Riemannian Optimization for Non-convex Euclidean Distance Geometry with Global Recovery Guarantees

Chandler Smith*[1], HanQin Cai[2], and Abiy Tasissa[1]

[1]Department of Mathematics, Tufts University, Medford, MA 02155, USA.
[2]Department of Statistics and Data Science and Department of Computer Science, University of Central Florida, Orlando, FL 32816, USA.

October 8, 2024

Abstract

The problem of determining the configuration of points from partial distance information, known as the Euclidean Distance Geometry (EDG) problem, is fundamental to many tasks in the applied sciences. In this paper, we propose two algorithms grounded in the Riemannian optimization framework to address the EDG problem. Our approach formulates the problem as a low-rank matrix completion task over the Gram matrix, using partial measurements represented as expansion coefficients of the Gram matrix in a non-orthogonal basis. For the first algorithm, under a uniform sampling with replacement model for the observed distance entries, we demonstrate that, with high probability, a Riemannian gradient-like algorithm on the manifold of rank-r matrices converges linearly to the true solution, given initialization via a one-step hard thresholding. This holds provided the number of samples, m, satisfies $m \geq \mathcal{O}(n^{7/4} r^2 \log(n))$. With a more refined initialization, achieved through resampled Riemannian gradient-like descent, we further improve this bound to $m \geq \mathcal{O}(nr^2 \log(n))$. Our analysis for the first algorithm leverages a non-self-adjoint operator and depends on deriving eigenvalue bounds for an inner product matrix of restricted basis matrices, leveraging sparsity properties for tighter guarantees than previously established. The second algorithm introduces a self-adjoint surrogate for the sampling operator. This algorithm demonstrates strong numerical performance on both synthetic and real data. Furthermore, we show that optimizing over manifolds of higher-than-rank-r matrices yields superior numerical results, consistent with recent literature on overparameterization in the EDG problem.

1 Introduction

The rapid advancement of technology across various scientific fields has greatly simplified data collection. In many practical applications, however, 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 [1,2]. 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 [3]. Similarly, in sensor networks, we may have mobile nodes with known distances only from fixed anchors [4,5]. In these and other scenarios, the fundamental problem is determining the configuration of points based on partial information about inter-point distances. This problem is known as the Euclidean distance geometry (EDG) problem, which has numerous applications throughout the applied sciences [6–15].

To formulate this problem mathematically, some notation is in order. Let $\{\boldsymbol{p}_i\}_{i=1}^n \subset \mathbb{R}^r$ denote a set of n points in $\mathbb{R}^r$. We define the $r \times n$ matrix $\boldsymbol{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, ..., \boldsymbol{p}_n]$, which has the points as columns. There are two essential mathematical objects related to $\boldsymbol{P}$. The first object is the Gram matrix $\boldsymbol{X} \in \mathbb{R}^{n \times n}$, defined as $\boldsymbol{X} = \boldsymbol{P}^\top \boldsymbol{P}$. By construction, $\boldsymbol{X}$ is symmetric and positive semi-definite. The second object is the squared distance matrix $\boldsymbol{D} \in \mathbb{R}^{n \times n}$, defined entry-wise as $D_{ij} = \|\boldsymbol{p}_i - \boldsymbol{p}_j\|_2^2$. The reason for working with the squared distance matrix instead of the distance matrix will become clear later. Computing $\boldsymbol{D}$ given $\boldsymbol{P}$ is conceptually straightforward. However, the inverse problem of determining $\boldsymbol{P}$ from $\boldsymbol{D}$ is not immediately straightforward. To address this problem, we need to precisely define what it means to identify $\boldsymbol{P}$. Since rigid motions and translations preserve distances, there is no unique $\boldsymbol{P}$

*Corresponding Author, Chandler.Smith@Tufts.edu corresponding to a given squared distance matrix $\boldsymbol{D}$. From here on, we assume the points are centered at the origin, i.e., for $\mathbf{1}$ as a column vector of ones, $\boldsymbol{P}\mathbf{1} = \mathbf{0}$. This implies that $\boldsymbol{X}\mathbf{1} = \boldsymbol{P}^\top \boldsymbol{P}\mathbf{1} = \mathbf{0}$. We refer to $\boldsymbol{P}$ and $\boldsymbol{X}$ with this property as centered point and centered Gram matrix, respectively. Since the Gram matrix is invariant under rigid motions, these assumptions allow for a one-to-one correspondence between $\boldsymbol{D}$ and $\boldsymbol{X}$.

When we have access to all the distances, a central result in [16] provides the following one-to-one correspondence between $\boldsymbol{D}$ and a centered $\boldsymbol{X}$:

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

$$\boldsymbol{D} = \operatorname{diag}(\boldsymbol{X})\mathbf{1}^\top + \mathbf{1}\operatorname{diag}(\boldsymbol{X})^\top - 2\boldsymbol{X}, \tag{2}$$

where $\operatorname{diag}(\cdot)$ inputs an $n \times n$ matrix and returns a column vector with the entries along the diagonal, and $\boldsymbol{J} = \boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$. Once $\boldsymbol{X}$ is reconstructed using the above formula, $\boldsymbol{P}$ can be computed from the r-truncated eigendecomposition of $\boldsymbol{X}$. It is important to note that, as previously mentioned, $\boldsymbol{P}$ is unique up to rigid motions. This procedure for computing $\boldsymbol{P}$ from a full squared distance matrix $\boldsymbol{D}$ is known as classical multidimensional scaling (Classical MDS) [16–19].

In many practical scenarios, the distance matrix may be incomplete, making classical MDS inapplicable for determining the point configuration. However, notice that $\operatorname{rank}(\boldsymbol{X}) \leq r$, and one can show that $\operatorname{rank}(\boldsymbol{D}) \leq r + 2$ [20]. This implies that when $r \ll n$, which is often the case in practice, $\boldsymbol{X}$ and $\boldsymbol{D}$ are low-rank. This allows us to utilize a rich library of tools from low-rank matrix completion. With that, one technique is to directly apply matrix completion techniques on $\boldsymbol{D}$ [21]. Let $\Omega \subset \{(i,j) \mid 1 \leq i < j \leq n\}$ denote the set of sampled indices corresponding to the strictly upper-triangular part of the distance matrix. Note that, since a distance matrix is hollow and symmetric, it suffices to consider the samples in the upper-triangular part; that is, if D_{ij} is sampled, D_{ji} is also assumed to be sampled. A matrix completion approach would consider the following optimization program to recover $\boldsymbol{D}$:

$$\begin{aligned} &\underset{\boldsymbol{Z}\in\mathbb{R}^{n\times n}}{\text{minimize}} && ||\boldsymbol{Z}||_* \\ &\text{subject to} && Z_{ij} = D_{ij} \quad \forall (i,j) \in \Omega, \end{aligned} \tag{3}$$

where $||\cdot||_*$ denotes the nuclear norm, which serves as a convex surrogate for rank [22]. The main idea of these tools is that, under some assumptions, the nuclear norm minimization program reconstructs the true low-rank squared distance matrix exactly with high probability from $\mathcal{O}(nr\log^2(n))$ randomly sampled entries [23–27]. Another set of techniques [28, 29] focus on recovering the point configuration by using the Gram matrix as an optimization variable, and using only partial information from the entries in $\boldsymbol{D}$. Specifically, these works consider the following optimization program for the EDG problem

$$\begin{aligned} &\underset{\boldsymbol{X}\in\mathbb{R}^{n\times n},\, \boldsymbol{X}=\boldsymbol{X}^\top,\, \boldsymbol{X}\succeq\mathbf{0}, \boldsymbol{X}\mathbf{1}=\mathbf{0}}{\text{minimize}} && ||\boldsymbol{X}||_* \\ &\text{subject to} && X_{ii} + X_{jj} - 2X_{ij} = D_{ij} \quad \forall (i,j) \in \Omega, \end{aligned} \tag{4}$$

where the constraints follow from the relation of $\boldsymbol{X}$ and $\boldsymbol{D}$ in (2) and (1). Due to the challenge of working with the constraints imposed by distance matrices, i.e., an entrywise triangle inequality that must be satisfied in order to remain a distance matrix, this work will follow the latter approach of optimizing over the Gram matrix. We note that, in contrast to completing the square distance matrix $\boldsymbol{D}$ which has rank at most $r + 2$, employing a minimization approach based on a Gram matrix that has rank at most r implicitly enforces the constraints of the Euclidean distances. Recent works have indicated that this approach can achieve better sampling complexity than direct distance matrix completion [28–30].

We note that theoretical guarantees for (4) have been established in [28, 31], but still suffer from the lack of scalability of convex techniques. A non-convex Lagrangian formulation was also proposed in [28], yielding strong numerical results but lacking local convergence guarantees. The work in [32] uses a Riemannian manifold approach to develop a conjugate gradient algorithm for estimating the underlying Gram matrix. The theoretical analysis therein shows that the squared distance matrix iterates globally converge to the true squared distance matrix at the sampled entries under three assumptions. However, the relationship between the problem parameters, such as the sampling scheme and sampled entries, and the third assumption remains unclear, as noted in Remark III.8 of the paper. In [30], the authors introduce a Riemannian conjugate gradient method with line search for the EDG problem. The paper provides a local convergence analysis for the case where the entries of the distance matrix are sampled according to the Bernoulli model given a suitable initialization. The initialization method used is known as rank reduction, which begins with initial points embedded in a higher-dimensional space than the target dimension. While [30] demonstrates strong empirical results for this initialization via tests on synthetic data for sensor localization, there are no provable guarantees provided for the initialization. In this manuscript, we aim to present a provable non-convex algorithm for the EDG problem, along with a provable initialization.

1.1 Contributions

The main contributions of this paper are as follows:

1. **Construction of two novel algorithms:** We propose two non-convex iterative algorithms in a Riemannian optimization framework for the Euclidean distance geometry problem. These algorithms are both smooth, first-order methods on the manifold of rank-r matrices for a fixed r, and have low computational complexity per iteration.

2. **Two different initialization methods:** We propose two different structured initializations from partial measurements, and prove an error bound between the initializations and true solution. Both initializations are relatively simple and require minimal a priori knowledge of the ground truth matrix, save from the measurements necessary to construct the algorithm.

3. **Convergence guarantees and sample complexity requirements:** We provide theoretical analysis that ensures high probability local convergence of one of the algorithms to the ground truth solution. Along with this characterization of an attractive basin, we prove sample complexity results for the initialization methods to guarantee the algorithm's starting point within the attractive basin.

1.2 Notation

We briefly summarize the notation used throughout this paper below. In general, uppercase boldface scripts, such as $\boldsymbol{A}$, will denote matrices, lowercase boldface scripts, such as $\boldsymbol{v}$, will denote vectors, calligraphic scripts, such as $\mathcal{A}$, will denote linear operators on matrices, and blackboard bold font, such as $\mathbb{V}$, will denote vector spaces and subspaces. $\boldsymbol{X}^\top$ denotes the transpose of $\boldsymbol{X}$, $\mathrm{Tr}(\boldsymbol{X})$ as the trace of $\boldsymbol{X}$, $\langle \boldsymbol{A}, \boldsymbol{B}\rangle = \mathrm{Tr}(\boldsymbol{A}^\top \boldsymbol{B})$ denotes the trace inner product, and δ_{ij} denotes the Kronecker delta. We denote the (i,j)-th entry of a matrix $\boldsymbol{X}$ by X_{ij}. By $\mathbf{1}$, we mean this to be a column vector of ones, of a size determined by the context, and by $\mathbf{0}$ we mean either a column vector of zeros or a matrix of zeros. $\boldsymbol{e}_i$ denotes a vector of zeros except a 1 at the i-th position. We denote $\|\boldsymbol{x}\|_2$ to be the standard l_2 norm on $\mathbb{R}^n$, $\|\boldsymbol{X}\|_\mathrm{F}$ to be the Frobenius norm on $\mathbb{R}^{n\times n}$, $\|\boldsymbol{X}\|$ to be the operator norm of a matrix, $\|\boldsymbol{X}\|_\infty$ to be the maximum element of $\boldsymbol{X}$, and $\|\boldsymbol{X}\|_\star$ to be the nuclear norm of $\boldsymbol{X}$. We denote $\|\mathcal{A}\| = \sup_{\|\boldsymbol{X}\|_\mathrm{F}=1} \|\mathcal{A}(\boldsymbol{X})\|_\mathrm{F}$ to be the operator norm of linear operators on matrices, and $\lambda_{\max}(\boldsymbol{X})/\lambda_{\min}(\boldsymbol{X})$ to the maximal/minimal eigenvalue of a matrix. We denote $\odot$ as the Hadamard product between two matrices. We denote the i-th row of a matrix $\boldsymbol{X}$ by $\boldsymbol{X}^{(i)}$, and the i-th column by $\boldsymbol{X}_{(i)}$. We denote the universal set of indices as $\mathbb{I}$ and random subsets of $\mathbb{I}$ by Ω. We denote the empty set as $\emptyset$. We denote the standard matrix basis as $\{\boldsymbol{e}_{ij}\}_{i,j=1}^n$, where $\boldsymbol{e}_{ij} = \boldsymbol{e}_i \boldsymbol{e}_j^\top$, which is zero everywhere except a 1 in the (i,j)-th entry. We denote the map $\mathrm{vec}(\cdot)$ as the operation that takes in a matrix $\boldsymbol{Y} \in \mathbb{R}^{n\times n}$ and returns a column vector, with each column of $\boldsymbol{Y}$ stacked in order, in $\mathbb{R}^{n^2}$. We define the thin spectral decomposition of a symmetric rank-r matrix as $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top$, where $\boldsymbol{U} \in \mathbb{R}^{n\times r}$ and $\boldsymbol{D} \in \mathbb{R}^{r\times r}$. We define $\mathcal{I}$ as the identity operator on matrices, and $\boldsymbol{I}$ as the identity matrix. We denote the condition number κ of a rank-r matrix $\boldsymbol{Y}$ as $\kappa = \frac{\|\boldsymbol{Y}\|}{\sigma_r(\boldsymbol{Y})}$, where $\sigma_r(\boldsymbol{Y})$ is the smallest non-zero singular value.

We denote the manifold of rank-r matrices as $\mathcal{M}_r$, and general smooth manifolds as $\mathcal{M}$. We denote the tangent space of the ground truth solution $\boldsymbol{X} \in \mathcal{M}_r$ to be $\mathbb{T}$, and the tangent space of the l-th iterate in the iterative sequences defined in Section 5 as $\mathbb{T}_l$. We denote the Euclidean gradient of a function $f \in C^1(\mathbb{R}^n)$ as ∇f, and the Riemannian gradient of a function $f \in C^1(\mathcal{M})$ as $\mathrm{grad}\, f$.

1.3 Organization

The organization of this paper is as follows. In Section 2, we discuss the requisite background information necessary to understand the work done in this paper. This consists of a brief discussion of dual bases of a vector space, first-order retraction-based Riemannian optimization methods, low-rank matrix completion, and discussion of EDG. Section 3 discusses related geometric approaches in matrix completion, relevant work done in EDG, and a more detailed discussion of geometric approaches to EDG. Section 4 is a discussion of our two proposed methodologies for solving the EDG problem using geometric low-rank matrix completion ideas in the developed dual basis framework. Section 5 discusses the underlying assumptions, convergence analysis, and initialization guarantees of one of the proposed algorithms, with most proofs deferred to the Appendices. The convergence analysis leverages the discussed dual basis structure, with properties proven in Appendix A, to get local convergence guarantees, discussed in more detail in Appendix C. We additionally provide initialization guarantees in this section, with relevant proofs in Appendix D. Section 6 discusses the numerical results of these algorithms. We conclude the paper in Section 7 with a brief discussion of the work and possible future research directions.

2 Background

In this section, we will provide some minor background necessary to understand the work done in the following sections.

2.1 Dual Basis

In a finite dimensional vector space of matrices $\mathbb{V}$, where $\dim(\mathbb{V}) = n$, a basis is a linearly independent set of matrices $B = \{\boldsymbol{X}_i\}_{i=1}^n$ that spans $\mathbb{V}$. Any basis for a finite dimensional vector space admits a dual, or bi-orthogonal, basis denoted $B^* = \{\boldsymbol{Y}_i\}_{i=1}^n$ that also spans $\mathbb{V}$, and admits a bi-orthogonality relationship

$$\langle \boldsymbol{X}_i, \boldsymbol{Y}_j \rangle = \delta_{ij}.$$

Additionally, B uniquely determines B^*. 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 $\boldsymbol{X}_i = \sum_{j=1}^n H_{ij} \boldsymbol{Y}_j$ [33].

2.2 Riemannian Optimization

The primary setting for this work is the Riemannian manifold of fixed-rank matrices. Throughout this work, we will only be considering square $n \times n$ matrices for simplicity and relevance to the problem of interest in this paper. For a fixed positive integer $r \le n$, we denote the set $\mathcal{M}_r = \{\boldsymbol{X} \in \mathbb{R}^{n\times n} \mid \operatorname{rank}(\boldsymbol{X}) = r\}$. Although not obvious at first glance, it is well-known that $\mathcal{M}_r$ is a smooth Riemannian manifold [34, 35]. To make this a Riemannian manifold, we equip it with the standard trace inner product as a metric, or $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \operatorname{Tr}(\boldsymbol{A}^\top \boldsymbol{B})$, restricted to the tangent bundle $T\mathcal{M}_r$, which is the disjoint union of tangent spaces [35].

Additionally, the tangent space at a point $\boldsymbol{X} \in \mathcal{M}_r$ is known and can be characterized [34–36]. For notational simplicity, and of relevance in the context of optimization, assume that $\boldsymbol{X}$ is the ground truth solution to an objective function. We additionally assume that $\boldsymbol{X} = \boldsymbol{X}^\top$, as all the matrices we consider are symmetric. The following ideas can be re-stated for rectangular matrices using a singular value decomposition, but these are not the subject of this paper. As such, we denote the tangent space at $\boldsymbol{X}$ as $\mathbb{T}$, and for a sequence of iterates $\{\boldsymbol{X}_l\}_{l\ge 0}$, we refer to their respective tangent spaces as $\mathbb{T}_l$. To characterize $\mathbb{T}$, let $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top$ be the thin spectral decomposition of $\boldsymbol{X}$. The tangent space $\mathbb{T}$ can be computed as follows:

$$\mathbb{T} = \{\boldsymbol{U}\boldsymbol{Z}^\top + \boldsymbol{Z}\boldsymbol{U}^\top \mid \boldsymbol{Z} \in \mathbb{R}^{n\times r}\}.$$

The tangent space can be described as the set of all possible rank-up-to-$2r$ perturbations, represented as the sum of a perturbation in the column and row space, and is computed by looking at first-order perturbations of the spectral decomposition of $\boldsymbol{X}$ [34]. Additionally, we can compute the orthogonal projection of an arbitrary $\boldsymbol{Y} \in \mathbb{R}^{n\times n}$ onto the tangent space at a point $T_{\boldsymbol{X}}\mathcal{M}_r$ as follows [34–36]:

$$\mathcal{P}_{\mathbb{T}}\boldsymbol{Y} = \mathcal{P}_U \boldsymbol{Y} + \boldsymbol{Y}\mathcal{P}_U - \mathcal{P}_U \boldsymbol{Y} \mathcal{P}_U$$

where $\mathcal{P}_U = \boldsymbol{U}\boldsymbol{U}^\top$ is the orthogonal projection onto the subspace spanned by the r columns of $\boldsymbol{U}$.

Optimization over $\mathcal{M}_r$ has been investigated in detail for quite some time, and in particular retraction-based methods are of particular interest to this work [34, 36–42]. First-order retraction-based methodologies rely on the general principle of taking a descent step in the tangent space, followed by a retraction onto the manifold. In the case of first-order optimization on $\mathcal{M}_r$, the retraction map $\mathcal{H}_r$ is given by the hard thresholding operator, which is a thin spectral decomposition that takes $\boldsymbol{Y} = \sum_{i=1}^n \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^\top \mapsto \sum_{i=1}^r \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^\top$, where $|\lambda_1| \ge ... \ge |\lambda_n|$ are the ordered eigenvalues of $\boldsymbol{Y}$ and $\boldsymbol{u}_i$ are the corresponding eigenvectors of $\boldsymbol{Y}$.

In order to construct a first-order method on $\mathcal{M}_r$, we need to define the notion of a Riemannian gradient. This object can be constructed in a greater degree of generality than our approach, but for simplicity, we will assume that a function $f : \mathcal{M}_r \to \mathbb{R}$ can be smoothly extended to all of $\mathbb{R}^{n\times n}$. That is to say, if we consider $f : \mathbb{R}^{n\times n} \to \mathbb{R}$, the Riemannian gradient of $f\big|_{\mathcal{M}_r}$, denoted grad f, for $\boldsymbol{X}_l \in \mathcal{M}_r$ is given by:

$$\operatorname{grad} f(\boldsymbol{X}_l) = \mathcal{P}_{\mathbb{T}_l} \nabla f(\boldsymbol{X}_l),$$

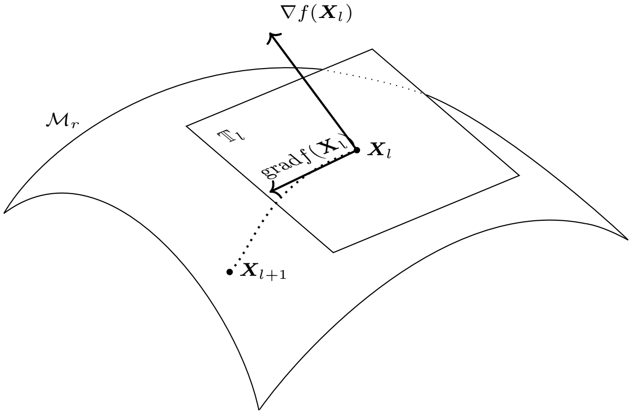

Figure 1: A diagram of a simple first-order retraction method on $\mathcal{M}_r$. Again, $\nabla f(\boldsymbol{X}_l)$ is the Euclidean gradient of f at $\boldsymbol{X}_l$, $\mathrm{grad}\, f(\boldsymbol{X}_l)$ is the Riemannian gradient at $\boldsymbol{X}_l$, and $\boldsymbol{X}_{l+1} = \mathcal{H}_r(\boldsymbol{X}_l - \alpha_l \mathrm{grad}\, f(\boldsymbol{X}_l))$, as in (5).

where ∇f is the Euclidean gradient of f. Using this approach, we can now define a Riemannian gradient descent iterate sequence using our retraction map, Riemannian gradient, and some step size sequence $\{\alpha_l\}_{l\geq 0}$ as follows:

$$\boldsymbol{X}_{l+1} = \mathcal{H}_r(\boldsymbol{X}_l - \alpha_l \mathcal{P}_{\mathbb{T}_l} \nabla f(\boldsymbol{X}_l)). \tag{5}$$

Intuitively, this algorithm seeks to look at changes in the objective function that lie, locally, along the manifold, followed by a retraction to stay on the desired manifold. An illustration can be seen in Figure 1.

This is a simple first pass to first-order optimization on Riemannian manifolds, and is not meant to be exhaustive. Interested readers should consult [35, 37] for further details on first-order methods on matrix (and other Riemannian) manifolds, along with convergence analysis for these algorithms.

2.3 Matrix Completion

One of the primary components this work relies on is the field of low-rank matrix completion, where a subset of the entries of a low-rank ground truth matrix $\boldsymbol{X}$ are observed. Consider $\boldsymbol{X}$ as an $n \times n$ matrix for simplicity, with $\Omega \subset [n] \times [n]$ representing the set of observed indices. Here, a sampling operator $\mathcal{P}_\Omega : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ is introduced, which aggregates the observed entries of $\boldsymbol{X}$ projected onto specific basis elements $\boldsymbol{e}_{ij}$:

$$\mathcal{P}_\Omega(\boldsymbol{X}) = \sum_{(i,j)\in\Omega} \langle \boldsymbol{X}, \boldsymbol{e}_{ij} \rangle \boldsymbol{e}_{ij}. \tag{6}$$

If Ω does not contain any repeated indices, $\mathcal{P}_\Omega$ is an orthogonal projection operator. The standard low-rank matrix completion problem can be phrased as follows:

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

As minimizing the rank directly is generally a challenging problem [25, 43], relaxations of this problem are often considered. For details on complexity class of rank constrained problems, we refer the reader to [44]. Exact recovery of $\boldsymbol{X}$ from $\mathcal{P}_\Omega(\boldsymbol{X})$ using a convex relaxation to the nuclear norm, such as the objective described in (3), is a well-studied problem [24, 45, 46] with strong convergence guarantees. This problem is at the core of matrix completion literature, and has inspired work in the completion of distance matrices [28, 29]. However, solving the convex problem is expensive for large matrices, which has led to the consideration of non-convex methodologies to solve the underlying problem. One approach that has received a great deal of attention is the Burer-Monteiro factorization approach, pioneered for semi-definite methods in [47], whereby a low rank matrix $\boldsymbol{X} \in \mathbb{R}^{n\times n}$ can be factored into a product $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{B}^\top$ for $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times r}$. Minimizing $\|\mathcal{P}_\Omega(\boldsymbol{X}) - \mathcal{P}_\Omega(\boldsymbol{A}\boldsymbol{B}^\top)\|_{\mathrm{F}}^2$ is a common approach, and is often dealt with using alternating minimization methods in both the noiseless and noisy case [48–51].

2.4 Dual Basis Approach to EDG

In the EDG problem, using the relation (2), we can relate each entry of the squared distance matrix to the Gram matrix as follows: $D_{ij} = X_{ii} + X_{jj} - X_{ij} - X_{ji}$. We describe here briefly the dual basis approach introduced in [28].

Given $\boldsymbol{\alpha} = (\alpha_1, \alpha_2), \alpha_1 < \alpha_2$, we define the matrix $\boldsymbol{w}_{\boldsymbol{\alpha}}$ as follows:

$$\boldsymbol{w}_{\boldsymbol{\alpha}} = \boldsymbol{e}_{\alpha_1\alpha_1} + \boldsymbol{e}_{\alpha_2\alpha_2} - \boldsymbol{e}_{\alpha_1\alpha_2} - \boldsymbol{e}_{\alpha_2\alpha_1}.$$

If we consider the set $\mathbb{I} = \{(\alpha_1, \alpha_2), 1 \leqslant \alpha_1 < \alpha_2 \leqslant n\}$, it can be checked that the set $\{\boldsymbol{w}_{\boldsymbol{\alpha}}\}$ is a non-orthogonal basis for the subspace of symmetric matrices with zero row sum, denoted $\mathbb{S} = \{\boldsymbol{Y} \in \mathbb{R}^{n\times n} \mid \boldsymbol{Y} = \boldsymbol{Y}^\top, \boldsymbol{Y}\boldsymbol{1} = \boldsymbol{0}\}$. In fact, for any two pairs of indices $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{I}$, we have:

$$\langle \boldsymbol{w}_{\boldsymbol{\alpha}}, \boldsymbol{w}_{\boldsymbol{\beta}} \rangle = \begin{cases} 4 & \boldsymbol{\alpha} = \boldsymbol{\beta}; \\ 1 & \boldsymbol{\alpha} \neq \boldsymbol{\beta},\ \boldsymbol{\alpha} \cap \boldsymbol{\beta} \neq \emptyset; \\ 0 & \boldsymbol{\alpha} \cap \boldsymbol{\beta} = \emptyset. \end{cases}$$

It can also easily be verified that the dimension of the linear space $\mathbb{S}$ is $L = n(n-1)/2$. Using this basis, we can realize each entry of the squared distance matrix as the trace inner product of the Gram matrix with the basis. Formally, $D_{ij} = \langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle$ for $\boldsymbol{\alpha} = (i, j)$. Further, we can introduce the dual basis to $\{\boldsymbol{w}_{\boldsymbol{\alpha}}\}$, denoted as $\{\boldsymbol{v}_{\boldsymbol{\alpha}}\}$, 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 EDG problem as a low-rank matrix recovery problem where we observe a subset of the expansion coefficients. In [28], this dual basis formulation has been used to provide theoretical guarantees for the convex program given in (4).

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. The direct form of the dual basis, based on its definition, relies on an inverse of a matrix of size $L \times L$ which requires the solution of a large linear system. In [52], it was shown that the dual basis admits a simple explicit form

$$\boldsymbol{v}_{\boldsymbol{\alpha}} = -\frac{1}{2}\left(\boldsymbol{a}\boldsymbol{b}^\top + \boldsymbol{b}\boldsymbol{a}^\top\right), \tag{7}$$

where $\boldsymbol{a} = \boldsymbol{e}_i - \frac{1}{n}\boldsymbol{1}$ and $\boldsymbol{b} = \boldsymbol{e}_j - \frac{1}{n}\boldsymbol{1}$ 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 [53], $\mathcal{R}_\Omega(\boldsymbol{X})$ is related to the sampling operator $\mathcal{P}_\Omega(\boldsymbol{D})$ as follows:

$$\mathcal{R}_\Omega(\boldsymbol{X}) = -\frac{1}{2}\boldsymbol{J}\mathcal{P}_\Omega(\boldsymbol{D})\boldsymbol{J}, \tag{8}$$

where $\boldsymbol{J}$ is as defined in Section 1. The next operator is the restricted frame operator $\mathcal{F}_\Omega : \mathbb{S} \to \mathbb{S}$, first studied in [28], and defined as

$$\mathcal{F}_\Omega(\cdot) = \sum_{\boldsymbol{\alpha}\in\Omega} \langle \cdot, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \boldsymbol{w}_{\boldsymbol{\alpha}}. \tag{9}$$

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 [30].

3 Related Work

3.1 A Riemannian Approach to Matrix Completion

A notable non-convex approach is to utilize prior knowledge regarding the rank of $\boldsymbol{X}$. This methodology centers around the fact that the set of fixed-rank matrices forms a Riemannian manifold, turning the problem into an unconstrained optimization task over a manifold. These methodologies lose convexity, however, and generally only local convergence guarantees can be established, done by proving the existence of attractive basins around solutions. Various retraction-based methodologies have been used with differing metrics and geometric structures [34,36,54–58]. The analysis conducted by [36] stands out for its interpretation of its first-order method as an iterative hard-thresholding algorithm with subspace projections and efficient numerical implementation. This implementation is done by reducing the hard thresholding step from a thin eigenvalue decomposition of an $n \times n$ matrix to a thin QR decomposition followed by a full eigenvalue decomposition of a far smaller $2r \times 2r$ matrix. The convergence analysis in this work builds on the analysis done in [36], and as such, a brief exposition of their work is provided.

In [36], the authors develop a gradient descent algorithm to solve the low-rank matrix completion problem leveraging this Riemannian structure. The objective function used in [36] is as follows:

$$\underset{\boldsymbol{Y}\in\mathbb{R}^{n\times n}}{\text{minimize}}\ \langle \boldsymbol{Y}-\boldsymbol{X}, \mathcal{P}_\Omega(\boldsymbol{Y}-\boldsymbol{X})\rangle \text{ subject to } \operatorname{rank}(\boldsymbol{Y}) = r. \tag{10}$$

The authors used a uniform sampling at random with replacement model for recovering a subset of the indices of the ground truth matrix. This is standard practice in existing matrix completion literature, as much of the analysis relies on concentration inequalities for sums of random matrices to get high probability guarantees. It follows that (10) is not equivalent to $\|\mathcal{P}_\Omega(\boldsymbol{X}-\boldsymbol{M})\|_{\mathrm{F}}^2$ when indices in Ω repeat, as $\mathcal{P}_\Omega^2 \neq \mathcal{P}_\Omega$ when this occurs. This is distinct from [34], which minimized the Frobenius norm difference between the observed entries of the low-rank matrices to solve the problem. Additionally, [34] demonstrates that the limit of their proposed algorithm agrees with the ground truth in the revealed entries when projected onto the tangent space of the ground truth. However, as the sampling operator has a non-trivial null space, noted in [34], this does not necessarily guarantee identification of the ground truth. In contrast, [36] establishes linear convergence to the ground truth solution in a local neighborhood of the ground truth, with high probability. After defining (10), [36] constructs a Riemannian gradient descent procedure similar to the retraction procedure described in Section 2.2 for its solution.

In addition to this approach, the work in [36] considered two initialization schemes. One is a simple one-step hard threshold onto $\mathcal{M}_r$, and is given by $\boldsymbol{X}_0 = \frac{n^2}{m}\mathcal{H}_r(\mathcal{P}_\Omega(\boldsymbol{M}))$. Additionally, a more delicate initialization can be considered by partitioning the set Ω into S equally sized subsets, and performing one Riemannian gradient descent step for each subset. This Riemannian resampling initialization breaks the dependence on each iterate from the previous, and provides a more reliable initialization for large enough sample sizes. A modification of this technique, applied to our scheme, can be seen in Algorithm 3.

3.2 Euclidean Distance Geometry Algorithms

To solve the EDG problem, various algorithms have been developed. Among them, one prominent family of algorithms is based on semi-definite programming (SDP), which leverages the connection between squared distance matrices and Gram matrices. To provide a concrete example of this approach, we briefly outline the method proposed in [59]. Consider the matrix $\boldsymbol{V} \in \mathbb{R}^{n\times(n-1)}$, whose columns form an orthonormal basis for the space $\{\boldsymbol{z} \in \mathbb{R}^n : \boldsymbol{z}^\top \boldsymbol{1} = 0\}$. The operator $\mathcal{K}$ is defined as:

$$\mathcal{K}(\boldsymbol{X}) = \operatorname{diag}(\boldsymbol{X})\boldsymbol{1}^\top + \boldsymbol{1}\operatorname{diag}(\boldsymbol{X})^\top - 2\boldsymbol{X}.$$

This definition of the operator $\mathcal{K}(\boldsymbol{X})$ is equivalent to the mapping of the Gram matrix to the squared Euclidean distance matrix, as expressed in (2). In [59], the optimization program is based on the operator $\mathcal{K}_{\boldsymbol{V}}(\boldsymbol{X})$, which is defined as $\mathcal{K}_{\boldsymbol{V}}(\boldsymbol{X}) = \boldsymbol{V}\boldsymbol{X}\boldsymbol{V}^\top$. The optimization problem in [59] can then formulated as follows:

$$\underset{\boldsymbol{X}\in\mathbb{R}^{(n-1)\times(n-1)},\ \boldsymbol{X}=\boldsymbol{X}^\top,\ \boldsymbol{X}\succeq\boldsymbol{0}}{\text{minimize}} \quad \sum_{(i,j)\in\Omega} \left[(\mathcal{K}_{\boldsymbol{V}}(\boldsymbol{V}\boldsymbol{X}\boldsymbol{V}^\top))_{ij} - D_{ij}\right]^2.$$

We refer the reader to [59] for theoretical and numerical aspects of the above optimization program. Given that standard SDP formulations can be computationally intensive, distributed and divide-and-conquer methods have also been explored. For additional SDP-based formulations of the EDG problem and their applications to molecular conformation and sensor network localization, we refer the reader to [6, 60–64].

In the context of protein structure determination, various algorithmic approaches to EDG have been developed. One notable example is the EMBED algorithm [65–67], which comprises three main steps [68]. The first step, known as bound smoothing, involves generating lower and upper bounds for all distances by extrapolating from the available limits of known distances. The second step is the embed step, where distances are sampled from these bounds to form a full distance matrix from which an initial estimate of the protein structure is obtained. The final step involves refining this initial structure by minimizing an energy function using non-convex optimization methods. Another approach to structure prediction is the discretizable molecular distance geometry framework, which can be formulated as a search in a discrete space and then uses a Branch-and-Prune method to estimate the structure [69, 70].

Another category of approaches to the EDG problem involves initially estimating a smaller portion of the point cloud and then using this initial estimate to incrementally reconstruct the rest of of the structure. These methods are referred to as geometric build-up algorithms [71–73]. The algorithm proposed in [74] addresses the molecular conformation problem by adopting a divide-and-conquer strategy, where a sequence of smaller optimization problems is solved instead of solving a single global optimization problem. Finally, we highlight algorithms that estimate the underlying points through non-convex optimization, utilizing a combination of methods such as majorization, alternating projection, and global continuation (transforming the optimization problem to a function with few local minimizers) [11, 75–77]. We note that the above discussion does not comprehensively cover all EDG algorithms, and we refer readers to [20, 78] for a more detailed overview.

3.2.1 Related Geometric Approaches to EDG

The main perspective taken in this paper is in line with low-rank matrix completion approach, albeit not one that employs the trace heuristic seen in [6, 28, 79]. This work is more in line with non-convex approaches based on optimizing over a Riemannian manifold [32, 80], and extends the Riemannian approach of [36] to the EDG basis case.

A recent work in [30] adopts a similar approach to us and considers solving the EDG problem through Riemannian methods as well. In this work, the authors use a Riemannian conjugate method paired with an inexact line search method to minimize the following s-stress objective function:

$$\underset{\boldsymbol{Y}\in\boldsymbol{R}^{n\times d}}{\text{minimize}}\ \frac{1}{2}\|\boldsymbol{W}\odot\mathcal{P}_{\Omega}(g(\boldsymbol{Y}\boldsymbol{Y}^{\top})-\boldsymbol{D}_e)\|_{\mathrm{F}}^2, \tag{11}$$

where g is the map defined by (2), $\boldsymbol{W}$ is a weight matrix to model noisy entries, and $\odot$ is the Hadamard product, and $\mathcal{P}_{\Omega}$ is defined as in (6). The analysis in [30] centers around the minimization of the s-stress function in (11) using a generalization of a Hager-Zhang line search method to a Riemannian quotient manifold. The main result in this work is that there exists an attractive basin for (11) that, with high probability, gives linear convergence to the ground truth provided an initialization in the basin. This result requires a Bernoulli sample complexity $p > C\frac{(\nu r)^3 log(n)}{n}$, where ν is the coherence of the ground truth matrix and r is the rank. Our method also describes two strong initialization candidates for the noiseless EDG recovery problem with provable high probability guarantees, with a sample complexity that only depends quadratically on the coherence and rank.

We provide a separate convergence analysis from [30], demonstrating a robust Restricted Isometry Property of a non-self-adjoint sampling operator, and prove local convergence for this non-orthogonal matrix completion problem. This novel approach requires a relaxation away from the minimization of a quadratic form over a manifold, instead considering a linearly contractive sequence in a neighborhood modeled after [36] and a surrogate step size, to be expanded on later. This approach requires novel analysis of the dual-basis framework discussed in [28, 29, 31], mainly centered around careful eigenvalue bounds in tandem with standard matrix completion tools, at a cost of slightly worse sample complexity. Additionally, we extend the initialization techniques of [36], and show that our modified approach can provide similar guarantees. The non-self adjoint nature of the EDG sampling operator provides a host of challenges that are resolved through careful analysis of the EDG basis and extensions of its properties beyond what has been discovered already. To the authors' knowledge, this is the first non-convex method that provides high probability guarantees on the initialization methods provided.

4 The Riemannian Dual Basis Approach to EDG

With the goal of translating the standard matrix completion problem to Gram matrix completion for EDG in mind, the most direct adaptation of the work conducted in [36] would be defining an objective function by analogy to (10) as follows:

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

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 $\mathcal{R}_{\Omega}^{\star}(\boldsymbol{X})$ as

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

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 $\mathcal{R}_\Omega$ [28, 53, 81], one of which we will expand upon shortly. These surrogates allow for the definition of an objective function in analogy to (10), but as of now lack the requisite theoretical properties for convergence.

The primary surrogate of interest in this work is the restricted frame operator $\mathcal{F}_\Omega$, defined in (9). 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 [30]. Using this operator, we can define the following objective function:

$$\underset{\boldsymbol{Y}\in\mathbb{S}}{\text{minimize}}\ \frac{1}{2}\langle \boldsymbol{Y}-\boldsymbol{X}, \mathcal{F}_\Omega(\boldsymbol{Y}-\boldsymbol{X})\rangle\ \text{subject to}\ \operatorname{rank}(\boldsymbol{Y})=r. \tag{12}$$

This is a true quadratic form, minimized over $\mathcal{M}_r$, and can be approached in an identical manner algorithmically as (10). This operator motivates Algorithm 1, where the hard thresholding operator $\mathcal{H}_r$ is again defined as the map from $\boldsymbol{Y}=\sum_{i=1}^n \lambda_i \boldsymbol{u}_i\boldsymbol{u}_i^\top \mapsto \sum_{i=1}^r \lambda_i \boldsymbol{u}_i\boldsymbol{u}_i^\top$, where $|\lambda_1|\geq ... \geq |\lambda_n|$:

Algorithm 1 Restricted Frame EDG Riemannian Gradient Descent

```
Initialization: $\boldsymbol{X}_0 = \boldsymbol{U}_0\boldsymbol{D}_0\boldsymbol{U}_0^\top$
for $l = 0, 1, ...$ do
    1. $\boldsymbol{G}_l = \mathcal{F}_\Omega(\boldsymbol{X}-\boldsymbol{X}_l)$
    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: $\boldsymbol{X}_\mathrm{rev}$
```

Algorithm 1 is easily implementable and gives strong numerical results, provided in Section 6, but proof of local convergence remains an open question. The missing analytical property that would yield local convergence is the Restricted Isometry Property (RIP), which states that a given operator does not distort a matrix too severely when projected on to the tangent space of the ground truth.

Remark 1. *More mathematically, let $\mathcal{K}_\Omega$ be a stochastic sampling operator, such as $\mathcal{P}_\Omega$, $\mathcal{R}_\Omega$, or $\mathcal{F}_\Omega$. RIP states that with high probability*

$$\|\mathcal{P}_\mathbb{T}\mathcal{K}_\Omega\mathcal{P}_\mathbb{T} - c\mathcal{P}_\mathbb{T}\| \leq \varepsilon_0,$$

for some $\varepsilon_0 > 0$ and some constant $c > 0$. In practice, this statement is proven using non-commutative concentration inequalities, first introduced in the matrix completion literature in [27, 45], requiring that $\mathcal{K}_\Omega$ is a sum of i.i.d. random operators and that the expectation of $\mathcal{P}_\mathbb{T}\mathcal{K}_\Omega\mathcal{P}_\mathbb{T} = c\mathcal{P}_\mathbb{T}$ for some $c > 0$. 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 5.4, 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 $\mathcal{P}_\Omega$. 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 $\mathbb{S}$, and the subsequent convergence analysis of Algorithm 1 will be the subject of future work.

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 Algorithm 1 but without any reference to an objective function. As we will show in Section 5, this algorithm will give us strong convergence guarantees with reasonable sample complexities at a cost of interpretability. As such, we define Algorithm 2 to reconstruct a ground truth matrix $\boldsymbol{X}$ as follows:

Unlike in Algorithm 1, 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 $\mathcal{M}_r$. When $\alpha_l = 0$ occurs, the algorithm terminates. Positive α_l is required for convergence, and the conditions are provided and characterized in Lemma C.1. This condition is satisfied in the high-sample regime, where ε_0 is small.

In both of the preceding approaches, the thin spectral decomposition in the gradient descent scheme is the most expensive, especially when n is large. As described previously, the authors in [36] found an efficient way to reduce the computational complexity of this decomposition from $\mathcal{O}(rn^2)$ to $\mathcal{O}(r^3) + \mathcal{O}(nr^2)$, substantially reducing the cost per iteration, which we implement as well.

Algorithm 2 Riemannian Pseudo-Gradient 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:** $\boldsymbol{X}_{\mathrm{rev}}$
```

Computation of $\mathcal{R}_\Omega(\boldsymbol{X})$ can be done efficiently, with a 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 (2), and through (8), $\mathcal{R}_\Omega(\boldsymbol{X})$ can be computed in $\mathcal{O}(m)$ operations, for $|\Omega| = m$. From [53], 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.

5 Theoretical Analysis

In this section, we will provide the main results of this work, which are the local convergence and recovery guarantees for Algorithm 2, presented in Theorems 5.5, 5.7, and 5.9. 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 5.1 (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 $\boldsymbol{\alpha} = (i,j) \in \mathbb{I}$:*

$$\|\mathcal{P}_U \boldsymbol{e}_{ij}\|_{\mathrm{F}} \leq \sqrt{\frac{\nu r}{128n}}, \qquad \|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \leq \sqrt{\frac{\nu r}{8n}}, \quad \text{and} \quad \|\mathcal{P}_U \boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \leq \sqrt{\frac{\nu r}{2n}}. \tag{13}$$

In addition to the above, we require that

$$\|\mathcal{P}_{\mathbb{T}} \boldsymbol{e}_{ij}\|_{\mathrm{F}} \leq \sqrt{\frac{\nu r}{128n}}, \qquad \|\mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \leq \sqrt{\frac{\nu r}{8n}}, \quad \text{and} \quad \|\mathcal{P}_{\mathbb{T}} \boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \leq \sqrt{\frac{\nu r}{2n}}. \tag{14}$$

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 (13) and (14) hold. As in [36], we note that the first condition above implies the following:

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

This indicates that $\left\|\boldsymbol{U}^{(i)}\right\|_2^2 \leq \frac{\nu r}{128n}$, which will be relevant when discussing the initialization of Algorithm 2 using a trimming step in Algorithm 4.

Remark 2. *We want to note that the first assumption in* (13) *actually implies the next two. That is to say, if $\|\mathcal{P}_U \boldsymbol{e}_{ij}\|_2 \leq \sqrt{\frac{\nu r}{128n}}$, 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 A.6, the claim follows. A similar proof shows the same relationship for the equations in (14)*.*

Additionally, we make an assumption in accordance with [36]:

Assumption 5.2. *Let $\boldsymbol{X} \in \mathbb{R}^{n \times n}$ be a rank-r matrix. We assume that an absolute numerical constant μ_1 such that*

$$\|\boldsymbol{X}\|_\infty \leq \mu_1 \sqrt{\frac{r}{n^2}} \|\boldsymbol{X}\|. \tag{15}$$

Remark 3. *Notice that this condition is equivalent up to scaling factors for a similar assumption in [57], which itself can be upper bounded by a similar coherence condition in [25]. In fact, we can relate μ_1 directly to ν as follows. For simplicity and relevance to our problem, let $\boldsymbol{X} \succeq \mathbf{0}$, and notice that*

$$\begin{aligned}\frac{\|\boldsymbol{X}\|_\infty}{\|\boldsymbol{X}\|} &= \frac{1}{\|\boldsymbol{X}\|} \max_{ij} |X_{ij}| \\ &= \frac{1}{\|\boldsymbol{X}\|} \max_{ij} \left| \sum_{kl} U_{ik} D_{kl} U_{jl} \right| \\ &\leq \max_{ij} \sum_{k=1}^{r} \left| U_{ik} \frac{\lambda_k}{\lambda_1} U_{jk} \right| \\ &\leq \max_{ij} \sum_{k=1}^{r} |U_{ik} U_{jk}| \\ &\leq \sqrt{\sum_{1 \leq k \leq r} |U_{ik}|^2} \sqrt{\sum_{1 \leq k \leq r} |U_{ik}|^2} \\ &\leq \frac{\nu r}{128n},\end{aligned}$$

where the penultimate inequality follows from Cauchy-Schwartz, and the final inequality from (13)*, indicating that in a worst-case scenario $\mu_1 \leq \frac{\nu\sqrt{r}}{128}$. This property is ultimately separate from the definition of ν, but at the very least it can be upper bounded as a function of ν.*

Additionally, 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 5.3. *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 both Assumptions 5.1 and 5.2 with $\mathcal{O}(1)$ constant factors ν and μ_1. As in [36], 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 Algorithm 2 is RIP. This theorem provides the conditions needed for RIP of $\mathcal{R}_\Omega$:

Theorem 5.4 (Restricted Isometry Property (RIP) for $\mathcal{P}_\mathbb{T}\mathcal{R}_\Omega\mathcal{P}_\mathbb{T}$)**.** *With probability at least $1 - 2n^{1-\beta}$,*

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

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

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

Proof sketch. This result is primarily a consequence of Theorem A.1, 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 is leveraging specific properties of the dual bases $\{\boldsymbol{w}_{\boldsymbol{\alpha}}\}$ and $\{\boldsymbol{v}_{\boldsymbol{\alpha}}\}$ to prove concentration of $\mathcal{P}_{\mathbb{T}}\mathcal{R}_{\Omega}\mathcal{P}_{\mathbb{T}}$ around its expectation. See Appendix B.1 for details. □

This leads us into the following theorem, which is the crux of this work. Theorem 5.5, 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 C.1.

Theorem 5.5 (Local Convergence of Algorithm 2)**.** *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*

$$\left\|\frac{L}{m}\mathcal{P}_{\mathbb{T}}\mathcal{R}_{\Omega}\mathcal{P}_{\mathbb{T}} - \mathcal{P}_{\mathbb{T}}\right\| \leq \varepsilon_0 \tag{16}$$

$$\frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\sigma_{min}(\boldsymbol{X})} \leq \frac{\sqrt{m}\varepsilon_0}{16n^{5/4}\sqrt{\beta\nu r\log n}} \tag{17}$$

$$\|\mathcal{R}_{\Omega}\| \leq \frac{m}{L} + 4\sqrt{\frac{8m\log(n)}{n}} \tag{18}$$

$$\|\mathcal{R}_{\Omega}\mathcal{P}_{\mathbb{T}}\| \leq \frac{m}{L} + \frac{m\sqrt{n}}{L}\sqrt{\frac{\beta\nu rn\log(n)}{3m}} \tag{19}$$

$$\|\mathcal{P}_{\mathbb{T}}\mathcal{R}_{\Omega}\| \leq \frac{m}{L} + \frac{4m\sqrt{n}}{L}\sqrt{\frac{\beta\nu rn\log(n)}{3m}}, \tag{20}$$

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 5.5. 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 [82]. Next, plugging in $\boldsymbol{W}_l = \boldsymbol{X}_l + \alpha_l\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l$, we see that

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

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 (17), a local form of RIP for $\mathcal{R}_\Omega$ holds if (16) is true. This proof leverages the assumptions made in (18), (19), and (20). I_2 follows from the neighborhood assumption of (17) in tandem with Lemma E.1, and I_3 follows from bounds on the step size (seen in Lemma C.1), the assumption in (20), and Lemma E.1. The assumptions in (16), (18), (19), and (20) are all proven via high probability guarantees using Theorem A.1. The technical details are deferred to the appendix, but Figure 5 highlights the main dependencies of each lemma and how they work into the overall convergence. See the full proof in C.1. □

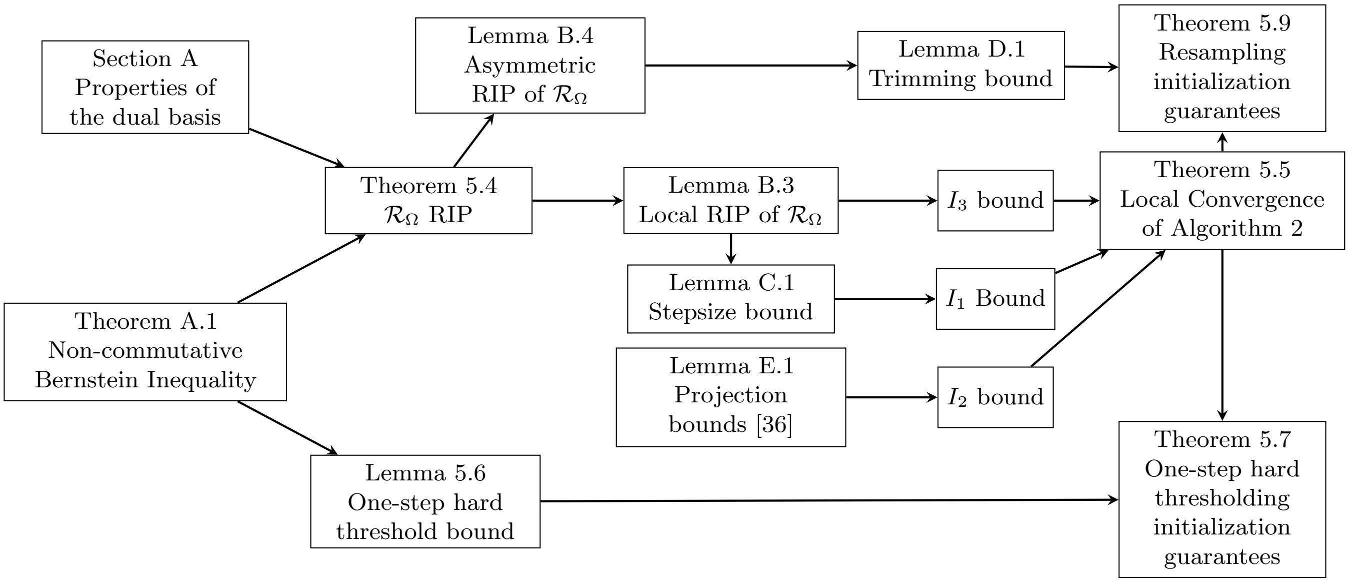

Figure 2: This diagram is a schematic of the overall proof of convergence. Arrows indicate how results depend on one another, and how they link together to form the overall proof of convergence. Not every exact dependency is shown in this figure for legibility purposes, instead focusing on the key pieces of the overall flow of the argument.

In Theorem 5.5, 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.

5.1 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 5.6 (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 $\Omega = m \geq \frac{40}{3}\beta n\log n$. It follows that with probability at least $1-2n^{1-\beta}$ that*

$$\|\boldsymbol{X}_0 - \boldsymbol{X}\|_{\mathrm{F}} \leq \sqrt{\frac{320r^2\mu_1^2 n\log(n)}{3m}}\,\|\boldsymbol{X}\|\,. \tag{21}$$

Proof. See Section D.1. □

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

$$m \geq 170\frac{\kappa\mu_1\sqrt{\beta\nu r^3}}{\varepsilon_0}n^{7/4}\log(n),$$

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 5.7 (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*

$$m \geq \max\left\{\frac{8}{3}\frac{\sqrt{\beta\nu^3 r}}{\varepsilon_0}, 170\kappa\mu_1 n^{3/4}\right\}\frac{\sqrt{\beta\nu r^3}}{\varepsilon_0} n\log(n),$$

with $\beta > 1$, then Algorithm 2 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 5.5. By setting the bound produced in Lemma 5.6 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 5.4. □

This naive one-step hard threshold initialization can be improved again following a construction in [36], using a resampling and trimming algorithm, both defined as follows:

Algorithm 3 Riemannian Resampling for Initialization

```
Partition $\Omega$ 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 4 `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.

We can analyze Algorithm 3 and get the following result:

Lemma 5.8 (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 3, 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 3 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 $\hat{m} \geq \max\left\{(1.61\times 10^5)\nu^2, (7.77\times 10^5)\kappa^4\mu_1^2\right\}\kappa^2 r^2 n\log(n)$.

Proof. See D.2. □

Assuming that $m \geq \beta\nu r n\log(n)$, which is relaxed from the requirement for RIP seen in Theorem 5.4, Lemma 5.8 shows that taking

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

the third condition in Lemma 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 5.9. *Let $\boldsymbol{X} \in \mathbb{R}^{n\times n}$ be the measured rank-r, ν-incoherent matrix with condition number κ, and suppose that $|\Omega| = m$ is a set of sampled indices from $\mathbb{I}$ uniformly at random with replacement. Let $\boldsymbol{X}_0$ be the output of Algorithm 3. Then for any $\beta > 1$, the iterates of Algorithm 2 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*

$$m \geq \max\left\{\frac{8\nu^2}{3\varepsilon_0^2}, (2.3 \times 10^5)\nu^2 \log\left(\frac{n^{3/4}}{16\varepsilon_0}\right), (1.2 \times 10^6)\kappa^4\mu_1^2 \log\left(\frac{n^{3/4}}{16\varepsilon_0}\right)\right\} \kappa^2 r^2 n \log(n).$$

Proof. As in Theorem 5.7, this proof follows from the local neighborhood conditions of Theorem 5.5, combined with the sample complexity results from Lemma 5.8 and Theorem 5.4. The constants are not optimized, and could be further improved. □

6 Numerical Results

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

6.1 Synthetic Data Experiments

To test Algorithms 1 and 2, various two and three dimensional datasets were used, and are referred to in Table 1 with their corresponding sizes. The goal of Algorithms 1 and 2 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 $\gamma \in [0, 1]$. Algorithm 2 reconstructs the Gram matrix $\boldsymbol{X} = \boldsymbol{P}^\top \boldsymbol{P}$, from which $\boldsymbol{P}$ can be recovered. The comparison referenced in Table 1 is the relative error between the recovered matrix $\boldsymbol{X}_{\text{rev}}$ 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.

Table 1: Relative recovery error $\|\boldsymbol{X} - \boldsymbol{X}_{\text{rev}}\|_{\text{F}} \,/\, \|\boldsymbol{X}\|_{\text{F}}$ between the recovered Gram matrix and the true Gram matrix averaged over 25 trials using Algorithms 1, 2, and the non-convex algorithm in [28].

γ Dataset	10%	7%	5%	3%	2%	1%
Algorithm 1						
Sphere (3D, $n = 1002$)	2.92e-05	3.69e-05	6.01e-05	1.28e-04	6.82e-03	9.11e-01
Cow (3D, $n = 2601$)	3.38e-05	4.06e-05	4.89e-05	7.61e-05	1.07e-04	8.60e-03
Swiss Roll (3D, $n = 2048$)	2.92e-05	3.97e-05	5.06-05	7.71e-05	1.21e-04	5.52e-02
U.S. Cities (2D, $n = 2920$)	4.10e-05	4.67e-05	2.01e-03	6.94e-03	1.63e-02	6.08e-02
Algorithm 2						
Sphere (3D, $n = 1002$)	1.53e-05	3.86e-05	2.11e-04	7.88e-02	2.29e-01	1.78e+00
Cow (3D, $n = 2601$)	4.15e-02	3.40e-02	6.28e-02	1.76e-01	3.60e+00	7.73e-01
Swiss Roll (3D, $n = 2048$)	8.34e-06	1.46e-05	3.00e-05	1.52e-03	2.82e-01	9.73e-01
U.S. Cities (2D, $n = 2920$)	3.02e-02	1.23e-01	1.46e-01	1.86e-01	3.13e-01	8.35e-01
Non-convex algorithm in [28]						
Sphere (3D, $n = 1002$)	6.14e-06	9.86e-06	1.36e-05	3.04e-05	6.18e-05	1.00e-01
Cow (3D, $n = 2601$)	5.73e-06	7.66e-06	1.06e-05	1.65e-05	2.11e-05	4.46e-05
Swiss Roll (3D, $n = 2048$)	2.19e-06	1.22e-06	1.01e-06	1.87e-06	1.06e-06	3.34e-05
U.S. Cities (2D, $n = 2920$)	4.09e-07	6.09e-07	8.19e-07	1.32e-06	2.30e-06	4.69e-06

In addition to the relative error comparison between the recovered Gram matrix and the ground truth Gram matrix, we compute the root mean square error (RMSE), defined as $\sqrt{\frac{1}{n}\|\boldsymbol{P}_{\text{rev}} - \boldsymbol{P}\|_{\text{F}}^2}$ between the recovered point cloud $\boldsymbol{P}_{\text{rev}}$ following a Procrustes realignment with the ground truth $\boldsymbol{P}$, under the same experimental parameters as with the Gram matrix recovery. The results are compiled in Table 2.

As indicated by these experiments, Algorithm 1 more reliably reconstructs the underlying datasets from distance samples than Algorithm 2, but both are outperformed by the non-convex algorithm in [28]. However, with the non-convex algorithm in [28] there is little hope of ever conducting local convergence analysis of this algorithm, whereas Algorithm 2 has been proven to exhibit local convergence. It remains to be seen if Algorithm 1 will be

Table 2: RMSE between $\boldsymbol{P}_{\text{rev}}$ and $\boldsymbol{P}$ averaged over 25 trials using Algorithms 1, 2, and the non-convex algorithm in [28].

γ Dataset	10%	7%	5%	3%	2%	1%
Algorithm 1						
Sphere (3D, $n = 1002$)	2.06e-05	2.61e-05	4.24e-05	9.07e-05	4.82e-03	7.14e-01
Cow (3D, $n = 2601$)	3.98e-05	4.89e-05	6.08e-05	9.25e-05	1.31e-04	1.24e-02
Swiss Roll (3D, $n = 2048$)	2.46e-04	3.36e-04	4.26e-04	6.48e-04	1.13e-03	4.92e-01
U.S. Cities (2D, $n = 2920$)	7.12e-04	8.13e-04	5.39e-02	1.86e-01	4.37e-01	1.70e+00
Algorithm 2						
Sphere (3D, $n = 1002$)	1.08e-05	2.72e-05	1.49e-04	5.37e-02	1.42e-02	1.24e+00
Cow (3D, $n = 2601$)	6.67e-02	7.81e-02	1.38e-01	2.49e-01	5.87e-01	5.52e-01
Swiss Roll (3D, $n = 2048$)	6.91e-05	1.21e-04	2.51e-04	1.24e-02	1.86e+00	1.34e+01
U.S. Cities (2D, $n = 2920$)	9.94e-01	3.89e+00	5.10e+00	6.25e+00	7.77e+00	1.13e+01
Non-convex algorithm in [28]						
Sphere (3D, $n = 1002$)	4.29e-06	6.94e-06	9.61e-06	2.14e-05	4.37e-05	7.10e-02
Cow (3D, $n = 2601$)	8.86e-06	1.19e-05	1.75e-05	2.70e-05	3.47e-05	7.08e-05
Swiss Roll (3D, $n = 2048$)	2.02e-05	1.10e-05	8.90e-06	1.66e-05	9.32e-06	2.96e-04
U.S. Cities (2D, $n = 2920$)	8.01e-06	1.40e-05	1.87e-05	3.11e-05	5.18e-05	1.13e-04

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 Algorithm 1 and [28] than in Algorithm 2, but the strong theoretical results provided by Algorithm 2 will guide future work for convergence analysis of Algorithm 1 and other self-adjoint surrogates for $\mathcal{R}_{\Omega}$.

6.2 Experiments on Real Data

Additional numerical experiments have been conducted on proteins, a common application of EDG, following the structured sampling method seen in [83]. The sampling method on n points has three different classes of points: m pseudoanchors, 1 central anchor, and $n - m - 1$ mobile nodes. The central anchor, corresponding to a row/column of the squared distance matrix, is fully known; that is, the distance between the central node and all n points is revealed in the masked square distance matrix. The pairwise distances between the pseudoanchors are sampled with a Bernoulli probability $\gamma \in [0, 1]$ for each entry, and each mobile node is connected uniformly at random to k of the pseudoanchors. None of the distances between mobile nodes are known. The Gram matrix $\boldsymbol{X}$ is then reordered into the following pattern: the first m rows/columns are the rows/columns corresponding to the pseudoanchors, the $m + 1$-th row/column corresponds to the central anchor, and the remaining $n - m - 1$ columns/rows correspond to the mobile nodes. This is illustrated in the figure below:

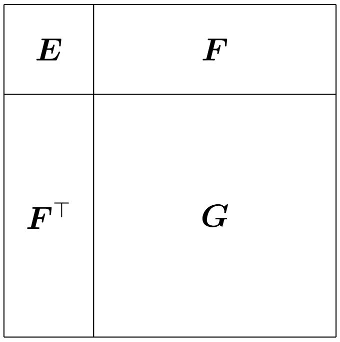

Figure 3: Structured sampling method for distance matrices proposed in [83] for the experiments.

More specifically, $\boldsymbol{E}$ is sampled according to an entry-wise Bernoulli distribution with parameter $\gamma \in [0, 1]$, and in each column of $\boldsymbol{F}$, k entries are sampled uniformly without replacement. $\boldsymbol{G}$ is not known at all for this experiment.

Three proteins were investigated in this experiment, identified as `1PTQ`, `1AX8`, and `1UBQ`. These proteins were downloaded from the Protein Data Bank [84]. For all three proteins, we select $m = 20$ anchors and space them uniformly throughout $[1, n]$ where n is the number of atoms in the protein. In practice, domain knowledge allows for better anchor selection which can improve algorithmic performance. For each of these proteins, we set the column sample number $k \in [6, 9]$, and we select the $\boldsymbol{E}$ block Bernoulli rate $\gamma \in [.1, .5]$. We additionally run this experiment with an assumed ground truth rank for $\boldsymbol{X}$ of both 3 and 4, as overparameterization has been shown previously to improve numerical performance for sensor network localization problems [85, 86].

We first show a table indicating that increasing the rank of $\mathcal{M}_r$ from 3 to 4 improves the numerical performance in this structured sampling setting using Algorithm 1. For this experiment, we will look at the protein `1PTQ` $(n = 402)$, and we fix the $\boldsymbol{E}$ block rate $\gamma = .3$, seen in Table 3. This experiment and all others following were averaged over 25 trials, each lasting for 10000 iterations or until a relative difference in Frobenius norm between iterates of 10^{-5} was achieved.

Table 3: RMSE between $\boldsymbol{P}_{\text{rev}}$ and $\boldsymbol{P}$ averaged over 25 trials for the protein `1PTQ` using Algorithm 1, each trial run for 10000 iterations or until a 10^{-5} relative difference in Frobenius norm is achieved.

Column samples	E block rate	Rank	RMSE	Column samples	E block rate	Rank	RMSE
6	.3	3	2.56	6	.3	4	0.324
7	.3	3	1.51	7	.3	4	0.211
8	.3	3	0.915	8	.3	4	0.134
9	.3	3	0.712	9	.3	4	0.102

This experiment is in line with existing literature on overparameterization aiding reconstruction, as this provides clear indication that the reconstruction of the ground truth improves with higher rank. As such, we will set the rank of $\mathcal{M}_r$ to 4 for the remainder of the experiments. Next, we test to see if the $\boldsymbol{E}$ block rate parameter γ exhibits a substantial performance effect on the final RMSE, seen in Table 4.

Table 4: RMSE between $\boldsymbol{P}_{\text{rev}}$ and $\boldsymbol{P}$ averaged over 25 trials for the protein `1PTQ` using Algorithm 1, each trial run for 10000 iterations or until a 10^{-5} relative difference in Frobenius norm is achieved.

Column samples	**_E_** block rate	RMSE	Column samples	**_E_** block rate	RMSE
6	.1	0.347	8	.1	0.142
6	.2	0.379	8	.2	0.143
6	.3	0.324	8	.3	0.134
6	.4	0.326	8	.4	0.134
6	.5	0.329	8	.5	0.121
7	.1	0.209	9	.1	0.107
7	.2	0.191	9	.2	0.104
7	.3	0.211	9	.3	0.102
7	.4	0.195	9	.4	0.0954
7	.5	0.200	9	.5	0.0987

From the experiment in Table 4, increasing the $\boldsymbol{E}$ block rate does not greatly improve in the final RMSE following reconstruction. From a total number of samples perspective, this is not surprising, as for $m = 20$, the expected number of samples in the $\boldsymbol{E}$ block for $\gamma = 0.1$ is 38, and for $\gamma = 0.5$ the expected number is 190. Given $n = 402$ for this dataset, $L = 80601$, and the relative difference in total number of visible samples is less than two tenths of a percent. Since this parameter does not demonstrate a strong effect on convergence of Algorithm 1, we will now just show the remaining experiments for the proteins `1AX8` and `1UBQ` with the $\boldsymbol{E}$ block rate $\gamma = 0.3$, seen in Table 5.

Table 5: RMSE between $\boldsymbol{P}_{\text{rev}}$ and $\boldsymbol{P}$ averaged over 25 trials for the proteins `1PTQ`, `1AX8`, `1UBQ` using Algorithm 1, each trial run for 10000 iterations or until a 10^{-5} relative difference in Frobenius norm is achieved.

`1PTQ` $(n = 402)$		`1AX8` $(n = 1003)$		`1UBQ` $(n = 660)$	
Column Samples	RMSE	Column Samples	RMSE	Column Samples	RMSE
6	0.324	6	0.915	6	0.409
7	0.211	7	0.435	7	0.266
8	0.134	8	0.269	8	0.205
9	0.102	9	0.201	9	0.177

These experiments indicate strong reconstruction ability with Algorithm 1 in this structured sampling setting. The dependence on the number of column samples in RMSE following reconstruction is visible as well across the

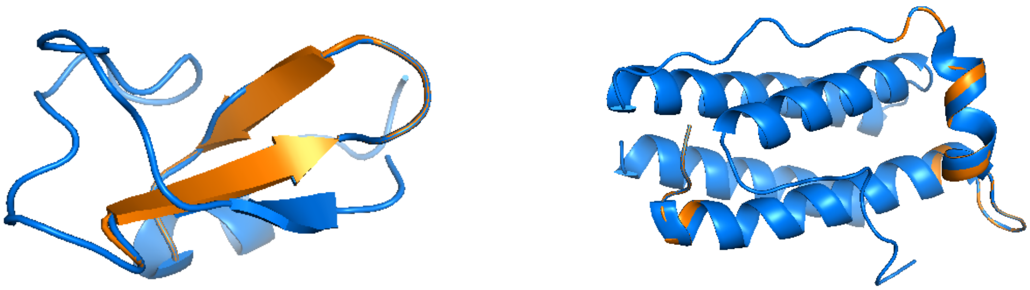

Figure 4: Target structure (in blue) and numerically estimated structure (in orange) following 100000 iterations of Algorithm 1. (Left) Target structure `1AX8`, $\gamma = 0.3$ and $k = 6$ (RMSE = 0.014). (Right): Target structure `1AX8`, $\gamma = 0.3$ and $k = 6$ (RMSE = 0.06).

experiments. This is expected from the perspective of total entries viewed in the underlying matrix, as the majority of the accessible connectivity information is stored in the $\boldsymbol{F}$ block in this sampling setup.

7 Conclusion and Future Work

In this work we proposed a novel approach for solving the EDG problem using a matrix completion approach on the manifold of rank-r matrices in Algorithm 2. We derived local linear convergence guarantees for this non-convex Riemannian gradient-like algorithm, and with this approach we provided two provably convergent initialization techniques when considering uniformly sampled distances. To the authors' knowledge, this is the first work to provide such initialization methods non-convex approaches to the EDG problem. The convergence analysis of this algorithm was predicated on understanding properties of a non-self-adjoint sampling operator, which required novel analysis of EDG-specific bases. We provided numerical results for this method to underline its efficacy in the high-sampling regime for the EDG problem. In addition to the provably convergent Algorithm 2, we provided an additional algorithm, Algorithm 1, that is a true first-order method on the manifold of rank-r matrices. This algorithm, although currently lacking in convergence guarantees, exhibited better numerical performance than the provably convergent one, performing nearly as well as some existing methods. Finally, we numerically investigated a structured sampling method relevant to the sensor network localization and protein structure problems, and studied how Algorithm 1 performed numerically in this setting on real-world data. We showed that it exhibited strong reconstruction performance in this new sampling framework, opening the door to future investigation.

One future goal will be a full characterization of the convergence of Algorithm 1, as this remains an open question. This will be important to investigate due to its stronger numerical performance. We are also interested in reconstruction of matrices expanded in more general non-orthogonal bases, and developing guarantees based on linear-algebraic properties similar to those investigated in this work. Additionally, this work relied on a uniform sampling with replacement model. Oftentimes, real world models for EDG or sensor network localization rely on different sampling models, such as nearest neighbor sampling. We are interested in seeing how we can extend this work and gain theoretical guarantees in the direction of non-uniform sampling models, alongside motivating other algorithmic developments.

8 Acknowledgment

HanQin Cai acknowledges partial support from the National Science Foundation through grant DMS-2304489. Abiy Tasissa and Chandler Smith acknowledge partial support from the National Science Foundation through grant DMS-2208392.

References

[1] M. Aldibaja, N. Suganuma, and K. Yoneda, "Improving localization accuracy for autonomous driving in snow-rain environments," in *2016 IEEE/SICE International Symposium on System Integration (SII)*. IEEE, 2016, pp. 212–217.

[2] J. V. Marti, J. Sales, R. Marin, and P. Sanz, “Multi-sensor localization and navigation for remote manipulation in smoky areas,” *International Journal of Advanced Robotic Systems*, vol. 10, no. 4, p. 211, 2013.

[3] G. M. Clore, M. A. Robien, and A. M. Gronenborn, “Exploring the limits of precision and accuracy of protein structures determined by nuclear magnetic resonance spectroscopy,” *Journal of molecular biology*, vol. 231, no. 1, pp. 82–102, 1993.

[4] A. Boukerche, H. A. Oliveira, E. F. Nakamura, and A. A. Loureiro, “Localization systems for wireless sensor networks,” *IEEE wireless Communications*, vol. 14, no. 6, pp. 6–12, 2007.

[5] J. Kuriakose, S. Joshi, R. Vikram Raju, and A. Kilaru, “A review on localization in wireless sensor networks,” *Advances in signal processing and intelligent recognition systems*, pp. 599–610, 2014.

[6] P. Biswas, T.-C. Lian, T.-C. Wang, and Y. Ye, “Semidefinite programming based algorithms for sensor network localization,” *ACM Transactions on Sensor Networks (TOSN)*, vol. 2, no. 2, pp. 188–220, 2006.

[7] Y. Ding, N. Krislock, J. Qian, and H. Wolkowicz, “Sensor network localization, euclidean distance matrix completions, and graph realization,” *Optimization and Engineering*, vol. 11, no. 1, pp. 45–66, 2010.

[8] N. Rojas, “Distance-based formulations for the position analysis of kinematic chains,” Ph.D. dissertation, Universitat Politècnica de Catalunya, 2012.

[9] J. M. Porta, N. Rojas, and F. Thomas, “Distance geometry in active structures,” *Mechatronics for Cultural Heritage and Civil Engineering*, pp. 115–136, 2018.

[10] J. B. Tenenbaum, V. De Silva, and J. C. Langford, “A global geometric framework for nonlinear dimensionality reduction,” *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[11] W. Glunt, T. Hayden, and M. Raydan, “Molecular conformations from distance matrices,” *Journal of Computational Chemistry*, vol. 14, no. 1, pp. 114–120, 1993.

[12] M. W. Trosset, “Applications of multidimensional scaling to molecular conformation,” 1997.

[13] X. Fang and K.-C. Toh, “Using a distributed sdp approach to solve simulated protein molecular conformation problems,” in *Distance Geometry.* Springer, 2013, pp. 351–376.

[14] L. Liberti, C. Lavor, and N. Maculan, “A branch-and-prune algorithm for the molecular distance geometry problem,” *International Transactions in Operational Research*, vol. 15, no. 1, pp. 1–17, 2008.

[15] T. Einav, Y. Khoo, and A. Singer, “Quantitatively visualizing bipartite datasets,” *Physical Review X*, vol. 13, no. 2, p. 021002, 2023.

[16] W. S. Torgerson, “Multidimensional scaling: I. theory and method,” *Psychometrika*, vol. 17, no. 4, pp. 401–419, 1952.

[17] G. Young and A. S. Householder, “Discussion of a set of points in terms of their mutual distances,” *Psychometrika*, vol. 3, no. 1, pp. 19–22, 1938.

[18] W. S. Torgerson, *Theory and methods of scaling.* Wiley, 1958.

[19] J. C. Gower, “Some distance properties of latent root and vector methods used in multivariate analysis,” *Biometrika*, vol. 53, no. 3-4, pp. 325–338, 1966.

[20] I. Dokmanic, R. Parhizkar, J. Ranieri, and M. Vetterli, “Euclidean distance matrices: essential theory, algorithms, and applications,” *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 12–30, 2015.

[21] N. Moreira, L. Duarte, C. Lavor, and C. Torezzan, “A novel low-rank matrix completion approach to estimate missing entries in euclidean distance matrices,” 2017.

[22] M. Fazel, H. Hindi, and S. P. Boyd, “A rank minimization heuristic with application to minimum order system approximation,” in *American Control Conference, 2001. Proceedings of the 2001*, vol. 6. IEEE, 2001, pp. 4734–4739.

[23] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE transactions on information theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[24] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on information theory*, vol. 52, no. 2, pp. 489–509, 2006.

[25] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[26] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM review*, vol. 52, no. 3, pp. 471–501, 2010.

[27] D. Gross and V. Nesme, "Note on sampling without replacing from a finite collection of matrices," *arXiv preprint arXiv:1001.2738*, 2010.

[28] A. Tasissa and R. Lai, "Exact reconstruction of euclidean distance geometry problem using low-rank matrix completion," *IEEE Transactions on Information Theory*, vol. 65, no. 5, pp. 3124–3144, 2018.

[29] R. Lai and J. Li, "Solving partial differential equations on manifolds from incomplete interpoint distance," *SIAM Journal on Scientific Computing*, vol. 39, no. 5, pp. A2231–A2256, 2017.

[30] Y. Li and X. Sun, "Sensor network localization via riemannian conjugate gradient and rank reduction," *IEEE Transactions on Signal Processing*, vol. 72, pp. 1910–1927, 2024.

[31] A. Tasissa and R. Lai, "Low-rank matrix completion in a general non-orthogonal basis," *Linear Algebra and its Applications*, vol. 625, pp. 81–112, 2021.

[32] L. T. Nguyen, J. Kim, S. Kim, and B. Shim, "Localization of iot networks via low-rank matrix completion," *IEEE Transactions on Communications*, vol. 67, no. 8, pp. 5833–5847, 2019.

[33] R. Bhatia, *Matrix Analysis*, ser. Graduate Texts in Mathematics. Springer New York, 2013. [Online]. Available: https://books.google.com/books?id=lh4BCAAAQBAJ

[34] B. Vandereycken, "Low-rank matrix completion by Riemannian optimization—extended version," 2012.

[35] N. Boumal, *An Introduction to Optimization on Smooth Manifolds.* Cambridge University Press, 3 2023.

[36] K. Wei, J.-F. Cai, T. F. Chan, and S. Leung, "Guarantees of riemannian optimization for low rank matrix completion." *Inverse Problems & Imaging*, vol. 14, no. 2, 2020.

[37] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds.* Princeton University Press, 2008. [Online]. Available: https://press.princeton.edu/absil

[38] U. Shalit, D. Weinshall, and G. Chechik, "Online learning in the embedded manifold of low-rank matrices," *J. Mach. Learn. Res.*, vol. 13, no. null, p. 429–458, feb 2012.

[39] K. Wei, J.-F. Cai, T. F. Chan, and S. Leung, "Guarantees of Riemannian optimization for low rank matrix recovery," *SIAM Journal on Matrix Analysis and Applications*, vol. 37, no. 3, pp. 1198–1222, 2016.

[40] H. Cai, J.-F. Cai, and K. Wei, "Accelerated alternating projections for robust principal component analysis," *Journal of Machine Learning Research*, vol. 20, no. 1, pp. 685–717, 2019.

[41] H. Cai, J.-F. Cai, T. Wang, and G. Yin, "Accelerated structured alternating projections for robust spectrally sparse signal recovery," *IEEE Transactions on Signal Processing*, vol. 69, pp. 809–821, 2021.

[42] K. Hamm, M. Meskini, and H. Cai, "Riemannian CUR decompositions for robust principal component analysis," in *Topological, Algebraic and Geometric Learning Workshops 2022.* PMLR, 2022, pp. 152–160.

[43] R. Meka, P. Jain, C. Caramanis, and I. S. Dhillon, "Rank minimization via online learning," in *Proceedings of the 25th International Conference on Machine learning*, 2008, pp. 656–663.

[44] D. Bertsimas, R. Cory-Wright, and J. Pauphilet, "Mixed-projection conic optimization: A new paradigm for modeling rank constraints," *Operations Research*, vol. 70, no. 6, pp. 3321–3344, 2022.

[45] B. Recht, "A simpler approach to matrix completion," *The Journal of Machine Learning Research*, vol. 12, pp. 3413–3430, 2011.

[46] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *Information Theory, IEEE Transactions on*, vol. 57, no. 3, pp. 1548–1566, 2011.

[47] S. Burer and R. D. Monteiro, "A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization," *Mathematical Programming*, vol. 95, no. 2, pp. 329–357, 2003.

[48] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, 2013, pp. 665–674.

[49] M. Hardt, "Understanding alternating minimization for matrix completion," *Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS*, pp. 651–660, 12 2014.

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

[51] S. J. Optim, Y. Chen, Y. Chi, J. Fan, and Y. Yan, "Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization," *SIAM J Optim*, vol. 30, pp. 3098–3121, 2020. [Online]. Available: https://doi.org/10.1137/19M1290000

[52] S. Lichtenberg and A. Tasissa, "A dual basis approach to multidimensional scaling: spectral analysis and graph regularity," 2023.

[53] C. Smith, S. Lichtenberg, H. Cai, and A. Tasissa, "Riemannian optimization for euclidean distance geometry," *OPT2023: 15th Annual Workshop on Optimization for Machine Learning*, 2023.

[54] B. Mishra, G. Meyer, S. Bonnabel, and R. Sepulchre, "Fixed-rank matrix factorizations and riemannian low-rank optimization," 2013.

[55] N. Boumal and P.-A. Absil, "Low-rank matrix completion via preconditioned optimization on the grassmann manifold," *Absil / Linear Algebra and its Applications*, vol. 475, p. 201, 2015. [Online]. Available: www.elsevier.com/locate/laahttp://dx.doi.org/10.1016/j.laa.2015.02.0270024-3795/

[56] W. Dai and O. Milenkovic, "Set: an algorithm for consistent matrix completion," 2010. [Online]. Available: https://arxiv.org/abs/0909.2705

[57] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Transactions on Information Theory*, vol. 56, no. 6, pp. 2980–2998, 2010.

[58] ——, "Matrix completion from noisy entries," *Journal of Machine Learning Research*, vol. 11, no. Jul, pp. 2057–2078, 2010.

[59] A. Y. Alfakih, A. Khandani, and H. Wolkowicz, "Solving euclidean distance matrix completion problems via semidefinite programming," *Computational optimization and applications*, vol. 12, no. 1-3, pp. 13–30, 1999.

[60] P. Biswas and Y. Ye, "Semidefinite programming for ad hoc wireless sensor network localization," in *Proceedings of the 3rd international symposium on Information processing in sensor networks*, 2004, pp. 46–54.

[61] P. Biswas, K.-C. Toh, and Y. Ye, "A distributed sdp approach for large-scale noisy anchor-free graph realization with applications to molecular conformation," *SIAM Journal on Scientific Computing*, vol. 30, no. 3, pp. 1251–1277, 2008.

[62] N.-H. Z. Leung and K.-C. Toh, "An sdp-based divide-and-conquer algorithm for large-scale noisy anchor-free graph realization," *SIAM Journal on Scientific Computing*, vol. 31, no. 6, pp. 4351–4372, 2010.

[63] B. Alipanahi, N. Krislock, A. Ghodsi, H. Wolkowicz, L. Donaldson, and M. Li, "Protein structure by semidefinite facial reduction," in *Research in Computational Molecular Biology: 16th Annual International Conference, RECOMB 2012, Barcelona, Spain, April 21-24, 2012. Proceedings 16.* Springer, 2012, pp. 1–11.

[64] S. Guo, H.-D. Qi, and L. Zhang, "Perturbation analysis of the euclidean distance matrix optimization problem and its numerical implications," *Computational Optimization and Applications*, vol. 86, no. 3, pp. 1193–1227, 2023.

[65] T. F. Havel, “An evaluation of computational strategies for use in the determination of protein structure from distance constraints obtained by nuclear magnetic resonance,” *Progress in biophysics and molecular biology*, vol. 56, no. 1, pp. 43–78, 1991.

[66] J. J. Moré and Z. Wu, “Distance geometry optimization for protein structures,” *Journal of Global Optimization*, vol. 15, pp. 219–234, 1999.

[67] G. M. Crippen, T. F. Havel *et al.*, *Distance geometry and molecular conformation.* Research Studies Press Taunton, 1988, vol. 74.

[68] T. F. Havel, “Distance geometry: Theory, algorithms, and chemical applications,” *Encyclopedia of Computational Chemistry*, vol. 120, pp. 723–742, 1998.

[69] C. Lavor, L. Liberti, N. Maculan, and A. Mucherino, “Recent advances on the discretizable molecular distance geometry problem,” *European Journal of Operational Research*, vol. 219, no. 3, pp. 698–706, 2012.

[70] ——, “The discretizable molecular distance geometry problem,” *Computational Optimization and Applications*, vol. 52, pp. 115–146, 2012.

[71] D. Wu and Z. Wu, “An updated geometric build-up algorithm for solving the molecular distance geometry problems with sparse distance data,” *Journal of Global Optimization*, vol. 37, pp. 661–673, 2007.

[72] Q. Dong and Z. Wu, “A geometric build-up algorithm for solving the molecular distance geometry problem with sparse distance data,” *Journal of Global Optimization*, vol. 26, pp. 321–333, 2003.

[73] A. Sit, Z. Wu, and Y. Yuan, “A geometric buildup algorithm for the solution of the distance geometry problem using least-squares approximation,” *Bulletin of mathematical biology*, vol. 71, no. 8, pp. 1914–1933, 2009.

[74] B. Hendrickson, “The molecule problem: Exploiting structure in global optimization,” *SIAM Journal on Optimization*, vol. 5, no. 4, pp. 835–857, 1995.

[75] D. LEEUW, “Application of convex analysis to multidimensional scaling,” *Recent developments in statistics*, pp. 133–145, 1977.

[76] J. J. Moré and Z. Wu, “Global continuation for distance geometry problems,” *SIAM Journal on Optimization*, vol. 7, no. 3, pp. 814–836, 1997.

[77] H.-r. Fang and D. P. O’Leary, “Euclidean distance matrix completion problems,” *Optimization Methods and Software*, vol. 27, no. 4-5, pp. 695–717, 2012.

[78] L. Liberti, C. Lavor, N. Maculan, and A. Mucherino, “Euclidean distance geometry and applications,” *SIAM review*, vol. 56, no. 1, pp. 3–69, 2014.

[79] A. Javanmard and A. Montanari, “Localization from incomplete noisy distance measurements,” *Foundations of Computational Mathematics*, vol. 13, no. 3, p. 297–345, Jul. 2012. [Online]. Available: http://dx.doi.org/10.1007/s10208-012-9129-5

[80] R. Parhizkar, A. Karbasi, S. Oh, and M. Vetterli, “Calibration using matrix completion with application to ultrasound tomography,” *IEEE Transactions on Signal Processing*, vol. 61, no. 20, pp. 4923–4933, 2013.

[81] A. Tasissa and R. Lai, “Low-rank matrix completion in a general non-orthogonal basis,” *Linear Algebra and its Applications*, vol. 625, pp. 81–112, 2021. [Online]. Available: www.elsevier.com/locate/laa

[82] C. Eckart and G. Young, “The approximation of one matrix by another of lower rank,” *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

[83] S. Lichtenberg and A. Tasissa, “Localization from structured distance matrices via low-rank matrix recovery,” *IEEE Transactions on Information Theory*, pp. 1–1, 2024.

[84] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, “The Protein Data Bank,” *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 01 2000. [Online]. Available: https://doi.org/10.1093/nar/28.1.235

[85] T. Tang, K.-C. Toh, N. Xiao, and Y. Ye, "A riemannian dimension-reduced second order method with application in sensor network localization," 2023. [Online]. Available: https://arxiv.org/abs/2304.10092

[86] M. Lei, J. Zhang, and Y. Ye, "Blessing of high-order dimensionality: from non-convex to convex optimization for sensor network localization," 2023. [Online]. Available: https://arxiv.org/abs/2308.02278

[87] K. Wei, J.-F. Cai, T. F. Chan, and S. Leung, "Guarantees of riemannian optimization for low rank matrix recovery," *SIAM Journal on Matrix Analysis and Applications*, vol. 37, no. 3, pp. 1198–1222, 2016. [Online]. Available: https://doi.org/10.1137/15M1050525

A Properties of the dual bases and Non-Commutative Bernstein Inequality

This section of the appendix details technical results about the specific dual bases, $\{\boldsymbol{w}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$ and $\{\boldsymbol{v}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$. These are needed to prove various technical lemmas throughout the work, but are particularly important in the proof of Theorem 5.4. Additionally, we provide a variant of the non-commutative Bernstein inequality leveraged throughout this work.

Theorem A.1 (Operator Bernstein Inequality [45]). *Let $\boldsymbol{X}_i$, $i = 1, ..., m$ be i.i.d, zero-mean, matrix-valued random variables, and let $\rho_i^2 \geq \max\{\mathbb{E}(\boldsymbol{X}_i\boldsymbol{X}_i^\star), \mathbb{E}(\boldsymbol{X}_i^\star\boldsymbol{X}_i))\}$. Assume there exists a $c \in \mathbb{R}$ such that $\|\boldsymbol{X}_i\| \leq c$ almost surely. Then for $t < \sum_{i=1}^m \frac{\rho_i}{c}$,*

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

If we assume that $\rho_1^2 = ... = \rho_m^2 = V_0$ and let $V = mV_0$, then for $t < \frac{V}{c}$ this simplifies to

$$\mathbb{P}\left(\left\|\sum_{i=1}^m \boldsymbol{X}_i\right\| > t\right) \leq 2n\exp\left(-\frac{3t^2}{8V}\right). \tag{22}$$

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 A.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, and let $\boldsymbol{G} = [\langle \boldsymbol{Z}_i, \boldsymbol{Z}_j\rangle] \in \mathbb{R}^{m\times m}$, and let $\boldsymbol{Z}_\mathbb{V} \in \mathbb{R}^{n^2\times m}$ be the matrix where the k-th column vector is 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 of Lemma A.2. We can see that

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

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 A.3 (Spectral norm of $\mathcal{R}_\Omega$)**.** *For $m \geq \frac{8}{3}\beta n \log(n)$ and with probability at least $1 - 2n^{1-\beta}$,*

$$\|\mathcal{R}_\Omega\| \leq \frac{m}{L} + 4\sqrt{\frac{8m\log(n)}{3n}}.$$

Proof of Lemma A.3. Compared to analogous sampling operators in matrix completion, $\mathcal{R}_\Omega$ is not self-adjoint. As such, it cannot be decomposed into a sum of orthogonal projection operators. This means that the operator norm $\|\mathcal{R}_\Omega\|$ cannot be bounded via a counting argument like in [45], as that would produce an upper bound for the maximum eigenvalue but not the maximum singular value. As such, we will proceed by using Theorem A.1 to prove a bound for $\left\|\mathcal{R}_\Omega - \frac{m}{L}\mathcal{I}\right\|$. To do so, let

$$\mathcal{T}_{\boldsymbol{\alpha}} = \langle \cdot, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \boldsymbol{v}_{\boldsymbol{\alpha}} - \frac{1}{L}\mathcal{I}.$$

This object is zero-mean, and $\mathcal{R}_\Omega - \frac{m}{L}\mathcal{I} = \sum_{\boldsymbol{\alpha}\in\Omega} \mathcal{T}_{\boldsymbol{\alpha}}$. We now need bounds on $\|\mathcal{T}_{\boldsymbol{\alpha}}\|$, $\|\mathbb{E}[\mathcal{T}_{\boldsymbol{\alpha}}\mathcal{T}_{\boldsymbol{\alpha}}^\star]\|$, and $\|\mathbb{E}[\mathcal{T}_{\boldsymbol{\alpha}}^\star\mathcal{T}_{\boldsymbol{\alpha}}]\|$.

For the first, notice that

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

where the third to last inequality follows from Lemma A.6 and the fact that $\|\boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} = 2$. Next, notice that

$$\mathbb{E}[\mathcal{T}_\alpha^\star \mathcal{T}_\alpha] = \frac{1}{L}\sum_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \cdot, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\alpha}} \rangle \boldsymbol{w}_{\boldsymbol{\alpha}} - \frac{1}{L^2}\mathcal{I}, \qquad \mathbb{E}[\mathcal{T}_\alpha \mathcal{T}_\alpha^\star] = \frac{1}{L}\sum_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \cdot, \boldsymbol{v}_{\boldsymbol{\alpha}} \rangle \langle \boldsymbol{w}_{\boldsymbol{\alpha}}, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \boldsymbol{v}_{\boldsymbol{\alpha}} - \frac{1}{L^2}\mathcal{I}.$$

Now, notice that

$$\begin{aligned}
\|\mathbb{E}[\mathcal{T}_\alpha \mathcal{T}_\alpha^\star]\| &= \left\| \frac{1}{L}\sum_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \cdot, \boldsymbol{v}_{\boldsymbol{\alpha}} \rangle \langle \boldsymbol{w}_{\boldsymbol{\alpha}}, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \boldsymbol{v}_{\boldsymbol{\alpha}} - \frac{1}{L^2}\mathcal{I} \right\| \\
&\leq \frac{1}{L} \max_{\|\boldsymbol{X}\|_{\mathrm{F}}=1} \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \boldsymbol{X}, \boldsymbol{v}_{\boldsymbol{\alpha}} \rangle^2 \langle \boldsymbol{w}_{\boldsymbol{\alpha}}, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle + \frac{1}{L^2} \\
&\leq \frac{4}{L} \max_{\|\boldsymbol{X}\|_{\mathrm{F}}=1} \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \boldsymbol{X}, \boldsymbol{v}_{\boldsymbol{\alpha}} \rangle^2 + \frac{1}{L^2} \\
&\leq \frac{4}{L}\lambda_{\max}(\boldsymbol{H}^{-1}) + \frac{1}{L^2} \\
&\leq \frac{4}{L},
\end{aligned}$$

where the first inequality follows from the triangle inequality, the second comes from $\|\boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} = 2$ in Lemma A.6, the third is an application of Lemma A.2, and the last comes from the fact that $\lambda_{\max}(\boldsymbol{H}^{-1}) = \frac{1}{2}$ from Lemma A.6. Next, we can see that

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

where the first inequality follows from the triangle inequality, the second comes from $\|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \leq \frac{1}{\sqrt{2}}$ in Lemma A.6, the third is an application of Lemma A.2, and the last comes from the fact that$\lambda_{\max}(\boldsymbol{H}) = 2n$ from Lemma A.6. As such, our variance estimate $V_0 = \frac{2n}{L}$. It follows that for any $t < \frac{mV_0}{c} = \frac{mn}{L}$, we have the following result from Theorem A.1:

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

and the proof statement follows from this. □

Lemma A.4 ($\lambda_{\max}(\tilde{\boldsymbol{H}})$ 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 of Lemma A.4 . First, by coherence we have that

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

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

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

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 coherence condition above. □

Lemma A.5. *For any $\boldsymbol{X} \in \mathbb{R}^{n\times n}$, $\boldsymbol{X} = \boldsymbol{X}^\top$, and any $\boldsymbol{w}_{\boldsymbol{\alpha}} \in \{\boldsymbol{w}_{\boldsymbol{\beta}}\}_{\boldsymbol{\beta}\in\mathbb{I}}$,*

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

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

$$\sum_{\boldsymbol{\alpha}\in\mathbb{I}}\langle \boldsymbol{X}\mathcal{P}_U, \mathcal{P}_U\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \leq \max_{\|\boldsymbol{X}\|_{\mathrm{F}}=1}\sum_{\boldsymbol{\alpha}\in\mathbb{I}}\langle \boldsymbol{X}\mathcal{P}_U, \mathcal{P}_U\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \leq \lambda_{\max}(\tilde{\boldsymbol{H}}).$$

Proof of Lemma A.5. 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

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

The second statement follows from Lemma A.2 and the fact that $\mathcal{P}_U$ is an orthogonal projection operator. This concludes the proof. □

Lemma A.6 (Eigenvalues of $\boldsymbol{H}$ and $\boldsymbol{H}^{-1}$, entries of $\boldsymbol{H}^{-1}$, and spectral norms of $\boldsymbol{w}_{\boldsymbol{\alpha}}$ and $\boldsymbol{v}_{\boldsymbol{\alpha}}$ [52])**.** *Let $\boldsymbol{H} = [\boldsymbol{w}_{\boldsymbol{\alpha}}, \boldsymbol{w}_{\boldsymbol{\beta}}] \in \mathbb{R}^{L\times L}$ be the Gram matrix for $\{\boldsymbol{w}_{\boldsymbol{\alpha}}\}$, and let $\boldsymbol{H}^{-1}$ be its inverse. Then*

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

Additionally,

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

Finally,

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

Lemma A.7. *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 of Lemma A.7. Recall that $\boldsymbol{v}_{\boldsymbol{\alpha}} = -\frac{1}{2}\left(\boldsymbol{a}\boldsymbol{b}^\top + \boldsymbol{b}\boldsymbol{a}^\top\right)$ where $\boldsymbol{a} = \boldsymbol{e}_i - \frac{1}{n}\boldsymbol{1}$ and $\boldsymbol{b} = \boldsymbol{e}_j - \frac{1}{n}\boldsymbol{1}$ 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

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

So it follows that

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

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

B Restricted Isometry Results

As RIP and its variants are critical to the analysis of Algorithm 2, this section is dedicated to the proofs of RIP and similar results.

B.1 Proof of Theorem 5.4

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

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

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 A.1 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 5.1,

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

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

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

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

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

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

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

where the first inequality follows from the triangle inequality, the second comes from the coherence conditions in Assumption 5.1, 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 A.2, and the sixth comes from the $\lambda_{\max}(\boldsymbol{H}^{-1})$ bound from Lemma A.6. Next for $\sum_{\boldsymbol{\alpha}\in\mathbb{I}}\mathcal{T}_{\boldsymbol{\alpha}}^{\star}\mathcal{T}_{\boldsymbol{\alpha}}$, we have that

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

where the first inequality follows from the triangle inequality, the second comes from the coherence conditions in Assumption 5.1, 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 A.5, the fourth comes from the definition of the max, the fifth comes from an application of Lemma A.2, and the sixth comes from the bound on $\lambda_{\max}(\tilde{\boldsymbol{H}})$ in Lemma A.4.

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 (22). 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 $\mathbb{E}(\mathcal{R}_{\Omega}^{\star}) = \frac{m}{L}\mathcal{I}$, the same proof can be repeated to show RIP for $\mathcal{R}_{\Omega}^{\star}$ 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 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 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 5.4 and Lemmas B.4 and 5.6, we will use Theorem A.1 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

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

where the first inequality follows from Assumption 5.1, the triangle inequality, and Lemma A.2. For the other term, we see that

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

where the third inequality follows from $\|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} < 1$, the fourth inequality follows from Lemma A.5, and fifth inequality from Lemma A.4. 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 A.1 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 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 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 E.1 and the assumptions of this lemma. This concludes the proof. □

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

$$\left\|\frac{L}{m}\mathcal{P}_{\mathbb{T}}\mathcal{R}_\Omega\mathcal{P}_{\mathbb{T}} - \mathcal{P}_{\mathbb{T}}\right\| \leq \varepsilon_0 < 1, \tag{23}$$

$$\frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\sigma_{min}(\boldsymbol{X})} \leq \frac{\sqrt{m}\varepsilon_0}{16n^{5/4}\sqrt{\beta\nu r\log n}}, \quad \|\mathcal{R}_\Omega\| \leq \frac{m}{L} + 4\sqrt{\frac{8m\log(n)}{n}}, \tag{24}$$

$$\|\mathcal{R}_\Omega\mathcal{P}_{\mathbb{T}}\| \leq \frac{m}{L} + \frac{m\sqrt{n}}{L}\sqrt{\frac{8\beta\nu r n\log(n)}{3m}}, \quad \text{and} \quad \|\mathcal{P}_{\mathbb{T}}\mathcal{R}_\Omega\| \leq \frac{m}{L} + \frac{4m\sqrt{n}}{L}\sqrt{\frac{\beta\nu r n\log(n)}{3m}}. \tag{25}$$

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

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

using the triangle inequality and the results gathered in Lemma E.1.

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

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

where the first inequality comes from the assumption on $\|\mathcal{R}_\Omega\mathcal{P}_{\mathbb{T}}\|$ in (25), the second inequality comes from the local neighborhood assumption in (24), the third inequality comes from term cancellation and the fact that $\beta, \nu, r, \log(n) \ge 1$, the fourth inequality comes from the fact that $m \le L$, 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 B.2 are satisfied, so

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

where the first inequality follows from the assumptions on $\|\mathcal{P}_{\mathbb{T}}\mathcal{R}_{\Omega}\|$ in (25), 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 (24), 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 (23), so $\varepsilon_0^2 \leq \varepsilon_0$. The desired statement follows from here, thus concluding the proof. □

For Algorithm 3, we will need what the authors of [36] call an asymmetric form of RIP for $\mathcal{R}_{\Omega}$. The statement and proof of this are below:

Lemma 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:

$$\left\| \frac{L}{m}\mathcal{P}_{\mathbb{T}_l}\mathcal{R}_{\Omega}(\mathcal{P}_U - \mathcal{P}_{U_l}) - \mathcal{P}_{\mathbb{T}_l}(\mathcal{P}_U - \mathcal{P}_{U_l}) \right\| \leq \sqrt{\frac{4\beta\nu^2 r^2 n \log(n)}{3m}}. \tag{26}$$

Proof of Lemma 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 $\mathcal{K}_{\boldsymbol{\alpha}}^{\star} = (\mathcal{P}_U - \mathcal{P}_{U_l})(\boldsymbol{w}_{\boldsymbol{\alpha}}) \otimes \mathcal{P}_{\mathbb{T}_l}(\boldsymbol{v}_{\boldsymbol{\alpha}})(\cdot) = \langle \cdot, \mathcal{P}_{\mathbb{T}_l}\boldsymbol{v}_{\boldsymbol{\alpha}}\rangle(\mathcal{P}_U - \mathcal{P}_{U_l})\boldsymbol{w}_{\boldsymbol{\alpha}}$. 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 $\mathbb{E}[\mathcal{K}_{\boldsymbol{\alpha}}^{\star}] = \frac{1}{L}(\mathcal{P}_U - \mathcal{P}_{U_l})\mathcal{P}_{\mathbb{T}_l}$. We note that the variance term $\mathbb{E}\left[\left(\mathcal{K}_{\boldsymbol{\alpha}} - \frac{1}{L}\mathcal{P}_{\mathbb{T}_l}(\mathcal{P}_U - \mathcal{P}_{U_l})\right)\left(\mathcal{K}_{\boldsymbol{\alpha}} - \frac{1}{L}\mathcal{P}_{\mathbb{T}_l}(\mathcal{P}_U - \mathcal{P}_{U_l})\right)^{\star}\right]$ can be bounded as follows:

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

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

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

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

For the first term, notice that

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

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

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

As such,

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

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 $\boldsymbol{\alpha}\cap\boldsymbol{\beta} = \emptyset$,

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

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

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

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

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

This gives a bound of

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

Next, we can see that

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

Now, it follows that

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

so $V_0 := \frac{4\nu^2 r^2}{nL}$. Now, for $t < \frac{mV_0}{c} = \frac{2m}{L}$, the result follows from Theorem A.1 with $t = \sqrt{\frac{4\beta\nu^2 r^2 n\log(n)}{3m}}$ with $m \geq \frac{4}{3}\nu^2 r^2 n \log(n)$. □

C Proof of Local Convergence (Theorem 5.5)

In this section, we will use the properties proven thus far to provide proof of local convergence of Algorithm 2.

We begin with the following technical lemmas:

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

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

Proof of Lemma 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 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 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 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*

$$\|\mathcal{P}_{\mathbb{T}_l} - \alpha_l\mathcal{P}_{\mathbb{T}_l}\mathcal{R}_\Omega\mathcal{P}_{\mathbb{T}_l}\| \leq \frac{8\varepsilon_0}{1-4\varepsilon_0}. \tag{27}$$

Proof of Lemma 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 5.5.

C.1 Proof of Theorem 5.5

Proof. First, it follows that

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

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

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

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

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

Next, notice that from Lemma E.1 and the fact that $\mathcal{P}_{\mathbb{T}_l}\boldsymbol{X}_l = \boldsymbol{X}_l$,

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

using Lemma E.1 and our initial assumption. Finally, we see that, following a similar argument as in the bound of I_2 and using Lemma B.2,

$$\begin{aligned}I_3 &\le 2|\alpha_l|\|\mathcal{P}_{\mathbb{T}_l}\mathcal{R}_\Omega\|\|(I - \mathcal{P}_{\mathbb{T}_l})\boldsymbol{X}\|_{\mathrm{F}} \\ &\le \frac{2L/m}{1 - 4\varepsilon_0}\left[\left(\frac{m}{L} + 4\sqrt{\frac{8m\log(n)}{n}}\right)\frac{2\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\sigma_{\min}(\boldsymbol{X})} + \frac{m}{L} + \frac{4m\sqrt{n}}{L}\sqrt{\frac{\beta\nu r n\log(n)}{3m}}\right]\left(\frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\sigma_{\min}(\boldsymbol{X})}\right)\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}} \\ &\le \frac{1}{1 - 4\varepsilon_0}\left(\frac{\varepsilon_0^2}{128} + \frac{\varepsilon_0}{16} + \frac{\varepsilon_0^2}{32} + \frac{\varepsilon_0}{\sqrt{48}}\right)\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}} \\ &\le \frac{\varepsilon_0}{1 - 4\varepsilon_0},\end{aligned}$$

where the second to last inequality follows from the same analysis conducted in Lemma B.3, just divided by 2. Collecting these results, we get

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

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 $l \ge 0$. This concludes the proof. □

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.

D.1 Proof of Lemma 5.6

We first start with a proof of the guarantee provided by one-step hard thresholding, detailed in Lemma 5.6:

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 [82]. 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

$$\begin{aligned}\|\boldsymbol{Z}_{\boldsymbol{\alpha}}\| &= \left\|\frac{L}{m}\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \boldsymbol{v}_{\boldsymbol{\alpha}} - \frac{1}{m}\boldsymbol{X}\right\| \\ &\leq \frac{L}{m}|\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle|\,\|\boldsymbol{v}_{\boldsymbol{\alpha}}\| + \frac{1}{m}\,\|\boldsymbol{X}\| \\ &\leq \frac{4L}{m}\,\|\boldsymbol{X}\|_\infty + \frac{n}{m}\,\|\boldsymbol{X}\|_\infty \\ &\leq \frac{5L}{m}\,\|\boldsymbol{X}\|_\infty =: c,\end{aligned}$$

as $\|\boldsymbol{v}_{\boldsymbol{\alpha}}\| < 1$ from Lemma A.6. 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 $\left\|\mathbb{E}[\boldsymbol{Z}_{\boldsymbol{\alpha}}^2]\right\| \leq \max\{\left\|\frac{L}{m^2}\sum_{\boldsymbol{\alpha}}\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \boldsymbol{v}_{\boldsymbol{\alpha}}^2\right\|, \left\|\frac{1}{m^2}\boldsymbol{X}^2\right\|\}$. It follows that

$$\begin{aligned}\left\|\frac{L}{m^2}\sum_{\boldsymbol{\alpha}}\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \boldsymbol{v}_{\boldsymbol{\alpha}}^2\right\| &= \max_{\boldsymbol{y}\in\mathbb{R}^n, \|\boldsymbol{y}\|_2=1} \frac{L}{m^2}\sum_{\boldsymbol{\alpha}}\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \boldsymbol{y}^\top \boldsymbol{v}_{\boldsymbol{\alpha}}^2 \boldsymbol{y} \\ &\leq \frac{16L}{m^2}\,\|\boldsymbol{X}\|_\infty^2\, \boldsymbol{y}^\top \left(\sum_{\boldsymbol{\alpha}} \boldsymbol{v}_{\boldsymbol{\alpha}}^2\right)\boldsymbol{y} \\ &= \frac{16L}{m^2}\,\|\boldsymbol{X}\|_\infty^2\, \lambda_{\max}\left(\sum_{\boldsymbol{\alpha}} \boldsymbol{v}_{\boldsymbol{\alpha}}^2\right).\end{aligned}$$

Now, from Lemma A.7, $\sum_{\boldsymbol{\alpha}} \boldsymbol{v}_{\boldsymbol{\alpha}}^2 = \frac{n^2-2n+2}{4n}\boldsymbol{J}$. 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,

$$\left\|\mathbb{E}\left(\boldsymbol{Z}_{\boldsymbol{\alpha}}^2\right)\right\| \leq \frac{5nL}{m^2}\,\|\boldsymbol{X}\|_\infty^2 =: V_0.$$

Now to determine t, we note that

$$\begin{aligned}\frac{V}{c} &= \frac{5nL}{m}\,\|\boldsymbol{X}\|_\infty^2\,\frac{m}{5L\,\|\boldsymbol{X}\|_\infty} \\ &= n\,\|\boldsymbol{X}\|_\infty \\ &\geq \sqrt{\frac{40\beta n^3 \log n}{3m}}\|\boldsymbol{X}\|_\infty,\end{aligned}$$

for $m \geq \frac{40}{3}\beta n \log n$. It follows that

$$\begin{aligned}\mathbb{P}\left(\|\boldsymbol{X}_0 - \boldsymbol{X}\| > \sqrt{\frac{40\beta n^3 \log n}{3m}}\|\boldsymbol{X}\|_\infty\right) &\leq 2n\exp\left(-\beta\log(n)\right) \\ &= 2n^{1-\beta},\end{aligned}$$

verifying the probabilistic bound. To complete the proof we use Assumption 5.2, and it follows that

$$\|\boldsymbol{X}_0 - \boldsymbol{X}\|_{\mathrm{F}} \leq \sqrt{2r}\|\boldsymbol{X}_0 - \boldsymbol{X}\| \leq \sqrt{\frac{320r\beta n^3\log(n)}{3m}}\|\boldsymbol{X}\|_\infty \leq \sqrt{\frac{320r^2\mu_1^2 n\log(n)}{3m}}\|\boldsymbol{X}\|,$$

thus concluding the proof. □

Now, we will prove a technical lemma about Algorithm 4:

Lemma 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 4 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 D.1. The proof of the first and fourth statements can be found in [36]. To see the second and third statements, we can apply the same analysis as in Remark 2. 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 A.6, it follows that $\sum_{\boldsymbol{\alpha}\in\mathbb{I}}|H^{\boldsymbol{\alpha}\boldsymbol{\beta}}| \leq 2$, so the last statement follows. □

D.2 Proof of Lemma 5.8

Proof. First, assume that at the l-th iteration of Algorithm 3,

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

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 5.5, we can decompose the error at the $l+1$-th iteration as follows:

$$\begin{aligned}
\|\boldsymbol{Z}_{l+1} - \boldsymbol{X}\|_{\mathrm{F}} \leq{} & \underbrace{2\left\|(\mathcal{P}_{\hat{\mathbb{T}}_l} - \frac{L}{m}\mathcal{P}_{\hat{\mathbb{T}}_l}\mathcal{R}_{\Omega_{l+1}}\mathcal{P}_{\hat{\mathbb{T}}_l})(\hat{\boldsymbol{Z}}_l - \boldsymbol{X})\right\|_{\mathrm{F}}}_{I_4} \\
& + \underbrace{2\left\|(\mathcal{I} - \mathcal{P}_{\hat{\mathbb{T}}_l})(\hat{\boldsymbol{Z}}_l - \boldsymbol{X})\right\|_{\mathrm{F}}}_{I_5} \\
& + \underbrace{2\left\|\frac{L}{m}\mathcal{P}_{\hat{\mathbb{T}}_l}\mathcal{R}_{\Omega_{l+1}}(\mathcal{I} - \mathcal{P}_{\hat{\mathbb{T}}_l})(\hat{\boldsymbol{Z}}_l - \boldsymbol{X})\right\|_{\mathrm{F}}}_{I_6}.
\end{aligned}$$

Now, as $\hat{\boldsymbol{Z}}_l$ and Ω_{l+1} are independent, we can use Theorem 5.4 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 E.1, the second inequality comes from Lemma 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 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

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

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

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

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

$$\|\boldsymbol{Z}_{l+1} - \boldsymbol{X}\|_{\mathrm{F}} \leq \left(\frac{1}{2} + 64\kappa\sqrt{\frac{\beta\nu^2 r^2 n\log(n)}{\hat{m}}}\right)\|\boldsymbol{Z}_l - \boldsymbol{X}\|_{\mathrm{F}}, \tag{28}$$

with probability at least $1 - 4n^{1-\beta}$. It follows that (28) 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 5.6 for $\hat{m} \geq (2\times 10^5)\kappa^6 r^2\mu_1^2 n\log(n)$, and the result follows from here. No attempts were made to optimize the constants. □

E Miscellaneous Results

Lemma E.1 (Bounds for Projections)**.** *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 the rank-r matrix manifold 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*

$$\|\boldsymbol{U}_l\boldsymbol{U}_l^\top - \boldsymbol{U}\boldsymbol{U}^\top\| \leq \frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\sigma_{\min}(\boldsymbol{X})}, \qquad \|\boldsymbol{U}_l\boldsymbol{U}_l^\top - \boldsymbol{U}\boldsymbol{U}^\top\|_{\mathrm{F}} \leq \frac{\sqrt{2}\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\sigma_{\min}(\boldsymbol{X})}$$

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

Proof of Lemma E.1. See [36, 87]. □