Provable Non-Convex Euclidean Distance Matrix Completion: Geometry, Reconstruction, and Robustness

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.

Abstract

The problem of recovering the configuration of points from their partial pairwise distances, referred to as the Euclidean Distance Matrix Completion (EDMC) problem, arises in a broad range of applications, including sensor network localization, molecular conformation, and manifold learning. In this paper, we propose a Riemannian optimization framework for solving the EDMC problem by formulating it as a low-rank matrix completion task over the space of positive semi-definite Gram matrices. The available distance measurements are encoded as expansion coefficients in a non-orthogonal basis, and optimization over the Gram matrix implicitly enforces geometric consistency through nonnegativity and the triangle inequality, a structure inherited from classical multidimensional scaling. Under a Bernoulli sampling model for observed distances, we prove that Riemannian gradient descent on the manifold of rank-r matrices locally converges linearly with high probability when the sampling probability satisfies $p \geq \mathcal{O}(\nu^2 r^2 \log(n)/n)$, where ν is an EDMC-specific incoherence parameter. Furthermore, we provide an initialization candidate using a one-step hard thresholding procedure that yields convergence, provided the sampling probability satisfies $p \geq \mathcal{O}(\nu r^{3/2} \log^{3/4}(n)/n^{1/4})$. A key technical contribution of this work is the analysis of a symmetric linear operator arising from a dual basis expansion in the non-orthogonal basis, which requires a novel application of the Hanson-Wright inequality to establish an optimal restricted isometry property in the presence of coupled terms. Empirical evaluations on synthetic data demonstrate that our algorithm achieves competitive performance relative to state-of-the-art methods. Moreover, we provide a geometric interpretation of matrix incoherence tailored to the EDMC setting and provide robustness guarantees for our method.

Keywords. *Euclidean Distance Matrix, Matrix Completion, Riemannian Optimization*

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 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 $n \times r$ matrix $\boldsymbol{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_n]^\top$, which has the points as rows. 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 $\boldsymbol{X} = \boldsymbol{P}\boldsymbol{P}^\top$. 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.

---

*Corresponding Author, Chandler.Smith@Tufts.edu

However, the inverse problem of determining $\boldsymbol{P}$ from $\boldsymbol{D}$ is less obvious. 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 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, $\boldsymbol{P}^\top \mathbf{1} = \mathbf{0}$. This implies that $\boldsymbol{X}\mathbf{1} = \boldsymbol{P}\boldsymbol{P}^\top \mathbf{1} = \mathbf{0}$. We refer to $\boldsymbol{P}$ and $\boldsymbol{X}$ with this relationship as the 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 $\boldsymbol{X}$:

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

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

where $\mathrm{diag}(\cdot)$ 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], and for the truncated eigendecomposition $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ with $\boldsymbol{U} \in \mathbb{R}^{n \times r}$ and $\boldsymbol{\Lambda} \in \mathbb{R}^{r \times r}$,

$$\boldsymbol{P} = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}. \tag{3}$$

We note that $\boldsymbol{X}\mathbf{1} = \mathbf{0}$ also implies that $\boldsymbol{U}^\top \mathbf{1} = \mathbf{0}$. In many practical scenarios, the distance matrix may be incomplete, making classical MDS inapplicable for determining the point configuration. However, notice that $\mathrm{rank}(\boldsymbol{X}) \leq r$, and one can show that $\mathrm{rank}(\boldsymbol{D}) \leq r + 2$ [20]. This implies that when $r \ll n$, which is often the case in practice (e.g., $r = 2$ or 3 in EDMC), $\boldsymbol{X}$ and $\boldsymbol{D}$ are low-rank matrices. This allows us to utilize a rich library of tools from low-rank matrix completion, and moves us to consider the problem of Euclidean Distance Matrix Completion (EDMC). With this in mind, 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 $\boldsymbol{D}$:

$$\begin{aligned} \underset{\boldsymbol{Z} \in \mathbb{R}^{n \times n}}{\text{minimize}} \quad & \|\boldsymbol{Z}\|_* \\ \text{subject to} \quad & Z_{ij} = D_{ij} \quad \forall (i,j) \in \Omega, \end{aligned} \tag{4}$$

where $\|\cdot\|_* = \sum_i \sigma_i$ 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 EDMC problem:

$$\underset{\boldsymbol{X} \in \mathbb{R}^{n \times n},\, \boldsymbol{X} = \boldsymbol{X}^\top,\, \boldsymbol{X} \succeq \mathbf{0},\, \boldsymbol{X}\mathbf{1} = \mathbf{0}}{\text{minimize}} \quad \|\boldsymbol{X}\|_*, \tag{5}$$

where the constraints follow from the relation of $\boldsymbol{X}$ and $\boldsymbol{D}$ in (1) and (2). 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 (5) 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 EDMC 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.

The work by [33] is thematically closer to ours, as it also considers a similar initialization, and proves linear convergence to the ground truth. Therein, the core approach is based on a Burer-Monteiro factorization of the Gram matrix, and this leads to a program where the optimization variable is in terms of points. To solve the resulting optimization problem, a simple gradient descent algorithm with a line search followed by a projection is proposed. The projection is done to ensure incoherence, but requires as input an incoherence parameter and maximum singular value of the true Gram matrix. We highlight few differences of our approach to this work. First, their empirical evaluations assume known incoherence and maximum singular value. We believe such sharp estimates may not be necessary, and the impact of not having this information is not clear. We also note that we provide robustness guarantees in contrast to [33].

The work in [34] proposes a non-convex algorithm for the EDMC problem based on the reweighted least squares framework. It considers the case where distance entries are observed uniformly at random and establishes that with $O(\nu r \log(n))$ distance entries, where ν is the incoherence parameter (see Section 3 for the definition of a weaker form of incoherence used in this paper), are sufficient for local convergence to the ground Gram matrix. However, [34] does not provide a provable initialization scheme or robustness guarantees for the proposed algorithm. We note that the analysis in [34] achieves optimal sample complexity, matching the lower bound established in [35]. However, their results rely on a stronger incoherence condition than ours. In fact, under our milder incoherence assumption, their sample complexity aligns with ours up to constant factors.

1.1 Contributions

The main contributions of this paper are as follows:

1. **Geometric Interpretation of EDMC Coherence:** We provide a geometric interpretation of coherence within the specific context of the EDMC problem. We derive both lower and upper bounds for this parameter and discuss the geometries that achieve these bounds. Under a random model of the underlying points, we show that the coherence scales logarithmically, which aligns with the scaling of standard coherence measures.

2. **Algorithmic Framework:** We propose a novel non-convex iterative algorithm for the Euclidean Distance Matrix Completion (EDMC) problem based on Riemannian optimization. The algorithm performs first-order updates on the manifold of fixed-rank matrices and enjoys low per-iteration computational complexity.

3. **Provable initialization scheme:** We develop a structured initialization procedure from partial distance measurements and establish an explicit error bound between the initialization and the ground truth. The method is simple to implement and only requires available measurements.

4. **Convergence guarantees, sample complexity requirements, and robustness guarantees:** We provide rigorous analysis establishing high-probability local convergence of the proposed algorithm to the ground truth configuration with near optimal sample complexity. We also derive sample complexity bounds to ensure that the initialization lies within the basin of attraction and to provide robustness guarantees against bounded noise perturbations of the underlying point cloud.

5. **Novel Analysis** We leverage statistical tools not common in the EDMC literature to analyze the local behavior of the algorithm, including a restricted isometry property for a symmetric operator with coupled structure.

To the best of our knowledge, this is the first non-convex algorithm for the EDMC problem that provides provable initialization, provable convergence guarantees, robustness guarantees under noise, and a geometric interpretation of incoherence in the EDMC context.

1.2 Notation

The notation used in this paper is summarized in Table 1. This table provides a general description of the conventions used throughout this paper, but not every assignment is a strict rule. For example, lowercase boldface, such as $\boldsymbol{x}$, is denoted as reserved for vectors; however, we extensively use the notation $\boldsymbol{w_\alpha}$ for certain matrices. If there is any contradiction with Table 1, the notation should be clear from context.

Symbol	Meaning
Matrices, Vectors, and Operators	
$\boldsymbol{A}$, $\boldsymbol{B}$	Matrices (uppercase boldface)
$\boldsymbol{v}$	Vectors (lowercase boldface)
$\mathcal{A}$	Linear operators on matrices (calligraphic)
$\mathbb{V}$	Vector spaces and subspaces (blackboard bold)
$\boldsymbol{X}^\top$	Transpose of matrix $\boldsymbol{X}$
$\mathrm{Trace}(\boldsymbol{X})$	Trace of matrix $\boldsymbol{X}$
$\langle \boldsymbol{A}, \boldsymbol{B} \rangle$	Trace inner product: $\mathrm{Trace}(\boldsymbol{A}^\top \boldsymbol{B})$
δ_{ij}	Kronecker delta
X_{ij}	(i,j)-th entry of matrix $\boldsymbol{X}$
$\mathcal{A}^*$	Adjoint of operator $\mathcal{A}$
$\boldsymbol{1}$	Column vector of ones (size determined by context)
$\boldsymbol{0}$	Zero vector or zero matrix (depending on context)
$\boldsymbol{e}_i$	Standard basis vector: 1 at i-th position, zeros elsewhere
$\boldsymbol{e}_{ij}$	Standard matrix basis: 1 at (i,j), zeros elsewhere
$\mathrm{vec}(\boldsymbol{Y})$	Column stack of matrix $\boldsymbol{Y}$ into $\mathbb{R}^{n^2}$
$\odot$	Hadamard (entrywise) product
$\mathcal{I}$	Identity operator on matrices
$\boldsymbol{I}$	Identity matrix
$\boldsymbol{A} \succeq \boldsymbol{B}$	Loewner ordering: $\boldsymbol{A} - \boldsymbol{B}$ is positive semi-definite
$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top$	Thin spectral decomposition of symmetric rank-r matrix
Norms and Spectral Quantities	
$\|\boldsymbol{x}\|_2$	Euclidean (ℓ_2) norm of vector $\boldsymbol{x}$
$\|\boldsymbol{X}\|_\mathrm{F}$	Frobenius norm of matrix $\boldsymbol{X}$
$\|\boldsymbol{X}\|$	Operator norm (largest singular value)
$\|\boldsymbol{X}\|_\infty$	Max absolute entry of $\boldsymbol{X}$
$\|\boldsymbol{X}\|_*$	Nuclear norm: $\sum_i \sigma_i(\boldsymbol{X})$
$\|\mathcal{A}\|$	Operator norm of $\mathcal{A}$: $\sup_{\|\boldsymbol{X}\|_\mathrm{F}=1} \|\mathcal{A}(\boldsymbol{X})\|_\mathrm{F}$
$\lambda_{\max}(\boldsymbol{X}), \lambda_{\min}(\boldsymbol{X})$	Max/min eigenvalues of $\boldsymbol{X}$
$\lambda_1(\boldsymbol{X}) \geq \cdots \geq \lambda_r(\boldsymbol{X})$	Ordered non-zero eigenvalues of rank-r matrix, when clear from context, $\boldsymbol{X}$ is omitted
$\sigma_r(\boldsymbol{Y})$	r-th singular value of matrix $\boldsymbol{Y}$
κ	Condition number: $\|\boldsymbol{Y}\|/\sigma_r(\boldsymbol{Y})$
Sets and Indexing	
$\mathbb{I}$	Universal set of indices $\{(i,j) : 1 \leq i < j \leq n\}$
Ω	Random subsets of $\mathbb{I}$
$\emptyset$	Empty set
$\boldsymbol{x}_i, \boldsymbol{x}^i$	i-th row and i-th column of $\boldsymbol{X}$, respectively, represented as column vectors.
Manifolds and Geometry	
$\mathcal{N}_r$	Manifold of rank-r matrices
$\mathcal{N}$	General smooth manifolds
$\mathbb{T}, \mathbb{T}_l$	Tangent space at $\boldsymbol{X} \in \mathcal{N}_r$ and at l-th iterate $\boldsymbol{X}_l \in \mathcal{N}_r$
∇f	Euclidean gradient of $f \in C^1(\mathbb{R}^{n \times n})$
$\mathrm{grad}\, f$	Riemannian gradient of $f \in C^1(\mathcal{N}_r)$

Table 1: Summary of notation used throughout the paper.

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 low-rank matrix completion and a discussion of EDMC, with further background on dual bases and first-order Riemannian methods found in Ap-

pendix G. Section 3 gives a detailed discussion of the EDMC-specific incoherence condition. Section 4 is a discussion of our proposed methodology for solving the EDMC problem using geometric low-rank matrix completion ideas in the developed dual basis framework. Section 5 discusses the underlying assumptions, convergence analysis, initialization guarantees, and robustness results of the proposed algorithm, 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 Appendices B and C. We additionally provide initialization and robustness guarantees in this section, with relevant proofs in Appendices D and E. Section 7 discusses related geometric approaches in matrix completion, relevant work done in EDMC, and a more detailed discussion of geometric approaches to EDMC. Section 8 discusses the numerical results of this algorithm, and compares its efficacy to another algorithm in the literature. We conclude the paper in Section 9 with a brief discussion of the work and possible future research directions.

2 Preliminary Material

In this section, we will provide some minor background necessary to understand the work done in the following sections. A discussion of dual bases in linear algebra and first-order Riemannian methods can be found in Appendix G.

2.1 Matrix Completion

This work is related to low-rank matrix completion, where a subset of the entries of a low-rank ground truth matrix $\boldsymbol{X}$ are observed. 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:

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

As minimizing the rank directly is generally a challenging problem [25, 36], relaxations of this problem are often considered. For details on the complexity class of rank constrained problems, we refer the reader to [37]. 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 (4), is a well-studied problem [24, 38, 39]. 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 [40], 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 [41–44]. Beyond standard matrix completion, these methods have also been applied to other structured problems [45–47].

One of the main statistical approaches to analyzing matrix completion problems is through studying the behavior of the sampling operator $\mathcal{P}_\Omega$ restricted to a feasible space for recovery. This is formalized by defining, for a rank-r ground truth matrix $\boldsymbol{X}$, the tangent space $\mathbb{T}$ at $\boldsymbol{X}$ on $\mathcal{N}_r$, the manifold of rank-r matrices. Explicitly, we have that

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

Intuitively, restricting $\mathcal{P}_\Omega$ to $\mathbb{T}$ and measuring the deviation of this operator from the identity measures how well $\mathcal{P}_\Omega$ preserves information associated to $\boldsymbol{X}$ upon measurement, and whether or not $\boldsymbol{X}$ is uniquely recoverable given the information accessed. Mathematically, this manifests in proving statements such as

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

for some constant $c > 0$ and some small $\varepsilon_0 > 0$, which depends on both the number of samples and intrinsic properties of the ground truth matrix $\boldsymbol{X}$ [38]. This property is known as the Restricted Isometry Property (RIP), and variants of this property have been critical to low-rank matrix completion and compressive sensing literature [48, 49].

2.2 Dual Basis Approach to EDMC

In the EDMC 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 $\boldsymbol{w_\alpha}$ as follows:

$$\boldsymbol{w_\alpha} = \boldsymbol{e}_{\alpha_1\alpha_1} + \boldsymbol{e}_{\alpha_2\alpha_2} - \boldsymbol{e}_{\alpha_1\alpha_2} - \boldsymbol{e}_{\alpha_2\alpha_1}. \tag{7}$$

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 $\{\boldsymbol{w_\alpha}\}$ is a non-orthogonal basis for the subspace of symmetric matrices with zero row sum, denoted $\mathbb{S} = \{\boldsymbol{Y} \in \mathbb{R}^{n \times n} \mid \boldsymbol{Y} = \boldsymbol{Y}^\top, \boldsymbol{Y1} = \boldsymbol{0}\}$. In fact, for any two pairs of indices $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{I}$, we have:

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

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, $D_{ij} = \langle \boldsymbol{X}, \boldsymbol{w_\alpha} \rangle$ for $\boldsymbol{\alpha} = (i, j)$. Further, we can introduce the dual basis to $\{\boldsymbol{w_\alpha}\}$, denoted as $\{\boldsymbol{v_\alpha}\}$, and represent any centered Gram matrix $\boldsymbol{X}$ using the following expansion:

$$\boldsymbol{X} = \sum_{\boldsymbol{\alpha}} \langle \boldsymbol{X}, \boldsymbol{w_\alpha} \rangle \boldsymbol{v_\alpha}.$$

The advantage of the dual basis representation is that it allows us to recast the EDMC problem as a low-rank matrix recovery problem where we observe a subset of the expansion coefficients. In [28], this dual basis formulation has been used to provide theoretical guarantees for the convex program given in (5).

To make use of the dual basis approach both in theory and applications, one of the first steps is to have a representation of the dual basis that is easier to use. 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 [50], it was shown that the dual basis admits a simple explicit form

$$\boldsymbol{v_\alpha} = -\frac{1}{2} \left(\boldsymbol{ab}^\top + \boldsymbol{ba}^\top \right), \tag{8}$$

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:

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

From the bi-orthogonality relationship of the dual basis, it follows that $\mathcal{R}_\Omega^2 = \mathcal{R}_\Omega$ if Ω does not have repeated indices. The adjoint of $\mathcal{R}_\Omega$ admits the following form:

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

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 [51], $\mathcal{R}_\Omega(\boldsymbol{X})$ is related to the sampling operator $\mathcal{P}_\Omega(\boldsymbol{D})$ as follows:

$$\mathcal{R}_\Omega(\boldsymbol{X}) = -\frac{1}{2}\boldsymbol{J}\mathcal{P}_\Omega(\boldsymbol{D})\boldsymbol{J}, \tag{9}$$

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

$$\mathcal{F}_\Omega(\cdot) = \sum_{\boldsymbol{\alpha} \in \Omega} \langle \cdot, \boldsymbol{w_\alpha} \rangle \boldsymbol{w_\alpha}. \tag{10}$$

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 Geometric Interpretation of EDMC Incoherence

In pathological cases, the ground truth matrix $\boldsymbol{X}$ may exhibit a sparse representation in the basis $\{\boldsymbol{w_\alpha}\}_{\alpha \in \mathbb{I}}$, which could lead to challenges in its recovery from sampled measurements. While the concept of incoherence is well-established in the standard matrix completion literature, the condition specific to the EDMC problem slightly differs in structure and admits a natural geometric interpretation. This section is devoted to a detailed examination of this geometric perspective. We will state more formally the incoherence assumptions in Section 5, but we will first introduce one of the conditions below. We say that a rank-r Gram matrix $\boldsymbol{X} \in \mathbb{R}^{n \times n}$ is ν-incoherent with respect to $\{\boldsymbol{w_\alpha}\}_{\alpha \in \mathbb{I}}$ if the following statement holds:

$$\max_{\boldsymbol{\alpha} \in \mathbb{I}} \| \mathcal{P}_U \boldsymbol{w_\alpha} \|_{\mathrm{F}} \leq \sqrt{\frac{4\nu r}{n}}. \tag{11}$$

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

$$\max_{i \in [n]} \| \mathcal{P}_U \boldsymbol{e}_i \|_2 \leq \sqrt{\frac{4\nu r}{n}}. \tag{12}$$

The standard incoherence assumption, shown in (12), is prevalent throughout matrix completion literature and is a measure of "entrywise diffuseness" in the ground truth matrix. Further discussion of standard matrix incoherence can be seen in [25].

The incoherence condition introduced in (11) can be interpreted in terms of the underlying point cloud data. For the specific case of the EDMC problem, (11) can be expanded as follows for $\boldsymbol{\alpha} = (i, j)$ with $i < j$:

$$\begin{aligned}
\|\mathcal{P}_U \boldsymbol{w_\alpha}\|_{\mathrm{F}}^2 &= \langle \mathcal{P}_U \boldsymbol{w_\alpha}, \mathcal{P}_U \boldsymbol{w_\alpha} \rangle \\
&= \mathrm{Trace}\left(2\boldsymbol{w_\alpha} \boldsymbol{U}\boldsymbol{U}^\top\right) \\
&= 2\left\langle \boldsymbol{e}_{ii} - \boldsymbol{e}_{ij} + \boldsymbol{e}_{jj} - \boldsymbol{e}_{ji}, \boldsymbol{U}\boldsymbol{U}^\top \right\rangle \\
&= 2\left((\boldsymbol{U}\boldsymbol{U}^\top)_{ii} + (\boldsymbol{U}\boldsymbol{U}^\top)_{jj} - (\boldsymbol{U}\boldsymbol{U}^\top)_{ij} - (\boldsymbol{U}\boldsymbol{U}^\top)_{ji}\right) \\
&= 2\left(\boldsymbol{u}_i{}^\top \boldsymbol{u}_i + \boldsymbol{u}_j{}^\top \boldsymbol{u}_j - \boldsymbol{u}_j{}^\top \boldsymbol{u}_i - \boldsymbol{u}_i{}^\top \boldsymbol{u}_j\right) \\
&= 2\left(\boldsymbol{u}_i{}^\top \left(\boldsymbol{u}_i - \boldsymbol{u}_j\right) + \boldsymbol{u}_j{}^\top \left(\boldsymbol{u}_j - \boldsymbol{u}_i\right)\right) \\
&= 2\left(\boldsymbol{u}_i - \boldsymbol{u}_j\right)^\top \left(\boldsymbol{u}_i - \boldsymbol{u}_j\right).
\end{aligned}$$

The incoherence condition can then equivalently be stated as

$$\max_{(i,j):i<j} \left(\boldsymbol{u}_i - \boldsymbol{u}_j\right)^\top \left(\boldsymbol{u}_i - \boldsymbol{u}_j\right) \leq 2\frac{\nu r}{n}. \tag{13}$$

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

Lemma 3.1. *For the incoherence condition in* (13), ν *is bounded below by* $1 + \frac{2}{n-1}$ *and above by* $2\frac{n}{r}$.

Proof. We consider $\sum_{(i,j):i<j} \left(\boldsymbol{u}_i - \boldsymbol{u}_j\right)^\top \left(\boldsymbol{u}_i - \boldsymbol{u}_j\right)$. Note that $\sum_i (\boldsymbol{u}_i)^\top \boldsymbol{u}_i = \mathrm{Trace}(\boldsymbol{U}\boldsymbol{U}^\top) = \mathrm{Trace}(\boldsymbol{U}^\top \boldsymbol{U}) = r$. Since we assume centered configurations, $\boldsymbol{U}^\top \mathbf{1} = \mathbf{0}$. It then follows that, for $i \neq j$, $\sum_{i,j}(\boldsymbol{u}_i)^\top \boldsymbol{u}_j = \left(\sum_i \boldsymbol{u}_i\right)^\top \sum_j \boldsymbol{u}_j = 0$. Using these two relations, we obtain:

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

The above equality notes that the sum of L terms is $(n+1)r$. Therefore, the maximum summand must be at least $\frac{(n+1)r}{L}$. In particular, we have:

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

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

Remark 1. *To show that the lower bound for the incoherence can be attained, we consider the following example:*

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

Up to the scaling factor of $\sqrt{\frac{2}{3}}$, the rows of $\boldsymbol{U}$ correspond to the vertices of an equilateral triangle inscribed in the unit circle. It can be easily verified that this attains the lower bound on incoherence. For the upper bound, a simple example is the matrix $\boldsymbol{U} \in \mathbb{R}^{n \times 3}$, where the first two columns are the standard basis vectors $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$, respectively, and the third column is a unit vector which is zero in its first two entries. Any set of points generated from this $\boldsymbol{U}$ lies entirely along the z-axis, except for two points, which lie on the x- and y-axes, respectively. Figures 1 and 2 provide a visual illustration of these examples.
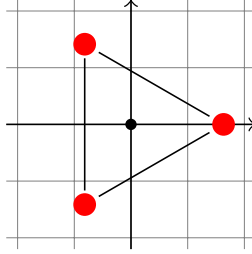

Figure 1: Visualizing the rows of $\boldsymbol{U}$ that lead to the lowest incoherence parameter.

Next, we aim to state the incoherence condition in terms of the points. Using (13) and noting the relation in (3), and recalling that $\boldsymbol{X} = \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^\top$ with matrix $\boldsymbol{\Lambda} := \mathrm{diag}\left(\lambda_1 \cdots \lambda_r\right)$, $\boldsymbol{u}_i = \boldsymbol{\Lambda}^{-1/2} \boldsymbol{p}_i$. Note that classical MDS only recovers a point cloud up to rotation, and that the vectors $\boldsymbol{p}_i$ referred to here are those recovered through MDS. As such, this exact relationship, $\boldsymbol{u}_i = \boldsymbol{\Lambda}^{-1/2} \boldsymbol{p}_i$, might not be held for any $\boldsymbol{P}$ that generates $\boldsymbol{X}$. However, as we discuss below, the relevant quantities of interest are invariant to an orthogonal transformation. We now consider $\boldsymbol{u}_i^\top \boldsymbol{u}_j$:

$$\left(\boldsymbol{u}_i\right)^\top \boldsymbol{u}_j = \left(\boldsymbol{\Lambda}^{-1/2} \boldsymbol{p}_i\right)^\top \left(\boldsymbol{\Lambda}^{-1/2} \boldsymbol{p}_j\right) = \boldsymbol{p}_i^\top \boldsymbol{\Lambda}^{-1} \boldsymbol{p}_j.$$

This indicates that our incoherence condition can be reinterpreted as

$$\max_{(i,j):i<j} \left(\boldsymbol{p}_i - \boldsymbol{p}_j\right)^\top \boldsymbol{\Lambda}^{-1} \left(\boldsymbol{p}_i - \boldsymbol{p}_j\right) \leq 2\frac{\nu r}{n}. \tag{14}$$

We first start our interpretation for the case where $\boldsymbol{\Lambda}$ is the identity matrix. In this setting, for any pair (i, j), the expression $\left(\boldsymbol{p}_i - \boldsymbol{p}_j\right)^\top \left(\boldsymbol{p}_i - \boldsymbol{p}_j\right)$ is the squared Euclidean distance between the points $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$. Hence, the incoherence can be directly linked to the maximum distance among the points. We now provide an interpretation of (14) in the general case. The quantity therein suggests that incoherence serves as a measure of how the displacement
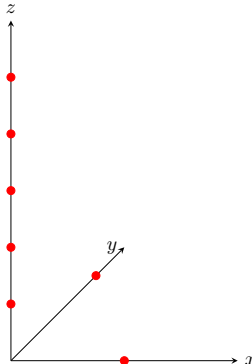

Figure 2: Example of a set of points with the highest incoherence parameter.

vectors $\boldsymbol{p}_i - \boldsymbol{p}_j$ align with the principal components of the embedding. In particular, for a fixed choice of $\boldsymbol{\Lambda}$, varying the matrix $\boldsymbol{U}$ leads to different sets of points. If the displacement vectors tend to align with directions corresponding to the smallest principal components (i.e., those with the lowest variance), the incoherence is expected to be high. Conversely, if they align more with the dominant components (those with the highest variance), the incoherence tends to be low. In essence, high incoherence indicates that certain pairs of points are stretching significantly in directions where the embedding space has low variance.

Using the variational characterization, note that $(\boldsymbol{p}_i - \boldsymbol{p}_j)^\top \boldsymbol{\Lambda}^{-1} (\boldsymbol{p}_i - \boldsymbol{p}_j) \leqslant \lambda_1(\boldsymbol{\Lambda}^{-1}) (\boldsymbol{p}_i - \boldsymbol{p}_j)^\top (\boldsymbol{p}_i - \boldsymbol{p}_j)$. Noting that $\lambda_1(\boldsymbol{\Lambda}^{-1}) = \frac{1}{\lambda_r}$, we can also state the incoherence condition in (14) as:

$$(\boldsymbol{p}_i - \boldsymbol{p}_j)^\top (\boldsymbol{p}_i - \boldsymbol{p}_j) \leq 2\frac{\nu r}{n}\lambda_r. \tag{15}$$

We note that that these statements are not equivalent, merely that this simpler statement implies the original incoherence condition. Continuing with the simplified incoherence condition in (15), we seek to derive an upper bound on ν in terms of other geometric properties of $\boldsymbol{P}$, or spectral properties of $\boldsymbol{X}$. First, notice that

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

As we seek a constant ν such that (15) is satisfied for all $(i, j) \in \mathbb{I}$, we can see that this will be satisfied if

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

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

$$\nu \leq \frac{n}{2r}\frac{\|\boldsymbol{D}\|_\infty}{\lambda_r(\boldsymbol{X})}. \tag{16}$$

Notice that if $\|\boldsymbol{D}\|_\infty = \mathcal{O}(r^2)$ and $\lambda_r(\boldsymbol{X}) = \mathcal{O}(n)$, then $\nu = \mathcal{O}(r)$. As r is most frequently either 2 or 3, this implies $\nu = \mathcal{O}(1)$ for relevant datasets, which is assumed throughout this work. We will now show that data drawn from bounded isotropic distributions exhibits this property.

Lemma 3.2. *[[52, Page 31]] Let $\{\boldsymbol{p}_i\}_{i=1}^n \sim \mu$ where μ is a probability measure defined on $\mathbb{R}^r$, and let $\boldsymbol{P} = [\boldsymbol{p}_1 \cdots \boldsymbol{p}_n]^\top \in \mathbb{R}^{n \times r}$. Define the covariance matrix of μ as $\boldsymbol{\Sigma}$. If $n \geq C(t/\varepsilon)^2 r$ for some constant $C > 0$, then with probability at least $1 - 2\exp(-t^2 n)$*

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

Let us now assume that μ is isotropic, that is $\boldsymbol{\Sigma} = \boldsymbol{I}$. Furthermore, as we are interested in point clouds satisfying $\boldsymbol{P}^\top \mathbf{1} = \mathbf{0}$, we consider mean zero distributions. As such, we can say that for isotropic distributions and for independent $\boldsymbol{p}_i$, $\boldsymbol{p}_j \sim \mu$ with $\mathbb{E}[\mu] = 0$ that

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

$$= \mathbb{E}\left[\|\boldsymbol{p}_i\|_2^2\right] + \mathbb{E}\left[\|\boldsymbol{p}_j\|_2^2\right] - 2\mathbb{E}\left[\boldsymbol{p}_i\right]^\top \mathbb{E}\left[\boldsymbol{p}_j\right]$$

$$= \mathbb{E}\left[\|\boldsymbol{p}_i\|_2^2\right] + \mathbb{E}\left[\|\boldsymbol{p}_j\|_2^2\right]$$

$$= 2\mathbb{E}\left[\|\boldsymbol{p}_i\|_2^2\right]$$

$$= 2\mathbb{E}\left[\text{Trace}\left(\boldsymbol{p}_i\boldsymbol{p}_i{}^\top\right)\right]$$

$$= 2\,\text{Trace}\left(\mathbb{E}\left[\boldsymbol{p}_i\boldsymbol{p}_i{}^\top\right]\right)$$

$$= 2\,\text{Trace}(\boldsymbol{I}) = 2r,$$

where the second and fourth lines follow from the independence of $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$, the third line follows from the fact that $\mathbb{E}[\mu] = 0$, and the seventh line follows from the fact that μ is isotropic, i.e. $\boldsymbol{\Sigma} = \boldsymbol{I}$.

Lemma 3.3. *Let $\{\boldsymbol{p}_i\}_{i=1}^n \subset \mathbb{R}^r$ be a collection of points drawn i.i.d. from an isotropic sub-Gaussian distribution μ. Furthermore, let $\mathbb{E}[\boldsymbol{p}_i] = \mathbf{0}$, and assume each coordinate of $\boldsymbol{p}_i$ is independent. Let $\|\boldsymbol{p}_i\|_{\psi_2} \leq K$, where $\|\cdot\|_{\psi_2}$ is the sub-Gaussian norm. Then with probability at least $1 - Cn^{-2}$,*

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

where $C > 0$ is an absolute constant.

Proof. This result is a simple application of the Hanson-Wright inequality, seen in Theorem A.3. First, given vectors $\boldsymbol{p}_i, \boldsymbol{p}_j$, notice that

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

Previously, we have shown that $\mathbb{E}[(\boldsymbol{p}_i - \boldsymbol{p}_j)^\top (\boldsymbol{p}_i - \boldsymbol{p}_j)] = 2r$. Furthermore, $\|\boldsymbol{A}\|_{\mathrm{F}}^2 = 4r$, and $\|\boldsymbol{A}\| \le 2$ by Gershgorin circle theorem. The result follows from an application of Theorem A.3. □

Next, we show that we can upper bound K by $C\lambda_1(\boldsymbol{\Sigma}) = C$ for an isotropic, sub-Gaussian μ and some absolute constant $C > 0$. We will use a moment generating function bound to prove this. First, from Definition 3.4.1 in [53], we have that $\|\boldsymbol{p}_i\|_{\psi_2} = \sup_{\|\boldsymbol{u}\|_2=1} \|\boldsymbol{u}^\top \boldsymbol{p}_i\|_{\psi_2}$. Using the moment-generating technique, we can see that

$$\begin{aligned}
\mathbb{E}\left[\exp\left(t^2(\boldsymbol{u}^\top \boldsymbol{p}_i)^2\right)\right] &= \mathbb{E}\left[\exp\left(t^2 \boldsymbol{u}^\top \boldsymbol{p}_i \boldsymbol{p}_i{}^\top \boldsymbol{u}\right)\right] \\
&\le \sup_u \mathbb{E}\left[\exp\left(t^2 \boldsymbol{u}^\top \boldsymbol{p}_i \boldsymbol{p}_i{}^\top \boldsymbol{u}\right)\right] \\
&\le \mathbb{E}\left[\sup_u \exp\left(t^2 \boldsymbol{u}^\top \boldsymbol{p}_i \boldsymbol{p}_i{}^\top \boldsymbol{u}\right)\right] \quad \le \exp\left(t^2 \lambda_1(\boldsymbol{I})\right).
\end{aligned}$$

This gives us the bound $K \le C$ for some absolute constant $C > 0$.

We can now see that $\|\boldsymbol{D}\|_\infty \le 2r + 4C^2\sqrt{r}\log n$ with high probability for a sub-Gaussian isotropic distribution. Furthermore, from Theorem 3.2, we know that $\lambda_r(\boldsymbol{X}) = cn$ for some $\mathcal{O}(1)$ constant $c > 0$. As such, we can see that the incoherence constant can be upper-bounded using (16) by

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

This indicates that, with high probability, the incoherence constant remains in a regime where it does not degrade the recovery guarantees established in Section 5 for data generated from sub-Gaussian distributions. We note that this result is very similar to the condition derived in [25] for the incoherence of matrices in the random orthogonal model. If it is further assumed that the distribution is bounded in such a way that $\|\boldsymbol{D}\|_\infty \le Cr$ for some $\mathcal{O}(1)$ constant C, e.g., if μ is supported in a ball of radius $r^{1/2}$, then this further reduces the incoherence constant to $\nu = \mathcal{O}(1)$.

We note that the analysis in this section exclusively pertained to data generated from isotropic measures. These techniques can be extended to centered and bounded anisotropic sub-Gaussian measures, and one can show the resulting bound for $\nu = \mathcal{O}(\kappa \log n/\sqrt{r})$, where κ is the condition number of $\boldsymbol{X}$. We provide a proof of this result in Theorem F.3.

Remark 2. *We now provide a geometric interpretation of* (12)*. Expanding* (12)*, we obtain:*

$$\|\mathcal{P}_U \boldsymbol{e}_i\|_2^2 = \boldsymbol{e}_i^\top \mathcal{P}_U \boldsymbol{e}_i = \mathrm{Trace}(\boldsymbol{e}_i^\top \mathcal{P}_U \boldsymbol{e}_i) = \mathrm{Trace}(\mathcal{P}_U \boldsymbol{e}_i \boldsymbol{e}_i^\top) = \langle \boldsymbol{U}\boldsymbol{U}^\top, \boldsymbol{e}_{ii}\rangle = \boldsymbol{u}_i{}^\top \boldsymbol{u}_i = \boldsymbol{p}_i{}^\top \boldsymbol{\Lambda}^{-1} \boldsymbol{p}_i.$$

As such, standard incoherence in the EDMC framework represents a re-scaled ℓ_2 norm maximum of the underlying point cloud.

3.1 Finer Interpretation of EDMC Incoherence and Applications

Throughout this work, we have treated incoherence as an index-by-index bound; that is to say that we only consider terms such as $\|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2$. We wish to investigate this in more detail now. The main technical problem that the incoherence assumption provides a solution for is in the variance estimations used in concentration inequalities, such as in Theorem 5.3, for example. This variance estimate comes from a Gershgorin style upper bound on the matrix $\tilde{\boldsymbol{H}} = [\langle \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}, \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\beta}}\rangle] \in \mathbb{R}^{L \times L}$, seen in Theorem A.6. The eigenvalue bound leverages the fact that, if $\boldsymbol{\alpha} \cap \boldsymbol{\beta} = \emptyset$, $\langle \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}, \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\beta}}\rangle = 0$, and the other terms we use Theorem 5.1 in tandem with Cauchy-Schwarz to get a uniform bound on the non-zero entries. This yields an upper bound that is used to estimate the variance term in the

concentration inequalities. We argue here that a more fine-grained representation of incoherence could potentially sharpen incoherence results and lead to more geometrically-optimal sampling strategies in the future.

For the Gershgorin estimate, we need to estimate $|\langle \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}, \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\beta}} \rangle|$ for all non-zero entries of $\tilde{\boldsymbol{H}}$. Without loss of generality, we assume that $\boldsymbol{\alpha} = (i, j)$ and $\boldsymbol{\beta} = (i, k)$ for $i, j, k \in [n]$. Following a nearly identical chain of computations as in Remark Remark 4, one can show that

$$|\langle \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}, \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\beta}} \rangle| = \left| (\boldsymbol{p}_i - \boldsymbol{p}_j)^\top \boldsymbol{\Lambda}^{-1} (\boldsymbol{p}_i - \boldsymbol{p}_k) \right|.$$

This interpretation indicates that what might be more relevant to variance minimization is sampling more orthogonal angles with respect to a whitened dataset, rather than just considering lengths. This could lead to more optimal non-uniform sampling techniques for solving the EDMC problem.

4 The Riemannian Dual Basis Approach to EDMC

With the goal of translating the standard matrix completion problem to Gram matrix completion of a ground truth matrix $\boldsymbol{X} \in \mathbb{S}$, where $\mathbb{S} = \{ \boldsymbol{Y} \in \mathbb{R}^{n \times n} | \boldsymbol{Y} = \boldsymbol{Y}^\top, \boldsymbol{Y1} = \boldsymbol{0} \}$, the most direct adaptation of the work conducted in [54] would be defining an objective function by analogy to (28) as follows:

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

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

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

where $\nabla_{\boldsymbol{Y}}$ denotes the gradient with respect to $\boldsymbol{Y}$. This is inaccessible given the problem statement, as each $\boldsymbol{v}_{\boldsymbol{\alpha}}$ depends on every $\boldsymbol{w}_{\boldsymbol{\alpha}}$ as $\boldsymbol{v}_{\boldsymbol{\alpha}} = \sum_{\boldsymbol{\alpha} \in \mathbb{I}} \boldsymbol{H}^{\boldsymbol{\alpha\beta}} \boldsymbol{w}_{\boldsymbol{\alpha}}$. To circumvent this difficulty, there has been exploration into self-adjoint alternatives to $\mathcal{R}_\Omega$ [28, 51, 55]. The novel surrogate introduced in this work, denoted $\mathcal{M}_\Omega$, allows for the definition of an objective function in analogy to (28).

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

$$\mathcal{M}_\Omega(\cdot) = \sum_{\boldsymbol{\alpha}, \boldsymbol{\beta} \in \Omega} C_{\boldsymbol{\alpha\beta}} \langle \cdot, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}} \rangle \boldsymbol{w}_{\boldsymbol{\beta}}, \tag{17}$$

where $C_{\boldsymbol{\alpha\alpha}} = p$ for all $\boldsymbol{\alpha}$, and $C_{\boldsymbol{\alpha\beta}} = 1$ for all $\boldsymbol{\alpha} \neq \boldsymbol{\beta}$. This diagonal re-scaling is introduced to make sure that $\mathbb{E}[\mathcal{M}_\Omega] = p^2 \mathcal{I}$. Previous literature introduced an unscaled form of this operator, i.e. $C_{\boldsymbol{\alpha\alpha}} = 1$, computed as $\mathcal{R}_\Omega^* \mathcal{R}_\Omega$ [51]. This operator does not concentrate around the identity operator, demonstrated in Theorem B.1, and as such a re-scaled form of the operator must be considered. The new operator is self-adjoint, and as such we can define the following objective function for the EDMC problem using this operator:

$$\underset{\boldsymbol{Y} \in \mathbb{S}}{\text{minimize}} \; \frac{1}{2} \langle \boldsymbol{Y} - \boldsymbol{X}, \mathcal{M}_\Omega (\boldsymbol{Y} - \boldsymbol{X}) \rangle \; \text{subject to } \text{rank}(\boldsymbol{Y}) = r. \tag{18}$$

This object is a true quadratic form with a symmetric operator, and its Euclidean gradient is given solely by $\mathcal{M}_\Omega(\boldsymbol{Y} - \boldsymbol{X})$. As such, it can be approached identically to (28) following the principles outlined in Appendix G. To perform this first-order retraction method from the tangent space at a point $\boldsymbol{X}_l \in \mathcal{N}_r$, we define the retraction map, known as the hard thresholding operator $\mathcal{H}_r : \mathbb{T}_l \to \mathcal{N}_r$, as follows:

$$\mathcal{H}_r(\boldsymbol{Y}) = \sum_{i=1}^{r} \lambda_i(\boldsymbol{Y}) \boldsymbol{U}_i \boldsymbol{U}_i^\top, \tag{19}$$

where $\boldsymbol{U}_i$ is the i-th eigenvector of $\boldsymbol{Y}$ corresponding to eigenvalue with the i-th largest magnitude $\lambda_i(\boldsymbol{Y})$. We note that for matrices $\boldsymbol{Y}$ with $\text{rank}(\boldsymbol{Y}) \geq r$ that $\text{rank}(\mathcal{H}_r(\boldsymbol{Y})) = r$. We can now define Algorithm 1, the main object of study in this work:

In the approach seen in Algorithm 1, the thin spectral decomposition in the gradient descent scheme is the most expensive, especially when n is large. As described previously, the authors in [54] found an efficient way to reduce the computational complexity of this decomposition from $\mathcal{O}(rn^2)$ to $\mathcal{O}(r^3 + r^2 n)$, substantially reducing the cost per iteration, which we implement as well. We note that in Algorithm 1 the reconstruction of the ground truth Gram matrix $\boldsymbol{X}$ is equivalent to the reconstruction of $\boldsymbol{D}$, as there is a one-to-one correspondence between $\boldsymbol{X}$ and $\boldsymbol{D}$ through (2).

Algorithm 1 Dual Basis Riemannian EDMC (DBRE)

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

Remark 3. *We wish to provide an interpretation of the operators $\mathcal{M}_\Omega$ and $\mathcal{R}_\Omega^* \mathcal{R}_\Omega$. First, if $\Omega = \mathbb{I}$, then the spectra of $\mathcal{F}_\Omega$ is known to be equivalent to the spectra of $\boldsymbol{H}$, and thus $\lambda_{\max}(\mathcal{F}_\Omega) = 2n$ [50]. As such, it is not the case that $\|\mathcal{F}_\Omega - \mathcal{I}\|$ is small. We can instead consider the following way to rescale the geometry of the linear space $\mathbb{S}$ that $\mathcal{F}_\Omega$ acts on through a preconditioner. First, define $\mathcal{S}_\Omega : \mathbb{R}^{n \times n} \to \mathbb{R}^L$ as $(\mathcal{S}_\Omega(\boldsymbol{X}))_{\boldsymbol{\alpha}} = \langle \boldsymbol{X}, \boldsymbol{w_\alpha} \rangle$ for $\boldsymbol{\alpha} \in \Omega$, and 0 otherwise. As such, one can show that $\mathcal{F}_\Omega = \mathcal{S}_\Omega^* \mathcal{S}_\Omega$. To re-scale $\mathcal{F}_\Omega$, one can instead consider $\mathcal{S}_\Omega^* \boldsymbol{H}^{-1} \mathcal{S}_\Omega$. This rescaling is done with $\boldsymbol{H}^{-1}$ to make it so that $\mathcal{S}_\Omega^* \boldsymbol{H}^{-1} \mathcal{S}_\Omega = \mathcal{I}$ when $\Omega = \mathbb{I}$. One can compute out $\mathcal{S}_\Omega^* \boldsymbol{H}^{-1} \mathcal{S}_\Omega$ and show that*

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

As $\mathcal{F}_\Omega$ exhibits the desired concentration properties (see Theorem B.6) but does not become the identity when $\Omega = \mathbb{I}$, this motivated the investigation into $\mathcal{R}_\Omega^* \mathcal{R}_\Omega$. Further investigation in Theorem B.1 validates the necessity of considering a rescaled variant of $\mathcal{R}_\Omega^* \mathcal{R}_\Omega$ to ensure concentration around $\mathcal{I}$, resulting in $\mathcal{M}_\Omega$.

4.1 Implementation Efficiency

We use recent advances in Riemannian optimization from [54] and [51] to develop an efficient implementation of the proposed algorithm. Computation of $\mathcal{R}_\Omega(\boldsymbol{X})$ and $\mathcal{M}_\Omega(\boldsymbol{X})$ can be done efficiently, with a minimal complexity per iteration. For $\mathcal{R}_\Omega$, a given iterate $\boldsymbol{X}_l$ can be easily translated to its distance matrix $\boldsymbol{D}_l$ via (2), and through (9), $\mathcal{R}_\Omega(\boldsymbol{X}_l)$ can be computed in $\mathcal{O}(m)$ operations, for $|\Omega| = m$. First, we note that $\mathcal{M}_\Omega(\boldsymbol{X}_l) = \mathcal{R}_\Omega^* \mathcal{R}_\Omega(\boldsymbol{X}_l) - \|\boldsymbol{v_\alpha}\|_{\mathrm{F}}^2 (1 - p) \mathcal{F}_\Omega(\boldsymbol{X}_l)$ and $\|\boldsymbol{v_\alpha}\|_{\mathrm{F}}$ is constant for all $\boldsymbol{\alpha} \in \mathbb{I}$ (Theorem A.8). This can be seen as follows:

$$
\begin{aligned}
\mathcal{M}_\Omega(\cdot) &= \sum_{\boldsymbol{\alpha}, \boldsymbol{\beta} \in \Omega} C_{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \boldsymbol{w_\alpha} \rangle \langle \boldsymbol{v_\alpha}, \boldsymbol{v_\beta} \rangle \boldsymbol{w_\beta} \\
&= \sum_{\substack{\boldsymbol{\alpha}, \boldsymbol{\beta} \in \Omega \\ \boldsymbol{\alpha} = \boldsymbol{\beta}}} C_{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \boldsymbol{w_\alpha} \rangle \langle \boldsymbol{v_\alpha}, \boldsymbol{v_\beta} \rangle \boldsymbol{w_\beta} + \sum_{\substack{\boldsymbol{\alpha}, \boldsymbol{\beta} \in \Omega \\ \boldsymbol{\alpha} \neq \boldsymbol{\beta}}} C_{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \boldsymbol{w_\alpha} \rangle \langle \boldsymbol{v_\alpha}, \boldsymbol{v_\beta} \rangle \boldsymbol{w_\beta} \\
&= \sum_{\boldsymbol{\alpha} \in \Omega} C_{\boldsymbol{\alpha}\boldsymbol{\alpha}} \langle \cdot, \boldsymbol{w_\alpha} \rangle \langle \boldsymbol{v_\alpha}, \boldsymbol{v_\alpha} \rangle \boldsymbol{w_\alpha} + \sum_{\substack{\boldsymbol{\alpha}, \boldsymbol{\beta} \in \Omega \\ \boldsymbol{\alpha} \neq \boldsymbol{\beta}}} C_{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \cdot, \boldsymbol{w_\alpha} \rangle \langle \boldsymbol{v_\alpha}, \boldsymbol{v_\beta} \rangle \boldsymbol{w_\beta} \\
&= p \|\boldsymbol{v_\alpha}\|_F^2 \mathcal{F}_\Omega(\cdot) + \sum_{\substack{\boldsymbol{\alpha}, \boldsymbol{\beta} \in \Omega \\ \boldsymbol{\alpha} \neq \boldsymbol{\beta}}} \langle \cdot, \boldsymbol{w_\alpha} \rangle \langle \boldsymbol{v_\alpha}, \boldsymbol{v_\beta} \rangle \boldsymbol{w_\beta} \\
&= p \|\boldsymbol{v_\alpha}\|_F^2 \mathcal{F}_\Omega(\cdot) + \sum_{\boldsymbol{\alpha}, \boldsymbol{\beta} \in \Omega} \langle \cdot, \boldsymbol{w_\alpha} \rangle \langle \boldsymbol{v_\alpha}, \boldsymbol{v_\beta} \rangle \boldsymbol{w_\beta} - \sum_{\boldsymbol{\alpha} \in \Omega} \langle \cdot, \boldsymbol{w_\alpha} \rangle \langle \boldsymbol{v_\alpha}, \boldsymbol{v_\alpha} \rangle \boldsymbol{w_\alpha} \\
&= p \|\boldsymbol{v_\alpha}\|_F^2 \mathcal{F}_\Omega(\cdot) + \mathcal{R}_\Omega^* \mathcal{R}_\Omega(\cdot) - \|\boldsymbol{v_\alpha}\|_F^2 \mathcal{F}_\Omega(\cdot), \quad (20)
\end{aligned}
$$

as expected. It is known that $\mathcal{R}_\Omega^* \mathcal{R}_\Omega(\boldsymbol{X}_l)$ is $\mathcal{O}(m)$ sparse and requires $\mathcal{O}(m)$ operations to compute [51]. The argument is outlined as follows. Let $\mathcal{T} : \mathbb{R}^{n \times n} \to \mathbb{R}^{n \times n}$ denote the map defined by (2) and let $\mathcal{T}^*$ denote its adjoint. It was shown in [51] that, up to a previously incorrect absence of a minus sign, for a Gram matrix $\boldsymbol{X}$,

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

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

Next, $\mathcal{F}_\Omega(\boldsymbol{X}_l)$ is efficiently computable in $\mathcal{O}(m)$ operations as each matrix $\boldsymbol{w}_{\boldsymbol{\alpha}}$ has 4 non-zero entries, allowing for easy computation given $\{\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle\}_{\boldsymbol{\alpha} \in \Omega}$. As such, $\mathcal{F}_\Omega(\boldsymbol{X}_l)$ is $\mathcal{O}(m)$ sparse. Using the fact that $\mathcal{R}_\Omega^* \mathcal{R}_\Omega(\cdot)$ and $\mathcal{F}_\Omega(\cdot)$ are sparse, it can be easily argued that the sum of the three terms in (20) preserves the a common sparsity pattern, and it can be computed in $\mathcal{O}(m)$ operations. Therefore $\mathcal{M}_\Omega(\boldsymbol{X}_l)$, and thus $\boldsymbol{G}_l$ in Step 4 of Algorithm 1, is computable in $\mathcal{O}(m)$ operations.

Step 5 can be computed in $\mathcal{O}(n^2)$ operations, as $\mathcal{P}_\mathbb{T} \boldsymbol{G}_l$ is a dense matrix. Some calculation yields that steps 6 and 7 can be computed with $n^2 r + \mathcal{O}(nr^2 + r^3)$ [54], giving a total cost per iteration of $n^2 r + \mathcal{O}(m + nr^2 + r^3)$. Note that the dominant cost is $n^2 r$, which is less expensive than computing Step 7 using the truncated singular value decomposition directly. Although both approaches have the same asymptotic complexity, the latter incurs a significantly higher constant factor (e.g., a factor of 6 or 14 depending on the choice of algorithm; see, for example, Figure 8.6.1 in [56]).

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 1, presented in Theorems 5.4 and 5.6. Prior to this, we formally state our incoherence assumptions, expanding upon the assumption first described in Section 3:

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 ν-incoherent to the basis $\{\boldsymbol{w}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathbb{I}}$ and ν-incoherent to its dual basis $\{\boldsymbol{v}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathbb{I}}$; that is, there exists a constant $\nu \geq 1$ such that for all $\boldsymbol{\alpha} = (i,j) \in \mathbb{I}$:*

$$\|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\|_\mathrm{F} \leq \sqrt{\frac{\nu r}{2n}}, \quad \text{and} \quad \|\mathcal{P}_U \boldsymbol{v}_{\boldsymbol{\alpha}}\|_\mathrm{F} \leq \sqrt{\frac{\nu r}{2n}}. \tag{21}$$

In addition to the above, we require that

$$\|\mathcal{P}_\mathbb{T} \boldsymbol{w}_{\boldsymbol{\alpha}}\|_\mathrm{F} \leq \sqrt{\frac{\nu r}{2n}}, \quad \text{and} \quad \|\mathcal{P}_\mathbb{T} \boldsymbol{v}_{\boldsymbol{\alpha}}\|_\mathrm{F} \leq \sqrt{\frac{\nu r}{2n}}. \tag{22}$$

Notice that the two definitions in (21) and (22) are equivalent up to a small constant, as

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

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

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

where the equality follows from the definition of $\mathcal{P}_U$ and $\mathcal{P}_\mathbb{T}$, and the inequality follows from Cauchy-Schwarz. As such, we pick a ν large enough such that the inequalities in (21) and (22) hold. We note that the constant difference in the condition stated above and in Section 3 is merely a matter of mathematical convenience. We also note that these incoherence conditions are similar to those seen in matrix completion with respect to the standard basis [38], as well as completion with respect to other bases [28, 39].

Remark 4. *We want to note that ν-incoherence with respect $\boldsymbol{w}_{\boldsymbol{\alpha}}$ in both (21) and (22) implies, at worst, 4ν-incoherence with respect to $\boldsymbol{v}_{\boldsymbol{\alpha}}$. As such, we choose a ν large enough so that both $\boldsymbol{X}$ is ν-incoherent with respect to $\boldsymbol{w}_{\boldsymbol{\alpha}}$ and $\boldsymbol{v}_{\boldsymbol{\alpha}}$. See Theorem F.1 for details.*

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

Assumption 5.2. *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 Theorem 5.1 with $\mathcal{O}(1)$ constant factor ν. As in [54], we identify a neighborhood in $\mathcal{N}_r$ around which any initial guess in this neighborhood converges linearly to the true solution with high probability using Algorithm 1.

5.1 Local Convergence Analysis

The most critical property for a sampling operator to possess in matrix completion theory is the restricted isometry property, briefly discussed in Section 2.1. This property roughly states that, when restricted to the local structure (or tangent space) around the true low-rank matrix, the partial observations preserve enough information to allow for faithful algorithmic recovery. We state this more formally with the following theorem:

Theorem 5.3 (RIP of $\mathcal{M}_\Omega$). *Let $\boldsymbol{X} \in \mathbb{S}$ be the ground truth, rank-r, ν-incoherent Gram matrix with tangent space $\mathbb{T}$ in $\mathcal{N}_r$. Let Ω be sampled from $\mathbb{I}$ via a Bernoulli sampling process with parameter $p \geq \frac{16}{3}\beta\frac{\log n}{n}$. If for some absolute numerical constant $C > 0$ and $\beta \geq 1$, then with probability at least $1 - 4n^{-\beta} - 2n^{1-\beta}$, we have that*

$$p^{-2}\|\mathcal{P}_\mathbb{T}\mathcal{M}_\Omega\mathcal{P}_\mathbb{T} - p^2\mathcal{P}_\mathbb{T}\| \leq 10\sqrt{\frac{\nu^2 r^2 \beta \log n}{pn}} + C\beta\nu r \frac{\log n}{pn}$$

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

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

Proof sketch. This proof works by decomposing $\mathcal{M}_\Omega$ into diagonal and off-diagonal components. We recognize that estimating off-diagonal terms in $\langle \boldsymbol{Y}, \mathcal{P}_\mathbb{T}\mathcal{M}_\Omega\mathcal{P}_\mathbb{T}(\boldsymbol{Y})\rangle$ can be written as a quadratic form with sub-Gaussian random vectors, allowing the application of the Hanson-Wright inequality (see Theorem A.3). The diagonal terms are equivalent to $p\|\boldsymbol{v}_\alpha\|_\text{F}^2\langle \boldsymbol{Y}, \mathcal{P}_\mathbb{T}\mathcal{F}_\Omega\mathcal{P}_\mathbb{T}(\boldsymbol{Y})\rangle$, and can be concentrated using a non-commutative Bernstein inequality, reproduced in Theorem A.1. See Section B.1 for details. □

Remark 5. *We note that this result is given in terms of the Bernoulli sampling probability, rather than the more traditional number of samples with replacement seen in the matrix completion literature. To provide a more direct comparison, and remarking that $\mathbb{E}[|\Omega|] = m$ and $p = \frac{m}{L}$, we have that for a sufficiently large constant $C > 0$ that*

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

gives ε_0-RIP of $\mathcal{M}_\Omega$. We again note that, due to using the weaker Theorem 5.1 instead of the incoherence assumption in [34], this is optimal up to constant factors and equivalent to the RIP established in [34].

Now that RIP is established, we can prove local convergence of Algorithm 1. This theorem describes a high-probability guarantee that Algorithm 1 exhibits linear convergence in an attractive basin near the solution, provided that $\mathcal{M}_\Omega$ exhibits RIP.

Theorem 5.4 (Local Convergence of Algorithm 1). *Let $\boldsymbol{X} \in \mathbb{R}^{n \times n}$ be the ground truth rank-r, ν-incoherent matrix and let $\mathbb{T}$ be the tangent space of $\mathcal{N}_r$ at $\boldsymbol{X}$. Suppose that $p \geq C\frac{\beta \log n}{n}$ for some absolute constant $C > 0$. Then*

$$p^{-2}\|\mathcal{P}_\mathbb{T}\mathcal{M}_\Omega\mathcal{P}_\mathbb{T} - p^2\mathcal{P}_\mathbb{T}\| \leq \varepsilon_0, \tag{23}$$

$$\|\mathcal{M}_\Omega\| \leq p^2\left(1 + 40\sqrt{\frac{\beta n \log n}{3p}}\right) + C'p\log n, \tag{24}$$

$$\|\mathcal{M}_\Omega\mathcal{P}_\mathbb{T}\| \leq p^{3/2}\sqrt{\frac{256\nu r \beta \log n}{3}}, \tag{25}$$

$$\|\mathcal{M}_\Omega\mathcal{P}_{\mathbb{T}_l}\| \leq 100p^{3/2}\sqrt{\beta n \log n}\frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_\text{F}}{\lambda_r(\boldsymbol{X})} + p^{3/2}\sqrt{\frac{256\nu r \beta \log n}{3}}, \tag{26}$$

$$\frac{\|\boldsymbol{X}_0 - \boldsymbol{X}\|_\text{F}}{\lambda_r(\boldsymbol{X})} \leq \frac{\varepsilon_0 p^{1/2}}{32\left(\beta n \log n\right)^{1/4}}, \tag{27}$$

where $C' > 0$ is an absolute numerical constant, $\beta > 1$, and where ε_0 is a constant satisfying

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

Then Algorithm 1 converges linearly as the iterates satisfy

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

Proof Sketch. We first note that each of the above assumptions, save for (27), holds with high probability for $p \geq C \frac{\nu^2 r^2}{\varepsilon_0^2} \frac{\beta \log n}{n}$, where $C > 0$ is an absolute constant. See Section C.1 for details.

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

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

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 (27), a local form of RIP for $\mathcal{M}_\Omega$ holds if (23) is true. This proof leverages the assumptions made in (25), and (26). I_2 follows from the neighborhood assumption of (27) in tandem with Theorem A.10, and I_3 follows from bounds on the step size (seen in Theorem C.1), the assumption in (25), and Theorem A.10. The assumptions in (23), (24), and (25) are all proven via high probability guarantees using Theorems A.1, A.2, and A.3. The technical details are deferred to the appendix, see Section C.1. See Figure 3 for a diagram of the main dependencies for the convergence proof. □

Figure 3: This diagram is a schematic of the overall proof of convergence for Algorithm 1. 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.

5.2 Initialization Results

In this section, we outline our initialization guarantees for Algorithm 1. 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 $\mathcal{N}_r$ of the measured information, provides a reasonable starting point and is described in

Algorithm 2, where $\mathcal{H}_r$ is as defined in (19). The following sections describe how close a one-step hard-thresholding initialization will be to the ground truth for Algorithm 1. Following this, and in tandem with Theorem 5.4, we show recovery guarantees for Algorithm 1.

Algorithm 2 1 Step Hard Thresholding Initialization

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

Lemma 5.5. *Under a Bernoulli sampling parameter $p \geq \frac{128\beta \log n}{3n}$, then with probability at least $1 - 2n^{1-\beta}$ we have for $\boldsymbol{X}_0 = p^{-1} \mathcal{H}_r(\mathcal{R}_\Omega(\boldsymbol{X}))$ that*

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

Proof. See Appendix D. □

Theorem 5.6 (Recovery Guarantee for Algorithm 1). *For $p \geq \max\left\{\frac{2\kappa\nu r^{3/2}}{\sqrt{3}\varepsilon_0} \frac{\beta^{3/4}\log^{3/4}(n)}{n^{1/4}}, C\frac{\nu^2 r^2}{\varepsilon_0^2} \frac{\beta \log n}{n}\right\}$, where κ is the condition number of $\boldsymbol{X}$, $\beta > 1$, and with $\varepsilon_0 < \frac{1}{22}$ for some sufficiently large constant $C > 0$, then with probability $1 - 8n^{1-\beta} - 14n^{-\beta}$, Algorithm 1 recovers the ground truth matrix $\boldsymbol{X}$ when initialized by Algorithm 2.*

Proof. This result is a consequence of Theorem 5.5 and the local neighborhood assumption in (39). We can see this by increasing the sample complexity p to a sufficiently large value such that the initialization is smaller than the local neighborhood assumption. □

Remark 6. *For Algorithm 1, we use a Bernoulli sampling model with parameter p, while other matrix completion methodologies use a uniform at random with replacement model. To provide a more direct sample complexity comparison, let $m = \mathbb{E}[|\Omega|]$ under a Bernoulli model. This implies that $p = \frac{m}{L}$. Theorem 5.6 therefore implies that, if*

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

for some sufficiently large constant $C > 0$, Algorithm 1 recovers $\boldsymbol{X}$.

Remark 7. *We note here that a more delicate initialization through a resampling technique, such as the one in [54], could likely reduce the sample complexity from $p \gtrsim \frac{\log^{3/4} n}{n^{1/4}}$ to $p \gtrsim \frac{\log n}{n}$. Further investigation of initialization has been omitted from this work due to space constraints, but is an area of interest for future research.*

6 Robustness Guarantees

In many applications, the distance matrix may be corrupted, and understanding the sources of this corruption is central to designing robust recovery algorithms [58–63]. Broadly, there are two main causes. First, even if distance measurements are perfectly accurate, the underlying point configuration may itself be perturbed due to physical factors. For instance, sensors placed in dynamic environments, such as the ocean, may drift over time. In such cases, the observed distances correspond to a perturbed version of the true point set. Second, the points themselves may be fixed, but the distance measurements are noisy. This can arise from various sources: sensor imprecision, environmental interference, or limited measurement resolution. In practice, both types of corruption may occur simultaneously. However, in this paper, we focus on the first scenario: perturbations in the point configuration. This assumption simplifies the analysis, since the resulting distance matrix remains a valid Euclidean distance matrix, and avoids challenges associated with arbitrary noise patterns that could violate geometric consistency. We believe that this setting is relevant to setting where environmental drift is more dominant than measurement noise. Moreover, the developed technical analysis for this setting could potentially serve as a foundation for future extensions to more general noise models.

In this section, we will provide robustness results for Algorithm 1. To begin, we assume the following: for a given point matrix $\boldsymbol{P}$, we denote $\hat{\boldsymbol{P}} = \boldsymbol{P} + \boldsymbol{N}$, where $\boldsymbol{N} \in \mathbb{R}^{n \times r}$ is a random matrix. We denote $\hat{\boldsymbol{X}} = \hat{\boldsymbol{P}}\hat{\boldsymbol{P}}^\top$. We make one more assumption on the ground truth matrix $\boldsymbol{X}$:

Assumption 6.1. *For a ground truth rank-r Gram matrix $\boldsymbol{X} \in \mathbb{R}^{n \times n}$, we assume that*

$$bn \leq \lambda_r(\boldsymbol{X}) \leq \cdots \leq \lambda_1(\boldsymbol{X}) \leq Bn$$

for some constants $b, B > 0$.

Remark 8. *We note here that for $\boldsymbol{P}$ generated from a sub-Gaussian distribution that each $\lambda_i(\boldsymbol{X})$ exhibits concentration around its expectation, per Theorem 3.2. For $\boldsymbol{P}$ generated from an isotropic distribution, $\mathbb{E}[\lambda_i(\boldsymbol{X})] = n \; \forall i \in [r]$, so it follows that $\lambda_r \approx \cdots \approx \lambda_1 \approx n$ with high probability, indicating $b \approx B$. We believe this assumption therefore only omits datasets that have ill-conditioned Gram matrices, or data that is scaled to be of a drastically different size than that of the unit ball in $\mathbb{R}^r$. We note that this latter condition is an artifact of the simplifying assumption presented above, and not a reflection of the non-scale-invariance of these techniques.*

To show robustness to noise, we first show that $\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_{\mathrm{F}}$ is small in Theorem E.1. Then, we show that for bounded noise the incoherence of the perturbed matrix is at most perturbed by an $\mathcal{O}(1)$ constant in Theorem E.2. We then show that, for a sufficiently large Bernoulli sample complexity p depending on the incoherence of $\hat{\boldsymbol{X}}$, that $\hat{\boldsymbol{X}}$ is recovered with Algorithm 1 with high probability, formally stated in the following theorem:

Theorem 6.2 (Robustness Guarantee for Algorithm 1). *Let $\hat{\boldsymbol{P}} = \boldsymbol{P} + \boldsymbol{N}$, where $\mathbb{E}[N_{ij}] = 0$ and $\|\boldsymbol{N}\|_\infty \leq \frac{\nu \gamma B^{1/2}}{16 n \beta \kappa \log n}$ for some $\gamma > 0$, $\beta > \max\left\{1, \frac{3r}{8 \log n}\right\}$, b, B are defined in accordance with Theorem 6.1, and κ is the condition number of $\boldsymbol{X}$. Assume that the measured distances are of the form $\langle \hat{\boldsymbol{X}}, \boldsymbol{w_\alpha} \rangle$ and are sampled in a Bernoulli scheme with parameter p with*

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

where $C > 484$ is an absolute constant. Assume furthermore that we initialize Algorithm 1 at a point $\boldsymbol{X}_0$ satisfying the assumptions of Theorem 5.4.

Then with probability at least $1 - 6n^{1-\beta} - 14n^{-\beta}$, Algorithm 1 recovers $\hat{\boldsymbol{X}}$, and

$$\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_{\mathrm{F}} \leq \frac{4b}{3}\gamma.$$

Proof of Theorem 6.2. This result follows first from Lemmas E.1 and E.2 with the selected constants to determine the incoherence parameter. From here, the sample complexity guarantee of Theorem 5.3, coupled with the high probability guarantees of the assumptions in Theorem 5.4 gives the desired result for an initialization satisfying (27). $\qquad \square$

This result indicates that the recovery of an object under noise is dependent on its underlying geometry. Highly degenerate objects with high condition numbers can only be perturbed by a small fraction of noise before the recovery becomes infeasible. Furthermore, the larger the noise, the higher the incoherence parameter can be perturbed by, which can result in a larger sample complexity necessary for recovery.

7 Related Work

7.1 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 [54, 64–69]. The analysis conducted by [54] 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 [54], and as such, a brief exposition of their work is provided.

In [54], the authors develop a gradient descent algorithm to solve the low-rank matrix completion problem, reconstructing a ground truth matrix $\boldsymbol{X}$ from partial measurements, leveraging this Riemannian structure. The objective function used in [54] is as follows:

$$\underset{\boldsymbol{Y} \in \mathbb{R}^{n \times n}}{\text{minimize}} \; \langle \boldsymbol{Y} - \boldsymbol{X}, \mathcal{P}_\Omega(\boldsymbol{Y} - \boldsymbol{X}) \rangle \; \text{subject to} \; \mathrm{rank}(\boldsymbol{Y}) = r. \tag{28}$$

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 (28) 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 [64], which minimized the Frobenius norm difference between the observed entries of the low-rank matrices to solve the problem. Additionally, [64] 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 [64], this does not necessarily guarantee identification of the ground truth. In contrast, [54] establishes linear convergence to the ground truth solution in a local neighborhood of the ground truth, with high probability. After defining (28), [54] constructs a Riemannian gradient descent procedure similar to the retraction procedure described in Section G.2 for its solution.

In addition to this approach, the work in [54] considered two initialization schemes. One is a simple one-step hard threshold onto $\mathcal{N}_r$, 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.

7.2   Euclidean Distance Matrix Completion Algorithms

To solve the EDMC 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 [70]. 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:

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

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 [70], 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 [70] 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 [70] 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 EDMC problem and their applications to molecular conformation and sensor network localization, we refer the reader to [6, 59, 71–74].

In the context of protein structure determination, various algorithmic approaches to EDMC have been developed. One notable example is the EMBED algorithm [75–77], which comprises three main steps [78]. 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 followed by a Branch-and-Prune method [79, 80].

Another category of approaches to the EDMC 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 [81–83]. The algorithm proposed in [84] 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.

Next, we highlight algorithms that estimate the underlying points through non-convex optimization. These utilize a combination of methods such as majorization, alternating projection, global continuation (transforming the optimization problem to a function with few local minimizers), and an asymmetric projected gradient descent scheme [11, 33, 34, 85–87]. One of particular interest is the an iteratively re-weighted least squares (IRLS) methodology. This technique relies on computing smoothed *log-det* objectives at each iterate of the continuous non-convex rank minimization problem, along with a least squares computation at each step. This algorithm relies on RIP of an

operator related to $\mathcal{R}_\Omega$, established for $|\Omega| \gtrsim \frac{\nu r}{\varepsilon_0^2} n \log n$ given a stronger incoherence assumption than used in this paper, and exhibits provable quadratic convergence in a local neighborhood around the solution provided RIP holds. No initialization guarantees are provided, however.

Certain nonconvex EDG algorithms have been shown to have better performance when the problem is formulated in a dimension higher than the true rank of the underlying points [28,87]. This overparameterization has previously been shown to enhance numerical performance in sensor network localization problems [88,89]. However, to the best of our knowledge, theoretical guarantees for such overparameterization in EDG problems remain largely unexplored. A recent study [90] conducts a landscape analysis of a nonconvex optimization problem for classical MDS and identifies dimensional regimes that lead to benign optimization landscapes.

We note that the above discussion does not comprehensively cover all EDMC algorithms, and we refer readers to [20, 91] for a more detailed overview.

7.2.1 Related Geometric Approaches to EDMC

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, 92]. This work is more in line with non-convex approaches based on optimizing over a Riemannian manifold [32, 93], and extends the Riemannian approach of [54] to the EDMC basis case.

A recent work in [30] adopts a similar approach to us and considers solving the EDMC 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:

$$\underset{\boldsymbol{Y} \in \boldsymbol{R}^{n \times d}}{\text{minimize}} \ \frac{1}{2} \|\boldsymbol{W} \odot \mathcal{P}_\Omega(g(\boldsymbol{Y}\boldsymbol{Y}^\top) - \boldsymbol{D}_e)\|_{\mathrm{F}}^2, \tag{29}$$

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 (29) 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 (29) 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 incoherence of the ground truth matrix and r is the rank. In contrast, our method also shows linear convergence in a local neighborhood and describes a strong initialization candidate for the noiseless EDMC recovery problem with provable high probability guarantees. We also provide robustness analysis for an EDMC problem perturbed by noise, and provide provable guarantees as well.

8 Numerical Results

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

8.1 Synthetic Data Experiment

In this section, we test the proposed algorithm on synthetic data. Various two and three dimensional datasets were used, and are referred to in Table 2 with their corresponding sizes. The goal of Algorithm 1 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 $\gamma \in [0, 1]$. The algorithm reconstructs the Gram matrix $\boldsymbol{X}_{\mathrm{rec}} = \boldsymbol{P}\boldsymbol{P}^\top$, from which $\boldsymbol{P}$ can be recovered using (3). The comparison referenced in Table 2 is the relative error between the recovered matrix $\boldsymbol{X}_{\mathrm{rec}}$ and the ground truth matrix $\boldsymbol{X}$ in Frobenius norm. Each run was terminated at either 1000 iterations or when a relative Frobenius norm difference between iterates of 10^{-5} was achieved. This experiment was initialized using the one-step-hard-thresholding method outlined in Section 5.

We note that the recovery completely fails for the sphere at 1% sampling, while recovery is partially successful for the other two datasets. This is because the other datasets are larger while maintaining the same rank, allowing for better scaling in the low sampling regime. In Figure 4, we show an image of the reconstruction of the figures described in Table 2.

Table 2: RELATIVE RECOVERY ERROR $\|\boldsymbol{X} - \boldsymbol{X}_{\text{rec}}\|_{\text{F}} / \|\boldsymbol{X}\|_{\text{F}}$ BETWEEN THE RECOVERED GRAM MATRIX AND THE TRUE GRAM MATRIX AVERAGED OVER 25 TRIALS USING ALGORITHM 1.

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

3% 2% 1%

Figure 4: Reconstruction of the synthetic datasets referenced in Table 2. From left to right, the Bernoulli parameter is 0.03, 0.02, and 0.01.

8.2 Comparison to Existing Methods

We provide an additional experiment to compare the efficacy of our algorithm DBRE to another provably convergent non-convex EDMC algorithm [34], which we refer as MatrixIRLS-EDMC . Let $r \in [2, 10]$, and consider $n = 100$ points sampled from $\text{Unif}(S^{r-1})$, the uniform distribution on the sphere embedded in r dimensions. As the number of degrees of freedom in a rank-r $n \times n$ matrix is $nr - \frac{r(r-1)}{2}$, define the oversampling ratio ρ as

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

as $\mathbb{E}[|\Omega|] = pL$ for Bernoulli random sampling with parameter p. In Figure 5, we compare the oversampling ratio versus the dimension of the sphere in a transition plot. Black indicates complete failure, classified as a relative Gram matrix error larger than 10^{-3}, and white indicates success. Each of these squares was run for 100 trials using DBRE and MatrixIRLS-EDMC. DBRE was initialized with 10 iterations of the Augmented Lagrangian algorithm in [28], and MatrixIRLS-EDMC was initialized using a least-squares methodology described in [34]. As these experiments indicate, performance between MatrixIRLS for EDMC and Algorithm 1 is comparable, with Algorithm 1 performing slightly better overall, but most noticeably in the higher rank regime.

8.3 Experiments on Noisy Distance Measurements

Finally, we also ran an experiment with noise following the model in Section 6 using Algorithm 1. Let $\{\boldsymbol{p}_i\}_{i=1}^{100} \sim \text{Unif}(S^2)$ be drawn i.i.d. and where $\boldsymbol{P} = [\boldsymbol{p}_1 \cdots \boldsymbol{p}_{100}]^\top \in \mathbb{R}^{100 \times 3}$. We perturb $\boldsymbol{P}$ with a bounded, centered noise matrix $\boldsymbol{N} \in \mathbb{R}^{100 \times 3}$ with $\|\boldsymbol{N}\|_\infty \leq 10^\gamma$ for $\gamma \in [-2, -1]$. Similar to the previous experiment, we set the oversampling

<div align="center">DBRE   MatrixIRLS-EDMC</div>

Figure 5: Oversampling ratio ρ versus dimension r for 100 points on the uniform distribution on S^{r-1}. Each parameter was tested 100 times.

Figure 6: Oversampling ratio ρ versus noise level 10^γ for 100 points drawn i.i.d. from $\mathrm{Unif}(S^2)$. Each parameter was tested 500 times.

ratio $\rho \in [1, 5]$. We set the success threshold at 10^{-2} relative difference, a relaxed value from previous experiments due to the addition of noise. Figure 6 shows the results over 500 trials.

Figure 6 indicates that recovery up to tolerance predominately gets worse with added noise, although there is some clear dependence on the size of the noise impacting the reconstruction of the ground truth. This is most likely due to an increase in the incoherence of the dataset, requiring higher sample complexities to reconstruct. However, the noise level is still the dominant factor, and after a large enough noise value, reconstruction up to a certain tolerance is no longer viable.

9   Conclusion and Future Work

In this work, we proposed a novel Riemannian gradient descent approach for solving the EDMC problem using a matrix completion approach on the manifold of rank-r matrices in Algorithm 1. In a local neighborhood, we proved that Algorithm 1 exhibits linear convergence with high probability. To the authors' knowledge, this is the first work to provide initialization guarantees for a non-convex approach to the EDMC problem. The convergence analysis of Algorithm 1 was predicated on a statistical understanding of coupled terms in a random operator, and required novel analysis to the matrix completion literature to our knowledge. For our method, we provided

numerical results to underline its efficacy, and Algorithm 1 performs comparably to other state-of-the-art non-convex methods. Additionally, we provided robustness analysis and corresponding convergence guarantees. Finally, we provided a novel interpretation of incoherence in the EDMC setting, highlighting potential areas for development of non-uniform sampling methods in this field. This is a primary avenue of future interest, as improving the sample complexity through geometrically-optimal sampling schemes would represent a noteworthy development in the EDMC literature.

Acknowledgments

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] Y. Li and X. Sun, "Euclidean distance matrix completion via asymmetric projected gradient descent," 2025. [Online]. Available: https://arxiv.org/abs/2504.19530

[34] I. Ghosh, A. Tasissa, and C. Kümmerle, "Sample-efficient geometry reconstruction from euclidean distances using non-convex optimization," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 77 226–77 268. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/8d57f138d14fdfdc520eb29804116d9e-Paper-Conference.pdf

[35] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.

[36] 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.

[37] 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.

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

[39] 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.

[40] 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.

[41] 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.

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

[43] 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.

[44] 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

[45] A. A. Abbasi, S. Moothedath, and N. Vaswani, "Fast federated low rank matrix completion," in *2023 59th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2023, pp. 1–6.

[46] A. A. Abbasi and N. Vaswani, "Efficient federated low rank matrix completion," *IEEE Transactions on Information Theory*, 2025.

[47] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[48] J. Wright and Y. Ma, *High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications*. Cambridge University Press, 2022.

[49] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*, ser. Applied and Numerical Harmonic Analysis. Springer New York, 2013. [Online]. Available: https://books.google.com/books?id=zb28BAAAQBAJ

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

[51] 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.

[52] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*. Cambridge University Press, 2012, p. 210–268.

[53] ——, *High-Dimensional Probability: An Introduction with Applications in Data Science*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[54] 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.

[55] 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

[56] G. H. Golub and C. F. Van Loan, *Matrix Computations - 4th Edition*. Philadelphia, PA: Johns Hopkins University Press, 2013. [Online]. Available: https://epubs.siam.org/doi/abs/10.1137/1.9781421407944

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

[58] U. A. Khan, S. Kar, and J. M. Moura, "Diland: An algorithm for distributed sensor localization with noisy distance measurements," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1940–1947, 2009.

[59] 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.

[60] P. Biswas, T.-C. Liang, K.-C. Toh, Y. Ye, and T.-C. Wang, "Semidefinite programming approaches for sensor network localization with noisy distance measurements," *IEEE transactions on automation science and engineering*, vol. 3, no. 4, pp. 360–371, 2006.

[61] A. Tasissa and W. Dargie, "Robust node localization for rough and extreme deployment environments," 2025. [Online]. Available: https://arxiv.org/abs/2507.03856

[62] C. Kundu, A. Tasissa, and H. Cai, "Structured sampling for robust euclidean distance geometry," in *2025 59th Annual Conference on Information Sciences and Systems (CISS)*. IEEE, Mar. 2025, p. 1–6. [Online]. Available: http://dx.doi.org/10.1109/CISS64860.2025.10944739

[63] ——, "A dual basis approach for structured robust euclidean distance geometry," 2025. [Online]. Available: https://arxiv.org/abs/2505.18414

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

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

[66] 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/

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

[68] 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.

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

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

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

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

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

[79] 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.

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

[81] 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.

[82] 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.

[83] 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.

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

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

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

[87] 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.

[88] 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

[89] 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

[90] C. Criscitiello, A. D. McRae, Q. Rebjock, and N. Boumal, "Sensor network localization has a benign landscape after low-dimensional relaxation," 2025. [Online]. Available: https://arxiv.org/abs/2507.15662

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

[92] 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

[93] 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.

[94] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Foundations of computational mathematics*, vol. 12, no. 4, pp. 389–434, 2012.

[95] M. Rudelson and R. Vershynin, "Hanson-wright inequality and sub-gaussian concentration," 2013. [Online]. Available: https://arxiv.org/abs/1306.2872

[96] C. Davis and W. M. Kahan, "The rotation of eigenvectors by a perturbation. iii," *SIAM Journal on Numerical Analysis*, vol. 7, no. 1, pp. 1–46, 1970. [Online]. Available: https://doi.org/10.1137/0707001

[97] 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

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

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

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

[101] 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.

[102] 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.

[103] 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.

[104] 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.

[105] 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.

A Properties of the Dual Bases and Concentration Inequalities

This section of the appendix details technical results about the specific dual bases, $\{\boldsymbol{w_\alpha}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$ and $\{\boldsymbol{v_\alpha}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$. These are needed to prove various technical lemmas throughout the work, but are particularly important in the proof of Theorem 5.3. Additionally, we provide the non-commutative and scalar Bernstein inequalities, as well as the Hanson-Wright inequality and the Davis-Kahan Theorem, all of which are leveraged throughout this work.

Theorem A.1 (Operator Bernstein Inequality [38,94])**.** *Let $\boldsymbol{X}_i \in \mathbb{R}^{n\times n}$, $i = 1, \cdots, m$ be independent, zero-mean, matrix-valued random variables, and let $\sigma^2 \geq \max\left\{\left\|\sum_{i=1}^m \mathbb{E}\left(\boldsymbol{X}_i\boldsymbol{X}_i^\top\right)\right\|, \left\|\sum_{i=1}^m \mathbb{E}\left(\boldsymbol{X}_i^\top\boldsymbol{X}_i\right)\right\|\right\}$. Assume there exists a $c \in \mathbb{R}$ such that $\|\boldsymbol{X}_i\| \leq c$ almost surely. Then for $t > 0$*

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

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

$$\mathbb{P}\left(\left\|\sum_{i=1}^m \boldsymbol{X}_i\right\| > t\right) \leq 2n\exp\left(-\frac{3t^2}{8\sigma^2}\right); \tag{30}$$

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

$$\mathbb{P}\left(\left\|\sum_{i=1}^m \boldsymbol{X}_i\right\| > t\right) \leq 2n\exp\left(-\frac{3t}{8c}\right). \tag{31}$$

Theorem A.2 (Scalar Bernstein Inequality [53])**.** *Let $Y_1, \cdots, Y_n$ be independent, mean zero random variables such that $|Y_i| \leq R$ for all i, and let $\sigma^2 = \mathbb{E}\left[\sum_{i=1}^n Y_i^2\right]$. Then*

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

and if $t \leq \frac{\sigma^2}{R}$ this simplifies to

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

Theorem A.3 (Hanson-Wright Inequality [95])**.** *Let* $\boldsymbol{Y} = (Y_1 \cdots Y_n) \in \mathbb{R}^n$ *be a random vector with* Y_i *independent and* $\mathbb{E}[Y_i] = 0$, *and* $\|Y_i\|_{\psi_2} \leq K$ *for some* $K \geq 0$, *where* $\|\cdot\|_{\psi_2}$ *is the sub-Gaussian norm. Additionally, let* $\boldsymbol{A} \in \mathbb{R}^{n \times n}$. *Then*

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

Theorem A.4 (Davis-Kahan $\sin\Theta$ Theorem [96])**.** *Let* $\boldsymbol{X}, \hat{\boldsymbol{X}} \in \mathbb{R}^{n \times n}$ *be symmetric matrices with eigenvalues* $\lambda_1 \geq \cdots \geq \lambda_n$ *and* $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_n$, *respectively. Fix* $1 \leq r \leq s \leq n$, *and let* $\boldsymbol{V}$ *and* $\hat{\boldsymbol{V}}$ *be* $n \times (s - r + 1)$ *matrices with orthonormal columns corresponding to eigenvectors with eigenvalues* $\{\lambda_j\}_{j=r}^s$ *and* $\{\hat{\lambda}_j\}_{j=r}^s$, *respectively, and let* $\mathbb{V}, \hat{\mathbb{V}}$ *be the subspaces spanned by the columns of* $\boldsymbol{V}$ *and* $\hat{\boldsymbol{V}}$. *Define the eigengap as*

$$\delta = \inf\left\{\left|\lambda - \hat{\lambda}\right| : \lambda \in [\lambda_s, \lambda_r], \hat{\lambda} \in \left(-\infty, \hat{\lambda}_{s+1}\right) \cup \left(\hat{\lambda}_{r-1}, \infty\right)\right\},$$

where $\hat{\lambda}_0 = \infty$ *and* $\hat{\lambda}_{n+1} = -\infty$. *If* $\delta > 0$, *then*

$$\|\sin\Theta(\mathbb{V}, \hat{\mathbb{V}})\|_{\mathrm{F}} \leq \frac{\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_{\mathrm{F}}}{\delta}.$$

In particular, for rank-r matrices $\boldsymbol{X}, \hat{\boldsymbol{X}} \succeq \boldsymbol{0}$ *with eigenvectors corresponding to non-zero eigenvalues forming the columns of* $\boldsymbol{V}, \hat{\boldsymbol{V}}$, $\delta = \lambda_r$ *and*

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

ne 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.5 (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. We can see that

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

$$= \max_{\|\boldsymbol{Y}\|_{\mathrm{F}}=1}\mathrm{vec}(\boldsymbol{Y})^\top \left(\sum_{k=1}^m \mathrm{vec}(\boldsymbol{Z}_k)\mathrm{vec}(\boldsymbol{Z}_k)^\top\right)\mathrm{vec}(\boldsymbol{Y})$$

$$= \max_{\|\boldsymbol{Y}\|_{\mathrm{F}}=1}\mathrm{vec}(\boldsymbol{Y})^\top \boldsymbol{Z}_{\mathbb{V}}\boldsymbol{Z}_{\mathbb{V}}^\top \mathrm{vec}(\boldsymbol{Y}).$$

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. $\square$

Lemma A.6 ($\lambda_{\max}(\tilde{\boldsymbol{H}})$ bound)**.** *Let* $\tilde{\boldsymbol{H}} = [\langle \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}, \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\beta}}\rangle] \in \mathbb{R}^{L \times L}$, *where* U *is the row/column space of the true solution* $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top$, *which is rank-r, and where* $\mathcal{P}_U$ *is the projection operator onto* U. *It follows that*

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

Proof. First, by incoherence we have that

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

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

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

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

Lemma A.7. *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 $\|\boldsymbol{X}\|_{\text{F}} = 1$,

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

Proof. 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 Theorem A.5 and the fact that $\mathcal{P}_U$ is an orthogonal projection operator. This concludes the proof. $\qquad\square$

Lemma A.8 (Eigenvalues of $\boldsymbol{H}$ and $\boldsymbol{H}^{-1}$, entries of $\boldsymbol{H}^{-1}$, and spectral norms of $\boldsymbol{w}_{\boldsymbol{\alpha}}$ and $\boldsymbol{v}_{\boldsymbol{\alpha}}$ [50]). *Let $\boldsymbol{H} \in \mathbb{R}^{L \times L}$ be the Gram matrix for $\{\boldsymbol{w}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathbb{I}}$ defined by $H_{\boldsymbol{\alpha}\boldsymbol{\beta}} = \langle \boldsymbol{w}_{\boldsymbol{\alpha}}, \boldsymbol{w}_{\boldsymbol{\beta}} \rangle$, and let $\boldsymbol{H}^{-1}$ be its inverse. Then*

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

Additionally,

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

Finally,

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

Lemma A.9. *Let $\{\boldsymbol{v}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathbb{I}}$ be the dual basis to $\{\boldsymbol{w}_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathbb{I}}$. It follows that*

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

Proof. Recall that $\boldsymbol{v}_{\boldsymbol{\alpha}} = -\frac{1}{2}\left(\boldsymbol{a}\boldsymbol{b}^\top + \boldsymbol{b}\boldsymbol{a}^\top\right)$ where $\boldsymbol{a} = \boldsymbol{e}_i - \frac{1}{n}\boldsymbol{1}$ 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

$$4\boldsymbol{v}_{\boldsymbol{\alpha}}^2 = \frac{n-1}{n}\left[\left(\boldsymbol{e}_{ii} - \frac{1}{n}\boldsymbol{e}_i\boldsymbol{1}^\top - \frac{1}{n}\boldsymbol{1}\boldsymbol{e}_i^\top + \frac{1}{n^2}\boldsymbol{1}\boldsymbol{1}^\top\right) + \left(\boldsymbol{e}_{jj} - \frac{1}{n}\boldsymbol{e}_j\boldsymbol{1}^\top - \frac{1}{n}\boldsymbol{1}\boldsymbol{e}_j^\top + \frac{1}{n^2}\boldsymbol{1}\boldsymbol{1}^\top\right)\right]$$

29

$$-\frac{1}{n}\left[\left(\boldsymbol{e}_{ij}-\frac{1}{n}\boldsymbol{e}_i\mathbf{1}^\top-\frac{1}{n}\mathbf{1}\boldsymbol{e}_j+\frac{1}{n^2}\mathbf{1}\mathbf{1}^\top\right)+\left(\boldsymbol{e}_{ji}-\frac{1}{n}\boldsymbol{e}_j\mathbf{1}^\top-\frac{1}{n}\mathbf{1}\boldsymbol{e}_i^\top+\frac{1}{n^2}\mathbf{1}\mathbf{1}^\top\right)\right]$$

$$=\frac{n-1}{n}\left(\boldsymbol{e}_{ii}+\boldsymbol{e}_{jj}\right)+\frac{2-n}{n^3}\left(\boldsymbol{e}_i\mathbf{1}^\top+\mathbf{1}\boldsymbol{e}_i^\top+\boldsymbol{e}_j\mathbf{1}^\top+\mathbf{1}\boldsymbol{e}_j^\top\right)+\frac{2(n-2)}{n^2}\mathbf{1}\mathbf{1}^\top-\frac{1}{n}\left(\boldsymbol{e}_{ij}+\boldsymbol{e}_{ji}\right).$$

So it follows that

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

$$=\frac{n^2-2n+2}{n}\boldsymbol{I}-\frac{n^2-2n+2}{n^2}\mathbf{1}\mathbf{1}^\top,$$

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

Lemma A.10 (Bounds for Projections [54,97]). *Let $\boldsymbol{X}_l=\boldsymbol{U}_l\boldsymbol{D}_l\boldsymbol{U}_l^\top$ be a rank-r matrix and $\mathbb{T}_l$ be the tangent space of $\mathcal{N}_r$ at $\boldsymbol{X}_l$. Let $\boldsymbol{X}=\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top$ be another rank-r matrix, and $\mathbb{T}$ be the corresponding tangent space. Then*

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

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

B   Restricted Isometry Results

As RIP and its variants are critical to the analysis of Algorithm 1 in this paper, this section is dedicated to the proofs of RIP and similar results. We begin with a demonstration that $\mathbb{E}\left[\mathcal{R}_\Omega^*\mathcal{R}_\Omega\right]\ne p^2\mathcal{I}$, and that $\mathbb{E}\left[\mathcal{M}_\Omega\right]=p^2\mathcal{I}$.

Lemma B.1 (Expectation of $\mathcal{M}_\Omega$). *Let $\mathcal{M}_\Omega$ be as defined in (17), and let $\mathcal{R}_\Omega^*\mathcal{R}_\Omega(\cdot)=\sum_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\Omega}\langle\cdot,\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\beta}}\rangle\boldsymbol{w}_{\boldsymbol{\beta}}$. Then $\mathbb{E}\left[\mathcal{R}_\Omega^*\mathcal{R}_\Omega\right]\ne p^2\mathcal{I}$, and $\mathbb{E}\left[\mathcal{M}_\Omega\right]=p^2\mathcal{I}$.*

Proof. First, notice that

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

$$=\sum_{\boldsymbol{\alpha}\in\Omega}\langle\cdot,\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\alpha}}\rangle\boldsymbol{w}_{\boldsymbol{\alpha}}+\sum_{\boldsymbol{\alpha}\in\Omega}\langle\cdot,\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\sum_{\boldsymbol{\beta}\in\Omega,\boldsymbol{\beta}\ne\boldsymbol{\alpha}}\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\beta}}\rangle\boldsymbol{w}_{\boldsymbol{\beta}}.$$

This was decomposed suggestively into diagonal and off-diagonal elements of the matrix $\boldsymbol{H}^{-1}$ for the following reason. Previously, it was believed that $\mathbb{E}[\mathcal{R}_\Omega^*\mathcal{R}_\Omega]=p^2\mathcal{I}$, as it is true that $\mathcal{R}_{\mathbb{I}}^*\mathcal{R}_{\mathbb{I}}=\mathcal{I}^2=\mathcal{I}$ [51]. Let us consider the problem of computing the expectation of $\mathcal{R}_\Omega^*\mathcal{R}_\Omega$ in more detail now. We will assume that each entry of $\mathbb{I}$ is sampled with Bernoulli probability p, contrasted to the uniformly at random with replacement sampling strategy employed in [51]. Now we see that

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

$$=p\sum_{\boldsymbol{\alpha}\in\mathbb{I}}\langle\cdot,\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\alpha}}\rangle\boldsymbol{w}_{\boldsymbol{\alpha}}+p^2\sum_{\boldsymbol{\alpha}\in\mathbb{I}}\langle\cdot,\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\sum_{\boldsymbol{\beta}\in\mathbb{I},\boldsymbol{\beta}\ne\boldsymbol{\alpha}}\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\beta}}\rangle\boldsymbol{w}_{\boldsymbol{\beta}}.$$

The reason for this difference is that the probability of each entry being selected in the diagonal terms is given by p, as per the Bernoulli definition. However, the probability of the off-diagonal elements requires looking at the probability of sampling two distinct entries at a time. This probability, for each of these options, is p^2. If we let $p=\frac{m}{L}$, where $\mathbb{E}|\Omega|=m$, p^2 is the exact scaling originally expected in [51] for the expectation of the operator to hold. This is incorrect however, and does not recognize that the diagonal entries are more likely to be sampled, as they only require the contributions of a single sample. As such, introducing a rescaling to de-bias the above operator leads to the definition of

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

where

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

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

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

It now follows that

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

as $\mathcal{R}_\mathbb{I} = \mathcal{I}$, thus concluding the proof. $\qquad\square$

Next, we define the following operator $\mathcal{S}_\Omega$:

Definition B.2 (Definition of $\mathcal{S}_\Omega$). *We define the map $\mathcal{S}_\Omega : \mathbb{R}^{n\times n} \to \mathbb{R}^L$ as*

$$
\begin{aligned}
(\mathcal{S}_\Omega(\boldsymbol{X}))_{\boldsymbol{\alpha}} &= \begin{cases} \langle \boldsymbol{X}, \boldsymbol{w_\alpha}\rangle & \boldsymbol{\alpha} \in \Omega \\ 0 & \boldsymbol{\alpha} \notin \Omega \end{cases} \\
&= \xi_{\boldsymbol{\alpha}} \langle \boldsymbol{X}, \boldsymbol{w_\alpha}\rangle,
\end{aligned}
\tag{32}
$$

where $\{\xi_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$ are i.i.d. Bernoulli random variables that are 1 with probability p and 0 with probability $1 - p$.

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

Definition B.3 (Definition of $\boldsymbol{H}_{\text{offdiag}}^{-1}$). *Let $\boldsymbol{H}^{-1} \in \mathbb{R}^{L\times L}$ be as defined previously. We define the following matrix $\boldsymbol{H}_{\text{offdiag}}^{-1}$ as follows:*

$$(H_{\text{offdiag}}^{-1})_{\boldsymbol{\alpha\beta}} = \begin{cases} H^{\alpha\beta} & \boldsymbol{\alpha} \neq \boldsymbol{\beta}; \\ 0 & \boldsymbol{\alpha} = \boldsymbol{\beta}. \end{cases} \tag{33}$$

For the application of Hanson-Wright to the proof of RIP, we need to compute the sub-Gaussian norm of $\mathcal{S}_\Omega(\boldsymbol{X})$.

Lemma B.4. *Let $\xi_{\boldsymbol{\alpha}}$ be a Bernoulli random variable that takes 1 with probability p, and 0 otherwise. Then for all $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$*

$$\|\xi_{\boldsymbol{\alpha}}\langle \boldsymbol{A}, \boldsymbol{B}\rangle\|_{\psi_2} \leq c\sqrt{p}\|\boldsymbol{A}\|_{\mathrm{F}}\|\boldsymbol{B}\|_{\mathrm{F}}$$

for some absolute constant c. Additionally, we have that

$$\|\xi_{\boldsymbol{\alpha}}\langle \boldsymbol{A}, \boldsymbol{B}\rangle - \mathbb{E}\left[\xi_{\boldsymbol{\alpha}}\langle \boldsymbol{A}, \boldsymbol{B}\rangle\right]\|_{\psi_2} \leq C\sqrt{p}\|\boldsymbol{A}\|_{\mathrm{F}}\|\boldsymbol{B}\|_{\mathrm{F}}$$

for some absolute constant $C > 0$.

Proof. We will use a moment generating function bound to prove this result. It is stated in [53] that a random variable is sub-Gaussian if there exists a constant K such that, for all $\lambda \leq \frac{1}{K}$,

$$\mathbb{E}\left[\exp\left(\lambda^2 Y^2\right)\right] \leq \exp(\lambda^2 K^2).$$

We note that this constant K is related to the sub-Gaussian norm, denoted $\|\cdot\|_{\psi_2}$ on an Orlicz space by an absolute constant c.

We will use this technique to bound the sub-Gaussian norm of $\xi_{\boldsymbol{\alpha}}\langle\boldsymbol{A},\boldsymbol{B}\rangle$. Notice that

$$
\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda^2\xi_{\boldsymbol{\alpha}}^2\langle\boldsymbol{A},\boldsymbol{B}\rangle^2\right)\right] &= \mathbb{E}\left[\exp\left(\lambda^2\xi_{\boldsymbol{\alpha}}\langle\boldsymbol{A},\boldsymbol{B}\rangle^2\right)\right] \\
&= \exp\left(\lambda^2 p\langle\boldsymbol{A},\boldsymbol{B}\rangle^2\right) \\
&\leq \exp\left(\lambda^2 p\|\boldsymbol{A}\|_{\mathrm{F}}^2\|\boldsymbol{B}\|_{\mathrm{F}}^2\right),
\end{aligned}
$$

where the second equality follows from the definition of $\xi_{\boldsymbol{\alpha}}$ as a Bernoulli random variable and the inequality follows from Cauchy-Schwarz and the monotonicity of the exponential. The result follows by setting $K = \sqrt{p}\|\boldsymbol{A}\|_{\mathrm{F}}\|\boldsymbol{B}\|_{\mathrm{F}}$. The final result follows from Lemma 2.6.8 in [53]. $\qquad\square$

Lemma B.5 (Reformulation of $\langle\boldsymbol{Y},\mathcal{M}_\Omega(\boldsymbol{Y})\rangle$). *For any $\boldsymbol{Y}\in\mathbb{R}^{n\times n}$, we have that*

$$
\langle\boldsymbol{Y},\mathcal{M}_\Omega(\boldsymbol{Y})\rangle = p\|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2\langle\boldsymbol{Y},\mathcal{F}_\Omega(\boldsymbol{Y})\rangle + \mathcal{S}_\Omega(\boldsymbol{Y})^\top\boldsymbol{H}_{\mathrm{offdiag}}^{-1}\mathcal{S}_\Omega(\boldsymbol{Y})
$$

Proof. Notice the following:

$$
\mathcal{S}_\Omega(\boldsymbol{X})^\top\boldsymbol{H}_{\mathrm{offdiag}}^{-1}\mathcal{S}_\Omega(\boldsymbol{X})
$$

$$
= \begin{pmatrix}\xi_1\langle\boldsymbol{X},\boldsymbol{w}_1\rangle & \cdots & \xi_L\langle\boldsymbol{X},\boldsymbol{w}_L\rangle\end{pmatrix}\begin{pmatrix}0 & \langle\boldsymbol{v}_1,\boldsymbol{v}_2\rangle & \cdots & \langle\boldsymbol{v}_1,\boldsymbol{v}_L\rangle \\ \langle\boldsymbol{v}_2,\boldsymbol{v}_1\rangle & 0 & & \vdots \\ \vdots & & \ddots & \langle\boldsymbol{v}_{L-1},\boldsymbol{v}_L\rangle \\ \langle\boldsymbol{v}_1,\boldsymbol{v}_L\rangle & \cdots & \langle\boldsymbol{v}_L,\boldsymbol{v}_{L-1}\rangle & 0\end{pmatrix}\begin{pmatrix}\xi_1\langle\boldsymbol{X},\boldsymbol{w}_1\rangle \\ \vdots \\ \xi_L\langle\boldsymbol{X},\boldsymbol{w}_L\rangle\end{pmatrix}
$$

$$
= \begin{pmatrix}\xi_1\langle\boldsymbol{X},\boldsymbol{w}_1\rangle & \cdots & \xi_L\langle\boldsymbol{X},\boldsymbol{w}_L\rangle\end{pmatrix}\begin{pmatrix}\sum_{\boldsymbol{\alpha}\in\Omega,\alpha\neq 1}\langle\boldsymbol{X},\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\langle\boldsymbol{v}_1,\boldsymbol{v}_{\boldsymbol{\alpha}}\rangle \\ \sum_{\boldsymbol{\alpha}\in\Omega,\alpha\neq 2}\langle\boldsymbol{X},\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\langle\boldsymbol{v}_2,\boldsymbol{v}_{\boldsymbol{\alpha}}\rangle \\ \vdots \\ \sum_{\boldsymbol{\alpha}\in\Omega,\alpha\neq L}\langle\boldsymbol{X},\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\langle\boldsymbol{v}_L,\boldsymbol{v}_{\boldsymbol{\alpha}}\rangle\end{pmatrix}
$$

$$
= \sum_{\substack{\alpha,\beta\in\Omega\\\alpha\neq\beta}}\langle\boldsymbol{X},\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle\boldsymbol{X},\boldsymbol{w}_{\boldsymbol{\beta}}\rangle.
$$

As such, we can see that

$$
\begin{aligned}
\langle\boldsymbol{X},\mathcal{M}_\Omega(\boldsymbol{X})\rangle &= \left\langle\boldsymbol{X},p\sum_{\boldsymbol{\alpha}\in\Omega}\langle\boldsymbol{X},\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\alpha}}\rangle\boldsymbol{w}_{\boldsymbol{\alpha}}\right\rangle + \left\langle\boldsymbol{X},\sum_{\boldsymbol{\alpha}\in\Omega}\langle\cdot,\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\sum_{\substack{\alpha,\beta\in\Omega\\\alpha\neq\beta}}\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\beta}}\rangle\boldsymbol{w}_{\boldsymbol{\beta}}\right\rangle \\
&= p\sum_{\boldsymbol{\alpha}\in\Omega}\langle\boldsymbol{X},\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\alpha}}\rangle + \sum_{\substack{\alpha,\beta\in\Omega\\\alpha\neq\beta}}\langle\boldsymbol{X},\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle\boldsymbol{X},\boldsymbol{w}_{\boldsymbol{\beta}}\rangle \\
&= p\|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2\langle\boldsymbol{X},\mathcal{F}_\Omega(\boldsymbol{X})\rangle + \mathcal{S}_\Omega(\boldsymbol{X})^\top\boldsymbol{H}_{\mathrm{offdiag}}^{-1}\mathcal{S}_\Omega(\boldsymbol{X}),
\end{aligned}
$$

thus concluding the proof. $\qquad\square$

We will need the following two results to compute the RIP of $\mathcal{M}_\Omega$:

Lemma B.6. *Let $\hat{\mathcal{F}}_\Omega$ be either $\mathcal{P}_{\mathbb{T}}\mathcal{F}_\Omega\mathcal{P}_{\mathbb{T}}$ or $\mathcal{F}_\Omega$, and let $\hat{\mathcal{F}}_{\mathbb{I}}$ be $\mathcal{P}_{\mathbb{T}}\mathcal{F}_{\mathbb{I}}\mathcal{P}_{\mathbb{T}}$ or $\mathcal{F}_{\mathbb{I}}$, respectively. Let $\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}$ be either $\mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}$ or $\boldsymbol{w}_{\boldsymbol{\alpha}}$, respectively. Let $\hat{\boldsymbol{H}} = [\langle\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}},\hat{\boldsymbol{w}}_{\boldsymbol{\beta}}\rangle]\in\mathbb{R}^{L\times L}$ be the corresponding correlation matrix. For a ground truth rank r ν-incoherent matrix $\boldsymbol{X}$ with tangent space $\mathbb{T}$ on $\mathcal{N}_r$, we have that for any $\beta > 1$, and with probability at least $1 - 2n^{1-\beta}$, and for $p\geq\frac{4\beta\log n}{3n}$, that*

$$
\|\hat{\mathcal{F}}_\Omega - p\hat{\mathcal{F}}_{\mathbb{I}}\| \leq p\sqrt{\frac{8\beta\left(\max_{\boldsymbol{\alpha}}\|\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2\right)\lambda_{\max}(\hat{\boldsymbol{H}})\log n}{3p}}.
$$

Proof. This proof will follow a standard Bernstein argument. In order to make this argument, we need to build out $\hat{\mathcal{F}}_\Omega$ as a sum of random operators, we need to bound each, and then to bound the variance term. First, let $\{\xi_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{I}}$ be i.i.d. Bernoulli random variables that are 1 with probability p and 0 with probability $1 - p$. It follows then that

$$
\hat{\mathcal{F}}_\Omega(\cdot) = \sum_{\boldsymbol{\alpha}\in\mathbb{I}}\xi_{\boldsymbol{\alpha}}\langle\cdot,\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\rangle\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}},
$$

and

$$\mathbb{E}[\hat{\mathcal{F}}_{\Omega}] = p \sum_{\boldsymbol{\alpha} \in \mathbb{I}} \langle \cdot, \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}} \rangle \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}} = p \hat{\mathcal{F}}_{\mathbb{I}}.$$

Now, to prove concentration of the desired sum, let

$$\boldsymbol{S}_{\boldsymbol{\alpha}} = (\xi_{\boldsymbol{\alpha}} - p) \langle \cdot, \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}} \rangle \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}},$$

and notice that $\sum_{\boldsymbol{\alpha} \in \mathbb{I}} \boldsymbol{S}_{\boldsymbol{\alpha}} = \hat{\mathcal{F}}_{\Omega} - p \hat{\mathcal{F}}_{\mathbb{I}}$. We can now use a Bernstein inequality to bound the deviation of the spectral norm from 0. Now, first notice that

$$\begin{aligned}
\|\boldsymbol{S}_{\boldsymbol{\alpha}}\| &= \|(\xi_{\boldsymbol{\alpha}} - p) \langle \cdot, \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}} \rangle \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\| \\
&\leq \|\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 \\
&\leq \max_{\boldsymbol{\alpha}} \|\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 =: c,
\end{aligned}$$

where the last inequality follows from Theorem 5.1. Next, we seek to bound the variance term, $\sigma^2 = \left\| \sum_{\boldsymbol{\alpha} \in \mathbb{I}} \mathbb{E} \left[\boldsymbol{S}_{\boldsymbol{\alpha}}^2 \right] \right\|$. To see this, first notice that

$$\begin{aligned}
\left\| \mathbb{E} \left[\sum_{\boldsymbol{\alpha} \in \mathbb{I}} \boldsymbol{S}_{\boldsymbol{\alpha}}^2 \right] \right\| &= \left\| \mathbb{E} \left[\sum_{\boldsymbol{\alpha} \in \mathbb{I}} (\xi_{\boldsymbol{\alpha}} - p)^2 \langle \cdot, \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}} \rangle \langle \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}, \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}} \rangle \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}} \right] \right\| \\
&= \left\| \mathbb{E} \left[\sum_{\boldsymbol{\alpha} \in \mathbb{I}} (\xi_{\boldsymbol{\alpha}} - 2\xi_{\boldsymbol{\alpha}} p + p^2) \langle \cdot, \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}} \rangle \langle \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}, \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}} \rangle \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}} \right] \right\| \\
&\leq p(1-p) \left(\max_{\boldsymbol{\alpha}} \|\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 \right) \left\| \sum_{\boldsymbol{\alpha} \in \mathbb{I}} \langle \cdot, \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}} \rangle \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}} \right\| \\
&\leq p \left(\max_{\boldsymbol{\alpha}} \|\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 \right) \lambda_{\max}(\hat{\boldsymbol{H}}).
\end{aligned}$$

Now, for $t \leq \frac{\sigma^2}{c} = \lambda_{\max}(\hat{\boldsymbol{H}})p$, we can see that for $p \geq \frac{8\beta \log n}{3n}$ that

$$\mathbb{P} \left(\left\| \sum_{\boldsymbol{\alpha} \in \mathbb{I}} \boldsymbol{S}_{\boldsymbol{\alpha}} \right\| \geq p \sqrt{\frac{8 \left(\max_{\boldsymbol{\alpha}} \|\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 \right) \lambda_{\max}(\hat{\boldsymbol{H}}) \beta \log n}{3p}} \right) \leq 2n \exp \left(-\beta \log n \right) = 2n^{1-\beta},$$

as stated previously. $\qquad \square$

Lemma B.7. *For a ground truth rank-r, ν-incoherent matrix $\boldsymbol{X}$ with tangent space $\mathbb{T}$ on $\mathcal{N}_r$, we have that for any $\beta > 1$ that if $p \geq \frac{4}{3} \frac{\beta \log n}{n}$, with probability at least $1 - 2n^{1-\beta}$*

$$\|\mathcal{F}_{\Omega} \mathcal{P}_{\mathbb{T}} - p \mathcal{F}_{\mathbb{I}} \mathcal{P}_{\mathbb{T}}\| \leq \sqrt{\frac{32 p \nu r \beta \log n}{3}}.$$

Proof. We will prove this result using Theorem A.1. First, notice that

$$\mathcal{F}_{\Omega} \mathcal{P}_{\mathbb{T}}(\cdot) - p \mathcal{F}_{\mathbb{I}} \mathcal{P}_{\mathbb{T}}(\cdot) = \sum_{\boldsymbol{\alpha} \in \mathbb{I}} (\xi_{\boldsymbol{\alpha}} - p) \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \boldsymbol{w}_{\boldsymbol{\alpha}},$$

so this is a sum of zero mean independent random variables and Bernstein's inequality holds. We define

$$\boldsymbol{J}_{\boldsymbol{\alpha}} = (\xi_{\boldsymbol{\alpha}} - p) \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \boldsymbol{w}_{\boldsymbol{\alpha}}.$$

Next, notice that

$$\begin{aligned}
|\boldsymbol{J}_{\boldsymbol{\alpha}}| &= |(\xi_{\boldsymbol{\alpha}} - p) \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \boldsymbol{w}_{\boldsymbol{\alpha}}| \\
&\leq |\xi_{\boldsymbol{\alpha}} \langle \cdot, \mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle \boldsymbol{w}_{\boldsymbol{\alpha}}| \\
&\leq \|\mathcal{P}_{\mathbb{T}} \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \|\boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \\
&\leq \sqrt{\frac{2\nu r}{n}},
\end{aligned}$$

33

where the first inequality follows from dropping the negative term, and the third inequality follows from Theorem 5.1 and $\|\boldsymbol{w_\alpha}\|_\mathrm{F} = 2$. Next, we note that

$$
\begin{aligned}
\mathbb{E}\left[\sum_{\boldsymbol{\alpha}\in\mathbb{I}} \boldsymbol{J_\alpha}\boldsymbol{J_\alpha^*}\right] &= \mathbb{E}\left[\sum_{\boldsymbol{\alpha}\in\mathbb{I}} (\xi_{\boldsymbol{\alpha}} - p)^2 \langle\cdot,\boldsymbol{w_\alpha}\rangle\langle\mathcal{P}_\mathbb{T}\boldsymbol{w_\alpha},\mathcal{P}_\mathbb{T}\boldsymbol{w_\alpha}\rangle\boldsymbol{w_\alpha}\right] \\
&= \mathbb{E}\left[\sum_{\boldsymbol{\alpha}\in\mathbb{I}} (\xi_{\boldsymbol{\alpha}}(1 - 2p) + p^2)\langle\cdot,\boldsymbol{w_\alpha}\rangle\langle\mathcal{P}_\mathbb{T}\boldsymbol{w_\alpha},\mathcal{P}_\mathbb{T}\boldsymbol{w_\alpha}\rangle\boldsymbol{w_\alpha}\right] \\
&= \sum_{\boldsymbol{\alpha}\in\mathbb{I}} p(1 - p)\langle\cdot,\boldsymbol{w_\alpha}\rangle\langle\mathcal{P}_\mathbb{T}\boldsymbol{w_\alpha},\mathcal{P}_\mathbb{T}\boldsymbol{w_\alpha}\rangle\boldsymbol{w_\alpha},
\end{aligned}
$$

so

$$
\begin{aligned}
\left\|\mathbb{E}\left[\sum_{\boldsymbol{\alpha}\in\mathbb{I}} \boldsymbol{J_\alpha}\boldsymbol{J_\alpha^*}\right]\right\| &= \left\|\sum_{\boldsymbol{\alpha}\in\mathbb{I}} p(1 - p)\langle\cdot,\boldsymbol{w_\alpha}\rangle\langle\mathcal{P}_\mathbb{T}\boldsymbol{w_\alpha},\mathcal{P}_\mathbb{T}\boldsymbol{w_\alpha}\rangle\boldsymbol{w_\alpha}\right\| \\
&\le p\frac{\nu r}{2n}\left\|\sum_{\boldsymbol{\alpha}\in\mathbb{I}} \langle\cdot,\boldsymbol{w_\alpha}\rangle\boldsymbol{w_\alpha}\right\| \\
&\le p\frac{\nu r}{2n}\lambda_{\max}(\boldsymbol{H}) \\
&= p\nu r =: \sigma_1^2,
\end{aligned}
$$

where the first inequality follows from Theorem 5.1, the second inequality comes from Theorem A.5, and the final line comes from Theorem A.8. Next, notice that

$$
\begin{aligned}
\mathbb{E}\left[\sum_{\boldsymbol{\alpha}\in\mathbb{I}} \boldsymbol{J_\alpha^*}\boldsymbol{J_\alpha}\right] &= \mathbb{E}\left[\sum_{\boldsymbol{\alpha}\in\mathbb{I}} (\xi_{\boldsymbol{\alpha}} - p)^2 \langle\cdot,\mathcal{P}_\mathbb{T}\boldsymbol{w_\alpha}\rangle\langle\boldsymbol{w_\alpha},\boldsymbol{w_\alpha}\rangle\mathcal{P}_\mathbb{T}\boldsymbol{w_\alpha}\right] \\
&= \mathbb{E}\left[\sum_{\boldsymbol{\alpha}\in\mathbb{I}} (\xi_{\boldsymbol{\alpha}}(1 - 2p) + p^2)\langle\cdot,\mathcal{P}_\mathbb{T}\boldsymbol{w_\alpha}\rangle\langle\boldsymbol{w_\alpha},\boldsymbol{w_\alpha}\rangle\mathcal{P}_\mathbb{T}\boldsymbol{w_\alpha}\right] \\
&= \sum_{\boldsymbol{\alpha}\in\mathbb{I}} p(1 - p)\langle\cdot,\mathcal{P}_\mathbb{T}\boldsymbol{w_\alpha}\rangle\langle\boldsymbol{w_\alpha},\boldsymbol{w_\alpha}\rangle\mathcal{P}_\mathbb{T}\boldsymbol{w_\alpha},
\end{aligned}
$$

so it follows that

$$
\begin{aligned}
\left\|\mathbb{E}\left[\sum_{\boldsymbol{\alpha}\in\mathbb{I}} \boldsymbol{J_\alpha^*}\boldsymbol{J_\alpha}\right]\right\| &= \left\|\sum_{\boldsymbol{\alpha}\in\mathbb{I}} p(1 - p)\langle\cdot,\mathcal{P}_\mathbb{T}\boldsymbol{w_\alpha}\rangle\langle\boldsymbol{w_\alpha},\boldsymbol{w_\alpha}\rangle\mathcal{P}_\mathbb{T}\boldsymbol{w_\alpha}\right\| \\
&\le 4p\left\|\sum_{\boldsymbol{\alpha}\in\mathbb{I}} \langle\cdot,\mathcal{P}_\mathbb{T}\boldsymbol{w_\alpha}\rangle\mathcal{P}_\mathbb{T}\boldsymbol{w_\alpha}\right\| \\
&\le 4p\nu r =: \sigma_2^2,
\end{aligned}
$$

where the first inequality comes from $\|\boldsymbol{w_\alpha}\|_F^2 = 4$, the second inequality comes from Theorem A.5, and the final line comes from $\lambda_{\max}(\tilde{\boldsymbol{H}}) \le \nu r$ in Theorem A.6. Taking $\sigma^2 = \max\{\sigma_1^2, \sigma_2^2\} = 4p\nu r$, we get that for $t = \sqrt{\frac{32p\nu r\beta\log n}{3}}$ and $p \ge \frac{4}{3}\frac{\beta\log n}{n}$,

$$
\begin{aligned}
\mathbb{P}\left[\|\mathcal{F}_\Omega\mathcal{P}_\mathbb{T} - p\mathcal{P}_\mathbb{T}\| \ge \sqrt{\frac{32p\nu r\beta\log n}{3}}\right] &\le 2n\exp\left(\frac{-3}{8(4p\nu r)}\frac{32p\nu r\beta\log n}{3}\right) \\
&= 2n^{1-\beta},
\end{aligned}
$$

thus concluding the proof. $\qquad\square$

Lemma B.8. *Let* $\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}$ *be either* $\mathcal{P}_\mathbb{T}\boldsymbol{w_\alpha}$ *or* $\boldsymbol{w_\alpha}$, *and let* $\hat{\boldsymbol{H}} = [\langle\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}},\hat{\boldsymbol{w}}_{\boldsymbol{\beta}}\rangle] \in \mathbb{R}^{L\times L}$. *Let* $\boldsymbol{Y} \in \mathbb{R}^{n\times n}$ *be any matrix where* $\|\boldsymbol{Y}\|_\mathrm{F} = 1$, *and let* $\hat{Y}$ *be either* $\mathcal{P}_\mathbb{T}\boldsymbol{Y}$ *or* $\boldsymbol{Y}$, *respectively. For a ground truth rank* r ν-*incoherent matrix* $\boldsymbol{X}$ *with tangent space* $\mathbb{T}$ *on* $\mathcal{N}_r$, *and for* $\mathcal{S}_\Omega$ *defined as in* (32) *and* $\boldsymbol{H}_{\mathrm{offdiag}}^{-1}$ *defined as in* (33), *we have that*

for any $\beta > 1$, and some absolute numerical constant $C > 0$, and with probability at least $1 - 4n^{-\beta}$, and for $p \geq \frac{32}{3}(\max_\alpha \|\hat{\boldsymbol{w}}_\alpha\|_{\mathrm{F}}^2)\beta \frac{\log n}{\lambda_{\max}(\tilde{H})} = C_1 \beta \frac{\log n}{n}$ for an $\mathcal{O}(1)$ constant $C_1 > 0$, that

$$\left| \mathcal{S}_\Omega^\top(\hat{\boldsymbol{Y}}) \boldsymbol{H}_{\mathrm{offdiag}}^{-1} \mathcal{S}_\Omega(\hat{\boldsymbol{Y}}) - \mathbb{E}\left[\mathcal{S}_\Omega^\top(\hat{\boldsymbol{Y}}) \boldsymbol{H}_{\mathrm{offdiag}}^{-1} \mathcal{S}_\Omega(\hat{\boldsymbol{Y}}) \right] \right| \leq Cp\|\hat{\boldsymbol{w}}_\alpha\|_{\mathrm{F}}^2 \beta \log n$$

$$+ \; p^2 \sqrt{\frac{128\left(\max_\alpha \|\hat{\boldsymbol{w}}_\alpha\|_{\mathrm{F}}^2\right)\lambda_{\max}(\hat{\boldsymbol{H}})\beta \log n}{3p}}.$$

Proof. To begin, we define $\mathcal{S}_\Omega^C = \mathcal{S}_\Omega - \mathbb{E}[\mathcal{S}_\Omega]$. For any $\boldsymbol{Y}$, we have that

$$\mathcal{S}_\Omega^C(\boldsymbol{Y})^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1} \mathcal{S}_\Omega^C(\boldsymbol{Y}) - \mathbb{E}\left[\mathcal{S}_\Omega^C(\boldsymbol{Y})^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1} \mathcal{S}_\Omega^C(\boldsymbol{Y}) \right]$$

$$= \mathcal{S}_\Omega(\boldsymbol{Y})^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1} \mathcal{S}_\Omega(\boldsymbol{Y}) - 2\mathcal{S}_\Omega(\boldsymbol{Y})^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1} \mathbb{E}\left[\mathcal{S}_\Omega(\boldsymbol{Y})\right] + \mathbb{E}\left[\mathcal{S}_\Omega(\boldsymbol{Y})\right]^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1} \mathbb{E}\left[\mathcal{S}_\Omega(\boldsymbol{Y})\right]$$

$$- \mathbb{E}\left[\mathcal{S}_\Omega^C(\boldsymbol{Y})^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1} \mathcal{S}_\Omega^C(\boldsymbol{Y}) \right]$$

$$= \mathcal{S}_\Omega(\boldsymbol{Y})^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1} \mathcal{S}_\Omega(\boldsymbol{Y}) \; - \; 2\mathcal{S}_\Omega(\boldsymbol{Y})^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1} \mathbb{E}\left[\mathcal{S}_\Omega(\boldsymbol{Y})\right] \; + \; \mathbb{E}\left[\mathcal{S}_\Omega(\boldsymbol{Y})\right]^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1} \mathbb{E}\left[\mathcal{S}_\Omega(\boldsymbol{Y})\right]$$

$$- \; \mathbb{E}\left[\mathcal{S}_\Omega(\boldsymbol{Y})^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1} \mathcal{S}_\Omega(\boldsymbol{Y}) \right] \; + \; 2\mathbb{E}[\mathcal{S}_\Omega(\boldsymbol{Y})]^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1} \mathbb{E}[\mathcal{S}_\Omega(\boldsymbol{Y})] \; - \; \mathbb{E}[\mathcal{S}_\Omega(\boldsymbol{Y})]^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1} \mathbb{E}[\mathcal{S}_\Omega(\boldsymbol{Y})]$$

$$= \mathcal{S}_\Omega(\boldsymbol{Y})^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1} \mathcal{S}_\Omega(\boldsymbol{Y}) - \mathbb{E}\left[\mathcal{S}_\Omega(\boldsymbol{Y})^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1} \mathcal{S}_\Omega(\boldsymbol{Y}) \right] + 2\left(\mathbb{E}\left[\mathcal{S}_\Omega(\boldsymbol{Y})\right] - \mathcal{S}_\Omega(\boldsymbol{Y}) \right)^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1} \mathbb{E}\left[\mathcal{S}_\Omega(\boldsymbol{Y})\right], \quad (34)$$

which implies that, by adding and subtracting $2\left(\mathbb{E}\left[\mathcal{S}_\Omega(\boldsymbol{Y})\right] - \mathcal{S}_\Omega(\boldsymbol{Y}) \right)^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1} \mathbb{E}\left[\mathcal{S}_\Omega(\boldsymbol{Y})\right]$,

$$\left| \mathcal{S}_\Omega(\hat{\boldsymbol{Y}})^\top \; \boldsymbol{H}_{\mathrm{offdiag}}^{-1} \mathcal{S}_\Omega(\hat{\boldsymbol{Y}}) - \mathbb{E}\left[\mathcal{S}_\Omega(\hat{\boldsymbol{Y}})^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1} \mathcal{S}_\Omega(\hat{\boldsymbol{Y}}) \right] \right|$$

$$= \left| \mathcal{S}_\Omega(\hat{\boldsymbol{Y}})^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1} \mathcal{S}_\Omega(\hat{\boldsymbol{Y}}) - \mathbb{E}\left[\mathcal{S}_\Omega(\hat{\boldsymbol{Y}})^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1} \mathcal{S}_\Omega(\hat{\boldsymbol{Y}}) \right] \right.$$

$$\left. + 2\left(\mathbb{E}\left[\mathcal{S}_\Omega(\hat{\boldsymbol{Y}})\right] - \mathcal{S}_\Omega(\hat{\boldsymbol{Y}}) \right)^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1} \mathcal{S}_\Omega(\hat{\boldsymbol{Y}}) - 2\left(\mathbb{E}\left[\mathcal{S}_\Omega(\hat{\boldsymbol{Y}})\right] - \mathcal{S}_\Omega(\hat{\boldsymbol{Y}}) \right)^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1} \mathbb{E}\left[\mathcal{S}_\Omega(\hat{\boldsymbol{Y}})\right] \right|$$

$$\leq \underbrace{\left| \mathcal{S}_\Omega^C(\hat{\boldsymbol{Y}})^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1} \mathcal{S}_\Omega^C(\hat{\boldsymbol{Y}}) - \mathbb{E}\left[\mathcal{S}_\Omega^C(\hat{\boldsymbol{Y}})^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1} \mathcal{S}_\Omega^C(\hat{\boldsymbol{Y}}) \right] \right|}_{B_1}$$

$$+ \underbrace{2\left| \left(\mathbb{E}\left[\mathcal{S}_\Omega(\hat{\boldsymbol{Y}})\right] - \mathcal{S}_\Omega(\hat{\boldsymbol{Y}}) \right)^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1} \mathbb{E}\left[\mathcal{S}_\Omega(\hat{\boldsymbol{Y}})\right] \right|}_{B_2},$$

where the inequality follows from (34) and the triangle inequality.

Bounding B_1: We will compute B_1 using the Hanson-Wright inequality, seen in Theorem A.3. We will first define $\boldsymbol{\xi} \in \mathbb{R}^L$ to be a Bernoulli random vector, where each entry is an i.i.d. Bernoulli random variable with parameter p, i.e. $\boldsymbol{\xi}_\alpha = \xi_\alpha$. Next, we define the following matrix $\boldsymbol{A}_{\hat{Y}} \in \mathbb{R}^{L \times L}$ as $(\boldsymbol{A}_{\hat{Y}})_{\alpha\beta} = \langle \boldsymbol{Y}, \hat{\boldsymbol{w}}_\alpha \rangle \langle \boldsymbol{Y}, \hat{\boldsymbol{w}}_\beta \rangle$. We first remark that

$$\boldsymbol{\xi}^\top(\boldsymbol{A}_{\hat{Y}} \circ \boldsymbol{H}_{\mathrm{offdiag}}^{-1})\boldsymbol{\xi} = \sum_{\substack{\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{I} \\ \boldsymbol{\alpha} \neq \boldsymbol{\beta}}} \xi_\alpha (\boldsymbol{A}_{\hat{Y}} \circ \boldsymbol{H}_{\mathrm{offdiag}}^{-1})_{\alpha\beta} \xi_\beta$$

$$= \sum_{\substack{\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{I} \\ \boldsymbol{\alpha} \neq \boldsymbol{\beta}}} \xi_\alpha \langle \boldsymbol{Y}, \hat{\boldsymbol{w}}_\alpha \rangle \langle \boldsymbol{v}_\alpha, \boldsymbol{v}_\beta \rangle \langle \boldsymbol{Y}, \hat{\boldsymbol{w}}_\beta \rangle \xi_\beta$$

$$= \mathcal{S}_\Omega(\hat{\boldsymbol{Y}})^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1} \mathcal{S}_\Omega(\hat{\boldsymbol{Y}}),$$

where $\circ$ denotes the Hadamard product. Similarly, we can write

$$(\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}])^\top (\boldsymbol{A}_{\hat{Y}} \circ \boldsymbol{H}_{\mathrm{offdiag}}^{-1})(\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}]) = \sum_{\substack{\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{I} \\ \boldsymbol{\alpha} \neq \boldsymbol{\beta}}} (\xi_\alpha - p)\left(\boldsymbol{A}_{\hat{Y}} \circ \boldsymbol{H}_{\mathrm{offdiag}}^{-1} \right)(\xi_\beta - p)$$

$$= \sum_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{I} \\ \boldsymbol{\alpha}\neq\boldsymbol{\beta}}} (\xi_{\boldsymbol{\alpha}} - p)\langle\boldsymbol{Y},\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\rangle\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle\boldsymbol{Y},\hat{\boldsymbol{w}}_{\boldsymbol{\beta}}\rangle(\xi_{\boldsymbol{\beta}} - p)$$

and

$$\begin{aligned}
\mathcal{S}_\Omega^C(\hat{\boldsymbol{Y}})^\top \boldsymbol{H}_{\text{offdiag}}^{-1} \mathcal{S}_\Omega^C(\hat{\boldsymbol{Y}}) &= (\mathcal{S}_\Omega(\hat{\boldsymbol{Y}}) - \mathbb{E}(\mathcal{S}_\Omega(\hat{\boldsymbol{Y}}))^\top \boldsymbol{H}_{\text{offdiag}}^{-1}(\mathcal{S}_\Omega(\hat{\boldsymbol{Y}}) - \mathbb{E}[\mathcal{S}_\Omega(\hat{\boldsymbol{Y}})]) \\
&= \mathcal{S}_\Omega(\hat{\boldsymbol{Y}})^\top \boldsymbol{H}_{\text{offdiag}}^{-1} \mathcal{S}_\Omega(\hat{\boldsymbol{Y}}) - \mathcal{S}_\Omega(\hat{\boldsymbol{Y}})^\top \boldsymbol{H}_{\text{offdiag}}^{-1} \mathbb{E}[\mathcal{S}_\Omega(\hat{\boldsymbol{Y}})] \\
&\quad - \mathbb{E}[\mathcal{S}_\Omega(\hat{\boldsymbol{Y}})]^\top \boldsymbol{H}_{\text{offdiag}}^{-1} \mathcal{S}_\Omega(\hat{\boldsymbol{Y}}) + \mathbb{E}[\mathcal{S}_\Omega(\hat{\boldsymbol{Y}})]^\top \boldsymbol{H}_{\text{offdiag}}^{-1} \mathbb{E}[\mathcal{S}_\Omega(\hat{\boldsymbol{Y}})] \\
&= \sum_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{I} \\ \boldsymbol{\alpha}\neq\boldsymbol{\beta}}} \xi_{\boldsymbol{\alpha}}\langle\boldsymbol{Y},\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\rangle\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle\boldsymbol{Y},\hat{\boldsymbol{w}}_{\boldsymbol{\beta}}\rangle\xi_{\boldsymbol{\beta}} - p\langle\boldsymbol{Y},\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\rangle\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle\boldsymbol{Y},\hat{\boldsymbol{w}}_{\boldsymbol{\beta}}\rangle\xi_{\boldsymbol{\beta}} \\
&\quad - \xi_{\boldsymbol{\alpha}}\langle\boldsymbol{Y},\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\rangle\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle\boldsymbol{Y},\hat{\boldsymbol{w}}_{\boldsymbol{\beta}}\rangle p + p\langle\boldsymbol{Y},\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\rangle\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle\boldsymbol{Y},\hat{\boldsymbol{w}}_{\boldsymbol{\beta}}\rangle p \\
&= \sum_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{I} \\ \boldsymbol{\alpha}\neq\boldsymbol{\beta}}} (\xi_{\boldsymbol{\alpha}} - p)\langle\boldsymbol{Y},\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\rangle\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle\boldsymbol{Y},\hat{\boldsymbol{w}}_{\boldsymbol{\beta}}\rangle(\xi_{\boldsymbol{\beta}} - p) \\
&= (\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}])^\top (\boldsymbol{A}_{\hat{\boldsymbol{Y}}} \circ \boldsymbol{H}_{\text{offdiag}}^{-1})(\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}]).
\end{aligned}$$

As such, we will now proceed to use Theorem A.3 using $(\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}])^\top (\boldsymbol{A}_{\hat{\boldsymbol{Y}}} \circ \boldsymbol{H}_{\text{offdiag}}^{-1})(\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}])$. We note that from Theorem B.4, setting $\boldsymbol{A} = \boldsymbol{B} = \frac{1}{\sqrt{n}}\boldsymbol{I}$ in the lemma statement, that $\|\boldsymbol{\xi}\|_{\psi_2} \leq C\sqrt{p}$ for some absolute constant $C > 0$. Next, we compute that

$$\begin{aligned}
\left\|\boldsymbol{A}_{\hat{\boldsymbol{Y}}} \circ \boldsymbol{H}_{\text{offdiag}}^{-1}\right\|_F^2 &= \sum_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{I} \\ \boldsymbol{\alpha}\neq\boldsymbol{\beta}}} \langle\boldsymbol{Y},\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\rangle^2\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\beta}}\rangle^2\langle\boldsymbol{Y},\hat{\boldsymbol{w}}_{\boldsymbol{\beta}}\rangle^2 \\
&\leq \frac{1}{n^2}\sum_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{I} \\ \boldsymbol{\alpha}\neq\boldsymbol{\beta}}} \langle\boldsymbol{Y},\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\rangle^2\langle\boldsymbol{Y},\hat{\boldsymbol{w}}_{\boldsymbol{\beta}}\rangle^2 \\
&\leq \frac{1}{n^2}\sum_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{I}} \langle\boldsymbol{Y},\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\rangle^2\langle\boldsymbol{Y},\hat{\boldsymbol{w}}_{\boldsymbol{\beta}}\rangle^2 \\
&= \frac{1}{n^2}\sum_{\boldsymbol{\alpha}\in\mathbb{I}} \langle\boldsymbol{Y},\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\rangle^2 \sum_{\boldsymbol{\beta}\in\mathbb{I}} \langle\boldsymbol{Y},\hat{\boldsymbol{w}}_{\boldsymbol{\beta}}\rangle^2 \\
&\leq \frac{1}{n^2}\lambda_{\max}(\hat{\boldsymbol{H}})^2,
\end{aligned}$$

where the first inequality follows from the largest off-diagonal element of $\boldsymbol{H}^{-1}$ from Theorem A.8, the second inequality follows from adding a positive term to the sum, and the third inequality follows from Theorem A.5. Next, we will use a Gershgorin estimate to compute $\left\|\boldsymbol{A}_{\hat{\boldsymbol{Y}}} \circ \boldsymbol{H}_{\text{offdiag}}^{-1}\right\|$ as follows:

$$\begin{aligned}
\left\|\boldsymbol{A}_{\hat{\boldsymbol{Y}}} \circ \boldsymbol{H}_{\text{offdiag}}^{-1}\right\| &\leq \max_{\boldsymbol{\alpha}} \sum_{\boldsymbol{\beta}\neq\boldsymbol{\alpha}} |\langle\boldsymbol{Y},\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\rangle\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle\boldsymbol{Y},\hat{\boldsymbol{w}}_{\boldsymbol{\beta}}\rangle| \\
&\leq (\max_{\boldsymbol{\alpha}} \|\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\|_{\text{F}}^2) \max_{\boldsymbol{\alpha}} \sum_{\boldsymbol{\beta}\neq\boldsymbol{\alpha}} |\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\beta}}\rangle| \\
&\leq 2\max_{\boldsymbol{\alpha}} \|\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\|_{\text{F}}^2,
\end{aligned}$$

where the first inequality follows from Cauchy-Schwarz and the second inequality follows from Theorem A.8. Now, as $\boldsymbol{A}_{\hat{\boldsymbol{Y}}} \circ \boldsymbol{H}_{\text{offdiag}}^{-1}$ is diagonal-free and $\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}]$ is a centered random vector, we can now say that for $t = Cp\max_{\boldsymbol{\alpha}} \|\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\|_{\text{F}}^2\beta\log n$ for some sufficiently large constant $C > 0$,

$$\begin{aligned}
&\mathbb{P}\bigg(\Big|(\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}])^\top (\boldsymbol{A}_{\hat{\boldsymbol{Y}}} \circ \boldsymbol{H}_{\text{offdiag}}^{-1})(\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}])\Big| > Cp\|\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\|_{\text{F}}^2\beta\log n\bigg) \\
&\qquad\qquad\qquad\qquad \leq 2\exp\left(-c\left\{\frac{Cp^2\|\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\|_{\text{F}}^4\beta^2n^2\log^2 n}{\lambda_{\max}(\hat{\boldsymbol{H}})^2}, \frac{Cp\max_{\boldsymbol{\alpha}} \|\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\|_{\text{F}}^2\beta\log n}{2p\max_{\boldsymbol{\alpha}} \|\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\|_{\text{F}}^2}\right\}\right)
\end{aligned}$$

$$\leq 2n^{-\beta},$$

as for both $\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}} = \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}$ or $\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}} = \boldsymbol{w}_{\boldsymbol{\alpha}}$ the minimum is achieved by the term on the right, from Theorem A.8.

Bounding B_2: The next step of this result requires bounding B_2. We will do this using the scalar Bernstein inequality, provided in Theorem A.2. To use this theorem, we need to decompose B_2 as a sum of independent random variables. To do this, notice that

$$\mathbb{E}\left[\mathcal{S}_{\Omega}(\hat{\boldsymbol{Y}})\right]^{\top} \boldsymbol{H}_{\text{offdiag}}^{-1} \mathbb{E}\left[\mathcal{S}_{\Omega}(\hat{\boldsymbol{Y}})\right] = p^2 \sum_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{I} \\ \boldsymbol{\alpha}\neq\boldsymbol{\beta}}} \langle \boldsymbol{Y}, \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}} \rangle \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}} \rangle \langle \boldsymbol{Y}, \hat{\boldsymbol{w}}_{\boldsymbol{\beta}} \rangle,$$

and that

$$\mathcal{S}_{\Omega}(\hat{\boldsymbol{Y}})^{\top} \boldsymbol{H}_{\text{offdiag}}^{-1} \mathbb{E}\left[\mathcal{S}_{\Omega}(\hat{\boldsymbol{Y}})\right] = p \sum_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{I} \\ \boldsymbol{\alpha}\neq\boldsymbol{\beta}}} \xi_{\boldsymbol{\alpha}} \langle \boldsymbol{Y}, \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}} \rangle \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}} \rangle \langle \boldsymbol{Y}, \hat{\boldsymbol{w}}_{\boldsymbol{\beta}} \rangle,$$

so it follows that

$$\left(\mathcal{S}_{\Omega}(\hat{\boldsymbol{Y}}) - \mathbb{E}\left[\mathcal{S}_{\Omega}(\hat{\boldsymbol{Y}})\right]\right)^{\top} \boldsymbol{H}_{\text{offdiag}}^{-1} \mathbb{E}\left[\mathcal{S}_{\Omega}(\hat{\boldsymbol{Y}})\right] = p \sum_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{I} \\ \boldsymbol{\alpha}\neq\boldsymbol{\beta}}} (\xi_{\boldsymbol{\alpha}} - p) \langle \boldsymbol{Y}, \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}} \rangle \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}} \rangle \langle \boldsymbol{Y}, \hat{\boldsymbol{w}}_{\boldsymbol{\beta}} \rangle.$$

Next, let $G_{\boldsymbol{\alpha}} = (\xi_{\boldsymbol{\alpha}} - p) \langle \boldsymbol{Y}, \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}} \rangle \sum_{\substack{\boldsymbol{\beta}\in\mathbb{I} \\ \boldsymbol{\beta}\neq\boldsymbol{\alpha}}} \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}} \rangle \langle \boldsymbol{Y}, \hat{\boldsymbol{w}}_{\boldsymbol{\beta}} \rangle$. Notice that $\mathbb{E}[G_{\boldsymbol{\alpha}}] = 0$ and that, for different indices $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \mathbb{I}$, that $G_{\boldsymbol{\alpha}_1}$ is independent of $G_{\boldsymbol{\alpha}_2}$, so what remains is to bound each term and compute the variance.

First, notice that

$$\begin{aligned}
|G_{\boldsymbol{\alpha}}| &= \left| (\xi_{\boldsymbol{\alpha}} - p) \langle \boldsymbol{Y}, \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}} \rangle \sum_{\substack{\boldsymbol{\beta}\in\mathbb{I} \\ \boldsymbol{\beta}\neq\boldsymbol{\alpha}}} \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}} \rangle \langle \boldsymbol{Y}, \hat{\boldsymbol{w}}_{\boldsymbol{\beta}} \rangle \right| \\
&\leq \left(\max_{\boldsymbol{\alpha}} \|\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\|_{\text{F}}^2 \right) \sum_{\substack{\boldsymbol{\beta}\in\mathbb{I} \\ \boldsymbol{\beta}\neq\boldsymbol{\alpha}}} |\langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}} \rangle| \\
&\leq \left(\max_{\boldsymbol{\alpha}} \|\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\|_{\text{F}}^2 \right) \left(\frac{L}{n^2} + \frac{2n}{n} \right) \\
&\leq 4 \left(\max_{\boldsymbol{\alpha}} \|\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\|_{\text{F}}^2 \right),
\end{aligned}$$

where the second inequality follows from Theorem A.8, and the final inequality is a numerical inequality.

Next, we seek to compute the variance. Notice that as

$$G_{\boldsymbol{\alpha}}^2 = (\xi_{\boldsymbol{\alpha}} - p)^2 \langle \boldsymbol{Y}, \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}} \rangle \left(\sum_{\substack{\boldsymbol{\beta}\in\mathbb{I} \\ \boldsymbol{\beta}\neq\boldsymbol{\alpha}}} \langle \boldsymbol{Y}, \hat{\boldsymbol{w}}_{\boldsymbol{\beta}} \rangle \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}} \rangle \right)^2,$$

we have that

$$\begin{aligned}
\sum_{\boldsymbol{\alpha}\in\mathbb{I}} G_{\boldsymbol{\alpha}}^2 &= \sum_{\boldsymbol{\alpha}\in\mathbb{I}} (\xi_{\boldsymbol{\alpha}} - p)^2 \langle \boldsymbol{Y}, \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}} \rangle^2 \left(\sum_{\substack{\boldsymbol{\beta}\in\mathbb{I} \\ \boldsymbol{\beta}\neq\boldsymbol{\alpha}}} \langle \boldsymbol{Y}, \hat{\boldsymbol{w}}_{\boldsymbol{\beta}} \rangle \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}} \rangle \right)^2 \\
&\leq \sum_{\boldsymbol{\alpha}\in\mathbb{I}} (\xi_{\boldsymbol{\alpha}} - p)^2 \langle \boldsymbol{Y}, \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}} \rangle^2 \left(\sum_{\substack{\boldsymbol{\beta}\in\mathbb{I} \\ \boldsymbol{\beta}\neq\boldsymbol{\alpha}}} |\langle \boldsymbol{Y}, \hat{\boldsymbol{w}}_{\boldsymbol{\beta}} \rangle \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}} \rangle| \right)^2 \\
&\leq \sum_{\boldsymbol{\alpha}\in\mathbb{I}} (\xi_{\boldsymbol{\alpha}} - p)^2 \langle \boldsymbol{Y}, \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}} \rangle^2 \left(\sum_{\substack{\boldsymbol{\beta}\in\mathbb{I} \\ \boldsymbol{\beta}\neq\boldsymbol{\alpha}}} \|\hat{\boldsymbol{w}}_{\boldsymbol{\beta}}\|_{\text{F}} |\langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}} \rangle| \right)^2
\end{aligned}$$

37

$$\leq \left(\max_{\boldsymbol{\beta}} \|\hat{\boldsymbol{w}}_{\boldsymbol{\beta}}\|_{\mathrm{F}}^2\right) \sum_{\boldsymbol{\alpha}\in\mathbb{I}} (\xi_{\boldsymbol{\alpha}} - p)^2 \langle \boldsymbol{Y}, \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\rangle^2 \left(\sum_{\substack{\boldsymbol{\beta}\in\mathbb{I}\\ \boldsymbol{\beta}\neq\boldsymbol{\alpha}}} |\langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle|\right)^2$$

$$\leq \left(\max_{\boldsymbol{\beta}} \|\hat{\boldsymbol{w}}_{\boldsymbol{\beta}}\|_{\mathrm{F}}^2\right) \sum_{\boldsymbol{\alpha}\in\mathbb{I}} (\xi_{\boldsymbol{\alpha}} - p)^2 \langle \boldsymbol{Y}, \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\rangle^2 \left(2 - \frac{15}{2n} + \frac{8}{n^2}\right)^2$$

$$\leq 4\left(\max_{\boldsymbol{\beta}} \|\hat{\boldsymbol{w}}_{\boldsymbol{\beta}}\|_{\mathrm{F}}^2\right) \sum_{\boldsymbol{\alpha}\in\mathbb{I}} (\xi_{\boldsymbol{\alpha}} - p)^2 \langle \boldsymbol{Y}, \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\rangle^2,$$

where the third inequality follows from Assumption Theorem 5.1, and the fourth inequality follows from Lemma 18 in [28], so using the monotonicity of expectation it follows that

$$\sum_{\boldsymbol{\alpha}\in\mathbb{I}} \mathbb{E}[G_{\boldsymbol{\alpha}}^2] \leq 4\left(\max_{\boldsymbol{\beta}} \|\hat{\boldsymbol{w}}_{\boldsymbol{\beta}}\|_{\mathrm{F}}^2\right) p(1-p) \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \boldsymbol{Y}, \hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\rangle^2$$

$$\leq 4\left(\max_{\boldsymbol{\beta}} \|\hat{\boldsymbol{w}}_{\boldsymbol{\beta}}\|_{\mathrm{F}}^2\right) p\|\boldsymbol{Y}\|_F^2 \lambda_{\max}(\hat{\boldsymbol{H}}),$$

where the second inequality follows from Theorem A.5. Letting $t = \frac{p}{2}\sqrt{\frac{128\left(\max \|\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2\right)\lambda_{\max}(\tilde{\boldsymbol{H}})\beta \log n}{3p}}$ for $\beta > 1$, it follows from the scalar Bernstein inequality that, using the specified restriction $p \geq \frac{32}{3}\max_{\alpha} \|\hat{\boldsymbol{w}}_{\alpha}\|_{\mathrm{F}}^2 \frac{\beta \log n}{\lambda_{\max}(\hat{\boldsymbol{H}})}$,

$$\mathbb{P}\left[\left|\sum_{\boldsymbol{\alpha}\in\mathbb{I}} G_{\boldsymbol{\alpha}}\right| \geq \frac{p}{2}\sqrt{\frac{128\left(\max \|\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2\right)\lambda_{\max}(\tilde{\boldsymbol{H}})\beta \log n}{3p}}\right] \leq 2\exp(-\beta \log n) = 2n^{-\beta},$$

and as $\left(\mathcal{S}_{\Omega}(\hat{\boldsymbol{Y}}) - \mathbb{E}\left[\mathcal{S}_{\Omega}(\hat{\boldsymbol{Y}})\right]\right)^{\top} \boldsymbol{H}_{\mathrm{offdiag}}^{-1} \mathbb{E}\left[\mathcal{S}_{\Omega}(\hat{\boldsymbol{Y}})\right] = p\sum_{\boldsymbol{\alpha}\in\mathbb{I}} G_{\boldsymbol{\alpha}}$, it follows that

$$\mathbb{P}\left(\left|\left(\mathcal{S}_{\Omega}(\hat{\boldsymbol{Y}}) - \mathbb{E}\left[\mathcal{S}_{\Omega}(\hat{\boldsymbol{Y}})\right]\right)^{\top} \boldsymbol{H}_{\mathrm{offdiag}}^{-1} \mathbb{E}\left[\mathcal{S}_{\Omega}(\hat{\boldsymbol{Y}})\right]\right| \geq \frac{p^2}{2}\sqrt{\frac{128\left(\max \|\hat{\boldsymbol{w}}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2\right)\lambda_{\max}(\hat{\boldsymbol{H}})\beta \log n}{3p}}\right) \leq 2n^{-\beta},$$

and the lemma statement follows. $\qquad\square$

B.1 Proof of Lemma 5.3

We are now ready to prove Theorem 5.3.

Proof. First, notice that since $\mathcal{M}_{\Omega} = \mathcal{M}_{\Omega}^*$,

$$\|\mathcal{P}_{\mathbb{T}}\mathcal{M}_{\Omega}\mathcal{P}_{\mathbb{T}} - p^2\mathcal{P}_{\mathbb{T}}\| = \max_{\|\boldsymbol{Y}\|_{\mathrm{F}}=1} \left|\langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\mathcal{M}_{\Omega}\mathcal{P}_{\mathbb{T}}\boldsymbol{Y}\rangle - \langle \boldsymbol{Y}, p^2\mathcal{P}_{\mathbb{T}}\boldsymbol{Y}\rangle\right|$$

$$= \max_{\|\boldsymbol{Y}\|_{\mathrm{F}}=1} \left|\sum_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\Omega} \xi_{\boldsymbol{\alpha}}\xi_{\boldsymbol{\beta}} C_{\boldsymbol{\alpha}\boldsymbol{\beta}} \langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\beta}}\rangle - p^2\sum_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{I}} \langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\beta}}\rangle\right|$$

$$= \max_{\|\boldsymbol{Y}\|_{\mathrm{F}}=1} \left|p\sum_{\boldsymbol{\alpha}\in\Omega} \xi_{\boldsymbol{\alpha}}\langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2\langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\alpha}}\rangle + \sum_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta}\in\Omega\\ \boldsymbol{\alpha}\neq\boldsymbol{\beta}}} \xi_{\boldsymbol{\alpha}}\xi_{\boldsymbol{\beta}}\langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\beta}}\rangle\right.$$

$$\left. - p^2\sum_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{I}} \langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\beta}}\rangle\right|$$

$$= \max_{\|\boldsymbol{Y}\|_{\mathrm{F}}=1} \left|p\|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2\langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\mathcal{F}_{\Omega}\mathcal{P}_{\mathbb{T}}\boldsymbol{Y}\rangle + \mathcal{S}_{\Omega}(\mathcal{P}_{\mathbb{T}}\boldsymbol{Y})\boldsymbol{H}_{\mathrm{offdiag}}^{-1}\mathcal{S}_{\Omega}(\mathcal{P}_{\mathbb{T}}\boldsymbol{Y}) - p^2\sum_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{I}} \langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\beta}}\rangle\right|$$

$$= \max_{\|\boldsymbol{Y}\|_{\mathrm{F}}=1} \Big| p\|\boldsymbol{v_\alpha}\|_{\mathrm{F}}^2 \langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\mathcal{F}_\Omega \mathcal{P}_{\mathbb{T}}\boldsymbol{Y}\rangle + \mathcal{S}_\Omega(\mathcal{P}_{\mathbb{T}}\boldsymbol{Y})\boldsymbol{H}_{\mathrm{offdiag}}^{-1}\mathcal{S}_\Omega(\mathcal{P}_{\mathbb{T}}\boldsymbol{Y})$$

$$- p^2 \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w_\alpha}\rangle^2 \|\boldsymbol{v_\alpha}\|_{\mathrm{F}}^2 - p^2 \sum_{\substack{\alpha,\beta\in\mathbb{I}\\ \alpha\neq\beta}} \langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w_\alpha}\rangle\langle \boldsymbol{v_\alpha}, \boldsymbol{v_\beta}\rangle\langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w_\beta}\rangle \Big|$$

$$= \max_{\|\boldsymbol{Y}\|_{\mathrm{F}}=1} \Big| p\|\boldsymbol{v_\alpha}\|_{\mathrm{F}}^2 \left(\langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\mathcal{F}_\Omega \mathcal{P}_{\mathbb{T}}\boldsymbol{Y}\rangle - p\langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\mathcal{F}_{\mathbb{I}}\mathcal{P}_{\mathbb{T}}(\boldsymbol{Y})\rangle\right)$$

$$+ \mathcal{S}_\Omega(\mathcal{P}_{\mathbb{T}}\boldsymbol{Y})^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1}\mathcal{S}_\Omega(\mathcal{P}_{\mathbb{T}}\boldsymbol{Y}) - \mathbb{E}\left[\mathcal{S}_\Omega(\mathcal{P}_{\mathbb{T}}\boldsymbol{Y})^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1}\mathcal{S}_\Omega(\mathcal{P}_{\mathbb{T}}\boldsymbol{Y})\right] \Big|$$

$$\leq \max_{\|\boldsymbol{Y}\|_{\mathrm{F}}=1} \Big| p\|\boldsymbol{v_\alpha}\|_{\mathrm{F}}^2 \left(\langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\mathcal{F}_\Omega \mathcal{P}_{\mathbb{T}}\boldsymbol{Y}\rangle - p\langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\mathcal{F}_{\mathbb{I}}\mathcal{P}_{\mathbb{T}}(\boldsymbol{Y})\rangle\right) \Big|$$

$$+ \max_{\|\boldsymbol{Y}\|_{\mathrm{F}}=1} \Big| \mathcal{S}_\Omega(\mathcal{P}_{\mathbb{T}}\boldsymbol{Y})^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1}\mathcal{S}_\Omega(\mathcal{P}_{\mathbb{T}}\boldsymbol{Y}) - \mathbb{E}\left[\mathcal{S}_\Omega(\mathcal{P}_{\mathbb{T}}\boldsymbol{Y})^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1}\mathcal{S}_\Omega(\mathcal{P}_{\mathbb{T}}\boldsymbol{Y})\right] \Big|$$

$$= \underbrace{\|\boldsymbol{v_\alpha}\|_{\mathrm{F}}^2 p \|\mathcal{P}_{\mathbb{T}}\mathcal{F}_\Omega\mathcal{P}_{\mathbb{T}} - p\mathcal{P}_{\mathbb{T}}\mathcal{F}_{\mathbb{I}}\mathcal{P}_{\mathbb{T}}\|}_{B_1} + \underbrace{\max_{\|\boldsymbol{Y}\|_{\mathrm{F}}=1} \Big| \mathcal{S}_\Omega(\mathcal{P}_{\mathbb{T}}\boldsymbol{Y})^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1}\mathcal{S}_\Omega(\mathcal{P}_{\mathbb{T}}\boldsymbol{Y}) - \mathbb{E}\left[\mathcal{S}_\Omega(\mathcal{P}_{\mathbb{T}}\boldsymbol{Y})^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1}\mathcal{S}_\Omega(\mathcal{P}_{\mathbb{T}}\boldsymbol{Y})\right] \Big|}_{B_2}.$$

The result follows from Lemmas B.6 and B.8, and the fact that $\|\boldsymbol{v_\alpha}\|_{\mathrm{F}}^2 \leq \frac{1}{2}$ from Theorem A.8. $\qquad\square$

Lemma B.9. *Let $\Omega \subset \mathbb{I}$ be sampled with uniform Bernoulli probability p. If $p \geq \frac{64}{3}\frac{\beta \log n}{n}$, then with probability at least $1 - 2n^{1-\beta} - 4n^{-\beta}$ we have that*

$$\|\mathcal{M}_\Omega\| \leq p^2\left(1 + 40\sqrt{\frac{\beta n \log n}{3p}}\right) + Cp\log n$$

for some absolute constant $C > 0$.

Proof. This proof follows directly from Lemmas B.6 and B.8. First, notice that

$$\|\mathcal{M}_\Omega\| = \|\mathcal{M}_\Omega - p^2\mathcal{I} + p^2\mathcal{I}\| \leq p^2 + \|\mathcal{M}_\Omega - p^2\mathcal{I}\|.$$

This second term can be analyzed in the same way as in the proof of Theorem 5.3, seen in Section B.1:

$$\|\mathcal{M}_\Omega - p^2\mathcal{I}\| \leq p\|\boldsymbol{v_\alpha}\|_{\mathrm{F}}^2\|\mathcal{F}_\Omega - p\mathcal{F}_{\mathbb{I}}\| + \max_{\|\boldsymbol{Y}\|_{\mathrm{F}}=1} \Big| \mathcal{S}_\Omega(\boldsymbol{Y})^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1}\mathcal{S}_\Omega(\boldsymbol{Y}) - \mathbb{E}\left[\mathcal{S}_\Omega(\boldsymbol{Y})^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1}\mathcal{S}_\Omega(\boldsymbol{Y})\right] \Big|.$$

Using the fact that $\|\boldsymbol{w_\alpha}\|_{\mathrm{F}}^2 = 4$ and that $\lambda_{\max}(\boldsymbol{H}) = 2n$ from Theorem A.8, the result follows. $\qquad\square$

Lemma B.10. *Let $\boldsymbol{Y}, \boldsymbol{Z} \in \mathbb{R}^{n\times n}$ be any matrix with $\|\boldsymbol{Y}\|_{\mathrm{F}} = \|\boldsymbol{Z}\|_{\mathrm{F}} = 1$. For a rank-$r$, ν-incoherent ground truth matrix $\boldsymbol{X}$ with tangent space $\mathbb{T}$ on $\mathcal{N}_r$, and for $\mathcal{S}_\Omega$ defined as in (32) and $\boldsymbol{H}_{\mathrm{offdiag}}^{-1}$ defined as in (33), we have that for any $\beta > 1$ and some absolute numerical constant $C > 0$, if $p \geq \frac{8}{3}\beta\frac{\log n}{n}$, that*

$$\Big| \mathcal{S}_\Omega(\boldsymbol{Z})\boldsymbol{H}_{\mathrm{offdiag}}^{-1}\mathcal{S}_\Omega(\mathcal{P}_{\mathbb{T}}\boldsymbol{Y}) - \mathbb{E}\left[\mathcal{S}_\Omega(\boldsymbol{Z})\boldsymbol{H}_{\mathrm{offdiag}}^{-1}\mathcal{S}_\Omega(\mathcal{P}_{\mathbb{T}}\boldsymbol{Y})\right] \Big| \leq Cp\sqrt{\nu r}\frac{\beta\log n}{\sqrt{n}} + p^{3/2}\sqrt{\frac{96\nu r\beta\log n}{3}}$$

with probability at least $1 - 6n^{-\beta}$.

Proof. This proof is similar to that of Theorem B.8, with some minor differences due to the asymmetry. Defining $\mathcal{S}_\Omega^C = \mathcal{S}_\Omega - \mathbb{E}[\mathcal{S}_\Omega]$, we have for any $\boldsymbol{Y}, \boldsymbol{Z}$ that

$$\mathcal{S}_\Omega^C(\boldsymbol{Z})^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1}\mathcal{S}_\Omega^C(\mathcal{P}_{\mathbb{T}}\boldsymbol{Y}) - \mathbb{E}\left[\mathcal{S}_\Omega^C(\boldsymbol{Z})^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1}\mathcal{S}_\Omega^C(\mathcal{P}_{\mathbb{T}}\boldsymbol{Y})\right]$$

$$= \mathcal{S}_\Omega(\boldsymbol{Z})\boldsymbol{H}_{\mathrm{offdiag}}^{-1}\mathcal{S}_\Omega(\mathcal{P}_{\mathbb{T}}\boldsymbol{Y}) - \mathcal{S}_\Omega(\boldsymbol{Z})\boldsymbol{H}_{\mathrm{offdiag}}^{-1}\mathbb{E}[\mathcal{S}_\Omega(\mathcal{P}_{\mathbb{T}}\boldsymbol{Y})] - \mathbb{E}[\mathcal{S}_\Omega(\boldsymbol{Z})]\boldsymbol{H}_{\mathrm{offdiag}}^{-1}\mathcal{S}_\Omega(\mathcal{P}_{\mathbb{T}}\boldsymbol{Y})$$

$$+ \mathbb{E}[\mathcal{S}_\Omega(\boldsymbol{Z})]\boldsymbol{H}_{\mathrm{offdiag}}^{-1}\mathbb{E}[\mathcal{S}_\Omega(\mathcal{P}_{\mathbb{T}}\boldsymbol{Y})] - \mathbb{E}\left[\mathcal{S}_\Omega^C(\boldsymbol{Z})^\top \boldsymbol{H}_{\mathrm{offdiag}}^{-1}\mathcal{S}_\Omega^C(\mathcal{P}_{\mathbb{T}}\boldsymbol{Y})\right]$$

$$= \mathcal{S}_\Omega(\boldsymbol{Z})\boldsymbol{H}_{\mathrm{offdiag}}^{-1}\mathcal{S}_\Omega(\mathcal{P}_{\mathbb{T}}\boldsymbol{Y}) - \mathcal{S}_\Omega(\boldsymbol{Z})\boldsymbol{H}_{\mathrm{offdiag}}^{-1}\mathbb{E}[\mathcal{S}_\Omega(\mathcal{P}_{\mathbb{T}}\boldsymbol{Y})] - \mathbb{E}[\mathcal{S}_\Omega(\boldsymbol{Z})]\boldsymbol{H}_{\mathrm{offdiag}}^{-1}\mathcal{S}_\Omega(\mathcal{P}_{\mathbb{T}}\boldsymbol{Y})$$

$$+ \mathbb{E}[\mathcal{S}_\Omega(\boldsymbol{Z})]\boldsymbol{H}_{\mathrm{offdiag}}^{-1}\mathbb{E}[\mathcal{S}_\Omega(\mathcal{P}_{\mathbb{T}}\boldsymbol{Y})] - \mathbb{E}[\mathcal{S}_\Omega(\boldsymbol{Z})\boldsymbol{H}_{\mathrm{offdiag}}^{-1}\mathcal{S}_\Omega(\mathcal{P}_{\mathbb{T}}\boldsymbol{Y})] + \mathbb{E}[\mathcal{S}_\Omega(\boldsymbol{Z})]\boldsymbol{H}_{\mathrm{offdiag}}^{-1}\mathbb{E}[\mathcal{S}_\Omega(\mathcal{P}_{\mathbb{T}}\boldsymbol{Y})]$$

$$
+ \mathbb{E}[\mathcal{S}_\Omega(\boldsymbol{Z})]\boldsymbol{H}_{\text{offdiag}}^{-1}\mathbb{E}[\mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y})] - \mathbb{E}[\mathcal{S}_\Omega(\boldsymbol{Z})]\boldsymbol{H}_{\text{offdiag}}^{-1}\mathbb{E}[\mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y})]
$$

$$
= \mathcal{S}_\Omega(\boldsymbol{Z})\boldsymbol{H}_{\text{offdiag}}^{-1}\mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y}) - \mathbb{E}[\mathcal{S}_\Omega(\boldsymbol{Z})\boldsymbol{H}_{\text{offdiag}}^{-1}\mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y})] + (\mathbb{E}[\mathcal{S}_\Omega(\boldsymbol{Z})] - \mathcal{S}_\Omega(\boldsymbol{Z}))\boldsymbol{H}_{\text{offdiag}}^{-1}\mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y})
$$

$$
+ \mathbb{E}[\mathcal{S}_\Omega(\boldsymbol{Z})]\boldsymbol{H}_{\text{offdiag}}^{-1}(\mathbb{E}[\mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y})] - \mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y})) . \tag{35}
$$

As such, it follows that

$$
\left| \mathcal{S}_\Omega(\boldsymbol{Z})\boldsymbol{H}_{\text{offdiag}}^{-1}\mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y}) - \mathbb{E}[\mathcal{S}_\Omega(\boldsymbol{Z})\boldsymbol{H}_{\text{offdiag}}^{-1}\mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y})] \right|
$$

$$
= \left| \mathcal{S}_\Omega(\boldsymbol{Z})\boldsymbol{H}_{\text{offdiag}}^{-1}\mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y}) - \mathbb{E}[\mathcal{S}_\Omega(\boldsymbol{Z})\boldsymbol{H}_{\text{offdiag}}^{-1}\mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y})] + (\mathbb{E}[\mathcal{S}_\Omega(\boldsymbol{Z})] - \mathcal{S}_\Omega(\boldsymbol{Z}))\boldsymbol{H}_{\text{offdiag}}^{-1}\mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y}) \right.
$$

$$
- (\mathbb{E}[\mathcal{S}_\Omega(\boldsymbol{Z})] - \mathcal{S}_\Omega(\boldsymbol{Z}))\boldsymbol{H}_{\text{offdiag}}^{-1}\mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y}) + \mathbb{E}[\mathcal{S}_\Omega(\boldsymbol{Z})]\boldsymbol{H}_{\text{offdiag}}^{-1}(\mathbb{E}[\mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y})] - \mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y}))
$$

$$
\left. - \mathbb{E}[\mathcal{S}_\Omega(\boldsymbol{Z})]\boldsymbol{H}_{\text{offdiag}}^{-1}(\mathbb{E}[\mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y})] - \mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y})) \right|
$$

$$
\leq \underbrace{\left| \mathcal{S}_\Omega^C(\boldsymbol{Z})^\top \boldsymbol{H}_{\text{offdiag}}^{-1}\mathcal{S}_\Omega^C(\mathcal{P}_\mathbb{T}\boldsymbol{Y}) - \mathbb{E}\left[\mathcal{S}_\Omega^C(\boldsymbol{Z})^\top \boldsymbol{H}_{\text{offdiag}}^{-1}\mathcal{S}_\Omega^C(\mathcal{P}_\mathbb{T}\boldsymbol{Y}) \right] \right|}_{T_1}
$$

$$
+ \underbrace{\left| (\mathbb{E}[\mathcal{S}_\Omega(\boldsymbol{Z})] - \mathcal{S}_\Omega(\boldsymbol{Z}))\boldsymbol{H}_{\text{offdiag}}^{-1}\mathbb{E}[\mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y})] \right|}_{T_2} + \underbrace{\left| \mathbb{E}[\mathcal{S}_\Omega(\boldsymbol{Z})]\boldsymbol{H}_{\text{offdiag}}^{-1}(\mathbb{E}[\mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y})] - \mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y})) \right|}_{T_3},
$$

where the inequality comes from the triangle inequality and (35). We will now seek to bound terms T_1, T_2, and T_3.
Bounding T_1: We will first bound T_1 using the Hanson-Wright inequality (Theorem A.3). First, we define the following matrix $\boldsymbol{A}_{\boldsymbol{Y},\boldsymbol{Z}} \in \mathbb{R}^{L \times L}$ as $(A_{\boldsymbol{Y},\boldsymbol{Z}})_{\boldsymbol{\alpha\beta}} = \langle \boldsymbol{Z}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\langle\mathcal{P}_\mathbb{T}\boldsymbol{Y}, \boldsymbol{w}_{\boldsymbol{\beta}}\rangle$. As such, we can write, for fixed $\boldsymbol{Y}, \boldsymbol{Z}$, the following:

$$
\boldsymbol{\xi}^\top(\boldsymbol{A}_{\boldsymbol{Y},\boldsymbol{Z}} \circ \boldsymbol{H}_{\text{offdiag}}^{-1})\boldsymbol{\xi} = \sum_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{I} \\ \boldsymbol{\alpha}\neq\boldsymbol{\beta}}} \xi_{\boldsymbol{\alpha}}(\boldsymbol{A}_{\boldsymbol{Y},\boldsymbol{Z}} \circ \boldsymbol{H}_{\text{offdiag}}^{-1})_{\boldsymbol{\alpha\beta}}\xi_{\boldsymbol{\beta}}
$$

$$
= \sum_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{I} \\ \boldsymbol{\alpha}\neq\boldsymbol{\beta}}} \xi_{\boldsymbol{\alpha}}\langle\boldsymbol{Z},\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle\boldsymbol{Y},\mathcal{P}_\mathbb{T}\boldsymbol{w}_{\boldsymbol{\beta}}\rangle\xi_{\boldsymbol{\beta}}
$$

$$
= \mathcal{S}_\Omega(\boldsymbol{Z})^\top\boldsymbol{H}_{\text{offdiag}}^{-1}\mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y}),
$$

where $\circ$ denotes the Hadamard product and $\boldsymbol{\xi} \in \mathbb{R}^L$ is a Bernoulli random vector with each component being an i.i.d. Bernoulli random variable with parameter p, i.e. $\boldsymbol{\xi}_{\boldsymbol{\alpha}} = \xi_{\boldsymbol{\alpha}}$ for all $\boldsymbol{\alpha} \in \mathbb{I}$. Similarly, we can write

$$
(\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}])^\top(\boldsymbol{A}_{\boldsymbol{Y},\boldsymbol{Z}} \circ \boldsymbol{H}_{\text{offdiag}}^{-1})(\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}]) = \sum_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{I} \\ \boldsymbol{\alpha}\neq\boldsymbol{\beta}}} (\xi_{\boldsymbol{\alpha}} - p)\left(\boldsymbol{A}_{\boldsymbol{Y},\boldsymbol{Z}} \circ \boldsymbol{H}_{\text{offdiag}}^{-1}\right)(\xi_{\boldsymbol{\beta}} - p)
$$

$$
= \sum_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{I} \\ \boldsymbol{\alpha}\neq\boldsymbol{\beta}}} (\xi_{\boldsymbol{\alpha}} - p)\langle\boldsymbol{Z},\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle\boldsymbol{Y},\mathcal{P}_\mathbb{T}\boldsymbol{w}_{\boldsymbol{\beta}}\rangle(\xi_{\boldsymbol{\beta}} - p),
$$

and

$$
\mathcal{S}_\Omega^C(\boldsymbol{Z})^\top\boldsymbol{H}_{\text{offdiag}}^{-1}\mathcal{S}_\Omega^C(\mathcal{P}_\mathbb{T}\boldsymbol{Y}) = (\mathcal{S}_\Omega(\boldsymbol{Z}) - \mathbb{E}(\mathcal{S}_\Omega(\boldsymbol{Z}))^\top\boldsymbol{H}_{\text{offdiag}}^{-1}(\mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y}) - \mathbb{E}[\mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y})])
$$

$$
= \mathcal{S}_\Omega(\boldsymbol{Z})^\top\boldsymbol{H}_{\text{offdiag}}^{-1}\mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y}) - \mathcal{S}_\Omega(\boldsymbol{Z})^\top\boldsymbol{H}_{\text{offdiag}}^{-1}\mathbb{E}[\mathcal{S}_\Omega\mathcal{P}_\mathbb{T}\boldsymbol{Y}]
$$

$$
- \mathbb{E}[\mathcal{S}_\Omega(\boldsymbol{Z})]^\top\boldsymbol{H}_{\text{offdiag}}^{-1}\mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y}) + \mathbb{E}[\mathcal{S}_\Omega(\boldsymbol{Z})]^\top\boldsymbol{H}_{\text{offdiag}}^{-1}\mathbb{E}[\mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y})]
$$

$$
= \sum_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{I} \\ \boldsymbol{\alpha}\neq\boldsymbol{\beta}}} \xi_{\boldsymbol{\alpha}}\langle\boldsymbol{Z},\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle\mathcal{P}_\mathbb{T}\boldsymbol{Y},\boldsymbol{w}_{\boldsymbol{\beta}}\rangle\xi_{\boldsymbol{\beta}} - p\langle\boldsymbol{Z},\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle\mathcal{P}_\mathbb{T}\boldsymbol{Y},\boldsymbol{w}_{\boldsymbol{\beta}}\rangle\xi_{\boldsymbol{\beta}}
$$

$$
- \xi_{\boldsymbol{\alpha}}\langle\boldsymbol{Z},\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle\mathcal{P}_\mathbb{T}\boldsymbol{Y},\boldsymbol{w}_{\boldsymbol{\beta}}\rangle p + p\langle\boldsymbol{Z},\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\beta}}\rangle\langle\mathcal{P}_\mathbb{T}\boldsymbol{Y},\boldsymbol{w}_{\boldsymbol{\beta}}\rangle p
$$

$$= \sum_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{I}\\ \boldsymbol{\alpha}\neq\boldsymbol{\beta}}} (\xi_{\boldsymbol{\alpha}} - p)\langle \boldsymbol{Z}, \boldsymbol{w_\alpha}\rangle\langle \boldsymbol{v_\alpha}, \boldsymbol{v_\beta}\rangle\langle \boldsymbol{Y}, \mathcal{P}_\mathbb{T}\boldsymbol{w_\beta}\rangle(\xi_{\boldsymbol{\beta}} - p)$$

$$= (\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}])^\top (\boldsymbol{A_{Y,Z}} \circ \boldsymbol{H}_{\text{offdiag}}^{-1})(\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}]).$$

With this equality established, we can now proceed with using Theorem A.3 to bound T_1. Using Theorem B.4 and setting $\boldsymbol{A} = \boldsymbol{B} = \frac{1}{\sqrt{n}}\boldsymbol{I}$ in the Lemma statement, we have that $\|\boldsymbol{\xi}\|_{\psi_2} \leq C\sqrt{p}$ for some absolute constant $C > 0$. Next, we need to bound the Frobenius norm of $\|\boldsymbol{A_{Y,Z}} \circ \boldsymbol{H}_{\text{offdiag}}^{-1}\|_\text{F}^2$. To do this, notice that, for $\|\boldsymbol{Z}\|_\text{F} = \|\boldsymbol{Y}\|_\text{F} = 1$,

$$\|\boldsymbol{A_{Y,Z}} \circ \boldsymbol{H}_{\text{offdiag}}^{-1}\|_\text{F}^2 = \sum_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{I}\\ \boldsymbol{\alpha}\neq\boldsymbol{\beta}}} (\langle \boldsymbol{Z}, \boldsymbol{w_\alpha}\rangle\langle \boldsymbol{v_\alpha}, \boldsymbol{v_\beta}\rangle\langle \boldsymbol{Y}, \mathcal{P}_\mathbb{T}\boldsymbol{w_\beta}\rangle)^2$$

$$= \sum_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{I}\\ \boldsymbol{\alpha}\neq\boldsymbol{\beta}}} \langle \boldsymbol{Z}, \boldsymbol{w_\alpha}\rangle^2\langle \boldsymbol{v_\alpha}, \boldsymbol{v_\beta}\rangle^2\langle \boldsymbol{Y}, \mathcal{P}_\mathbb{T}\boldsymbol{w_\beta}\rangle^2$$

$$\leq \frac{1}{n^2} \sum_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{I}\\ \boldsymbol{\alpha}\neq\boldsymbol{\beta}}} \langle \boldsymbol{Z}, \boldsymbol{w_\alpha}\rangle^2\langle \boldsymbol{Y}, \mathcal{P}_\mathbb{T}\boldsymbol{w_\beta}\rangle^2$$

$$\leq \frac{1}{n^2} \sum_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{I}} \langle \boldsymbol{Z}, \boldsymbol{w_\alpha}\rangle^2\langle \boldsymbol{Y}, \mathcal{P}_\mathbb{T}\boldsymbol{w_\beta}\rangle^2$$

$$= \frac{1}{n^2} \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \boldsymbol{Z}, \boldsymbol{w_\alpha}\rangle^2 \sum_{\boldsymbol{\beta}\in\mathbb{I}} \langle \boldsymbol{Y}, \mathcal{P}_\mathbb{T}\boldsymbol{w_\beta}\rangle^2$$

$$\leq \frac{1}{n^2}\lambda_{\max}(\boldsymbol{H})\lambda_{\max}(\tilde{\boldsymbol{H}})$$

$$\leq \frac{2\nu r}{n},$$

where the first inequality follows from Theorem A.8, the third inequality follows from Theorem A.5, and the final inequality follows from Lemmas A.8 and A.6. Next, to bound $\|\boldsymbol{A_{Y,Z}} \circ \boldsymbol{H}_{\text{offdiag}}^{-1}\|$, we will use a Gershgorin estimate as follows:

$$\|\boldsymbol{A_{Y,Z}} \circ \boldsymbol{H}_{\text{offdiag}}^{-1}\| \leq \max_{\boldsymbol{\alpha}} \sum_{\boldsymbol{\beta}\in\mathbb{I}} \left|(\boldsymbol{A_{Y,Z}} \circ \boldsymbol{H}_{\text{offdiag}}^{-1})_{\boldsymbol{\alpha\beta}}\right|$$

$$= \max_{\boldsymbol{\alpha}} \sum_{\boldsymbol{\beta}\neq\boldsymbol{\alpha}} |\langle \boldsymbol{Z}, \boldsymbol{w_\alpha}\rangle\langle \boldsymbol{v_\alpha}, \boldsymbol{v_\beta}\rangle\langle \boldsymbol{Y}, \mathcal{P}_\mathbb{T}\boldsymbol{w_\beta}\rangle|$$

$$\leq \max_{\boldsymbol{\alpha}} 2\sqrt{\frac{\nu r}{2n}} \sum_{\boldsymbol{\beta}} \left|(\boldsymbol{H}_{\text{offdiag}}^{-1})_{\boldsymbol{\alpha\beta}}\right|$$

$$\leq 2\sqrt{\frac{2\nu r}{n}},$$

where the first inequality follows from Gershgorin circle theorem, the second inequality follows from Cauchy Schwarz and Theorem 5.1, and the final inequality comes from Theorem A.8. Furthermore, as $\text{Trace}(\boldsymbol{H}_{\text{offdiag}}^{-1}) = 0$, $(\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}])^\top (\boldsymbol{A_{Y,Z}} \circ \boldsymbol{H}_{\text{offdiag}}^{-1})(\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}]) = 0$. Taking $\beta > 1$, for some sufficiently large constant $C > 0$ we have from Theorem A.3 that

$$\mathbb{P}\left(\left|(\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}])^\top (\boldsymbol{A_{Y,Z}} \circ \boldsymbol{H}_{\text{offdiag}}^{-1})(\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}])\right| > Cp\sqrt{\nu r}\frac{\beta \log n}{\sqrt{n}}\right)$$

$$\leq 2\exp\left(-c\min\left\{\frac{n}{2p^2\nu r}C^2p^2\frac{\nu r\beta^2\log^2 n}{n}, \frac{\sqrt{n}}{2p\sqrt{2\nu r}}Cp\sqrt{\nu r}\frac{\beta \log n}{\sqrt{n}}\right\}\right)$$

$$\leq 2n^{-\beta},$$

completing the bound for T_1.

Bounding T_2: To bound T_2 and T_3, we will use the scalar Bernstein inequality seen in Theorem A.2. We will bound T_2 first. Defining

$$L_{\boldsymbol{\alpha}} = (\xi_{\boldsymbol{\alpha}} - p)\langle \boldsymbol{Z}, \boldsymbol{w_\alpha}\rangle \sum_{\substack{\boldsymbol{\beta}\in\mathbb{I}\\ \boldsymbol{\beta}\neq\boldsymbol{\alpha}}} \langle \boldsymbol{v_\alpha}, \boldsymbol{v_\beta}\rangle\langle \boldsymbol{Y}, \mathcal{P}_\mathbb{T}\boldsymbol{w_\beta}\rangle,$$

we can see that

$$\sum_{\boldsymbol{\alpha}\in\mathbb{I}} L_{\boldsymbol{\alpha}} = \sum_{\boldsymbol{\alpha}\in\mathbb{I}} (\xi_{\boldsymbol{\alpha}} - p)\langle \boldsymbol{Z}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \sum_{\substack{\boldsymbol{\beta}\in\mathbb{I}\\\boldsymbol{\beta}\neq\boldsymbol{\alpha}}} \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle \langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\beta}}\rangle$$

$$= p(\mathcal{S}_{\Omega}(\boldsymbol{Z}) - \mathbb{E}(\mathcal{S}_{\Omega}(\boldsymbol{Z}))^{\top} \boldsymbol{H}_{\mathrm{offdiag}}^{-1} \mathcal{S}_{\Omega}(\mathcal{P}_{\mathbb{T}}\boldsymbol{Y}).$$

As $L_{\boldsymbol{\alpha}}$ is a zero-mean bounded random variable, we can proceed with the proof using Bernstein's inequality. First, notice that

$$|L_{\boldsymbol{\alpha}}| = \left| (\xi_{\boldsymbol{\alpha}} - p)\langle \boldsymbol{Z}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \sum_{\substack{\boldsymbol{\beta}\in\mathbb{I}\\\boldsymbol{\beta}\neq\boldsymbol{\alpha}}} \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle \langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\beta}}\rangle \right|$$

$$\leq |(\xi_{\boldsymbol{\alpha}} - p)\langle \boldsymbol{Z}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle| \sum_{\substack{\boldsymbol{\beta}\in\mathbb{I}\\\boldsymbol{\beta}\neq\boldsymbol{\alpha}}} |\langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle \langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\beta}}\rangle|$$

$$\leq \|\boldsymbol{Z}\|_{\mathrm{F}} \|\boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} \|\boldsymbol{Y}\|_{\mathrm{F}} \sum_{\substack{\boldsymbol{\beta}\in\mathbb{I}\\\boldsymbol{\beta}\neq\boldsymbol{\alpha}}} |\langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle| \|\mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\beta}}\|_{\mathrm{F}}$$

$$\leq 2\sqrt{\frac{\nu r}{2n}} \sum_{\substack{\boldsymbol{\beta}\in\mathbb{I}\\\boldsymbol{\beta}\neq\boldsymbol{\alpha}}} |\langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle|$$

$$\leq 2\sqrt{\frac{2\nu r}{n}} =: R,$$

where the first inequality follows from the triangle inequality, the second follows from Cauchy-Schwarz, the third follows from Theorem 5.1, and the final inequality follows from Theorem A.8. Next, notice that

$$\sum_{\boldsymbol{\alpha}\in\mathbb{I}} L_{\boldsymbol{\alpha}}^2 = \sum_{\boldsymbol{\alpha}} (\xi_{\boldsymbol{\alpha}} - p)^2 \langle \boldsymbol{Z}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \left(\sum_{\substack{\boldsymbol{\beta}\in\mathbb{I}\\\boldsymbol{\beta}\neq\boldsymbol{\alpha}}} \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle \langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\beta}}\rangle \right)^2$$

$$= \sum_{\boldsymbol{\alpha}\in\mathbb{I}} (\xi_{\boldsymbol{\alpha}} - 2p\xi_{\boldsymbol{\alpha}} + p^2)\langle \boldsymbol{Z}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \left(\sum_{\substack{\boldsymbol{\beta}\in\mathbb{I}\\\boldsymbol{\beta}\neq\boldsymbol{\alpha}}} \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle \langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\beta}}\rangle \right)^2,$$

so

$$\mathbb{E}\left[\sum_{\boldsymbol{\alpha}\in\mathbb{I}} L_{\boldsymbol{\alpha}}^2 \right] = \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \mathbb{E}\left[(\xi_{\boldsymbol{\alpha}} - 2p\xi_{\boldsymbol{\alpha}} + p^2) \right] \langle \boldsymbol{Z}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \left(\sum_{\substack{\boldsymbol{\beta}\in\mathbb{I}\\\boldsymbol{\beta}\neq\boldsymbol{\alpha}}} \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle \langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\beta}}\rangle \right)^2$$

$$= \sum_{\boldsymbol{\alpha}\in\mathbb{I}} p(1-p)\langle \boldsymbol{Z}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \left(\sum_{\substack{\boldsymbol{\beta}\in\mathbb{I}\\\boldsymbol{\beta}\neq\boldsymbol{\alpha}}} \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle \langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\beta}}\rangle \right)^2$$

$$\leq \sum_{\boldsymbol{\alpha}\in\mathbb{I}} p(1-p)\langle \boldsymbol{Z}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \left(\sum_{\substack{\boldsymbol{\beta}\in\mathbb{I}\\\boldsymbol{\beta}\neq\boldsymbol{\alpha}}} |\langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle \langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\beta}}\rangle| \right)^2$$

$$\leq p\frac{\nu r}{2n} \sum_{\boldsymbol{\alpha}\in\mathbb{I}} \langle \boldsymbol{Z}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \left(\sum_{\substack{\boldsymbol{\beta}\in\mathbb{I}\\\boldsymbol{\beta}\neq\boldsymbol{\alpha}}} |\langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle| \right)^2$$

$$\leq 2p\frac{\nu r}{n}\sum_{\boldsymbol{\alpha}\in\mathbb{I}}\langle\boldsymbol{Z},\boldsymbol{w_\alpha}\rangle^2$$

$$\leq 2p\frac{\nu r}{n}\lambda_{\max}(\boldsymbol{H})$$

$$= 4p\nu r =: \sigma^2,$$

where the second inequality follows from Theorem 5.1, the third inequality follows from Theorem A.8, the fourth inequality follows from Theorem A.5, and the final line follows from Theorem A.8. As such, we have that for $p \geq \frac{4}{3}\beta\frac{\log n}{n}$ that

$$\mathbb{P}\left(\left|\sum_{\boldsymbol{\alpha}\in\mathbb{I}}L_{\boldsymbol{\alpha}}\right| > \sqrt{\frac{32p\nu r\beta\log n}{3}}\right) \leq 2\exp\left(\frac{-3}{8(4p\nu r)}\frac{32}{3}p\nu r\beta\log n\right)$$

$$= 2n^{-\beta},$$

thus completing the bound for T_2.

Bounding T_3:

We conclude this proof with a bound on T_3. We first remark that, due to $(\boldsymbol{H}_{\text{offdiag}}^{-1})^\top = \boldsymbol{H}_{\text{offdiag}}^{-1}$, $(\mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y}) - \mathbb{E}(\mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y}))^\top\boldsymbol{H}_{\text{offdiag}}^{-1}\mathcal{S}_\Omega(\boldsymbol{Z}) = \mathcal{S}_\Omega(\boldsymbol{Z})^\top\boldsymbol{H}_{\text{offdiag}}^{-1}(\mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y}) - \mathbb{E}[\mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y})]$. We will work with the first term for simplicity. Next, we define

$$N_{\boldsymbol{\alpha}} = (\xi_{\boldsymbol{\alpha}} - p)\langle\boldsymbol{Y},\mathcal{P}_\mathbb{T}\boldsymbol{w_\alpha}\rangle\sum_{\substack{\boldsymbol{\beta}\in\mathbb{I}\\\boldsymbol{\beta}\neq\boldsymbol{\alpha}}}\langle\boldsymbol{v_\alpha},\boldsymbol{v_\beta}\rangle\langle\boldsymbol{Z},\boldsymbol{w_\beta}\rangle,$$

noticing that

$$\sum_{\boldsymbol{\alpha}\in\mathbb{I}}N_{\boldsymbol{\alpha}} = \sum_{\boldsymbol{\alpha}\in\mathbb{I}}(\xi_{\boldsymbol{\alpha}} - p)\langle\mathcal{P}_\mathbb{T}\boldsymbol{Y},\boldsymbol{w_\alpha}\rangle\sum_{\substack{\boldsymbol{\beta}\in\mathbb{I}\\\boldsymbol{\beta}\neq\boldsymbol{\alpha}}}\langle\boldsymbol{v_\alpha},\boldsymbol{v_\beta}\rangle\langle\boldsymbol{Z},\boldsymbol{w_\beta}\rangle$$

$$= p(\mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y}) - \mathbb{E}(\mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y}))^\top\boldsymbol{H}_{\text{offdiag}}^{-1}\mathcal{S}_\Omega(\boldsymbol{Z}).$$

As before, we see that $\mathbb{E}[N_{\boldsymbol{\alpha}}] = 0$ and we can proceed using Bernstein's inequality.

First, notice that

$$|N_{\boldsymbol{\alpha}}| = \left|(\xi_{\boldsymbol{\alpha}} - p)\langle\boldsymbol{Y},\mathcal{P}_\mathbb{T}\boldsymbol{w_\alpha}\rangle\sum_{\substack{\boldsymbol{\beta}\in\mathbb{I}\\\boldsymbol{\beta}\neq\boldsymbol{\alpha}}}\langle\boldsymbol{v_\alpha},\boldsymbol{v_\beta}\rangle\langle\boldsymbol{Z},\boldsymbol{w_\beta}\rangle\right|$$

$$\leq |(\xi_{\boldsymbol{\alpha}} - p)\langle\boldsymbol{Y},\mathcal{P}_\mathbb{T}\boldsymbol{w_\alpha}\rangle|\sum_{\substack{\boldsymbol{\beta}\in\mathbb{I}\\\boldsymbol{\beta}\neq\boldsymbol{\alpha}}}|\langle\boldsymbol{v_\alpha},\boldsymbol{v_\beta}\rangle\langle\boldsymbol{Z},\boldsymbol{w_\beta}\rangle|$$

$$\leq \|\boldsymbol{Y}\|_\text{F}\|\mathcal{P}_\mathbb{T}\boldsymbol{w_\alpha}\|_\text{F}\|\boldsymbol{Z}\|_\text{F}\sum_{\substack{\boldsymbol{\beta}\in\mathbb{I}\\\boldsymbol{\beta}\neq\boldsymbol{\alpha}}}|\langle\boldsymbol{v_\alpha},\boldsymbol{v_\beta}\rangle|\|\boldsymbol{w_\beta}\|_\text{F}$$

$$\leq 2\sqrt{\frac{\nu r}{2n}}\sum_{\substack{\boldsymbol{\beta}\in\mathbb{I}\\\boldsymbol{\beta}\neq\boldsymbol{\alpha}}}|\langle\boldsymbol{v_\alpha},\boldsymbol{v_\beta}\rangle|$$

$$\leq 2\sqrt{\frac{2\nu r}{n}} =: R,$$

where the first inequality follows from the triangle inequality, the second follows from Cauchy-Schwarz, the third follows from Theorem 5.1, and the final inequality follows from Theorem A.8. Next, notice that

$$\sum_{\boldsymbol{\alpha}\in\mathbb{I}}N_{\boldsymbol{\alpha}}^2 = \sum_{\boldsymbol{\alpha}}(\xi_{\boldsymbol{\alpha}} - p)^2\langle\boldsymbol{Y},\mathcal{P}_\mathbb{T}\boldsymbol{w_\alpha}\rangle^2\left(\sum_{\substack{\boldsymbol{\beta}\in\mathbb{I}\\\boldsymbol{\beta}\neq\boldsymbol{\alpha}}}\langle\boldsymbol{v_\alpha},\boldsymbol{v_\beta}\rangle\langle\boldsymbol{Z},\boldsymbol{w_\beta}\rangle\right)^2$$

$$= \sum_{\boldsymbol{\alpha} \in \mathbb{I}} (\xi_{\boldsymbol{\alpha}} - 2p\xi_{\boldsymbol{\alpha}} + p^2)\langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \left(\sum_{\substack{\boldsymbol{\beta} \in \mathbb{I} \\ \boldsymbol{\beta} \neq \boldsymbol{\alpha}}} \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle \langle \boldsymbol{Z}, \boldsymbol{w}_{\boldsymbol{\beta}}\rangle \right)^2,$$

so

$$\mathbb{E}\left[\sum_{\boldsymbol{\alpha} \in \mathbb{I}} N_{\boldsymbol{\alpha}}^2\right] = \sum_{\boldsymbol{\alpha} \in \mathbb{I}} \mathbb{E}\left[(\xi_{\boldsymbol{\alpha}} - 2p\xi_{\boldsymbol{\alpha}} + p^2)\right] \langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \left(\sum_{\substack{\boldsymbol{\beta} \in \mathbb{I} \\ \boldsymbol{\beta} \neq \boldsymbol{\alpha}}} \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle \langle \boldsymbol{Z}, \boldsymbol{w}_{\boldsymbol{\beta}}\rangle \right)^2$$

$$= \sum_{\boldsymbol{\alpha} \in \mathbb{I}} p(1-p)\langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \left(\sum_{\substack{\boldsymbol{\beta} \in \mathbb{I} \\ \boldsymbol{\beta} \neq \boldsymbol{\alpha}}} \langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle \langle \boldsymbol{Z}, \boldsymbol{w}_{\boldsymbol{\beta}}\rangle \right)^2$$

$$\leq \sum_{\boldsymbol{\alpha} \in \mathbb{I}} p(1-p)\langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \left(\sum_{\substack{\boldsymbol{\beta} \in \mathbb{I} \\ \boldsymbol{\beta} \neq \boldsymbol{\alpha}}} |\langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle \langle \boldsymbol{Z}, \boldsymbol{w}_{\boldsymbol{\beta}}\rangle| \right)^2$$

$$\leq 4p \sum_{\boldsymbol{\alpha} \in \mathbb{I}} \langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \left(\sum_{\substack{\boldsymbol{\beta} \in \mathbb{I} \\ \boldsymbol{\beta} \neq \boldsymbol{\alpha}}} |\langle \boldsymbol{v}_{\boldsymbol{\alpha}}, \boldsymbol{v}_{\boldsymbol{\beta}}\rangle| \right)^2$$

$$\leq 8p \sum_{\boldsymbol{\alpha} \in \mathbb{I}} \langle \boldsymbol{Y}, \mathcal{P}_{\mathbb{T}}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2$$

$$\leq 8p\lambda_{\max}(\tilde{\boldsymbol{H}})$$

$$\leq 8p\nu r =: \sigma^2,$$

where the second inequality follows from Theorem 5.1, the third inequality follows from Theorem A.8, the fourth inequality follows from Lemmas A.5 and A.7, and the final line follows from Theorem A.6. As such, we have that for $p \geq \frac{8}{3}\beta\frac{\log n}{n}$ that

$$\mathbb{P}\left(\left|\sum_{\boldsymbol{\alpha} \in \mathbb{I}} L_{\boldsymbol{\alpha}}\right| > \sqrt{\frac{64p\nu r\beta\log n}{3}}\right) \leq 2\exp\left(\frac{-3}{8(8p\nu r)}\frac{64}{3}p\nu r\beta\log n\right)$$

$$= 2n^{-\beta},$$

thus completing the bound for T_3, and in sum completing the proof. $\qquad\square$

Lemma B.11. *Let $\Omega \subset \mathbb{I}$ be sampled with uniform Bernoulli probability p, and let $\mathbb{T}$ be the tangent space on $\mathcal{N}_r$ for a rank-r, ν-incoherent ground truth matrix $\boldsymbol{X}$. If $p \geq \frac{8}{3}\frac{\beta\log n}{n}$, then with probability at least $1 - 2n^{1-\beta} - 6n^{-\beta}$ we have that, for some absolute constant $c > 0$,*

$$\|\mathcal{M}_{\Omega}\mathcal{P}_{\mathbb{T}}\| \leq p^2 + cp\sqrt{\nu}r\frac{\beta\log n}{\sqrt{n}} + p^{3/2}\sqrt{\frac{128\nu r\beta\log n}{3}}.$$

$$\|\mathcal{M}_{\Omega}\mathcal{P}_{\mathbb{T}_l}\| \leq 2\|\mathcal{M}_{\Omega}\|\frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\lambda_r(\boldsymbol{X})} + \|\mathcal{M}_{\Omega}\mathcal{P}_{\mathbb{T}}\|.$$

Furthermore, with probability at least $1 - 4n^{1-\beta} - 10n^{-\beta}$, for some sufficiently large constant $C > 0$ independent of ν and r, if $p \geq C\frac{\beta\log n}{n}$, then

$$\|\mathcal{M}_{\Omega}\mathcal{P}_{\mathbb{T}}\| \leq p^{3/2}\sqrt{\frac{256\nu r\beta\log n}{3}} \qquad \text{and} \qquad \|\mathcal{M}_{\Omega}\mathcal{P}_{\mathbb{T}_l}\| \leq 100p^{3/2}\sqrt{\beta n\log n}\frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\lambda_r(\boldsymbol{X})} + p^{3/2}\sqrt{\frac{256\nu r\beta\log n}{3}}.$$

Proof. We first notice that

$$\|\mathcal{M}_\Omega \mathcal{P}_\mathbb{T}\| \le \|\mathcal{M}_\Omega \mathcal{P}_\mathbb{T} - \mathbb{E}[\mathcal{M}_\Omega \mathcal{P}_\mathbb{T}]\| + \|\mathbb{E}[\mathcal{M}_\Omega \mathcal{P}_\mathbb{T}]\|$$
$$= \|\mathcal{M}_\Omega \mathcal{P}_\mathbb{T} - \mathbb{E}[\mathcal{M}_\Omega \mathcal{P}_\mathbb{T}]\| + p^2.$$

Similarly to the proof of Theorem 5.3, seen in Section B.1, and in the proof of Theorem B.9, we can decompose the difference between $\mathcal{M}_\Omega \mathcal{P}_\mathbb{T}$ and $\mathbb{E}[\mathcal{M}_\Omega \mathcal{P}_\mathbb{T}]$ as concentration of $\mathcal{F}_\Omega \mathcal{P}_\mathbb{T} - \mathbb{E}[\mathcal{F}_\Omega]\mathcal{P}_\mathbb{T}$ and the off-diagonal quadratic form term. As such, we can see that

$$\|\mathcal{M}_\Omega \mathcal{P}_\mathbb{T}\| \le p^2 + p\|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_F^2 \|\mathcal{F}_\Omega \mathcal{P}_\mathbb{T} - p\mathcal{F}_\mathbb{I} \mathcal{P}_\mathbb{T}\|$$

$$+ \left\| \sum_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{I} \\ \boldsymbol{\alpha}\ne\boldsymbol{\beta}}} \xi_{\boldsymbol{\alpha}}\xi_{\boldsymbol{\beta}}\langle\cdot,\mathcal{P}_\mathbb{T}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\beta}}\rangle\boldsymbol{w}_{\boldsymbol{\beta}} - \mathbb{E}\left[\sum_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{I} \\ \boldsymbol{\alpha}\ne\boldsymbol{\beta}}} \xi_{\boldsymbol{\alpha}}\xi_{\boldsymbol{\beta}}\langle\cdot,\mathcal{P}_\mathbb{T}\boldsymbol{w}_{\boldsymbol{\alpha}}\rangle\langle\boldsymbol{v}_{\boldsymbol{\alpha}},\boldsymbol{v}_{\boldsymbol{\beta}}\rangle\boldsymbol{w}_{\boldsymbol{\beta}} \right] \right\|$$

$$= p^2 + p\|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_F^2 \|\mathcal{F}_\Omega \mathcal{P}_\mathbb{T} - p\mathcal{F}_\mathbb{I} \mathcal{P}_\mathbb{T}\| + \max_{\substack{\|\boldsymbol{Y}\|_F=1 \\ \|\boldsymbol{Z}\|_F=1}} \left| \mathcal{S}_\Omega(\boldsymbol{Z})\boldsymbol{H}_{\text{offdiag}}^{-1}\mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y}) - \mathbb{E}\left[\mathcal{S}_\Omega(\boldsymbol{Z})\boldsymbol{H}_{\text{offdiag}}^{-1}\mathcal{S}_\Omega(\mathcal{P}_\mathbb{T}\boldsymbol{Y})\right] \right|.$$

From Lemmas B.7 and B.10, the first result follows. For the second result, notice that

$$\|\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l}\| = \|\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l} - \mathcal{M}_\Omega \mathcal{P}_\mathbb{T} + \mathcal{M}_\Omega \mathcal{P}_\mathbb{T}\|$$
$$\le \|\mathcal{M}_\Omega(\mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_\mathbb{T})\| + \|\mathcal{M}_\Omega \mathcal{P}_\mathbb{T}\|$$
$$\le \|\mathcal{M}_\Omega\| \frac{2\|\boldsymbol{X}_l - \boldsymbol{X}\|_F}{\lambda_r(\boldsymbol{X})} + \|\mathcal{M}_\Omega \mathcal{P}_\mathbb{T}\|.$$

For the final result, if $p \ge \frac{C'\beta\log n}{n}$ for some sufficiently large constant $C' > 0$, the conditions of Theorem B.9 hold with high probability and the expression can be simplified to

$$\|\mathcal{M}_\Omega\| \le 50p^{3/2}\sqrt{\beta n\log n}.$$

Similarly, for a sufficiently large constant $C'' > 0$ independent of ν and r, with $p \ge \frac{C''\beta\log n}{n}$, the derived expression for $\|\mathcal{M}_\Omega \mathcal{P}_\mathbb{T}\|$ can be simplified to

$$\|\mathcal{M}_\Omega \mathcal{P}_\mathbb{T}\| \le p^{3/2}\sqrt{\frac{256\beta\nu r\log n}{3}}.$$

Choosing $C = \max\{C', C''\}$ concludes the proof. $\qquad\square$

Lemma B.12 (Local RIP of $\mathcal{M}_\Omega$). *Assume that*

$$p^{-2}\|\mathcal{P}_\mathbb{T}\mathcal{M}_\Omega \mathcal{P}_\mathbb{T} - p^2\mathcal{P}_\mathbb{T}\| \le \varepsilon_0, \tag{36}$$

$$\|\mathcal{M}_\Omega \mathcal{P}_\mathbb{T}\| \le p^{3/2}\sqrt{\frac{256\beta\nu r\log n}{3}}, \tag{37}$$

$$\|\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l}\| \le 100p^{3/2}\sqrt{\beta n\log n}\frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_F}{\lambda_r(\boldsymbol{X})} + p^{3/2}\sqrt{\frac{256\nu r\beta\log n}{3}}, \tag{38}$$

$$\frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_F}{\lambda_r(\boldsymbol{X})} \le \frac{\varepsilon_0 p^{1/2}}{32(\beta n\log n)^{1/4}}. \tag{39}$$

Then

$$p^{-2}\|\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l} - p^2\mathcal{P}_{\mathbb{T}_l}\| \le 4\varepsilon_0.$$

Proof.

$$\|\mathcal{P}_{\mathbb{T}_l} - p^{-2}\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l}\| = \|\mathcal{P}_{\mathbb{T}_l} - p^{-2}\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l}$$
$$+ \mathcal{P}_\mathbb{T} - \mathcal{P}_\mathbb{T} + p^{-2}\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega \mathcal{P}_\mathbb{T} - p^{-2}\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega \mathcal{P}_\mathbb{T} + p^{-2}\mathcal{P}_\mathbb{T}\mathcal{M}_\Omega \mathcal{P}_\mathbb{T} - p^{-2}\mathcal{P}_\mathbb{T}\mathcal{M}_\Omega \mathcal{P}_\mathbb{T}\|$$
$$\le \|\mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_\mathbb{T}\| + p^{-2}\|\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega \mathcal{P}_\mathbb{T}\| + p^{-2}\|\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega \mathcal{P}_\mathbb{T} - \mathcal{P}_\mathbb{T}\mathcal{M}_\Omega \mathcal{P}_\mathbb{T}\| + \|\mathcal{P}_\mathbb{T} - p^{-2}\mathcal{P}_\mathbb{T}\mathcal{M}_\Omega \mathcal{P}_\mathbb{T}\|$$
$$\le \|\mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_\mathbb{T}\| + p^{-2}\|\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega\|\|\mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_\mathbb{T}\|$$

45

$$+ p^{-2}\|\mathcal{M}_\Omega \mathcal{P}_\mathbb{T}\|\|\mathcal{P}_{\mathbb{T}_l} - \mathcal{P}_\mathbb{T}\| + \|\mathcal{P}_\mathbb{T} - p^{-2}\mathcal{P}_\mathbb{T}\mathcal{M}_\Omega \mathcal{P}_\mathbb{T}\|$$

$$\leq \frac{2\|\boldsymbol{X}_l - \boldsymbol{X}\|_\mathrm{F}}{\lambda_r(\boldsymbol{X})}\left(1 + p^{-2}\|\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega\| + p^{-2}\|\mathcal{M}_\Omega \mathcal{P}_\mathbb{T}\|\right) + \|\mathcal{P}_\mathbb{T} - p^{-2}\mathcal{P}_\mathbb{T}\mathcal{M}_\Omega \mathcal{P}_\mathbb{T}\|$$

$$\leq 2\varepsilon_0 + p^{-2}\frac{2\|\boldsymbol{X}_l - \boldsymbol{X}\|_\mathrm{F}}{\lambda_r(\boldsymbol{X})}\left(100p^{3/2}\sqrt{\beta n \log n}\frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_\mathrm{F}}{\lambda_r(\boldsymbol{X})} + 2p^{3/2}\sqrt{\frac{256\nu r\beta \log n}{3}}\right)$$

$$= 2\varepsilon_0 + 200p^{-1/2}\sqrt{\beta n \log n}\left(\frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_\mathrm{F}}{\lambda_r(\boldsymbol{X})}\right)^2 + 32p^{-1/2}\sqrt{\frac{\nu r\beta \log n}{3}}\frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_\mathrm{F}}{\lambda_r(\boldsymbol{X})}$$

$$\leq 4\varepsilon_0,$$

where the first inequality is the triangle inequality, the second inequality is Cauchy-Schwarz, the third inequality is due to Theorem A.10 and (36), the fourth inequality is a result of (37) and (38), and the final inequality is due to (39). □

C   Local Convergence Results

We begin with the following technical lemmas used in the proof of local convergence.

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

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

Proof. We will prove this by leveraging the local RIP assumption. Notice the following:

$$\langle \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l, \mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l\rangle = \langle \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l, \mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l\rangle$$

$$= \langle \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l, \mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l - p^2\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l\rangle + p^2\langle \mathcal{P}_\mathbb{T}\boldsymbol{G}_l, \mathcal{P}_\mathbb{T}\boldsymbol{G}_l\rangle.$$

We can now leverage the variational characterization of the spectral norm and local RIP, proven in Theorem B.12, to bound the following:

$$-p^2(4\varepsilon_0)\|\mathcal{P}_\mathbb{T}\boldsymbol{G}_l\|_\mathrm{F}^2 \leq \langle \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l, \mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l - p^2\mathcal{P}_{\mathbb{T}_l}\boldsymbol{G}_l\rangle \leq p^2(4\varepsilon_0)\|\mathcal{P}_\mathbb{T}\boldsymbol{G}_l\|_\mathrm{F}^2.$$

As such, we can now bound the denominator as

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

Rearrangement of this last expression yields the upper and lower bounds on the step size derived above. □

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

$$\|\mathcal{P}_{\mathbb{T}_l} - \alpha_l\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l}\| \leq \frac{8\varepsilon_0}{1 - 4\varepsilon_0}. \tag{40}$$

Proof. From direct calculation, it follows that

$$\|\mathcal{P}_{\mathbb{T}_l} - \alpha_l\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l}\| \leq \|\mathcal{P}_{\mathbb{T}_l} - p^{-2}\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l}\| + |\alpha_l - p^{-2}|\|\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l}\|$$

$$\leq 4\varepsilon_0 + |\alpha_l - p^{-2}|\left(\|\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l} - p^2\mathcal{P}_{\mathbb{T}_l}\| + p^2\|\mathcal{P}_{\mathbb{T}_l}\|\right)$$

$$\leq 4\varepsilon_0 + \left(\frac{p^{-2}}{1 - 4\varepsilon_0} - \frac{p^{-2}(1 - 4\varepsilon_0)}{1 - 4\varepsilon_0}\right)\left(\|\mathcal{P}_{\mathbb{T}_l}\mathcal{M}_\Omega \mathcal{P}_{\mathbb{T}_l} - p^2\mathcal{P}_{\mathbb{T}_l}\| + p^2\|\mathcal{P}_{\mathbb{T}_l}\|\right)$$

$$\leq 4\varepsilon_0 + \left(\frac{p^{-2}}{1 - 4\varepsilon_0} - \frac{p^{-2}(1 - 4\varepsilon_0)}{1 - 4\varepsilon_0}\right)\left(4\varepsilon_0 p^2 + p^2\right)$$

$$= 4\varepsilon_0 + \frac{4\varepsilon_0}{1 - 4\varepsilon_0}(1 + 4\varepsilon_0)$$

$$= \frac{8\varepsilon_0}{1 - 4\varepsilon_0},$$

where the first inequality comes from the triangle inequality, the second inequality comes from Local RIP in Theorem B.12, the third inequality comes from the stepsize bound in Theorem C.1, the fourth inequality again comes from Theorem B.12, and the remainder comes from algebraic simplification of terms. This finishes the proof. $\qquad\square$

C.1 Proof of Theorem 5.4

We can now prove Theorem 5.4.

Proof. First, it follows that

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

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

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

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

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

Next, notice that from Theorem A.10 and the fact that $\mathcal{P}_{\mathbb{T}_l}\boldsymbol{X}_l = \boldsymbol{X}_l$,

$$
\begin{aligned}
I_2 &= 2\|(I - \mathcal{P}_{\mathbb{T}_l})\boldsymbol{X}_l - (I - \mathcal{P}_{\mathbb{T}_l})\boldsymbol{X}\|_{\mathrm{F}} \\
&= 2\|(I - \mathcal{P}_{\mathbb{T}_l})\boldsymbol{X}\|_{\mathrm{F}} \\
&\leq \frac{2\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}^2}{\lambda_r(\boldsymbol{X})} \\
&\leq \frac{\varepsilon_0 p^{1/2}}{32\left(\beta n \log n\right)^{1/4}}\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}} \\
&\leq \varepsilon_0\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}} \\
&\leq \frac{\varepsilon_0}{1 - 4\varepsilon_0}\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}},
\end{aligned}
$$

using Theorem A.10 and our initial local neighborhood assumption. Finally, we see that, following a similar argument as in the bound of I_2,

$$
\begin{aligned}
I_3 &\leq 2|\alpha_l| \|\mathcal{P}_{\mathbb{T}_l} \mathcal{M}_\Omega\| \|(I - \mathcal{P}_{\mathbb{T}_l})\boldsymbol{X}\|_{\mathrm{F}} \\
&\leq \frac{2p^{-2}}{1 - 4\varepsilon_0}\left[100 p^{3/2}\sqrt{\beta n \log n}\frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\lambda_r(\boldsymbol{X})} + p^{3/2}\sqrt{\frac{256\nu r \beta \log n}{3}}\right]\left(\frac{\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}}{\lambda_r(\boldsymbol{X})}\right)\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}} \\
&\leq \frac{2}{1 - 4\varepsilon_0}\left[100 p^{-1/2}\sqrt{\beta n \log n}\left(\frac{\varepsilon_0 p^{1/2}}{32\left(\beta n \log n\right)^{1/4}}\right)^2 + p^{-1/2}\sqrt{\frac{256\nu r \beta \log n}{3}}\frac{\varepsilon_0 p^{1/2}}{32\left(\beta n \log n\right)^{1/4}}\right]\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}} \\
&\leq \frac{\varepsilon_0}{1 - 4\varepsilon_0}\|\boldsymbol{X}_l - \boldsymbol{X}\|_{\mathrm{F}}
\end{aligned}
$$

where the second to last inequality follows from the same analysis conducted in Theorem B.12, just divided by 2. Collecting these results, we get

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

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

D Initialization Results (Proof of Lemma 5.5)

Proof. First, notice that for $\boldsymbol{W}_0 = p^{-1}\mathcal{R}_\Omega(\boldsymbol{X})$, we get

$$\|\boldsymbol{X}_0 - \boldsymbol{X}\| \le \|\boldsymbol{W}_0 - \boldsymbol{X}\| + \|\boldsymbol{W}_0 - \boldsymbol{X}_0\|$$
$$\le 2\|\boldsymbol{W}_0 - \boldsymbol{X}\|,$$

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 [57]. We now need a bound for this last term. Notice that $\boldsymbol{W}_0 - \boldsymbol{X} = \sum_{\boldsymbol{\alpha} \in \mathbb{I}}(p^{-1}\xi_{\boldsymbol{\alpha}} - 1)\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \boldsymbol{v}_{\boldsymbol{\alpha}}$ is a sum of zero-mean i.i.d random matrices, opening up use of Bernstein's inequality. In order to use this, define $\boldsymbol{Z}_{\boldsymbol{\alpha}} = (p^{-1}\xi_{\boldsymbol{\alpha}} - 1)\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \boldsymbol{v}_{\boldsymbol{\alpha}}$. We need a bound on $\|\boldsymbol{Z}_{\boldsymbol{\alpha}}\|$ and $\left\|\mathbb{E}\left[\sum_{\boldsymbol{\alpha} \in \mathbb{I}}\boldsymbol{Z}_{\boldsymbol{\alpha}}\right]^2\right\|$. First, notice that

$$\|\boldsymbol{Z}_{\boldsymbol{\alpha}}\| = \left\|(p^{-1}\xi_{\boldsymbol{\alpha}} - 1)\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle \boldsymbol{v}_{\boldsymbol{\alpha}}\right\|$$
$$\le (p^{-1} + 1)|\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle|\,\|\boldsymbol{v}_{\boldsymbol{\alpha}}\|$$
$$\le 2p^{-1}\left(\max_{\boldsymbol{\alpha} \in \mathbb{I}}|\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle|\right) =: c,$$

where the second inequality comes from the fact that $p \le 1$ and $\|\boldsymbol{v}_{\boldsymbol{\alpha}}\| < 1$ from Theorem A.8. Next, notice that

$$\left\|\mathbb{E}\left[\sum_{\boldsymbol{\alpha} \in \mathbb{I}}\boldsymbol{Z}_{\boldsymbol{\alpha}}^2\right]\right\| = \left\|\sum_{\boldsymbol{\alpha} \in \mathbb{I}}\mathbb{E}\left[p^{-2}\xi_{\boldsymbol{\alpha}} - 2\xi_{\boldsymbol{\alpha}}p^{-1} + 1\right]\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \boldsymbol{v}_{\boldsymbol{\alpha}}^2\right\|$$
$$= \left\|\sum_{\boldsymbol{\alpha} \in \mathbb{I}}\left(p^{-1} - 1\right)\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle^2 \boldsymbol{v}_{\boldsymbol{\alpha}}^2\right\|$$
$$\le p^{-1}\left(\max_{\boldsymbol{\alpha} \in \mathbb{I}}|\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle|\right)^2 \lambda_{\max}\left(\sum_{\boldsymbol{\alpha}}\boldsymbol{v}_{\boldsymbol{\alpha}}^2\right).$$

Now, as Theorem A.9, $\sum_{\boldsymbol{\alpha}}\boldsymbol{v}_{\boldsymbol{\alpha}}^2 = \frac{n^2 - 2n + 2}{4n}\boldsymbol{J}$. It follows that $\lambda_{\max}\left(\sum_{\boldsymbol{\alpha}}\boldsymbol{v}_{\boldsymbol{\alpha}}^2\right) = \frac{n^2 - 2n + 2}{4n} \le \frac{n}{4}$ as $\boldsymbol{J}$ is an orthogonal projection matrix. Thus,

$$\left\|\mathbb{E}\left[\sum_{\boldsymbol{\alpha} \in \mathbb{I}}\boldsymbol{Z}_{\boldsymbol{\alpha}}^2\right]\right\| \le \frac{np^{-1}}{4}\left(\max_{\boldsymbol{\alpha} \in \mathbb{I}}|\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle|\right) =: \sigma^2$$

Now to determine t, we note that

$$\frac{\sigma^2}{c} = \frac{np^{-1}\left(\max_{\boldsymbol{\alpha} \in \mathbb{I}}|\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle|\right)^2}{8p^{-1}\left(\max_{\boldsymbol{\alpha} \in \mathbb{I}}|\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle|\right)}$$
$$= \frac{n}{8}\left(\max_{\boldsymbol{\alpha} \in \mathbb{I}}|\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle|\right)$$
$$\ge \sqrt{\frac{2\beta n \log n}{3p}}\left(\max_{\boldsymbol{\alpha} \in \mathbb{I}}|\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}}\rangle|\right),$$

for $p \ge \frac{128\beta \log n}{3n}$. It follows that

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

verifying the probabilistic bound. To complete the proof, we use Theorem F.2, from which it follows that

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

This concludes the proof. $\qquad\square$

E   Robustness Guarantees

In this section, we will prove Theorem 6.2. To begin, we will prove a result highlighting the dependencies of the size of the noise on the reconstruction of an object

Lemma E.1. *Let $\hat{\boldsymbol{P}} = \boldsymbol{P} + \boldsymbol{N}$ where $\boldsymbol{N} \in \mathbb{R}^{n \times r}$ is a matrix with independent mean-zero entries, and $\hat{\boldsymbol{X}} = \hat{\boldsymbol{P}}\hat{\boldsymbol{P}}^\top$. Let $\lambda_1 \geq \cdots \geq \lambda_r > 0$ be the non-zero eigenvalues of $\boldsymbol{X}$ with corresponding eigenvectors $\boldsymbol{U}_i$, similarly $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_r \geq 0$ be the non-zero eigenvalues of $\hat{\boldsymbol{X}}$ with corresponding eigenvectors $\hat{\boldsymbol{U}}_i$. Assume that $\|\boldsymbol{N}\|_\infty \leq \frac{3\delta\|\boldsymbol{X}\|_{\mathrm{F}}}{8\beta r n^{1/2}\lambda_1^{1/2}\log n}$ for some $\delta \in (0,1)$ and some sufficiently large $\beta > \max\left\{1, \frac{3r}{8\log n}\right\}$. Additionally, let $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{n}^i \boldsymbol{n}^{i^\top}]$ be the covariance matrix of the columns of $\boldsymbol{N}$. Then*

$$\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_{\mathrm{F}} \leq 3\delta\|\boldsymbol{X}\|_{\mathrm{F}},$$

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

Proof. First, notice that

$$\hat{\boldsymbol{P}}\hat{\boldsymbol{P}}^\top = \boldsymbol{P}\boldsymbol{P}^\top + \boldsymbol{N}\boldsymbol{P}^\top + \boldsymbol{P}\boldsymbol{N}^\top + \boldsymbol{N}\boldsymbol{N}^\top,$$

so

$$\begin{aligned}
\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_{\mathrm{F}} &= \|\boldsymbol{N}\boldsymbol{P}^\top + \boldsymbol{P}\boldsymbol{N}^\top + \boldsymbol{N}\boldsymbol{N}^\top\|_{\mathrm{F}} \\
&\leq 2\|\boldsymbol{N}\boldsymbol{P}^\top\|_{\mathrm{F}} + \|\boldsymbol{N}\boldsymbol{N}^\top\|_{\mathrm{F}}.
\end{aligned}$$

We will first bound the term $\|\boldsymbol{N}\boldsymbol{N}^\top\|_{\mathrm{F}}$. Notice that

$$\begin{aligned}
\|\boldsymbol{N}\boldsymbol{N}^\top\|_{\mathrm{F}}^2 &= \sum_{i,j=1}^n \left(\sum_{k=1}^r N_{ik}N_{jk}\right)^2 \\
&\leq \sum_{i,j=1}^n \left(\sum_{k=1}^r \left(\frac{3\delta\|\boldsymbol{X}\|_{\mathrm{F}}}{8\beta r n^{1/2}\lambda_1^{1/2}\log n}\right)^2\right)^2 \\
&= n^2 \left(\frac{3\delta\|\boldsymbol{X}\|_{\mathrm{F}}}{8\beta n^{1/2}r^{1/2}\lambda_1^{1/2}\log n}\right)^4 \\
&= \left(\frac{3\delta}{8\beta r^{1/2}\log n}\right)^4 \frac{\|\boldsymbol{X}\|_{\mathrm{F}}^2}{\lambda_1^2}\|\boldsymbol{X}\|_{\mathrm{F}}^2 \\
&\leq \left(\frac{3\delta}{8\beta r^{1/2}\log n}\right)^4 r\|\boldsymbol{X}\|_{\mathrm{F}}^2 \\
&= \left(\frac{3\delta}{8\beta n^{1/2}r^{1/4}\log n}\right)^4 \|\boldsymbol{X}\|_{\mathrm{F}}^2 \\
&\leq \delta^2\|\boldsymbol{X}\|_{\mathrm{F}}^2,
\end{aligned}$$

where the first inequality follows from the bound on $\|\boldsymbol{N}\|_\infty$, the second inequality follows from the definition of the Frobenius norm, and the final inequality follows from $\beta n^{1/2}r^{1/4}\log n > 1$ and $\delta < 1$. Now, notice that $\mathbb{E}\left[\boldsymbol{N}\boldsymbol{P}^\top\right] = \boldsymbol{0}$, and can be decomposed as the sum of independent random matrices as $\boldsymbol{N}\boldsymbol{P}^\top = \sum_{i=1}^r \boldsymbol{n}^i \boldsymbol{p}^{i^\top}$, where $\boldsymbol{n}^i$ and $\boldsymbol{p}^i$ are the i-th columns of $\boldsymbol{N}$ and $\boldsymbol{P}$, respectively. As such, we will use Theorem A.1. Now, using the bound on $\|\boldsymbol{N}\|_\infty$, we have that

$$\begin{aligned}
\|\boldsymbol{n}^i \boldsymbol{p}^{i^\top}\| &= \|\boldsymbol{n}^i\|_2\|\boldsymbol{p}^i\|_2 \\
&= \left(\sum_{j=1}^n N_{ij}^2\right)^{1/2}\left(\sum_{j=1}^n P_{ji}^2\right)^{1/2} \\
&= \left(n\left(\frac{3\delta\|\boldsymbol{X}\|_{\mathrm{F}}}{8\beta r n^{1/2}\lambda_1^{1/2}\log n}\right)^2\right)^{1/2}\left(\sum_{j=1}^n P_{ji}^2\right)^{1/2}
\end{aligned}$$

49

$$= \frac{3\delta\|\boldsymbol{X}\|_{\mathrm{F}}}{8\beta r\lambda_1^{1/2}\log n}\left(\sum_{j=1}^{n}U_{ji}^2\lambda_i\right)^{1/2}$$

$$= \frac{3\delta\|\boldsymbol{X}\|_{\mathrm{F}}}{8\beta r\lambda_1^{1/2}\log n}\lambda_i^{1/2}\left(\sum_{j=1}^{n}U_{ji}^2\right)^{1/2}$$

$$\leq \frac{3\delta\|\boldsymbol{X}\|_{\mathrm{F}}}{8\beta r\log n}\left(\sum_{j=1}^{n}U_{ji}^2\right)^{1/2}$$

$$= \frac{3\delta\|\boldsymbol{X}\|_{\mathrm{F}}}{8\beta r\log n} := c,$$

where the third line comes from (3). Next, we need to estimate

$$\max\left\{\left\|\mathbb{E}\left[\boldsymbol{n}^i\boldsymbol{p}^{i\top}\boldsymbol{p}^i\boldsymbol{n}^{i\top}\right]\right\|, \left\|\mathbb{E}\left[\boldsymbol{p}^i\boldsymbol{n}^{i\top}\boldsymbol{n}^i\boldsymbol{p}^{i\top}\right]\right\|\right\}.$$

Looking at the first term first, we see that

$$\|\mathbb{E}(\boldsymbol{n}^i\boldsymbol{p}^{i\top}\boldsymbol{p}^i\boldsymbol{n}^{i\top})\| = \left\|\mathbb{E}\left[\boldsymbol{n}^i\left(\sum_{j=1}^{n}P_{ji}^2\right)\boldsymbol{n}^{i\top}\right]\right\|$$

$$= \left\|\mathbb{E}\left[\boldsymbol{n}^i\left(\sum_{j=1}^{n}U_{ji}^2\lambda_i\right)\boldsymbol{n}^{i\top}\right]\right\|$$

$$= \lambda_i\left\|\mathbb{E}\left[\boldsymbol{n}^i\boldsymbol{n}^{i\top}\right]\right\|$$

$$= \lambda_i\lambda_{\max}(\boldsymbol{\Sigma}).$$

Looking at the second term, we see that

$$\left\|\mathbb{E}\left[\boldsymbol{p}^i\boldsymbol{n}^{i\top}\boldsymbol{n}^i\boldsymbol{p}^{i\top}\right]\right\| = \left\|\boldsymbol{p}^i\mathbb{E}\left[\boldsymbol{n}^{i\top}\boldsymbol{n}^i\right]\boldsymbol{p}^{i\top}\right\|$$

$$= \mathbb{E}\left[\boldsymbol{n}^{i\top}\boldsymbol{n}^i\right]\left\|\boldsymbol{p}^i\boldsymbol{p}^{i\top}\right\|$$

$$= \mathrm{Trace}(\boldsymbol{\Sigma})\|\boldsymbol{p}^i\boldsymbol{p}^{i\top}\|$$

$$= \mathrm{Trace}(\boldsymbol{\Sigma})\|\boldsymbol{p}^i\|_2^2$$

$$= \lambda_i\mathrm{Trace}(\boldsymbol{\Sigma})$$

$$\leq n\lambda_i\lambda_{\max}(\boldsymbol{\Sigma}).$$

As the entries of $\boldsymbol{N}$ are independent, $\boldsymbol{\Sigma}$ is diagonal, and as we have a bound on $\|\boldsymbol{N}\|_\infty$ it follows that

$$\lambda_{\max}(\boldsymbol{\Sigma}) \leq \left(\frac{3\delta\|\boldsymbol{X}\|_{\mathrm{F}}}{8\beta r n^{1/2}\lambda_1^{1/2}\log n}\right)^2.$$

As such, the variance parameter $\sigma^2 = n\sum_i\lambda_i\lambda_{\max}(\boldsymbol{\Sigma}) = n\lambda_{\max}(\boldsymbol{\Sigma})\|\boldsymbol{X}\|_*$. As

$$\frac{\sigma^2}{c} = \frac{n\lambda_{\max}(\boldsymbol{\Sigma})\|\boldsymbol{X}\|_*}{\frac{3\delta\|\boldsymbol{X}\|_{\mathrm{F}}}{8\beta r\log n}}$$

$$\leq n\left(\frac{3\delta\|\boldsymbol{X}\|_{\mathrm{F}}}{8\beta r n^{1/2}\lambda_1^{1/2}\log n}\right)^2\frac{\|\boldsymbol{X}\|_*}{\|\boldsymbol{X}\|_{\mathrm{F}}}\frac{8\beta r\log n}{3\delta}$$

$$= \frac{3\delta\|\boldsymbol{X}\|_{\mathrm{F}}}{8\beta r\log n}\frac{\|\boldsymbol{X}\|_*}{\lambda_1}$$

$$\leq \frac{3\delta\|\boldsymbol{X}\|_{\mathrm{F}}}{8\beta\log n}$$

50

$$\leq \frac{\delta \|\boldsymbol{X}\|_{\mathrm{F}}}{r},$$

where the last inequality follows from the fact that $\frac{8}{3}\beta \log n > r$ for sufficiently large β, as stipulated in the Lemma statement. As such, for $t = \frac{\delta}{r}\|\boldsymbol{X}\|_{\mathrm{F}}$ we have that

$$\mathbb{P}\left[\|\boldsymbol{N}\boldsymbol{P}\| \geq t\right] \leq 2n \exp\left(\frac{-3t}{8c}\right)$$

$$= 2n \exp\left(\frac{-3\frac{\delta}{r}\|\boldsymbol{X}\|_{\mathrm{F}}}{8\frac{3\delta\|\boldsymbol{X}\|_{\mathrm{F}}}{8\beta r \log n}}\right)$$

$$= 2n \exp\left(-\beta \log n\right).$$

The proof statement now follows from the fact that

$$\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_{\mathrm{F}} \leq \sqrt{r}\|\boldsymbol{X} - \hat{\boldsymbol{X}}\| \leq r\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|.$$

This finishes the proof. $\qquad\square$

Next we will prove the following lemma showing that bounded noise on the points does not change the incoherence of a Gram matrix substantially.

Lemma E.2. *For $\hat{\boldsymbol{P}} = \boldsymbol{P} + \boldsymbol{N}$, where $\boldsymbol{N}$ is a mean-zero random matrix. Let $\boldsymbol{X} = \boldsymbol{P}\boldsymbol{P}^{\top}$ and $\hat{\boldsymbol{X}} = \hat{\boldsymbol{P}}\hat{\boldsymbol{P}}^{\top}$, and let $\lambda_1 \geq \cdots \geq \lambda_r > 0$ be the eigenvalues of $\boldsymbol{X}$. If $\|\boldsymbol{N}\|_{\infty} \leq \frac{\nu\lambda_1^{1/2}\gamma}{16n^{3/2}\beta\kappa \log n}$ for some $\gamma > 0$, $\beta > 1$, where κ is the condition number of $\boldsymbol{X}$, then*

$$\left\|\mathcal{P}_{\hat{U}}\boldsymbol{w}_{\boldsymbol{\alpha}}\right\| \leq \frac{(2+\gamma)\nu r}{2n},$$

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

Proof. This result will follow from the classic Davis-Kahan $\sin\Theta$ Theorem, seen in Theorem A.4. Let $\mathbb{U}$, $\hat{\mathbb{U}}$ be the subspace spanned by the columns of $\boldsymbol{U}, \hat{\boldsymbol{U}}$ respectively. First, as $\left\|\mathcal{P}_U - \mathcal{P}_{\hat{U}}\right\|_{\mathrm{F}} = \|\sin\Theta(\mathbb{U}, \hat{\mathbb{U}})\|_{\mathrm{F}}$ from [98], we can see that

$$\left\|\mathcal{P}_U - \mathcal{P}_{\hat{U}}\right\|_{\mathrm{F}} = \|\sin\Theta(U, \hat{U})\|_{\mathrm{F}} \leq \frac{\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_{\mathrm{F}}}{\lambda_r}.$$

Next, notice that

$$\left\|\mathcal{P}_{\hat{U}}\boldsymbol{w}_{\boldsymbol{\alpha}}\right\|_{\mathrm{F}} \leq \left\|\left(\mathcal{P}_{\hat{U}} - \mathcal{P}_U\right)\right\|_{\mathrm{F}} \|\boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}} + \|\mathcal{P}_U\boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}$$

$$\leq \frac{2\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_{\mathrm{F}}}{\lambda_r(\boldsymbol{X})} + \frac{\nu r}{2n}$$

$$\leq 2\frac{\nu r \gamma}{2n} + \frac{\nu r}{2n}$$

$$\leq \frac{(2+\gamma)\nu r}{2n},$$

where the third inequality follows from Theorem E.1, thus ending the proof. $\qquad\square$

F Incoherence Results

In this section, we provide proofs for the statements in Section 3.

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

Proof. To see this result, notice that

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

$$\leq \sum_{\boldsymbol{\beta} \in \mathbb{I}} \left| H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \right| \left\| \mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\beta}} \right\|_{\mathrm{F}}$$

$$\leq \sqrt{\frac{\nu r}{8n}} \sum_{\boldsymbol{\beta} \in \mathbb{I}} \left| H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \right|,$$

and as $\sum_{\boldsymbol{\alpha} \in \mathbb{I}} \left| H^{\boldsymbol{\alpha}\boldsymbol{\beta}} \right| \leq 2$ from Theorem A.8, the claim follows. An identical proof shows the second result, with $\mathcal{P}_{\mathbb{T}}$ in place of $\mathcal{P}_U$. $\qquad\square$

Lemma F.2. *Let $\boldsymbol{X} \succeq \boldsymbol{0}$ be a rank-r, ν-incoherent matrix satisfying (21) with constant ν. Then*

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

Proof. To see the above statement, notice that

$$
\begin{aligned}
\frac{(\max_{\boldsymbol{\alpha} \in \mathbb{I}} |\langle \boldsymbol{X}, \boldsymbol{w}_{\boldsymbol{\alpha}} \rangle|)}{\|\boldsymbol{X}\|} &= \frac{1}{\|\boldsymbol{X}\|} \max_{i,j} |X_{ii} + X_{jj} - 2X_{ij}| \\
&= \frac{1}{\|\boldsymbol{X}\|} \max_{i,j} \left| \sum_{kl} U_{ik} D_{kl} U_{il} + \sum_{kl} U_{jk} D_{kl} U_{jl} - 2 \sum_{kl} U_{ik} D_{kl} U_{jl} \right| \\
&= \frac{1}{\|\boldsymbol{X}\|} \max_{i,j} \left| \sum_{k=1}^{r} U_{ik} \lambda_k U_{ik} + U_{jk} \lambda_k U_{jk} - 2 U_{ik} \lambda_k U_{jk} \right| \\
&= \max_{i,j} \left| \sum_{k=1}^{r} U_{ik} \frac{\lambda_k}{\lambda_1} U_{ik} + U_{jk} \frac{\lambda_k}{\lambda_1} U_{jk} - 2 U_{ik} \frac{\lambda_k}{\lambda_1} U_{jk} \right| \\
&\leq \max_{ij} \sum_{k=1}^{r} \left| U_{ik}^2 + U_{jk}^2 - 2 U_{ik} U_{jk} \right| \\
&= \max_{i,j} |\boldsymbol{u}_i^\top \boldsymbol{u}_i + \boldsymbol{u}_j^\top \boldsymbol{u}_j - 2 \boldsymbol{u}_i^\top \boldsymbol{u}_j| \\
&= \max_{i,j} (\boldsymbol{u}_i - \boldsymbol{u}_j)^\top (\boldsymbol{u}_i - \boldsymbol{u}_j) \\
&= \max_{\boldsymbol{\alpha} \in \mathbb{I}} \frac{1}{2} \|\mathcal{P}_U \boldsymbol{w}_{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 \\
&\leq \frac{\nu r}{4n},
\end{aligned}
$$

where the first inequality comes from the definition of the spectral norm, the penultimate line follows from a rescaled form of (13) in accordance with Theorem 5.1, and the final line follows from (21), thus concluding the proof. $\qquad\square$

Lemma F.3. *Let μ be an a.s. bounded, mean-zero, sub-Gaussian distribution with positive definite covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{r \times r}$. Let n points $\{\boldsymbol{p}_i\}_{i=1}^{n} \sim \mu$ be sampled i.i.d., and let $\boldsymbol{P} = [\boldsymbol{p}_1 \ldots \boldsymbol{p}_n]^\top \in \mathbb{R}^{n \times d}$ be the corresponding point matrix with Gram matrix $\boldsymbol{X}$, which has condition number κ. Let $\|\boldsymbol{p}_i\|_{\psi_2} \leq K$ for some $K > 0$. Then with probability at least $1 - Cn^{-2}$ for some absolute constant $C > 0$, the incoherence parameter of $\boldsymbol{X}$ is bounded by*

$$\nu \leq \mathcal{O}\left(\frac{\kappa \log n}{\sqrt{r}} \right)$$

Proof. This proof is much the same as the proof in Section 3. First, we remark that

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

52

$$= 2 \operatorname{Trace} \left(\mathbb{E} \left[\boldsymbol{p}_i {\boldsymbol{p}_i}^\top \right] \right)$$
$$= 2 \operatorname{Trace}(\boldsymbol{\Sigma}) \leq 2r\lambda_1(\boldsymbol{\Sigma}),$$

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

Next, following the argument of Theorem 3.3 but replacing $2r$ with $\mathbb{E}\left[(\boldsymbol{p}_i - \boldsymbol{p}_j)^\top (\boldsymbol{p}_i - \boldsymbol{p}_j) \right]$, we have that, with probability at least $1 - Cn^{-2}$

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

Next, we show that we can upper bound K by $\lambda_1(\boldsymbol{\Sigma})$ for sub-Gaussian μ. We will use a moment generating function bound to prove this. First, from Definition 3.4.1 in [53], we have that $\|\boldsymbol{p}_i\|_{\psi_2} = \sup_{\|\boldsymbol{u}\|_2 = 1} \|\boldsymbol{u}^\top \boldsymbol{p}_i\|_{\psi_2}$. Using the moment-generating technique, we can see that

$$
\begin{aligned}
\mathbb{E}\left[\exp \left(t^2 (\boldsymbol{u}^\top \boldsymbol{p}_i)^2 \right) \right] &= \mathbb{E}\left[\exp \left(t^2 \boldsymbol{u}^\top \boldsymbol{p}_i {\boldsymbol{p}_i}^\top \boldsymbol{u} \right) \right] \\
&\leq \sup_u \mathbb{E}\left[\exp \left(t^2 \boldsymbol{u}^\top \boldsymbol{p}_i {\boldsymbol{p}_i}^\top \boldsymbol{u} \right) \right] \\
&\leq \mathbb{E}\left[\sup_u \exp \left(t^2 \boldsymbol{u}^\top \boldsymbol{p}_i {\boldsymbol{p}_i}^\top \boldsymbol{u} \right) \right] \quad \leq \exp \left(t^2 \lambda_1(\boldsymbol{\Sigma}) \right).
\end{aligned}
$$

This gives us the bound $K \leq C\lambda_1(\boldsymbol{\Sigma})^{1/2}$ for some absolute constant $C > 0$. Leveraging this, along with the fact that from Theorem 3.2 that $\lambda_r(\boldsymbol{X}) \approx n\lambda_r(\boldsymbol{\Sigma})$, we have that for some $c > 0$ with high probability that

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

This concludes the proof. $\qquad \square$

G   Further Background

G.1   Dual Bases

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 $\boldsymbol{X}_i = \sum_{j=1}^n H_{ij} \boldsymbol{Y}_j$ [98].

G.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 $r \leq n$, we denote the set $\mathcal{N}_r = \{\boldsymbol{X} \in \mathbb{R}^{n \times n} \mid \operatorname{rank}(\boldsymbol{X}) = r\}$. Although not obvious at first

glance, it is well-known that $\mathcal{N}_r$ is a smooth Riemannian manifold [64, 99]. To make this a Riemannian manifold, we equip it with the standard trace inner product as a metric, or $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \text{Trace}(\boldsymbol{A}^\top \boldsymbol{B})$, restricted to the tangent bundle $T\mathcal{N}_r$, which is the disjoint union of tangent spaces [99].

Additionally, the tangent space at a point $\boldsymbol{X} \in \mathcal{N}_r$ is known and can be characterized [54, 64, 99]. 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 $\{\boldsymbol{X}_l\}_{l \geq 0}$, 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 $\boldsymbol{X}$ [64]. 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 $T_{\boldsymbol{X}}\mathcal{N}_r$ as follows [54, 64, 99]:

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

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 $\mathcal{N}_r$ has been investigated in detail for quite some time, and retraction-based methods are of particular interest to this work [54, 64, 100–105]. 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 $\mathcal{N}_r$, 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 $|\lambda_1| \geq \cdots \geq |\lambda_n|$ 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 $\mathcal{N}_r$, 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 $f : \mathcal{N}_r \to \mathbb{R}$ 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 $f|_{\mathcal{N}_r}$, denoted $\text{grad } f$, for $\boldsymbol{X}_l \in \mathcal{N}_r$ is given by:

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

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:

$$\boldsymbol{X}_{l+1} = \mathcal{H}_r(\boldsymbol{X}_l - \alpha_l \mathcal{P}_{\mathbb{T}_l} \nabla f(\boldsymbol{X}_l)). \tag{41}$$

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 7.

This is a simple first pass to first-order optimization on Riemannian manifolds, and is not meant to be exhaustive. Interested readers should consult [99, 100] for further details on first-order methods on matrix (and other Riemannian) manifolds, along with convergence analysis for these algorithms.
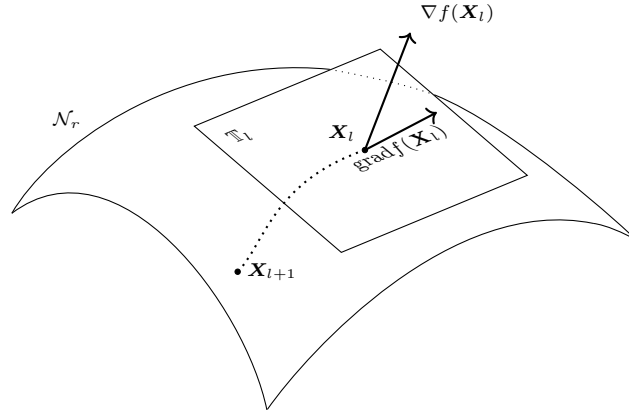
Figure 7: A diagram of a simple first-order retraction method on $\mathcal{N}_r$. Again, $\nabla f(\boldsymbol{X}_l)$ is the Euclidean gradient of f at $\boldsymbol{X}_l$, grad $f(\boldsymbol{X}_l)$ is the Riemannian gradient at $\boldsymbol{X}_l$, and $\boldsymbol{X}_{l+1} = \mathcal{H}_r(\boldsymbol{X}_l - \alpha_l \text{grad} f(\boldsymbol{X}_l))$, as in (41).