$

Assuming that $\|\boldsymbol{X}_l - \boldsymbol{X}\| \leq \frac{1}{2}\lambda_{\min}(\boldsymbol{X})$, which holds from the assumptions stated in (A.1), we

have that $\frac{1}{\lambda_{\min}(\boldsymbol{X}_l)} \leq \frac{2}{\lambda_{\min}(\boldsymbol{X})}$. We can now bound this as

$$\begin{aligned}
\frac{1}{\lambda_{\min}(\boldsymbol{X}_l)}\|\boldsymbol{G}_l\|_{\mathrm{F}}^2 &\leq \frac{2}{\lambda_{\min}(\boldsymbol{X})}\|\alpha_l^\star \mathcal{P}_{\mathbb{T}_l}\mathcal{F}_\Omega(\boldsymbol{X}_l - \boldsymbol{X})\|_{\mathrm{F}}^2 \\
&\leq \frac{2\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\lambda_{\min}(\boldsymbol{X})}\|\alpha_l^\star \mathcal{P}_{\mathbb{T}_l}\mathcal{F}_\Omega\|^2\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}} \\
&\leq 4n\left(\frac{2L/m}{36\nu r + 2}\right)^2 \frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\lambda_{\min}(\boldsymbol{X})}\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}} \\
&\leq \frac{1}{3(36\nu r + 2)},
\end{aligned}$$

using the definitions of $\alpha_l^\star$, (A.1), and the fact that $\|\mathcal{F}_\Omega\| \leq 2n$. Lastly, we have that

$$\begin{aligned}
c\|\boldsymbol{G}_l\|_{\mathrm{F}}^3 &= c\|\alpha_l \mathcal{P}_{\mathbb{T}_l}\mathcal{F}_\Omega(\boldsymbol{X}_l - \boldsymbol{X})\|_{\mathrm{F}}^3 \\
&= c\left(\frac{2L/m}{36\nu r + 2}\right)^3 8n^3\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}^3 \\
&\leq \frac{64cn^3L^3}{m^3(36\nu r + 2)^3}\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}^2\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}} \\
&\leq \frac{1}{3(36\nu r + 2)}\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}.
\end{aligned}$$

Collecting these results, we have now that

$$\begin{aligned}
\|\boldsymbol{X}_{l+1} - \boldsymbol{X}\|_{\mathrm{F}} &< \left(\frac{36\nu r - 2}{36\nu r + 2} + \frac{5}{3(36\nu r + 2)}\right)\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}} \\
&= \left(1 - \frac{1}{3(36\nu r + 2)}\right)\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}.
\end{aligned}$$

Since the assumptions of (A.1) are satisfied for $l = 0$ and $\|\boldsymbol{X}_0 - \boldsymbol{X}\|_{\mathrm{F}}$, the assumptions are satisfied for any $l \geq 0$. As such, we have defined a contractive sequence, and the algorithm linearly converges with the specified contraction rate. This concludes the proof. □

We note that this analysis is far from optimal. First, one could follow a similar approach as in Theorem 2.B.1 and Theorem 2.B.2 to bound the expressions $\|\mathcal{P}_{\mathbb{T}_l}\mathcal{F}_\Omega\|$ and $\|\mathcal{P}_{\mathbb{T}}\mathcal{F}_\Omega\|$. The tighter bound in this concentration framework would be $\approx \sqrt{n\nu r \log n}$, which would tighten all of the local neighborhood assumptions.

Another way to approach the analysis of this algorithm is to let the stepsize $\alpha_l \ll \frac{L}{m}$. As we get contraction for any $\alpha \in \left(0, \frac{2L/m}{36\nu r+2}\right)$, we could choose a very small α_l to attenuate the worst of the local neighborhood assumptions in (A.1), as the tightest neighborhood restrictions come from the quadratic and cubic terms from the approximation of $R_{\boldsymbol{X}_l}$ with $R^{(2)}_{\boldsymbol{X}_l}$. This would decrease the size of the contraction factor from I_4, but the linear growth in that expression would be counteracted by the quadratic/cubic growth from the approximation terms.

A.1.2 Supplementary Results

Lemma A.1.3. *Let $\boldsymbol{X} \in \mathbb{R}^{n\times n}$ be a ν-incoherent rank-r Gram matrix. Let $\boldsymbol{X}_l \in \mathbb{R}^{n\times n}$ be a matrix such that*

$$\frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\lambda_{\min}(\boldsymbol{X})} \leq \sqrt{\frac{\nu r}{n}}.$$

Then the matrix $\boldsymbol{X}_l$ is at most 36ν-incoherent.

Proof. This theorem follows directly from Theorem 1.8.1. Let $\boldsymbol{X}_l = \boldsymbol{U}_l \boldsymbol{\Lambda}_l \boldsymbol{U}_l^\top$. First, notice that

$$\begin{aligned}
\max_{\boldsymbol{\alpha}\in\mathbb{I}} \|\mathcal{P}_{U_l}\boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} &= \max_{\boldsymbol{\alpha}\in\mathbb{I}} \|(\mathcal{P}_{U_l} - \mathcal{P}_U)\boldsymbol{w}_{\boldsymbol{\alpha}} + \mathcal{P}_U\boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \\
&\leq \max_{\boldsymbol{\alpha}\in\mathbb{I}} \|\mathcal{P}_U\boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} + \|(\mathcal{P}_{U_l} - \mathcal{P}_U)\boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \\
&\leq \sqrt{\frac{\nu r}{n}} + \frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\lambda_{\min}(\boldsymbol{X})} \\
&\leq \sqrt{\frac{4\nu r}{n}},
\end{aligned}$$

where the second inequality follows from Theorem 1.8.1 and the final inequality follows from the assumption in the Lemma statement. The final statement of the theorem comes from the fact that

$$\|\mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \leq 3\|\mathcal{P}_U\boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}.$$

□

Lemma A.1.4 (Local Concentration of $\mathcal{F}_\Omega$)**.** *Let*

$$\frac{L}{m}\left\|\mathcal{P}_{\mathbb{T}}\mathcal{F}_\Omega\mathcal{P}_{\mathbb{T}} - \frac{m}{L}\mathcal{P}_{\mathbb{T}}\mathcal{F}_\Omega\mathcal{P}_{\mathbb{T}}\right\| \leq \varepsilon_0.$$

Then for $\frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\lambda_{\min}(\boldsymbol{X})} \leq C$*, we have that*

$$\frac{L}{m}\left\|\mathcal{P}_{\mathbb{T}_l}\mathcal{F}_\Omega\mathcal{P}_{\mathbb{T}_l} - \frac{m}{L}\mathcal{P}_{\mathbb{T}_l}\mathcal{F}_\mathbb{I}\mathcal{P}_{\mathbb{T}_l}\right\| \leq 4\varepsilon_0.$$

Proof. This proof will follow a near-identical strategy to Theorem 3.B.11, Theorem 2.B.3 and Theorem 4.B.8. Since we recognize that

$$\|\mathcal{P}_{\mathbb{T}_l}\mathcal{F}_\Omega\| \leq \|\mathcal{F}_\Omega\| \leq \|\mathcal{F}_\mathbb{I}\| = 2n$$

and

$$\|\mathcal{P}_{\mathbb{T}}\mathcal{F}_\Omega\| \leq \|\mathcal{F}_\Omega\| \leq \|\mathcal{F}_\mathbb{I}\| = 2n,$$

the result will follow under the same analysis given the local neighborhood assumption provided above. □

A.2 Dual Bases

In a finite dimensional vector space of matrices $\mathbb{V}$, where $\dim(\mathbb{V}) = n$, any basis admits a dual, or bi-orthogonal, basis denoted $B^* = \{\boldsymbol{Y}_i\}_{i=1}^n$ that also spans $\mathbb{V}$. Importantly, this dual basis B^* and admits a bi-orthogonality relationship

$$\langle \boldsymbol{X}_i, \boldsymbol{Y}_j \rangle = \delta_{ij}.$$

Additionally, B uniquely determines B^* and vice versa. The bi-orthogonality relationship allows for the decomposition of any matrix $\boldsymbol{Z} \in \mathbb{V}$ as follows:

$$\boldsymbol{Z} = \sum_{i=1}^n \langle \boldsymbol{Z}, \boldsymbol{Y}_i \rangle \boldsymbol{X}_i = \sum_{i=1}^n \langle \boldsymbol{Z}, \boldsymbol{X}_i \rangle \boldsymbol{Y}_i.$$

We define the Gram, or correlation matrix, $\boldsymbol{H} \in \mathbb{R}^{n\times n}$, for B as $H_{ij} = \langle \boldsymbol{X}_i, \boldsymbol{X}_j \rangle$, and let $H^{ij} = (\boldsymbol{H}^{-1})_{ij}$. It is straightforward to show that $\boldsymbol{Y}_i = \sum_{j=1}^n H^{ij}\boldsymbol{X}_j$ generates B^*, and similarly that $\boldsymbol{X}_i = \sum_{j=1}^n H_{ij}\boldsymbol{Y}_j$ [Bha13].

A.3 Geometry of Low-Rank Matrices

One of the primary motivations for constructing geometric algorithms on manifolds is the fact that many natural sets of matrices form Riemannian manifolds. The set of fixed-rank matrices are one of these sets, and will naturally appear in our discussion of matrix completion. The following lemma showing that matrices with a fixed rank form a smooth manifold uses the regular level set theorem from [Tu10]. A statement has been reproduced below:

Theorem A.3.1. *Let $\mathcal{N}$ and $\mathcal{Q}$ be two manifolds with $\dim \mathcal{N} = n$ and $\dim \mathcal{Q} = m$, and let $F : \mathcal{N} \to \mathcal{Q}$ be a C^∞ map between these manifolds. Then for any regular $c \in \mathcal{Q}$, the preimage $F^{-1}(c)$ is a regular submanifold of $\mathcal{N}$ with dimension $n - m$.*

We note that we have not defined regularity in this context. Essentially, regularity just requires that the differential of F is surjective at a point c. More details can be found in [Tu10].

Lemma A.3.2. *The set $\mathcal{N}_r = \{\boldsymbol{A} \in \mathbb{R}^{m\times n} \mid \operatorname{rank}(\boldsymbol{A}) = r\}$ is a smooth manifold with dimension $r(m + n - r)$*

Proof. We will prove that $\mathcal{N}_r$ is a smooth submanifold of $\mathbb{R}^{m\times n}$ of dimension $r(m + n - r)$ utilizing Theorem A.3.1. Let $\boldsymbol{A}_0 \in \mathcal{N}_r$. Since $\operatorname{rank}(\boldsymbol{A}_0) = r$, there exists a non-zero $r \times r$ minor. By applying permutation matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$ to permute the rows and columns, we can assume without loss of generality that the leading principal $r \times r$ submatrix of $\boldsymbol{A}_0$ is invertible. That is, we can partition the matrix as

$$\boldsymbol{P}\boldsymbol{A}_0\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{X}_0 & \boldsymbol{Y}_0 \\ \boldsymbol{Z}_0 & \boldsymbol{W}_0 \end{pmatrix}, \tag{A.4}$$

where $\boldsymbol{X}_0 \in \mathrm{GL}(r, \mathbb{R})$.

Let $U \subset \mathbb{R}^{m\times n}$ be the open neighborhood of $\boldsymbol{A}_0$ consisting of all matrices $\boldsymbol{A}$ that are identically partitioned as

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{X} & \boldsymbol{Y} \\ \boldsymbol{Z} & \boldsymbol{W} \end{pmatrix},$$

such that $\boldsymbol{X} \in \mathrm{GL}(r, \mathbb{R})$. The set $\mathcal{U}$ is open because the general linear group is an open subset of the space of square matrices.

For any $\boldsymbol{A} \in \mathcal{U}$, we can perform block Gaussian elimination by left-multiplying $\boldsymbol{A}$ with a non-singular block lower-triangular matrix:

$$\begin{pmatrix} \boldsymbol{I}_r & \boldsymbol{0} \\ -\boldsymbol{Z}\boldsymbol{X}^{-1} & \boldsymbol{I}_{m-r} \end{pmatrix} \begin{pmatrix} \boldsymbol{X} & \boldsymbol{Y} \\ \boldsymbol{Z} & \boldsymbol{W} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X} & \boldsymbol{Y} \\ \boldsymbol{0} & \boldsymbol{W} - \boldsymbol{Z}\boldsymbol{X}^{-1}\boldsymbol{Y} \end{pmatrix}.$$

As multiplication by an invertible matrix preserves rank, it follows that $\operatorname{rank}(\boldsymbol{A}) = r$ if and only if:

$$\operatorname{rank}\left(\begin{pmatrix} \boldsymbol{X} & \boldsymbol{Y} \\ \boldsymbol{0} & \boldsymbol{W} - \boldsymbol{Z}\boldsymbol{X}^{-1}\boldsymbol{Y} \end{pmatrix}\right) = r.$$

Because $\boldsymbol{X}$ is an $r \times r$ invertible matrix, the first r rows of this resulting matrix are linearly independent. For the total rank of the matrix to be exactly r, the remaining $m - r$ rows must be identically zero. This yields the necessary and sufficient condition:

$$\boldsymbol{W} - \boldsymbol{Z}\boldsymbol{X}^{-1}\boldsymbol{Y} = \boldsymbol{0}.$$

We define a map $\Phi : \mathcal{U} \to \mathbb{R}^{(m-r)\times(n-r)}$ by:

$$\Phi(\boldsymbol{A}) = \boldsymbol{W} - \boldsymbol{Z}\boldsymbol{X}^{-1}\boldsymbol{Y},$$

where $\boldsymbol{A}$ has the same block structure as in (A.4). This map is smooth because it is composed of matrix multiplication, addition, and inversion, all of which are smooth operations on the domain where $\boldsymbol{X}$ is invertible.

We now show that $\boldsymbol{0}$ is a regular value of Φ. We compute the differential of Φ at an arbitrary point $\boldsymbol{A} \in \mathcal{U}$. The partial derivative of Φ with respect to the block $\boldsymbol{W}$ is the identity map on $\mathbb{R}^{(m-r)\times(n-r)}$:

$$D_{\boldsymbol{W}}\Phi(\boldsymbol{A})[\boldsymbol{H}] = \boldsymbol{H},$$

for any $\boldsymbol{H} \in \mathbb{R}^{(m-r)\times(n-r)}$.

Since the differential with respect to a subset of the coordinates is surjective, the total differential $d\Phi_{\boldsymbol{A}}$ is surjective for all $\boldsymbol{A} \in \mathcal{U}$. As such, Φ is a smooth submersion.

By Theorem A.3.1, the preimage $\Phi^{-1}(\mathbf{0}) = \mathcal{N}_r \cap U$ is a smooth submanifold of $\mathcal{U}$, and consequently a smooth submanifold of $\mathbb{R}^{m\times n}$. The codimension of this submanifold in $\mathbb{R}^{m\times n}$ is equal to the dimension of the codomain of the submersion Φ, which is $(m-r)(n-r)$. Therefore, the dimension of the manifold $\mathcal{N}_r$ is:

$$mn - (m-r)(n-r) = r(m+n-r).$$

Because $\boldsymbol{A}_0$ was chosen as an arbitrary element of $\mathcal{N}_r$, this local coordinate construction holds in an open neighborhood around every point in the set. Thus, $\mathcal{N}_r$ is a smooth manifold of dimension $r(m+n-r)$. □

We note that there has been a tremendous amount of work conducted into the geometry of matrix manifolds. See [AMS08] for more details.

Bibliography

[AMS08] P.-A. Absil, R. Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

[ASY16] Mohammad Aldibaja, Noaki Suganuma, and Keisuke Yoneda. Improving localization accuracy for autonomous driving in snow-rain environments. In *2016 IEEE/SICE International Symposium on System Integration (SII)*, pages 212–217. IEEE, 2016.

[AW02] Rudolf Ahlswede and Andreas Winter. Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 48(3):569–579, 2002.

[BCWP22] Dimitris Bertsimas, Ryan Cory-Wright, and Jean Pauphilet. Mixed-projection conic optimization: A new paradigm for modeling rank constraints. *Operations Research*, 70(6):3321–3344, 2022.

[Bha13] R. Bhatia. *Matrix Analysis*. Graduate Texts in Mathematics. Springer New York, 2013.

[BJ14] Srinadh Bhojanapalli and Prateek Jain. Universal matrix completion. In *International Conference on Machine Learning*, pages 1881–1889. PMLR, 2014.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.

[BLT+06] Pratik Biswas, T-C Liang, K-C Toh, Yinyu Ye, and T-C Wang. Semidefinite programming approaches for sensor network localization with noisy distance measurements. *IEEE transactions on automation science and engineering*, 3(4):360–371, 2006.

[BLWY06] Pratik Biswas, Tzu-Chen Lian, Ta-Chung Wang, and Yinyu Ye. Semidefinite programming based algorithms for sensor network localization. *ACM Transactions on Sensor Networks (TOSN)*, 2(2):188–220, 2006.

[BM03] 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.

[BONL07] Azzedine Boukerche, Horacio ABF Oliveira, Eduardo F Nakamura, and Antonio AF Loureiro. Localization systems for wireless sensor networks. *IEEE wireless Communications*, 14(6):6–12, 2007.

[Bou23] Nicolas Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 3 2023.

[BV20] Shantanu Prasad Burnwal and Mathukumalli Vidyasagar. Deterministic completion of rectangular matrices using asymmetric ramanujan graphs: Exact and stable recovery. *IEEE Transactions on Signal Processing*, 68:3834–3848, 2020.

[BVB16] Nicolas Boumal, Vlad Voroninski, and Afonso Bandeira. The non-convex burer-monteiro approach works on smooth semidefinite programs. *Advances in Neural Information Processing Systems*, 29, 2016.

[CBSW14] Yudong Chen, Srinadh Bhojanapalli, Sujay Sanghavi, and Rachel Ward. Completing any low-rank matrix, provably, 2014.

[CP10] Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[CR09] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

[CR12] Inc. CVX Research. CVX: Matlab software for disciplined convex programming, version 2.0. `https://cvxr.com/cvx`, August 2012.

[CRG93] G Marius Clore, Mark A Robien, and Angela M Gronenborn. Exploring the limits of precision and accuracy of protein structures determined by nuclear magnetic resonance spectroscopy. *Journal of molecular biology*, 231(1):82–102, 1993.

[DK70] Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[DKQW10] Yichuan Ding, Nathan Krislock, Jiawei Qian, and Henry Wolkowicz. Sensor network localization, euclidean distance matrix completions, and graph realization. *Optimization and Engineering*, 11(1):45–66, 2010.

[DlPG12] Victor De la Pena and Evarist Giné. *Decoupling: from dependence to independence*. Springer Science & Business Media, 2012.

[DLY25] Lijun Ding, Haihao Lu, and Jinwen Yang. New understandings and computation on augmented lagrangian methods for low-rank semidefinite programming. *arXiv preprint arXiv:2505.15775*, 2025.

[EKS23] Tal Einav, Yuehaw Khoo, and Amit Singer. Quantitatively visualizing bipartite datasets. *Physical Review X*, 13(2):021002, 2023.

[EWW18] Armin Eftekhari, Michael B Wakin, and Rachel A Ward. Mc2: A two-phase algorithm for leveraged matrix completion. *Information and Inference: A Journal of the IMA*, 7(3):581–604, 2018.

[EY36] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[FT13] Xingyuan Fang and Kim-Chuan Toh. Using a distributed sdp approach to solve simulated protein molecular conformation problems. In *Distance Geometry*, pages 351–376. Springer, 2013.

[GB08] M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited, 2008. `http://stanford.edu/~boyd/graph_dcp.html`.

[GHR93] W Glunt, TL Hayden, and M Raydan. Molecular conformations from distance matrices. *Journal of Computational Chemistry*, 14(1):114–120, 1993.

[GN10] David Gross and Vincent Nesme. Note on sampling without replacing from a finite collection of matrices. *arXiv preprint arXiv:1001.2738*, 2010.

[Gow85] J. C. Gower. Properties of euclidean and non-euclidean distance matrices. *Linear Algebra and its Applications*, 67:81–97, 6 1985.

[GQZ23] Shaoyan Guo, Hou-Duo Qi, and Liwei Zhang. Perturbation analysis of the euclidean distance matrix optimization problem and its numerical implications. *Computational Optimization and Applications*, 86(3):1193–1227, 2023.

[Gro11] David Gross. Recovering low-rank matrices from few coefficients in any basis. *Information Theory, IEEE Transactions on*, 57(3):1548–1566, 2011.

[GTK24] Ipsita Ghosh, Abiy Tasissa, and Christian Kümmerle. Sample-efficient geometry reconstruction from euclidean distances using non-convex optimization. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 77226–77268. Curran Associates, Inc., 2024.

[GVL13] Gene H. Golub and Charles F. Van Loan. *Matrix Computations - 4th Edition*. Johns Hopkins University Press, Philadelphia, PA, 2013.

[JM12] Adel Javanmard and Andrea Montanari. Localization from incomplete noisy distance measurements. *Foundations of Computational Mathematics*, 13(3):297–345, July 2012.

[KJVRK14] Jeril Kuriakose, Sandeep Joshi, R Vikram Raju, and Aravind Kilaru. A review on localization in wireless sensor networks. *Advances in signal processing and intelligent recognition systems*, pages 599–610, 2014.

[KKM09] Usman A Khan, Soummya Kar, and José MF Moura. Diland: An algorithm for distributed sensor localization with noisy distance measurements. *IEEE Transactions on Signal Processing*, 58(3):1940–1947, 2009.

[KTC25a] Chandra Kundu, Abiy Tasissa, and HanQin Cai. A dual basis approach for structured robust euclidean distance geometry, 2025.

[KTC25b] Chandra Kundu, Abiy Tasissa, and HanQin Cai. Structured sampling for robust euclidean distance geometry. In *2025 59th Annual Conference on Information Sciences and Systems (CISS)*, page 1–6. IEEE, March 2025.

[Led01] Michel Ledoux. *The concentration of measure phenomenon*. American Mathematical Soc., 2001.

[Lee19] A J Lee. *U-statistics: Theory and Practice*. Routledge, 2019.

[LL17] Rongjie Lai and Jia Li. Solving partial differential equations on manifolds from incomplete interpoint distance. *SIAM Journal on Scientific Computing*, 39(5):A2231–A2256, 2017.

[LLM08] Leo Liberti, Carlile Lavor, and Nelson Maculan. A branch-and-prune algorithm for the molecular distance geometry problem. *International Transactions in Operational Research*, 15(1):1–17, 2008.

[LLMM14] Leo Liberti, Carlile Lavor, Nelson Maculan, and Antonio Mucherino. Euclidean distance geometry and applications. *SIAM review*, 56(1):3–69, 2014.

[LNPR23] Alan Lew, Eran Nevo, Yuval Peled, and Orit E Raz. Sharp threshold for rigidity of random graphs. *Bulletin of the London Mathematical Society*, 55(1):490–501, 2023.

[LS24] Yicheng Li and Xinghua Sun. Sensor network localization via riemannian conjugate gradient and rank reduction. *IEEE Transactions on Signal Processing*, 72:1910–1927, 2024.

[LS25] Yicheng Li and Xinghua Sun. Euclidean distance matrix completion via asymmetric projected gradient descent, 2025.

[LT23] Samuel Lichtenberg and Abiy Tasissa. A dual basis approach to multidimensional scaling: spectral analysis and graph regularity, 2023.

[Mah18] Prasanta Chandra Mahalanobis. On the generalized distance in statistics. *Sankhyā: The Indian Journal of Statistics, Series A (2008-)*, 80:S1–S7, 2018.

[MJCD08] Raghu Meka, Prateek Jain, Constantine Caramanis, and Inderjit S Dhillon. Rank minimization via online learning. In *Proceedings of the 25th International Conference on Machine learning*, pages 656–663, 2008.

[MSMS13] Jose Vicente Marti, Jorge Sales, Raul Marin, and Pedro Sanz. Multi-sensor localization and navigation for remote manipulation in smoky areas. *International Journal of Advanced Robotic Systems*, 10(4):211, 2013.

[NW06] Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 2006.

[PRT18] Josep M Porta, Nicolás Rojas, and Federico Thomas. Distance geometry in active structures. *Mechatronics for Cultural Heritage and Civil Engineering*, pages 115–136, 2018.

[Rec11a] Benjamin Recht. A simpler approach to matrix completion. *The Journal of Machine Learning Research*, 12:3413–3430, 2011.

[Rec11b] Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, 2011.

[Roj12] Nicolàs Rojas. *Distance-based formulations for the position analysis of kinematic chains*. PhD thesis, Universitat Politècnica de Catalunya, 2012.

[RV13] Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration, 2013.

[Sat21] Hiroyuki Sato. *SPRINGER BRIEFS IN ELECTRICAL AND COMPUTER ENGINEERING CONTROL, AUTOMATION AND ROBOTICS Riemannian Optimization and Its Applications*. Springer, 2021.

[SCT24] Chandler Smith, HanQin Cai, and Abiy Tasissa. Riemannian optimization for non-convex euclidean distance geometry with global recovery guarantees, 2024.

[SCT25] Chandler Smith, HanQin Cai, and Abiy Tasissa. Provable non-convex euclidean distance matrix completion: Geometry, reconstruction, and robustness. *arXiv preprint arXiv:2508.00091*, 2025.

[SKM19] Hiroyuki Sato, Hiroyuki Kasai, and Bamdev Mishra. Riemannian stochastic variance reduced gradient algorithm with retraction and vector transport. *SIAM Journal on Optimization*, 29:1444–1472, 2019.

[SLCT23] Chandler Smith, Samuel Lichtenberg, Hanqin Cai, and Abiy Tasissa. Riemannian optimization for euclidean distance geometry. *OPT2023: 15th Annual Workshop on Optimization for Machine Learning*, 2023.

[T+15] Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

[TD25] Abiy Tasissa and Waltenegus Dargie. Robust node localization for rough and extreme deployment environments, 2025.

[TDSL00] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[TL18] Abiy Tasissa and Rongjie Lai. Exact reconstruction of euclidean distance geometry problem using low-rank matrix completion. *IEEE Transactions on Information Theory*, 65(5):3124–3144, 2018.

[TL21] Abiy Tasissa and Rongjie Lai. Low-rank matrix completion in a general non-orthogonal basis. *Linear Algebra and its Applications*, 625:81–112, 2021.

[Tor52] Warren S Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419, 1952.

[Tor58] Warren S Torgerson. *Theory and methods of scaling.* Wiley, 1958.

[Tro97] Michael W Trosset. Applications of multidimensional scaling to molecular conformation, 1997.

[Tro12] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

[Tu10] Loring Tu. *Graduate Texts in Mathematics Dierential Geometry Connections, Curvature, and Characteristic Classes*. Springer, 2010.

[Van12] Bart Vandereycken. Low-rank matrix completion by Riemannian optimization—extended version, 2012.

[Ver12] Roman Vershynin. *Introduction to the non-asymptotic analysis of random matrices*, page 210–268. Cambridge University Press, 2012.

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[Wal21] Irène Waldspurger. Lecture notes on non-convex algorithms for low-rank matrix recovery. *arXiv preprint arXiv:2105.10318*, 2021.

[WCCL16] Ke Wei, Jian-Feng Cai, Tony F. Chan, and Shingyu Leung. Guarantees of riemannian optimization for low rank matrix recovery. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1198–1222, 2016.

[WCCL20] Ke Wei, Jian-Feng Cai, Tony F Chan, and Shingyu Leung. Guarantees of riemannian optimization for low rank matrix completion. *Inverse Problems & Imaging*, 14(2), 2020.

[WM22] John Wright and Yi Ma. *High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications*. Cambridge University Press, 2022.

[XLL24] Hang Xu, Song Li, and Junhong Lin. Nonconvex deterministic matrix completion by projected gradient descent methods. *arXiv preprint arXiv:2401.06592*, 2024.

[YWS15] Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.