The optimization landscape of Spectral neural network

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2024

Abstract

There is a large variety of machine learning methodologies that are based on the extraction of spectral geometric information from data. However, the implementations of many of these methods often depend on traditional eigensolvers, which present limitations when applied in practical online big data scenarios. To address some of these challenges, researchers have proposed different strategies for training neural networks as alternatives to traditional eigensolvers, with one such approach known as Spectral Neural Network (SNN). In this paper, we initiate a theoretical exploration of the optimization landscape of SNN's objective to shed light on the training dynamics of SNN. Unlike typical studies of convergence to global solutions of NN training dynamics, SNN presents an additional complexity due to its non-convex ambient loss function, a feature that is common in unsupervised learning settings. We show that the ambient optimization landscape is benign in a quotient geometry. Furthermore, we use the experimental results to see that the parameterized optimization landscape inherits from the benignness of the ambient landscape if the neural network is appropriately overparameterized.

1. Main

In the past decades, researchers from a variety of disciplines have studied the use of spectral geometric methods to process, analyze, and learn from data. These methods have been used in supervised learning [2, 5, 24], clustering [21, 27], dimensionality reduction [4, 9], and contrastive learning [14]. While the aforementioned methods have strong theoretical foundations, their algorithmic implementations often depend on traditional eigensolvers. These eigensolvers tend to underperform in practical big data scenarios due to high computational demands and memory constraints. Moreover, they are particularly vulnerable in online settings since the introduction of new data typically necessitates a full computation from scratch.

Spectralnet [23] and Spectral Neural Network (SNN) [14] have been proposed to overcome these issues. In these approaches, the goal is to find neural networks that can approximate the spectrum of a large target matrix, and the differences among the approaches lie mostly in the specific loss functions used for training. Here we focus on SNN. A SNN is trained by minimizing the *spectral constrastive* loss function:

$$\min_{\theta\in\Theta} L(\theta) \stackrel{\text{def}}{=} \ell(\mathbf{Y}_\theta), \quad \text{where} \quad \ell(\mathbf{Y}) \stackrel{\text{def}}{=} \left\|\mathbf{Y}\mathbf{Y}^\top - \mathcal{A}_{\mathbf{n}}\right\|_{\mathrm{F}}^2, \quad \mathbf{Y}\in\mathbb{R}^{n\times r}, \tag{1.1}$$

through first-order optimization methods. In the above and in the sequel, θ represents the vector of parameters of the neural network $f_\theta : \mathbb{R}^d \to \mathbb{R}^r$, the matrix $\mathbf{Y}_\theta$ is the $n \times r$ matrix whose rows are the outputs $f_\theta(x_1), \ldots, f_\theta(x_n)$, and $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm. The mapping f_θ can be

interpreted as a feature or representation map for the input data. In the remainder, $\mathbf{Y}^*$ will denote a minimizer of ℓ and we will use f_{θ^*} to denote the neural network that minimizes $L(\theta)$. We will make a minimum assumption on $\mathcal{A}_\mathbf{n}$ in the sequel. We only assume $\mathcal{A}_\mathbf{n}$ is symmetric and PSD, which are mild assumptions in spectral embedding setup.

In this paper, we describe the optimization landscape of ℓ under an eigengap assumption of $\mathcal{A}_\mathbf{n}$. In a *quotient* geometry, the optimization landscape of ℓ is shown benign. We remark that the ambient landscape of Spectralnet is recently discussed in [1]. We further observe that the parameterized optimization landscape inherits from the ambient landscape by using some experimental results.

Related work Other types of NN-based Eigensolvers have been considered in [22] and [11]. [22] uses a bi-level optimization algorithm to solve a constrained optimization problem. This algorithm's computational complexity is typically higher than the one of SNN training and it requires keeping certain covariance matrices in memory during updates. [11] takes a similar approach as [22] but can avoid the bi-level optimization of the latter. This, however, comes at the expense of having an intractable theoretical computational complexity.

One of our main objects of study in this paper is the ambient problem Equation 1.1. This formulation of the problem is related to linear networks. Linear networks are neural networks with identity activation. A variety of prior works have studied many different aspects of shallow linear networks such as their loss landscape and their associated optimization dynamics [3, 7, 20, 26]. Of relevance are also other works in the literature studying optimization problems very closely related to Equation 2.1 [8, 17, 19]. For example, in Section 3 in [17], there is a landscape analysis for Equation 2.1 when the matrix $\mathcal{A}_\mathbf{n}$ is assumed to have rank smaller than or equal to r. That setting is typically referred to as overparameterized or exactly parameterized, whereas here, our focus is on the underparameterized setting.

2. Preliminary Results

In this section, we briefly provide some technical results that will serve as foundations for the analysis presented in this paper.

Rotational Invariance Recall the ambient optimization problem defined in Equation 1.1 as

$$\min_{\mathbf{Y}\in\mathbb{R}^{n\times r}} \ell(\mathbf{Y}), \quad \text{where} \quad \ell(\mathbf{Y}) \stackrel{\text{def}}{=} \left\|\mathbf{Y}\mathbf{Y}^\top - \mathcal{A}_\mathbf{n}\right\|_{\mathrm{F}}^2. \tag{2.1}$$

Suppose $\mathbf{Y}$ is a stationary point of Equation 2.1. Then $\mathbf{Y}\boldsymbol{O}$ is also a stationary point for any $r \times r$ orthogonal matrix $\boldsymbol{O} \in \mathbb{O}_r$. This implies that the loss function ℓ is non-convex in any neighborhood of a stationary point [18]. Hence we shall consider a quotient geometry to attempt to remove the local non-convexity induced by the action of the orthogonal group. Let $\overline{\mathcal{N}}^n_{r+}$ be the space of $n \times r$ matrices with full column rank. To define the quotient manifold, we encode the invariance mapping, i.e., $\mathbf{Y} \to \mathbf{Y}\boldsymbol{O}$, by defining the equivalence classes $[\mathbf{Y}] = \{\mathbf{Y}\boldsymbol{O} : \boldsymbol{O} \in \mathbb{O}_r\}$. From [16], $\mathcal{N}^n_{r+} \stackrel{\text{def}}{=} \overline{\mathcal{N}}^n_{r+}/\mathbb{O}_r$ is a quotient manifold of $\overline{\mathcal{N}}^n_{r+}$. For a detailed introduction to Riemannian optimization see [6]. Since ℓ is invariant within each the equivalence classes of $\overline{\mathcal{N}}^n_{r_+}$, one obtains the following optimization problem on the quotient manifold $\mathcal{N}^n_{r_+}$:

$$\min_{[\mathbf{Y}]\in\mathcal{N}^n_{r_+}} H([\mathbf{Y}]) \stackrel{\text{def}}{=} \frac{1}{2}\left\|\mathbf{Y}\mathbf{Y}^\top - \mathcal{A}_\mathbf{n}\right\|_{\mathrm{F}}^2, \tag{2.2}$$

which can be approached via Riemannian first order methods. In this case, however, the Riemannian gradient descent is equivalent to standard gradient descent in the Euclidean geometry. This is the reason why the results in Section 3 directly justify the ability of the gradient based algorithm in the Euclidean geometry to find global optimizers of the ambient space problem Equation 2.2.

First Order Stationary Points Since $\mathcal{A}_{\mathbf{n}}$ is a PSD matrix, the Eckart–Young–Mirsky theorem (see [12]) implies that the global optimizers of Equation 2.1 are the matrices $\mathbf{Y}$ of the form $\mathbf{Y} = \mathbf{Y}^* \boldsymbol{O}$, where $\boldsymbol{O} \in \mathbb{O}_r$ and

$$\mathbf{Y}^* \stackrel{\text{def}}{=} \begin{bmatrix} | & & | \\ \sqrt{\lambda_1(\mathcal{A}_{\mathbf{n}})} v_1 & \dots & \sqrt{\lambda_r(\mathcal{A}_{\mathbf{n}})} v_r \\ | & & | \end{bmatrix}.$$

In the above, $\lambda_l(\mathcal{A}_{\mathbf{n}})$ represents the l-th largest eigenvalue of $\mathcal{A}_{\mathbf{n}}$ and v_l is a corresponding eigenvector with Euclidean norm one. In case there are repeated eigenvalues, the corresponding v_l need to be chosen to be orthogonal to each other.

Next, we need to understand the other first-order stationary points (FOSP) of Equation 2.2. We use SVD to get $\mathbf{Y} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{n\times k}$, $\mathbf{D} \in \mathbb{R}^{k\times k}$ is a diagonal matrix, and $\mathbf{V} \in \mathbb{R}^{k\times r}$. In addition, $\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V} = \mathbf{I}_k$, where k is the rank of $\mathbf{Y}$ and $\mathbf{I}_k$ is the $k \times k$ identity matrix.

Theorem 2.1 (FOSP of Equation 2.2) *Let $\overline{\mathbf{U}}\boldsymbol{\Sigma}\overline{\mathbf{U}}^\top$ be $\mathcal{A}_{\mathbf{n}}$'s SVD factorization, and let $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{1/2}$. Then for any S subset of $[n]$ we have that $\left[\overline{\mathbf{U}}_S\boldsymbol{\Lambda}_S\right]$ is a Riemannian FOSP of Equation 2.2. Further, these are the only Riemannian FOSPs.*

Theorem 2.1 shows that linear combinations of eigenvectors can be used to construct Riemannian first-order stationary points (FOSP) of Equation 2.2. This theorem also shows that there are many FOSPs of Equation 2.2. This is quite different from the regime studied in [19]. In general, gradient descent is known to converge to a FOSP. Hence one might expect that if we initialized near one of the saddle points, then we might converge to that saddle point.

3. Landscape of SNN's Ambient Optimization Problem

In this section, we focus on the non-convexity due to the loss function, and show that gradient descent converges to the global minimum of ℓ. We do this by characterizing the optimization landscape of ℓ. To analyze the landscape for Equation 2.2, we need expressions for the Riemannian gradient, the Riemannian Hessian, as well as the geodesic distance d on this quotient manifold. By Lemma 2 from [19], we have that

$$d\left([\mathbf{Y}_1], [\mathbf{Y}_2]\right) = \min_{\mathbf{Q}\in\mathbb{O}_r} \|\mathbf{Y}_2\mathbf{Q} - \mathbf{Y}_1\|_{\mathrm{F}}$$

and from Lemma 3 from [19], we have that

$$\begin{aligned} \overline{\operatorname{grad} H([\mathbf{Y}])} &= 2\left(\mathbf{Y}\mathbf{Y}^\top - \mathcal{A}_{\mathbf{n}}\right)\mathbf{Y}, \\ \overline{\operatorname{Hess} H([\mathbf{Y}])}\left[\theta_{\mathbf{Y}}, \theta_{\mathbf{Y}}\right] &= \left\|\mathbf{Y}\theta_{\mathbf{Y}}^\top + \theta_{\mathbf{Y}}\mathbf{Y}^\top\right\|_{\mathrm{F}}^2 + 2\left\langle \mathbf{Y}\mathbf{Y}^\top - \mathcal{A}_{\mathbf{n}}, \theta_{\mathbf{Y}}\theta_{\mathbf{Y}}^\top\right\rangle. \end{aligned} \tag{3.1}$$

Finally, by the classical theory on low-rank approximation (Eckart–Young–Mirsky theorem [12]), $[\mathbf{Y}^*]$ is the unique global minimizer of Equation 2.2. Let $\kappa^* = \sigma_1(\mathbf{Y}^*)/\sigma_r(\mathbf{Y}^*)$ be the condition

number of $\mathbf{Y}^*$. Here, $\sigma_i(A)$ is the i^{th} largest singular value of A, and $\|A\| = \sigma_1(A)$ is its spectral norm. Our precise assumption on the matrix $\mathcal{A}_n$ for this section is as follows.

Assumption 1 (Eigengap) *$\sigma_{r+1}(\mathcal{A}_\mathbf{n})$ is strictly smaller than $\sigma_r(\mathcal{A}_\mathbf{n})$.*

Let $\mu, \alpha, \beta, \gamma \geqslant 0$. We then split the landscape of $H([\mathbf{Y}])$ into the following five regions (not necessarily non-overlapping).

$$
\begin{aligned}
\mathcal{R}_1 &\overset{\text{def}}{=} \left\{ \mathbf{Y} \in \mathbb{R}_*^{n\times r} \middle| d\left([\mathbf{Y}],[\mathbf{Y}^*]\right) \leqslant \mu\sigma_r\left(\mathbf{Y}^*\right)/\kappa^* \right\}, \\
\mathcal{R}_2 &\overset{\text{def}}{=} \left\{ \mathbf{Y} \in \mathbb{R}_*^{n\times r} \middle| \begin{array}{l} d\left([\mathbf{Y}],[\mathbf{Y}^*]\right) > \mu\sigma_r\left(\mathbf{Y}^*\right)/\kappa^*, \|\overline{\operatorname{grad} H([\mathbf{Y}])}\|_{\mathrm{F}} \leqslant \alpha\mu\sigma_r^3\left(\mathbf{Y}^*\right)/\left(4\kappa^*\right), \\ \|\mathbf{Y}\| \leqslant \beta\left\|\mathbf{Y}^*\right\|, \left\|\mathbf{Y}\mathbf{Y}^\top\right\|_{\mathrm{F}} \leqslant \gamma\left\|\mathbf{Y}^*\mathbf{Y}^{*\top}\right\|_{\mathrm{F}} \end{array} \right\}, \\
\mathcal{R}_3' &\overset{\text{def}}{=} \left\{ \mathbf{Y} \in \mathbb{R}_*^{n\times r} \middle| \begin{array}{l} \|\overline{\operatorname{grad} H([\mathbf{Y}])}\|_{\mathrm{F}} > \alpha\mu\sigma_r^3\left(\mathbf{Y}^*\right)/\left(4\kappa^*\right), \|\mathbf{Y}\| \leqslant \beta\left\|\mathbf{Y}^*\right\|, \\ \left\|\mathbf{Y}\mathbf{Y}^\top\right\|_{\mathrm{F}} \leqslant \gamma\left\|\mathbf{Y}^*\mathbf{Y}^{*\top}\right\|_{\mathrm{F}} \end{array} \right\}, \\
\mathcal{R}_3'' &\overset{\text{def}}{=} \left\{ \mathbf{Y} \in \mathbb{R}_*^{n\times r} \middle| \|\mathbf{Y}\| > \beta\|\mathbf{Y}^*\|, \|\mathbf{Y}\mathbf{Y}^\top\|_{\mathrm{F}} \leqslant \gamma\|\mathbf{Y}^*\mathbf{Y}^{*\top}\|_{\mathrm{F}} \right\}, \\
\mathcal{R}_3''' &\overset{\text{def}}{=} \left\{ \mathbf{Y} \in \mathbb{R}_*^{n\times r} \middle| \|\mathbf{Y}\mathbf{Y}^\top\|_{\mathrm{F}} > \gamma\left\|\mathbf{Y}^*\mathbf{Y}^{*\top}\right\|_{\mathrm{F}} \right\},
\end{aligned}
$$

We show that for small values of μ, the *loss function is geodesically convex* in $\mathcal{R}_1$. $\mathcal{R}_2$ is then defined as the region outside of $\mathcal{R}_1$ such that the Riemannian gradient is small relative to μ. Hence this is the region in which we are close to the saddle points. We show that for this region there is *always an escape direction* (i.e., directions where the Hessian is strictly negative). $\mathcal{R}_3'$, $\mathcal{R}_3''$, and $\mathcal{R}_3'''$ are the remaining regions. We show that the *Riemannian gradient is large* (relative to μ) in these regions. Finally, it is easy to see that $\mathcal{R}_1 \bigcup \mathcal{R}_2 \bigcup \mathcal{R}_3' \cup \mathcal{R}_3'' \bigcup \mathcal{R}_3''' = \mathbb{R}_*^{n\times r}$.

Theorem 3.1 (Local Geodesic Strong Convexity and Smoothness of Equation 2.2) *Suppose $0 \leqslant \mu \leqslant \kappa^*/3$. Given that Assumption 1 holds, for any $\mathbf{Y} \in \mathcal{R}_1$,*

$$
\begin{aligned}
\sigma_{\min}(\overline{\operatorname{Hess} H([\mathbf{Y}])}) &\geqslant \left(2\left(1-\mu/\kappa^*\right)^2 - (14/3)\mu\right)\sigma_r\left(\mathcal{A}_\mathbf{n}\right) - 2\sigma_{r+1}(\mathcal{A}_\mathbf{n}), \\
\sigma_{\max}(\overline{\operatorname{Hess} H([\mathbf{Y}])}) &\leqslant 4\left(\sigma_1\left(\mathbf{Y}^*\right) + \mu\sigma_r\left(\mathbf{Y}^*\right)/\kappa^*\right)^2 + 14\mu\sigma_r^2\left(\mathbf{Y}^*\right)/3
\end{aligned}
$$

In particular, if μ is further chosen such that $\left(2\left(1-\mu/\kappa^*\right)^2 - (14/3)\mu\right)\sigma_r\left(\mathcal{A}_\mathbf{n}\right) - 2\sigma_{r+1}(\mathcal{A}_\mathbf{n}) > 0$, we have $H([\mathbf{Y}])$ is geodesically strongly convex and smooth in $\mathcal{R}_1$.

Theorem 3.1 guarantees that the optimization problem Equation 2.2 is geodesically strongly convex and smooth in a neighborhood of $[\mathbf{Y}^*]$. It also shows that if $\mathbf{Y}$ is close to the global minimizer, then Riemannian gradient descent stays in $\mathcal{R}_1$ and converges to the global minimizer of the quotient space linearly following the proof of 6, Theorem 11.29. Without quotient out the rotation invariance property, the Riemannian strong convexity cannot be guaranteed.

In general, gradient descent is known to converge to a FOSP. Hence one might expect that if we initialized near one of the saddle points, then we might converge to that saddle point. However, our next main result of the section shows that even if we initialize near the saddle, there always exist escape directions.

Theorem 3.2 (Region with Negative Eigenvalue in the Riemannian Hessian of Equation 2.1) *Assume that Assumption 1 holds and α is small enough. Given any $\mathbf{Y} \in \mathbb{R}_*^{n\times r}$ such that $\mathbf{Y} \in \mathcal{R}_2$, there exist two explicit escaping directions $\theta_{\mathbf{Y}}^1$ and $\theta_{\mathbf{Y}}^2$ such that*

$$\overline{\operatorname{Hess} H([\mathbf{Y}])}\left[\theta_{\mathbf{Y}}^j, \theta_{\mathbf{Y}}^j\right] \leq -C_1 \|\theta_{\mathbf{Y}}^j\|_{\mathrm{F}}^2 \tag{3.2}$$

for some constant $C_1 > 0$ depending on α, μ and $\mathcal{A}_{\mathbf{n}}$ and for either $j = 1$ or 2.

Finally, the next result says that if we are not close to a FOSP, then we have large gradients.

Theorem 3.3 ((Regions with Large Riemannian Gradient of Equation 2.1)

1. $\|\overline{\operatorname{grad} H([\mathbf{Y}])}\|_{\mathrm{F}} > \alpha\mu\sigma_r^3(\mathbf{Y}^*)/(4\kappa^*), \forall \mathbf{Y} \in \mathcal{R}_3'$;
2. $\|\overline{\operatorname{grad} H([\mathbf{Y}])}\|_{\mathrm{F}} \geqslant 2\left(\|\mathbf{Y}\|^3 - \|\mathbf{Y}\|\,\|\mathbf{Y}^*\|^2\right) > 2\left(\beta^3 - \beta\right)\|\mathbf{Y}^*\|^3, \quad \forall \mathbf{Y} \in \mathcal{R}_3''$;
3. $\langle \overline{\operatorname{grad} H([\mathbf{Y}])}, \mathbf{Y}\rangle > 2(1-1/\gamma)\left\|\mathbf{Y}\mathbf{Y}^\top\right\|_{\mathrm{F}}^2, \quad \forall \mathbf{Y} \in \mathcal{R}_3'''$.

In particular, if $\beta > 1$ and $\gamma > 1$, we have the Riemannian gradient of $H([\mathbf{Y}])$ has large magnitude in all regions $\mathcal{R}_3', \mathcal{R}_3''$ and $\mathcal{R}_3'''$.

The behavior, implied by our theorems, of gradient descent in Euclidean space as it goes through the regions $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3$ is illustrated in Figures 1 and 2. See a discussion in Section 4.

These results can be seen as an under-parameterized generalization to the regression problem of Section 5 in [19]. The proof in [19] is simpler because, in their setting, there are no saddle points or local minima that are not global in $\mathbb{R}_*^{n\times r}$. Conceptually, [26] proves that in the setting $r \geq n$, the gradient flow for Equation 2.1 converges to a global minimum linearly; in particular, in their setting there aren't any saddle points. We complement this result by studying the case $r < n$.

Theorem 3.1, 3.2 and 3.3 guarantee the benignness of the ambient optimization problem, which is necessary condition of the benignness of parameterized optimization problem Equation 1.1. Also, these landscape results imply that perturbed gradient descent is guaranteed to converge to the global minima in polynomial time for Equation 2.1 10, 13, 15, 25.

4. Parameterized Loss Landscape

Finally, answering whether gradient descent converges to the global minimum for the parameterized problem (i.e., the NN training problem) is quite challenging. Hence, for this piece, we explore the question experimentally. Specifically, we present some numerical experiments where we consider different initializations for the training of SNN. Here we take 100 data points from MNIST and let $\mathcal{A}_n$ be the $n \times n$ gram matrix for the data points for simplicity. We remark that while we care about a $\mathcal{A}_n$ with a specific form for our approximation theory results, our analysis of the loss landscape described below holds for an arbitrary positive semi-definite matrix. In Figure 1, we plot the norm of the gradient during training when initialized in two different regions of parameter space. Concretely, in a region of parameters for which $\mathbf{Y}_\theta$ is close to a solution $\mathbf{Y}^*$ to problem 2.1 and a region of parameters for which $\mathbf{Y}_\theta$ is close to a saddle point of the ambient loss ℓ. We compare these plots to the ones we produce from the gradient descent dynamics for the ambient problem 2.1, which are shown in Figure 2. We notice a similar qualitative behavior with the training dynamics of the NN, suggesting that the landscape of problem 1.1, if the NN is properly overparameterized, inherits properties of the landscape of ℓ.

References

[1] Foivos Alimisis and Bart Vandereycken. Geodesic convexity of the symmetric eigenvalue problem and convergence of riemannian steepest descent. *arXiv preprint arXiv:2209.03480*, 2022.

[2] Rie Ando and Tong Zhang. Learning on graph with laplacian regularization. *Advances in neural information processing systems*, 19, 2006.

[3] Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2(1):53–58, 1989.

[4] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in neural information processing systems*, 14, 2001.

[5] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(11), 2006.

[6] Nicolas Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023.

[7] Pierre Bréchet, Katerina Papagiannouli, Jing An, and Guido Montúfar. Critical points and convergence analysis of generative deep linear networks trained with bures-wasserstein loss. *arXiv preprint arXiv:2303.03027*, 2023.

[8] Yuejie Chi, Yue M. Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

[9] Ronald R Coifman, Stephane Lafon, Ann B Lee, Mauro Maggioni, Boaz Nadler, Frederick Warner, and Steven W Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the national academy of sciences*, 102(21):7426–7431, 2005.

[10] Christopher Criscitiello and Nicolas Boumal. Efficiently escaping saddle points on manifolds. *Advances in Neural Information Processing Systems*, 32, 2019.

[11] Zhijie Deng, Jiaxin Shi, and Jun Zhu. NeuralEF: Deconstructing kernels by deep neural networks. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 4976–4992. PMLR, 17–23 Jul 2022.

[12] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[13] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on learning theory*, pages 797–842. PMLR, 2015.

[14] Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34, 2021.

[15] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International conference on machine learning*, pages 1724–1732. PMLR, 2017.

[16] John M Lee. *Introduction to Riemannian manifolds*, volume 176. Springer, 2018.

[17] Qiuwei Li and Gongguo Tang. The nonconvex geometry of low-rank matrix optimizations with general objective functions. In *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1235–1239, 2017.

[18] Xingguo Li, Junwei Lu, Raman Arora, Jarvis Haupt, Han Liu, Zhaoran Wang, and Tuo Zhao. Symmetry, saddle points, and global optimization landscape of nonconvex matrix factorization. *IEEE Transactions on Information Theory*, 65(6):3489–3514, 2019.

[19] Yuetian Luo and Nicolás García Trillos. Nonconvex matrix factorization is geodesically convex: Global landscape analysis for fixed-rank matrix optimization from a riemannian perspective. *arXiv preprint arXiv:2209.15130*, 2022.

[20] Hancheng Min, Salma Tarmoun, Rene Vidal, and Enrique Mallada. On the explicit role of initialization on the convergence and implicit bias of overparametrized linear networks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7760–7768. PMLR, 18–24 Jul 2021.

[21] Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14, 2001.

[22] David Pfau, Stig Petersen, Ashish Agarwal, David G. T. Barrett, and Kimberly L. Stachenfeld. Spectral inference networks: Unifying deep and spectral learning. In *International Conference on Learning Representations*, 2019.

[23] Uri Shaham, Kelly Stanton, Henry Li, Ronen Basri, Boaz Nadler, and Yuval Kluger. Spectralnet: Spectral clustering using deep neural networks. In *International Conference on Learning Representations*, 2018.

[24] Alexander J Smola and Risi Kondor. Kernels and regularization on graphs. In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings*, pages 144–158. Springer, 2003.

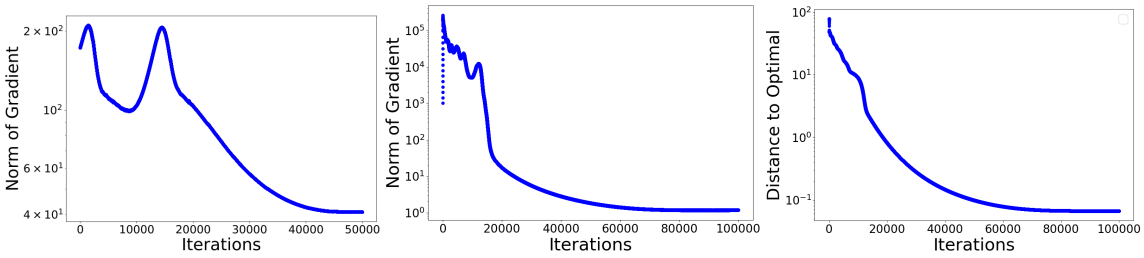
[25] Yue Sun, Nicolas Flammarion, and Maryam Fazel. Escaping from saddle points on riemannian manifolds. *Advances in Neural Information Processing Systems*, 32, 2019.

[26] Salma Tarmoun, Guilherme Franca, Benjamin D Haeffele, and Rene Vidal. Understanding the dynamics of gradient flow in overparameterized linear models. In Marina Meila and Tong

Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10153–10161. PMLR, 18–24 Jul 2021.

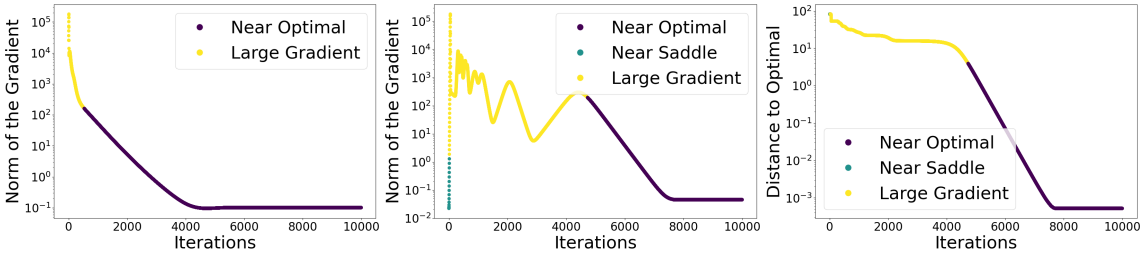
[27] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Appendix A. Parameterized optimization landscape



(a) Initialized Near Optimal (b) Initialized Near Saddle (c) Initialized Near Saddle

Figure 1: (a) and (b) Sum of the norms of the gradients for a two-layer ReLU Neural Network. In (a), the network is initialized near the global optimal solution and in (b) the network is initialized near a saddle point. (c) shows the distance between the current outputs of the neural network and the optimal solution for the case when it was initialized near a saddle point.



(a) Initialized Near Optimal (b) Initialized Near Saddle (c) Initialized Near Saddle

Figure 2: Norms of the gradients for the ambient problem and the distance to the optimal solution. In (a), $\mathbf{Y}$ is initialized near the global optimal solution, and in (b) $\mathbf{Y}$ is initialized near a saddle point. c) shows the distance between $\mathbf{Y}$ and the optimal solution for the case when it was initialized near a saddle point.

Appendix B. Full Paper

Spectral Neural Networks: Approximation Theory and Optimization Landscape

May 28, 2024

Abstract

There is a large variety of machine learning methodologies that are based on the extraction of spectral geometric information from data. However, the implementations of many of these methods often depend on traditional eigensolvers, which present limitations when applied in practical online big data scenarios. To address some of these challenges, researchers have proposed different strategies for training neural networks as alternatives to traditional eigensolvers, with one such approach known as Spectral Neural Network (SNN). In this paper, we investigate key theoretical aspects of SNN. First, we present quantitative insights into the tradeoff between the number of neurons and the amount of spectral geometric information a neural network learns. Second, we initiate a theoretical exploration of the optimization landscape of SNN's objective to shed light on the training dynamics of SNN. Unlike typical studies of convergence to global solutions of NN training dynamics, SNN presents an additional complexity due to its non-convex ambient loss function, a feature that is common in unsupervised learning settings.

1 Introduction

In the past decades, researchers from a variety of disciplines have studied the use of spectral geometric methods to process, analyze, and learn from data. These methods have been used in supervised learning [1, 2, 3], clustering [4, 5], dimensionality reduction [6, 7], and contrastive learning [8]. While the aforementioned methods have strong theoretical foundations, their algorithmic implementations often depend on traditional eigensolvers. These eigensolvers tend to underperform in practical big data scenarios due to high computational demands and memory constraints. Moreover, they are particularly vulnerable in online settings since the introduction of new data typically necessitates a full computation from scratch.

To overcome some of the drawbacks of traditional eigensolvers, new frameworks for learning from spectral geometric information that are based on the training of neural networks have emerged. To begin discussing some of popular training strategies, consider a data set $\mathcal{X}_n = \{x_1, \ldots, x_n\}$ in $\mathbb{R}^d$ and a $n \times n$

adjacency matrix $\mathcal{A}_{\mathbf{n}}$ describing similarity among points in $\mathcal{X}_n$. One could start by computing the eigendecomposition of $\mathcal{A}_n$ using traditional eigensolvers and get eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_r$. Then, to generalize these eigenvectors to points outside of $\mathcal{X}_n$, one can minimize the following ℓ_2 loss:

$$\min_{\theta} \|f_\theta(\mathcal{X}_n) - \mathbf{v}\|^2, \tag{1}$$

where $\mathbf{v} = [\mathbf{v}_1, \mathbf{v}_2 \ldots, \mathbf{v}_r]$ and θ denotes the parameters of a neural network. This approach, referred to as Eigensolver net, is a natural way to extend the geometric information contained in the similarity matrix of a finite collection of points to out-of-sample data and can be used even when the matrix $\mathcal{A}_n$ is not PSD. On the other hand, the Eigensolver net has some drawbacks. Specifically, one still needs to compute the eigendecomposition using traditional eigensolvers, which is precisely what one may want to avoid.

Spectralnet [9] and Spectral Neural Network (SNN) [8] have been proposed to overcome this issue. In these approaches, the goal is to find neural networks that can approximate the spectrum of a large target matrix, and the differences among the approaches lie mostly in the specific loss functions used for training; here we focus on SNN, and provide some details on Spectralnet in Appendix A.2. A SNN is trained by minimizing the *spectral constrastive* loss function:

$$\min_{\theta \in \Theta} L(\theta) \stackrel{\text{def}}{=} \ell(\mathbf{Y}_\theta), \quad \text{where} \quad \ell(\mathbf{Y}) \stackrel{\text{def}}{=} \left\|\mathbf{Y}\mathbf{Y}^\top - \mathcal{A}_{\mathbf{n}}\right\|_{\mathrm{F}}^2, \quad \mathbf{Y} \in \mathbb{R}^{n \times r}, \tag{2}$$

through first-order optimization methods. Figure 3 illustrates that SNNs can well approximate the desired eigenvectors associated to a proximity based similarity matrix. In the above and in the sequel, θ represents the vector of parameters of the neural network $f_\theta : \mathbb{R}^d \to \mathbb{R}^r$, the matrix $\mathbf{Y}_\theta$ is the $n \times r$ matrix whose rows are the outputs $f_\theta(x_1), \ldots, f_\theta(x_n)$, and $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm. The mapping f_θ can be interpreted as a feature or representation map for the input data. In the remainder, $\mathbf{Y}^*$ will denote a minimizer of ℓ and we will use f_{θ^*} to denote the neural network that minimizes $L(\theta)$.

In this paper, we investigate some of SNN's theoretical underpinnings. To make our setting more precise, through our discussion we adopt the *manifold hypothesis* and assume the data set $\mathcal{X}_n = \{x_1, \ldots, x_n\}$ to be supported on a low dimensional manifold $\mathcal{M}$ embedded in $\mathbb{R}^d$. Specifically, we make the following assumption on the generation process of the data $\mathcal{X}_n$.

Assumption 1. *The points $x_1, \ldots, x_n$ are assumed to be sampled from a distribution supported on an m-dimensional manifold $\mathcal{M}$ that is assumed to be smooth, compact, orientable, connected, and without a boundary. We assume that this sampling distribution has a density $\rho : \mathcal{M} \to \mathbb{R}_+$ with respect to $\mathcal{M}'s$ volume form and that ρ is bounded away from zero and bounded above by a constant. In addition, ρ is assumed $C^2(\mathcal{M})$.*

We also assume that $\mathcal{X}_n$ is endowed with a similarity matrix $\boldsymbol{G}^\tau$ with entries

$$\boldsymbol{G}^\tau_{ij} = \eta\left(\frac{\|x_i - x_j\|}{\tau}\right), \tag{3}$$

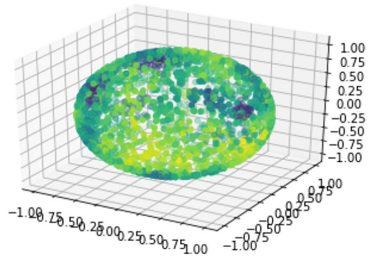


Figure 1: (A)

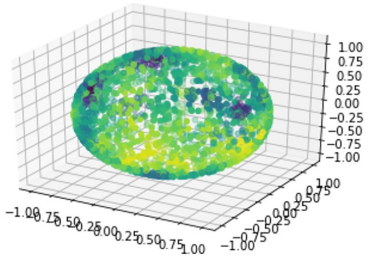


Figure 2: (B)

Figure 3: (B) shows the first eigenvector for the Laplacian of a proximity graph from data points sampled from S^2 obtained using an eigensolver. (A) shows the same eigenvector but obtained using SNN. The difference between the two figures is minor, showing that the neural network learns the eigenvector of the graph Laplacian well. See details in Appendix B.1.

where $\|x-y\|$ denotes the Euclidean distance between x and y, τ is a proximity parameter (which, for theoretical reasons stated below, will be assumed to scale like $1 \gg \tau \gg n^{-1/(m+4)}$), and η is a decreasing, non-negative function. In short, $\boldsymbol{G}^\tau$ measures the similarity between points according to their Euclidean proximity. Examples of functions η include the indicator function $\mathbb{1}_{[0,1]}$ and the Gaussian kernel. Associated to η we define the following normalization factor

$$c_\eta \stackrel{\text{def}}{=} \int_{\mathbb{R}^m} |y_1|^2 \, \eta(|y|) dy, \tag{4}$$

where y_1 is the first coordinate of y.

From $\boldsymbol{G}^\tau$ we define the adjacency matrix $\mathcal{A}_{\mathbf{n}}$ that we'll use within Equation 2 by

$$\mathcal{A}_{\mathbf{n}} \stackrel{\text{def}}{=} \mathbf{D}_{\boldsymbol{G}}^{-\frac{1}{2}} \boldsymbol{G} \mathbf{D}_{\boldsymbol{G}}^{-\frac{1}{2}} + a\mathbf{I}, \tag{5}$$

where $(\mathbf{D}_{\boldsymbol{G}})_{ii} = \sum_{j=1}^n \boldsymbol{G}_{ij}$ is the degree matrix associated to $\boldsymbol{G}$, and $a > 1$ is a fixed quantity. Here, we distance ourselves slightly from the choice made in the original SNN paper [8], where $\mathcal{A}_{\mathbf{n}}$ is taken to be $\boldsymbol{G}$ itself, and instead consider a normalized version. This is due to the following key properties satisfied by our choice of $\mathcal{A}_{\mathbf{n}}$.

The matrix $\mathcal{A}_{\mathbf{n}}$ defined in Equation 5 satisfies the following properties:

1. $\mathcal{A}_{\mathbf{n}}$ is symmetric positive definite.
2. $\mathcal{A}_{\mathbf{n}}$'s r *top* eigenvectors (the ones corresponding to the r largest eigenvalues) coincide with the eigenvectors of the r *smallest* eigenvalues of the symmetric normalized graph Laplacian matrix (see [5]):

$$\Delta_n \stackrel{\text{def}}{=} \mathbf{I} - \mathbf{D}_{\boldsymbol{G}}^{-1/2} \boldsymbol{G} \mathbf{D}_{\boldsymbol{G}}^{-1/2}. \tag{6}$$

When n is large and τ scales with n appropriately, Δ_n's spectrum is known to be closely connected to that of the weighted Laplace-Beltrami operator Δ_ρ defined as

$$\Delta_\rho f \stackrel{\text{def}}{=} -\frac{1}{\rho^{3/2}}\text{div}\left(\rho^2 \nabla\left(\frac{f}{\sqrt{\rho}}\right)\right)$$

for all smooth enough $f : \mathcal{M} \to \mathbb{R}$; see Section 1.4 in [10]. In the above, div stands for the divergence operator on $\mathcal{M}$, and ∇ for the gradient in $\mathcal{M}$. Δ_ρ can be easily seen to be a positive semi-definite operator with respect to the $L^2(\mathcal{M}, \rho)$ inner product and its eigenvalues (repeated according to multiplicity) can be listed in increasing order as

$$0 = \lambda_1^{\mathcal{M}} \leq \lambda_2^{\mathcal{M}} \leq \dots.$$

We will use $f_1, f_2, \dots$ to denote the associated normalized (in the $L^2(\mathcal{M}, \rho)$-sense) eigenfuntions.

We explore the following three questions:

Q1 How large do neural networks need to be to approximate the eigenvectors of Δ_n and eigenfunctions of Δ_ρ simultaneously?
Q2 Is it possible to use Equation 2 to build an approximating neural network?
Q3 What can be said about the landscape of the objective function in Equation 2?

Contributions We provide answers to the above three questions. We also formulate and discuss open problems that, while motivated by our current investigation, we believe are of interest in their own right. In summary, the main contributions of our work are the following:

- We provide precise tradeoffs between the size of the neural network and the error in simultaneously approximating the eigenvectors of a large adjacency matrix and the eigenfunctions of the Laplace Beltrami operator on the manifold supporting the data; see Theorem 3.1 and Corollary 3. In this way, we present an example of a setting where we can rigorously quantify the error of approximation of a solution to a PDE on a manifold with NNs.
- We show that by solving Equation 2 one can *construct* such approximation provided the parameter space of the NN is rich enough; see Theorem 3.2. Specifically, we show that the global minimizer of the loss function in Equation 2 well approximates the eigenvectors when the neural network is sufficiently expressive.
- Motivated by numerical evidence, we begin an exploration of the optimization landscape of SNN and, in particular, provide a full description of SNN's associated ambient space optimization landscape. This landscape

is proved to be benign; see discussion in Section 4. This observation opens up a series of interesting future research directions that we briefly describe in Section 6.

1.1 Related work

Other NN-based Eigensolvers: Other types of NN-based Eigensolvers have been considered in [11] and [12]. [11] uses a bi-level optimization algorithm to solve a constrained optimization problem. This algorithm's computational complexity is typically higher than the one of SNN training and it requires keeping certain covariance matrices in memory during updates. [12] takes a similar approach as [11] but can avoid the bi-level optimization of the latter. This, however, comes at the expense of having an intractable theoretical computational complexity.

Spectral clustering and manifold learning Several works have attempted to establish precise mathematical connections between the spectra of graph Laplacian operators over proximity graphs and the spectrum of weighted Laplace-Beltrami operators over manifolds. Some examples include [13, 14, 15, 16, 17, 18, 19, 20, 21]. In this paper, we use adaptations of the results in [18] to infer that, with very high probability, the eigenvectors of the normalized graph Laplacian matrix Δ_n defined in Equation 6 are essentially Lipschitz continuous functions. These regularity estimates are one of the crucial tools for proving our Theorem 3.1.

Contrastive Learning Contrastive learning is a self-supervised learning technique that has gained considerable attention in recent years due to its success in computer vision, natural language processing, and speech recognition [22, 23, 24, 25]. Theoretical properties of contrastive representation learning were first studied by [26, 27, 28] where they assumed conditional independence. [8] relaxes the conditional independence assumption by imposing the manifold assumption. With the spectral contrastive loss Equation 2 crucially in use, [8] provides an error bound for downstream tasks. In this work, we analyze how a neural network can approximate and optimize the spectral loss function Equation 2, which is the pertaining step of [8].

Neural Network Approximations. Given a function f with certain amount of regularity, many works have studied the tradeoff between width, depth, and total number of neurons needed for a neural network to approximate it [29, 30]. Specifically, [31] looks at the problem of Hölder continuous functions on the unit cube, [32, 33] studies the class of continuous functions on the unit cube, and [29, 34, 8] consider the case when the function is defined on a manifold. A related area is that of neural network memorization of a finite number of data points [35]. In this paper, we use these results to show that for our specific type of regularity, we can prove similar results.

Neural Networks and Partial Differential Equations [36] introduced Physics Informed Neural Networks as a method for solving PDEs using neural networks. Specifically, [37, 38, 36] use neural networks to parameterize the solution of a PDE which is trained by optimizing a loss function that is designed to be minimized when the equation is satisfied. Other works such as [39, 40, 41, 38] use neural networks to parameterize the solution operator on a given mesh on the PDE's domain. Finally, the search for eigenfunctions of operators on function spaces has deep connections to PDEs. Recent works such as [42, 43, 44] demonstrate how to learn these operators. In this work we show that we can approximate eigenfunctions to a weighted Laplace-Beltrami operator using neural networks by minimizing the spectral loss L.

Shallow Linear Networks and Non-convex Optimization in Linear Algebra Problems One of our main objects of study in this paper is the ambient problem Equation 2. This formulation of the problem is related to linear networks. Linear networks are neural networks with identity activation. A variety of prior works have studied many different aspects of shallow linear networks such as their loss landscape and their associated optimization dynamics [45, 46, 47, 48], and generalization for one layer networks [49, 50, 51, 52, 53]. Of relevance are also other works in the literature studying optimization problems very closely related to Equation 7. For example, in Section 3 in [54], there is a landscape analysis for Equation 7 when the matrix $\mathcal{A}_{\mathbf{n}}$ is assumed to have rank smaller than or equal to r. That setting is typically referred to as overparameterized or exactly parameterized, whereas here our focus is on the underparameterized setting. On the other hand, the case studied in Section 3 in [55] is the simplest case we could consider for our problem and corresponds to $r = 1$. In this simpler case, the non-convexity of the objective is completely due to a sign ambiguity, which makes the analysis more straightforward and the need to introduce quotient geometries less pressing. [56] describes the global optimization landscape of Equation 7 under the assumption that $\mathcal{A}_{\mathbf{n}}$ is rank r. In contrast, we recall that here $\mathcal{A}_{\mathbf{n}}$ is assumed to be full rank.

Notation Throughout the paper, we use $C > 1$ and $c < 1$ to denote constants which depend on ρ and the intrinsic properties of $\mathcal{M}$ including the embedded dimension of $\mathcal{M}$, the lower bound of the injectivity radius of $\mathcal{M}$, the reach of $\mathcal{M}$, and the upper bound on the absolute values of the sectional curvature of $\mathcal{M}$.

2 Preliminary Results

In this section, we briefly provide some technical results that will serve as foundations for the analysis presented in this paper.These preliminary results also provide some context for our discussion in the rest of the paper.

2.1 Rotational Invariance

Recall the ambient optimization problem defined in Equation 2 as

$$\min_{\mathbf{Y}\in\mathbb{R}^{n\times r}} \ell(\mathbf{Y}), \quad \text{where} \quad \ell(\mathbf{Y}) \stackrel{\text{def}}{=} \left\|\mathbf{Y}\mathbf{Y}^\top - \mathcal{A}_\mathbf{n}\right\|_{\text{F}}^2. \tag{7}$$

Suppose $\mathbf{Y}$ is a stationary point of Equation 7. Then $\mathbf{Y}\boldsymbol{O}$ is also a stationary point for any $r\times r$ orthogonal matrix $\boldsymbol{O}\in\mathbb{O}_r$. This implies that the loss function ℓ is non-convex in any neighborhood of a stationary point [57]. Hence we shall consider a quotient geometry to attempt to remove the local non-convexity induced by the action of the orthogonal group. Let $\overline{\mathcal{N}}_{r+}^n$ be the space of $n\times r$ matrices with full column rank. To define the quotient manifold, we encode the invariance mapping, i.e., $\mathbf{Y}\to\mathbf{Y}\boldsymbol{O}$, by defining the equivalence classes $[\mathbf{Y}] = \{\mathbf{Y}\boldsymbol{O} : \boldsymbol{O}\in\mathbb{O}_r\}$. From [58], $\mathcal{N}_{r+}^n \stackrel{\text{def}}{=} \overline{\mathcal{N}}_{r+}^n/\mathbb{O}_r$ is a quotient manifold of $\overline{\mathcal{N}}_{r+}^n$. For a detailed introduction to Riemannian optimization see [59]. Since ℓ is invariant within each the equivalence classes of $\overline{\mathcal{N}}_{r+}^n$, one obtains the following optimization problem on the quotient manifold $\mathcal{N}_{r+}^n$:

$$\min_{[\mathbf{Y}]\in\mathcal{N}_{r+}^n} H([\mathbf{Y}]) \stackrel{\text{def}}{=} \frac{1}{2}\left\|\mathbf{Y}\mathbf{Y}^\top - \mathcal{A}_\mathbf{n}\right\|_{\text{F}}^2, \tag{8}$$

which can be approached via Riemannian first order methods. In this case, however, the Riemannian gradient descent is equivalent to standard gradient descent in the Euclidean geometry. This is the reason why the results in Section 4 directly justify the ability of the gradient based algorithm in the Euclidean geometry to find global optimizers of the ambient space problem Equation 8.

2.2 First Order Stationary Points

Since $\mathcal{A}_\mathbf{n}$ is a PSD matrix, the Eckart–Young–Mirsky theorem (see [60]) implies that the global optimizers of Equation 7 are the matrices $\mathbf{Y}$ of the form $\mathbf{Y} = \mathbf{Y}^*\boldsymbol{O}$, where $\boldsymbol{O}\in\mathbb{O}_r$ and

$$\mathbf{Y}^* \stackrel{\text{def}}{=} \begin{bmatrix} | & & | \\ \sqrt{\lambda_1(\mathcal{A}_\mathbf{n})}v_1 & \dots & \sqrt{\lambda_r(\mathcal{A}_\mathbf{n})}v_r \\ | & & | \end{bmatrix}.$$

In the above, $\lambda_l(\mathcal{A}_\mathbf{n})$ represents the l-th largest eigenvalue of $\mathcal{A}_\mathbf{n}$ and v_l is a corresponding eigenvector with Euclidean norm one. In case there are repeated eigenvalues, the corresponding v_l need to be chosen to be orthogonal to each other. For convenience, we rescale the vectors v_l as follows:

$$u_l \stackrel{\text{def}}{=} \sqrt{n}v_l.$$

In this way we guarantee that

$$\|u_l\|_{L^2(\mathcal{X}_n)}^2 \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n (u_l(x_i))^2 = 1, \tag{9}$$

i.e., the rescaled eigenvectors u_l are normalized in the L^2-norm with respect to the empirical measure $\frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. In terms of the rescaled eigenvectors u_l, we can rewrite $\mathbf{Y}^*$ as follows:

$$\mathbf{Y}^* = \begin{bmatrix} | & & | \\ \sqrt{\frac{\lambda_1(\mathcal{A}_\mathbf{n})}{n}}u_1 & \cdots & \sqrt{\frac{\lambda_r(\mathcal{A}_\mathbf{n})}{n}}u_r \\ | & & | \end{bmatrix}. \tag{10}$$

Remark 1. *As discussed in Remark 2 below, under Assumptions 1 we can assume that all the $\lambda_s(\mathcal{A}_\mathbf{n})$ are quantities of order* 1.

Next, we need to understand the other first-order stationary points (FOSP) of Equation 8. We use SVD to get $\mathbf{Y} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{n\times k}$, $\mathbf{D} \in \mathbb{R}^{k\times k}$ is a diagonal matrix, and $\mathbf{V} \in \mathbb{R}^{k\times r}$. In addition, $\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V} = \mathbf{I}_k$, where k is the rank of $\mathbf{Y}$ and $\mathbf{I}_k$ is the $k \times k$ identity matrix.

Theorem 2.1 (FOSP of Equation 8)**.** *Let $\overline{\mathbf{U}}\mathbf{\Sigma}\overline{\mathbf{U}}^\top$ be $\mathcal{A}_\mathbf{n}$'s SVD factorization, and let $\mathbf{\Lambda} = \mathbf{\Sigma}^{1/2}$. Then for any S subset of $[n]$ we have that $\left[\overline{\mathbf{U}}_S\mathbf{\Lambda}_S\right]$ is a Riemannian FOSP of Equation 8. Further, these are the only Riemannian FOSPs.*

Theorem 2.1 shows that linear combinations of eigenvectors can be used to construct Riemannian first-order stationary points (FOSP) of Equation 8. This theorem also shows that there are many FOSPs of Equation 8. This is quite different from the regime studied in [56]. In general, gradient descent is known to converge to a FOSP. Hence one might expect that if we initialized near one of the saddle points, then we might converge to that saddle point. A careful landscape analysis will be presented in section 4.

2.3 Spectral Convergence of Δ_n to Δ_ρ

The first result for this section, whose proof we omit as it is a straightforward adaptation of the proof of Theorem 2.4 in [17] –which considers the *unnormalized* graph Laplacian case–, relates the eigenvalues of Δ_n and Δ_ρ.

Theorem 1 (Convergence of eigenvalues of graph Laplacian; Adapted from Theorem 2.4 in [17])**.** *Let $l \in \mathbb{N}$ be fixed. Under Assumptions 1, with probability at least $1 - Cn\exp\left(-cn\tau^{m+4}\right)$ over the sampling of the x_i, we have:*

$$\left| c_\eta \lambda_s^{\mathcal{M}} - \frac{\hat{\lambda}_s}{\tau^2} \right| \leq C_l \tau, \quad \forall s = 1, \ldots, l.$$

In the above, $\hat{\lambda}_1 \leq \cdots \leq \hat{\lambda}_l$ are the first eigenvalues of Δ_n in increasing order, C_l is a deterministic constant that depends on $\mathcal{M}$'s geometry and on l, and c_η is a scaling constant that depends on the kernel η determining the graph weights (see Equation 4). We also recall that m denotes the intrinsic dimension of the manifold $\mathcal{M}$.

Remark 2. *From Theorem 1 and Equation 29 we see that the top l eigenvalues of $\mathcal{A}_{\mathbf{n}}$ (for l fixed), i.e., $\lambda_1(\mathcal{A}_{\mathbf{n}}), \ldots, \lambda_l(\mathcal{A}_{\mathbf{n}})$, can be written as*

$$\lambda_s(\mathcal{A}_{\mathbf{n}}) = 1 + a - c_\eta \lambda_s^{\mathcal{M}} \tau^2 + O(\tau^3)$$

with very high probability. In particular, although each individual $\lambda_s(\mathcal{A}_{\mathbf{n}})$ is an order one quantity, the difference between any two of them is an order τ^2 quantity.

The above remark is an important observation. In particular, it implies that if $\mathbf{Y}$ is any saddle point of ℓ, then $\|\mathbf{Y} - \mathbf{Y}^*\|_{\mathrm{F}} = O(\sqrt{r}\tau)$.

Next we discuss the convergence of eigenvectors of Δ_n toward eigenfunctions of Δ_ρ. For the purposes of this paper (see some discussion below), we follow a strong, almost $C^{0,1}$ convergence result established in [18] for the case of unnormalized graph Laplacians. A straightforward adaptation of Theorem 2.6 in [18] implies the following.

Theorem 2 (Almost $C^{0,1}$ convergence of graph Laplacian eigenvectors; Adapted from Theorem 2.6 in [18])**.** *Let $r \in \mathbb{N}$ be fixed and let $u_1, \ldots, u_r$ be normalized eigenvectors of Δ_n as in Equation 9. Under Assumptions 1, with probability at least $1 - C\tau^{-6m} \exp\left(-cn\tau^{m+4}\right)$ over the sampling of the x_i, we have:*

$$\|f_s - u_s\|_{L^\infty(\mathcal{X}_n)} + [f_s - u_s]_{\tau, \mathcal{X}_n} \leq C_r \tau. \quad \forall s = 1, \ldots, r, \tag{11}$$

for normalized (in the $L^2(\mathcal{M}, \rho)$ sense) eigenfuctions $f_i : \mathcal{M} \to \mathbb{R}$ of Δ_ρ. In the above, $\|\cdot\|_{L^\infty(\mathcal{X}_n)}$ is the norm $\|v\|_{L^\infty(\mathcal{X}_n)} \stackrel{def}{=} \max_{x_i \in \mathcal{X}_n} |v(x_i)|$, $d_{\mathcal{M}}(\cdot, \cdot)$ denotes the geodesic distance on $\mathcal{M}$, and $[\cdot]_{\tau, \mathcal{X}_n}$ is the seminorm

$$[v]_{\tau, \mathcal{X}_n} \stackrel{def}{=} \max_{x_i, x_j \in \mathcal{X}_n} \frac{|v(x_i) - v(x_j)|}{d_{\mathcal{M}}(x_i, x_j) + \tau}.$$

2.4 Regularity of the Eigenvectors of Δ_n

An essential corollary of Theorem 2 is the following set of regularity estimates satisfied by eigenvectors of the normalized graph Laplacian Δ_n.

Corollary 1. *Under the same setting, notation, and assumptions as in Theorem 2, the functions u_s satisfy*

$$|u_s(x_i) - u_s(x_j)| \leq L_s(d_{\mathcal{M}}(x_i, x_j) + \tau^2), \quad \forall x_i, x_j \in \mathcal{X}_n \tag{12}$$

for some deterministic constant L_s which depends on s and some intrinsic properties of $\mathcal{M}$.

Proof. From Equation 11 we have

$$|(u_s(x_i) - f_s(x_i)) - (u_s(x_j) - f_s(x_j))| \leq C_s \tau (d_{\mathcal{M}}(x_i, x_j) + \tau), \quad \forall x_i, x_j \in \mathcal{X}_n.$$

It follows from the triangle inequality that

$$\begin{aligned}|u_s(x_i) - u_s(x_j)| &\leq |u_s(x_i) - f_s(x_i) - (u_s(x_j) - f_s(x_j))| + |f_s(x_i) - f_s(x_j)| \\ &\leq C_s\tau(d_{\mathcal{M}}(x_i, x_j) + \tau) + C_s' d_{\mathcal{M}}(x_i, x_j) \\ &\leq L_s(d_{\mathcal{M}}(x_i, x_j) + \tau^2).\end{aligned}$$

In the above, the second inequality follows from inequality 11 and the fact that f_s, being a normalized eigenfunction of the elliptic operator Δ_ρ, is Lipschitz continuous with some Lipschitz constant C_s'. □

Remark 3. *We observe that the τ^2 term on the right hand side of Equation 12 is strictly better than the term that appears in the explicit regularity estimates in Remark 2.4 in [18], where the dependence on τ is $O(\tau)$. It turns out that in the proof of Theorem 3.2 it is essential to have a correction term for the distance that is $o(\tau)$, as we have thanks to the above corollary.*

3 Spectral Approximation with neural networks

Question Q1 is a particular example of the more general problem below.

Problem 1. *Given a family of neural networks $\mathcal{F}$, a compact set $\mathcal{K}$ with density ρ, a Lipschitz function $f : \mathcal{K} \to \mathbb{R}$, n potentially noisy data points (x_i, y_i) sampled IID according to ρ and $\varepsilon > 0$, does there exist $f_\theta \in \mathcal{F}$ such that f_θ nearly interpolates the data*

$$\max_{i=1,\ldots,n} \|y_i - f_\theta(x_i)\| = O(\varepsilon)$$

and the network is a good approximation of the true function $\|f - f_\theta\|_{L^\infty(\mathcal{K})} = O(\varepsilon)$?

Prior work on universal approximation has focused on two different types of questions. First, given a compact set $\mathcal{K}$ and a function $f : \mathcal{K} \to \mathbb{R}$ with some form of regularity and $\varepsilon > 0$, we are interested in determining if there is a neural network f_θ of constrained size such that $\|f - f_\theta\|_{L_\infty(\mathcal{K})} < \varepsilon$. This problem is known as universal approximation. Second, given finite amount of data $(x_1, y_1), \ldots, (x_n, y_n)$, we are interested in the smallest network f_θ such that $y_i = f_\theta(x_i)$. This problem is known as memorization. However, both of these problems have drawbacks. The problem with universal approximation is that it ignores the data. Hence universal approximation is disconnected from the experimental procedures. The problem with memorization is that it ignores the behavior of the neural network away from the given data points.

Recent works have studied a new problem known as benign overfitting [51]. Given a neural network that perfectly fits the training data we are interested in whether the network exactly fits true data-generating function. This notion has been further refined to notions of benign, tempered, and catastrophic overfitting

[61]. Building on this, researchers have developed an interest in understanding near interpolators [62]. These are networks that do not exactly fit the data but nearly fit the data.

In this context, Problem 1 can be interpreted as asking for the existence of a near interpolator that benignly overfits the data. Problem 1 is an interesting interplay between the two types of questions – universal approximation and benign overfitting. Prior work has provided rates for the neural network approximation problem. However, the existence of an (nearly) interpolating network does not imply that the existence of one that does so in a benign or tempered manner. Hence understanding how big networks need to be before we can guarantee the existence of a network that benignly overfits is in its own right an interesting question.

In what follows we make precise the setting where we can answer Question Q1, which we recall is concerned with approximation of eigenvectors of Δ_n and eigenfunctions of $\Delta_{\mathcal{M}}$. Here, the former can be interpreted as noisy versions of the latter.

3.1 Multilayer ReLU Neural Network Family

For concreteness, in this work we use multi-layer ReLU neural networks. To be precise, our neural networks are parameterized functions $f : \mathbb{R}^d \to \mathbb{R}^r$ of the form:

$$f(\mathbf{x}) = \mathbf{W}_L \cdot \mathrm{ReLU}\left(\mathbf{W}_{L-1} \cdots \mathrm{ReLU}\left(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1\right) \cdots + \mathbf{b}_{L-1}\right) + \mathbf{b}_L, \quad \mathbf{x} \in \mathbb{R}^d. \tag{13}$$

More specifically, for a given choice of parameters r, κ, L, p, N we will consider the family of functions:

$$\begin{aligned} \mathcal{F}(r, \kappa, L, p, N) = \Bigg\{ & f \mid f(\mathbf{x}) \text{ has the form 13, where:} \\ & \mathbf{W}_l \in \mathbb{R}^{p \times p}, \mathbf{b}_l \in \mathbb{R}^p \text{ for } l = 2, \ldots, L-1, \\ & \mathbf{W}_1 \in \mathbb{R}^{p \times d}, \mathbf{b}_1 \in \mathbb{R}^p, \mathbf{W}_L \in \mathbb{R}^{r \times p}, \mathbf{b}_L \in \mathbb{R}^r. \\ & \|\mathbf{W}_l\|_{\infty,\infty} \leq \kappa, \|\mathbf{b}_l\|_\infty \leq \kappa \text{ for } l = 1, \ldots, L, \\ & \sum_{l=1}^{L} \|W_l\|_0 + \|\mathbf{b}_l\|_0 \leq N \Bigg\} \end{aligned} \tag{14}$$

where $\|\cdot\|_0$ denotes the number of nonzero entries in a vector or a matrix, $\|\cdot\|_\infty$ denotes the ℓ_∞ norm of a vector. For a matrix M, we use $\|M\|_{\infty,\infty} = \max_{i,j} |M_{ij}|$. For convenience, after specifying the quantities r, κ, L, p, N, we denote by Θ the space of admissible parameters $\theta = (\mathbf{W}_1, \mathbf{b}_1, \ldots, \mathbf{W}_L, \mathbf{b}_L)$ in the function class $\mathcal{F}(r, \kappa, L, p, N)$, and we use f_θ to represent the function in Equation 13.

3.2 Spectral approximation with multilayer ReLU NNs

In this section, we answer Question Q1 by providing Theorem 3.1 and Corollary 3. Specifically, we provide upper bounds on the size of the neural network size so that with high probability there exists a neural network that nearly interpolates the data and well approximates the true eigenfunctions of Δ_ρ. We provide bounds in terms of the width, the depth, the number of non-zero parameters, and the size of the parameters.

Lemma 1. *Let $u : \mathcal{X}_n \to \mathbb{R}$ be a function satisfying*

$$|u(x) - u(\tilde{x})| \leq L(d_{\mathcal{M}}(x, \tilde{x}) + \tau^2), \quad \forall x, \tilde{x} \in \mathcal{X}_n \tag{15}$$

for some L and $\tau > 0$. Then there exists a $3L$-Lipschitz function $\tilde{g} : \mathcal{M} \to \mathbb{R}$ such that

$$\|u - \tilde{g}\|_{L^\infty(\mathcal{X}_n)} \leq 5L\tau^2. \tag{16}$$

Remark 4. *Lemma 1 guarantees that if a function u, defined in any given metric space, is (L, τ^2)-almost Lipschitz, then we can find a function $\tilde{g}$ that is L-Lipschitz continuous in the same space and is within uniform distance τ^2 from u.*

By combining Lemma 1 and Theorem 2, and using some universal approximation theory results for neural networks, we will be able to prove the following result, whose proof can be found in Appendix E.

Theorem 3.1 (Spectral approximation of normalized Laplacians with neural networks)**.** *Let $r \in \mathbb{N}$ be fixed. Under Assumptions 1, there are constants c, C that depend on $\mathcal{M}, \rho$, and the embedding dimension r, such that, with probability at least*

$$1 - C\tau^{-6m} \exp\left(-cn\tau^{m+4}\right),$$

for every $\varepsilon \in (0, 1)$ there are κ, L, p, N and a ReLU neural network $f_\theta \in \mathcal{F}(r, \kappa, L, p, N)$ (defined in Equation 14), such that:

1. *$\sqrt{n}\|\mathbf{Y}_\theta - \mathbf{Y}^*\|_{\infty,\infty} \leq C(\varepsilon + \tau^2)$, and thus also $\|\mathbf{Y}_\theta - \mathbf{Y}^*\|_{\mathrm{F}} \leq C\sqrt{r}(\varepsilon + \tau^2)$.*
2. *The depth of the network, L, satisfies: $L \leq C\left(\log \frac{1}{\varepsilon} + \log d\right)$, and its width, p, satisfies $p \leq C\left(\varepsilon^{-m} + d\right)$.*
3. *The number of neurons of the network, N, satisfies: $N \leq Cr\left(\varepsilon^{-m} \log \frac{1}{\varepsilon} + d \log \frac{1}{\varepsilon} + d \log d\right)$, and the range of weights, κ, satisfies $\kappa \leq \frac{C}{n^{1/(2L)}}$.*

Theorem 3.1 can be interpreted as follows. First, we note that the high probability statement is with respect to the sampling of the data points. Here we see that the probability depends on the number of data points and our scale parameter τ. Specifically, if our scale parameter doesn't decay too quickly ($\tau \gg n^{-\frac{1}{m+4}}$), then as $n \to \infty$, the probability goes to 1. Theorem 3.1 also bounds the error in terms of two different norms. Additionally, we provide explicit rates of the width, depth, as well as the magnitude of the parameters. Note these rates *do not* depend on the connectivity parameter τ.

Remark 5. *Notice that the term $\sqrt{n}\|\mathbf{Y}^*\|_{\infty,\infty}$ is of order one. Consequently, the estimate in Theorem 3.1 is a non-trivial error bound.*

The bound in $\|\cdot\|_{\infty,\infty}$ between $\mathbf{Y}_\theta$ and $\mathbf{Y}^*$ in Theorem 3.1 can be used to bound the difference between $\mathbf{Y}_\theta\mathbf{Y}_\theta^\top$ and $\mathbf{Y}^*\mathbf{Y}^{*\top}$ in $\|\cdot\|_{\infty,\infty}$.

Corollary 2. *For f_θ as in Theorem 3.1 we have*

$$\sqrt{n}\|\mathbf{Y}_\theta\mathbf{Y}_\theta^\top - \mathbf{Y}^*\mathbf{Y}^{*\top}\|_{\infty,\infty} \leq C_r(\varepsilon + \tau^2), \tag{17}$$

and thus also

$$\|\mathbf{Y}_\theta\mathbf{Y}_\theta^\top - \mathbf{Y}^*\mathbf{Y}^{*\top}\|_{\mathrm{F}} \leq \sqrt{r}C_r(\varepsilon + \tau^2),$$

for some deterministic constant C_r.

The τ^2 term that appears in the bound for $\|\mathbf{Y}_\theta - \mathbf{Y}^*\|_{\mathrm{F}}$ in Theorem 3.1 *cannot* be obtained simply from convergence of eigenvectors of Δ_n toward eigenfunctions of Δ_ρ in L^∞. More concretely, if we use a standard universal approximation result such as Theorem 3 from [63] to approximate the eigenfunctions $f_1, \ldots, f_r$ and then the convergence result for Δ_n to Δ_ρ (such as Theorem 1 from [17]), we would get that $\|\mathbf{Y}_\theta - \mathbf{Y}^*\|_{\mathrm{F}} = O(\sqrt{r}\tau)$. However, we know from Remark 2, that for any saddle point $\mathbf{Y}$, we have that $\|\mathbf{Y} - \mathbf{Y}^*\|_{\mathrm{F}}$ is order $O(\sqrt{r}\tau)$. Hence being distance $O(\sqrt{r}\tau)$ is not useful, as it does not tell us whether we are near a global minimizer of our problem or to a saddle point. To prove our results, we thus need to use a stronger notion of convergence (almost $C^{0,1}$) that in particular implies sharper regularity estimates for eigenvectors of Δ_n (see Corollary 1 and Remark 3 below it). In turn, the sharper τ^2 term is essential for our proof of Theorem 3.2 below to work, which formalizes our answer to Question Q2.

So far we have discussed approximations of the eigenvectors of $\mathcal{A}_\mathbf{n}$ (and thus also of Δ_n) with neural networks, but more can be said about generalization of these NNs. In particular, the NN in our proof of Theorem 3.1 can be shown to approximate eigenfunctions of the weighted Laplace-Beltrami operator Δ_ρ. Precisely, we have the following result.

Corollary 3. *Under the same setting, notation, and assumptions as in Theorem 3.1, the neural network $f_\theta : \mathbb{R}^d \to \mathbb{R}^r$ can be chosen to satisfy*

$$\left\| \sqrt{\frac{n}{1+a}} f_\theta^i - f_i \right\|_{L^\infty(\mathcal{M})} \leq C(\varepsilon + \tau), \quad \forall i = 1, \ldots, r.$$

In the above, $f_\theta^1, \ldots, f_\theta^r$ are the coordinate functions of the vector-valued neural network f_θ, and the functions $f_1, \ldots, f_r$ are the normalized eigenfunctions of the Laplace-Beltrami operator Δ_ρ that are associated to Δ_ρ's r smallest eigenvalues.

With Theorem 3.1 and Corollary 3 we thus provide an answer to Problem 1. We notice that, while one could use existing *memorization* results (e.g., Theorem 3.1 in [35]) to show that there is a neural network with ReLU activation function and $\mathcal{O}(\sqrt{n})$ neurons that fits $\mathbf{Y}^*$ perfectly, this does not constitute an

improvement over our results in Theorem 3.1 and Corollary 3. Indeed, by using this type of memorization result, we can not state any bounds on the size of the parameters of the network, and none of the out-of-sample generalization properties that we have discussed before (i.e., approximation of eigenfunctions of Δ_ρ) can be guaranteed.

On the other hand, we see that the size of neural network are the same as the size for universal approximation. That is, once we can guarantee the existence of a network that well approximates the eigenfunctions, we can guarantee the existence of a network that simultaneously well approximates the given data and the eigenfunctions. In short, τ does not affect the size of the network.

3.3 Spectral approximation with global minimizers of SNN's objective

After discussing the *existence* of approximating NNs, we turn our attention to *constructive* ways to approximate $\mathbf{Y}^*$ using neural networks. This is non-trivial, and is further complicated by the fact that the we measure the distance to optimal solutions via the convex function $\|\mathbf{Y}_\theta - \mathbf{Y}^*\|_{\mathrm{F}}$. However, this is not the loss function SNNs use to train the network. Instead, the nonconvex loss function $\|\mathbf{Y}_\theta \mathbf{Y}_\theta^T - \mathcal{A}_n\|_{\mathrm{F}}$ is used to train the network. Our next result shows that this loss function is still relevant for the type of approximation we are after. Specifically, Theorem 3.2 shows that the global minimizer of the loss function has good properties. With this we provide an answer to question Q2. The proof for Theorem 3.2 can be found in Appendix F.

Theorem 3.2 (Optimizing SNN approximates eigenvectors up to rotation)**.** *Let $r \in \mathbb{N}$ be fixed and suppose that Δ_ρ is such that Δ_ρ has a spectral gap between its r and $r+1$ smallest eigenvalues, assume that $\lambda_r^{\mathcal{M}} < \lambda_{r+1}^{\mathcal{M}}$. For given κ, L, p, N (to be chosen below), let f_{θ^*} be such that*

$$f_{\theta^*} \in \underset{f_\theta \in \mathcal{F}(r,\kappa,L,p,N)}{\arg\min} \|\mathbf{Y}_\theta \mathbf{Y}_\theta^\top - \mathcal{A}_{\mathbf{n}}\|_{\mathrm{F}}^2. \tag{18}$$

Under Assumptions 1, there are constants c, C that depend on $\mathcal{M}, \rho$, and the embedding dimension r, such that, with probability at least $1 - C\tau^{-6m}\exp\left(-cn\tau^{m+4}\right)$, for every $\tilde{\delta} \in (0, c)$ (i.e., $\tilde{\delta}$ sufficiently small) and for $\kappa = \frac{C}{n^{1/(2L)}}$, $L = C\left(\log \frac{1}{\tilde{\delta}\tau} + \log d\right)$, and $p = C\left((\tilde{\delta}\tau)^{-m} + d\right)$, we have

$$\min_{\boldsymbol{O} \in \mathbb{O}_r} \|\mathbf{Y}_{\theta^*} - \mathbf{Y}^* \boldsymbol{O}\|_{\mathrm{F}} \leq C\tau(\tilde{\delta} + \tau). \tag{19}$$

In the above theorem $\tilde{\delta}\tau$ can be understood as ε in Theorem 3.1 and Corollary 3 with an extra assumption $\tilde{\delta} \leq c\tau$ that guarantees that the solution found by minimizing L is energetically better than any saddle point of the ambient loss function ℓ.

Remark 6. *Equation 19 says that $\mathbf{Y}_{\theta^*}$ approximates a minimizer of the ambient problem 7 and that $\mathbf{Y}_{\theta^*}$ can be recovered but only up to rotation. This*

is unavoidable, since the loss function ℓ is invariant under multiplication on the right by a $r \times r$ orthogonal matrix. On the other hand, we do not enforce sparsity constraints in the optimization of the NN parameters. This is convenient in practical settings and this is the reason why we state the theorem in this way. However, we can also set $N = r\left((\tilde{\delta}\tau)^{-m}\log\frac{1}{\tilde{\delta}\tau} + d\log\frac{1}{\tilde{\delta}\tau} + d\log d\right)$ without affecting the conclusion of the theorem.

We want to highlight that Theorem 3.1,a s stated, is needed in our proof of Theorem 3.2. To see this, note that the following

$$\tau^2 C_r(\tilde{\delta}+\tau)^2 < \sigma_r^2(\mathcal{A}_n) - \sigma_{r+1}^2(\mathcal{A}_n)$$

is satisfied under the assumptions in the statement of Theorem 3.2. Indeed, notice that $\sigma_r^2(\mathcal{A}_\mathbf{n}) - \sigma_{r+1}^2(\mathcal{A}_\mathbf{n}) \sim \tau^2$ according to Remark 2 and the fact that $\lambda_r^{\mathcal{M}} < \lambda_{r+1}^{\mathcal{M}}$. Thus, taking $\tilde{\delta}$ to be sufficiently small, we get the needed inequality. Further, if the correction term in the Lipschitz estimate for graph Laplacian eigenvectors had been τ, and not τ^2, the term $\tau^2 C_r(\tilde{\delta}+\tau)^2$ would have to be replaced with the term $(C_r\tau\tilde{\delta} + C_r\varepsilon)^2$, but the latter cannot be guaranteed to be smaller than $\sigma_r^2(\mathcal{A}_\mathbf{n}) - \sigma_{r+1}^2(\mathcal{A}_\mathbf{n})$. Guaranteeing this is important as the energy gap (i.e., $\left\|\mathbf{Y}^*\mathbf{Y}^{*\top} - \mathcal{A}_\mathbf{n}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{Y}_\theta\mathbf{Y}_\theta^\top - \mathcal{A}_\mathbf{n}\right\|_{\mathrm{F}}^2$) is of order $\tau^2 C_r(\tilde{\delta}+\tau)^2$ (Corollary 5), whereas the energy gap between $\mathbf{Y}^*$ and any other critical point of ℓ that is not a global optimizer is in the order of τ^2, as it follows from Remark 2. Continuing the discussion from the previous section, it was thus relevant to use estimates that could guarantee that, at least energetically, our constructed $\mathbf{Y}_\theta$ was closer to $\mathbf{Y}^*$ than any other saddle of ℓ.

4 Landscape of SNN's Ambient Optimization Problem

From the discussion in the previous sections we now know that the global minimizer of the loss function L well approximates the eigenvectors of Δ_n. However, the loss is non-convex – both due to the non-convexity of the loss function and the parameterization of the neural network. Hence it is not immediate that gradient descent converges to the desired global minimum. In this section, we focus on the non-convexity due to the loss function and we show that gradient descent converges to the global minimum of ℓ. We do this by characterizing the optimization landscape of ℓ.

While in prior sections we considered a specific $\mathcal{A}_n$, the analysis in this section only relies on $\mathcal{A}_\mathbf{n}$ being positive definite with an eigengap between its r-th and $(r+1)$th top eigenvalues. We analyze the global optimization landscape of the non-convex Problem 7 under a suitable Riemannian *quotient geometry* [64, 59].

To analyze the landscape for Equation 8, we need expressions for the Riemannian gradient, the Riemannian Hessian, as well as the geodesic distance d

on this quotient manifold. By Lemma 2 from [56], we have that

$$d\left([\mathbf{Y}_1],[\mathbf{Y}_2]\right) = \min_{\mathbf{Q}\in\mathbb{O}_r} \|\mathbf{Y}_2\mathbf{Q}-\mathbf{Y}_1\|_{\mathrm{F}}$$

and from Lemma 3 from [56], we have that

$$\begin{aligned} \overline{\operatorname{grad} H([\mathbf{Y}])} &= 2\left(\mathbf{Y}\mathbf{Y}^\top - \mathcal{A}_{\mathbf{n}}\right)\mathbf{Y}, \\ \overline{\operatorname{Hess} H([\mathbf{Y}])}\left[\theta_{\mathbf{Y}},\theta_{\mathbf{Y}}\right] &= \left\|\mathbf{Y}\theta_{\mathbf{Y}}^\top + \theta_{\mathbf{Y}}\mathbf{Y}^\top\right\|_{\mathrm{F}}^2 + 2\left\langle \mathbf{Y}\mathbf{Y}^\top - \mathcal{A}_{\mathbf{n}}, \theta_{\mathbf{Y}}\theta_{\mathbf{Y}}^\top\right\rangle. \end{aligned} \tag{20}$$

Finally, by the classical theory on low-rank approximation (Eckart–Young–Mirsky theorem [60]), $[\mathbf{Y}^*]$ is the unique global minimizer of Equation 8. Let $\kappa^* = \sigma_1(\mathbf{Y}^*)/\sigma_r(\mathbf{Y}^*)$ be the condition number of $\mathbf{Y}^*$. Here, $\sigma_i(A)$ is the i^{th} largest singular value of A, and $\|A\| = \sigma_1(A)$ is its spectral norm. Our precise assumption on the matrix $\mathcal{A}_n$ for this section is as follows.

Assumption 2 (Eigengap)**.** *$\sigma_{r+1}(\mathcal{A}_{\mathbf{n}})$ is strictly smaller than $\sigma_r(\mathcal{A}_{\mathbf{n}})$.*

See Remark 12 for a discussion of the potential relaxation of the Eigengap assumption.

Let $\mu,\alpha,\beta,\gamma \geqslant 0$. We then split the landscape of $H([\mathbf{Y}])$ into the following five regions (not necessarily non-overlapping).

$$\begin{aligned} \mathcal{R}_1 &\overset{\text{def}}{=} \left\{\mathbf{Y}\in\mathbb{R}_*^{n\times r} \middle| d\left([\mathbf{Y}],[\mathbf{Y}^*]\right) \leqslant \mu\sigma_r(\mathbf{Y}^*)/\kappa^*\right\}, \\ \mathcal{R}_2 &\overset{\text{def}}{=} \left\{\mathbf{Y}\in\mathbb{R}_*^{n\times r} \middle| \begin{array}{l} d\left([\mathbf{Y}],[\mathbf{Y}^*]\right) > \mu\sigma_r(\mathbf{Y}^*)/\kappa^*, \|\overline{\operatorname{grad} H([\mathbf{Y}])}\|_{\mathrm{F}} \leqslant \alpha\mu\sigma_r^3(\mathbf{Y}^*)/(4\kappa^*), \\ \|\mathbf{Y}\| \leqslant \beta\|\mathbf{Y}^*\|, \left\|\mathbf{Y}\mathbf{Y}^\top\right\|_{\mathrm{F}} \leqslant \gamma\left\|\mathbf{Y}^*\mathbf{Y}^{*\top}\right\|_{\mathrm{F}} \end{array}\right\}, \\ \mathcal{R}_3' &\overset{\text{def}}{=} \left\{\mathbf{Y}\in\mathbb{R}_*^{n\times r} \middle| \begin{array}{l} \|\overline{\operatorname{grad} H([\mathbf{Y}])}\|_{\mathrm{F}} > \alpha\mu\sigma_r^3(\mathbf{Y}^*)/(4\kappa^*), \|\mathbf{Y}\| \leqslant \beta\|\mathbf{Y}^*\|, \\ \left\|\mathbf{Y}\mathbf{Y}^\top\right\|_{\mathrm{F}} \leqslant \gamma\left\|\mathbf{Y}^*\mathbf{Y}^{*\top}\right\|_{\mathrm{F}} \end{array}\right\}, \\ \mathcal{R}_3'' &\overset{\text{def}}{=} \left\{\mathbf{Y}\in\mathbb{R}_*^{n\times r} \middle| \|\mathbf{Y}\| > \beta\|\mathbf{Y}^*\|, \|\mathbf{Y}\mathbf{Y}^\top\|_{\mathrm{F}} \leqslant \gamma\|\mathbf{Y}^*\mathbf{Y}^{*\top}\|_{\mathrm{F}}\right\}, \\ \mathcal{R}_3''' &\overset{\text{def}}{=} \left\{\mathbf{Y}\in\mathbb{R}_*^{n\times r} \middle| \left\|\mathbf{Y}\mathbf{Y}^\top\right\|_{\mathrm{F}} > \gamma\left\|\mathbf{Y}^*\mathbf{Y}^{*\top}\right\|_{\mathrm{F}}\right\}, \end{aligned} \tag{21}$$

We show that for small values of μ, the *loss function is geodesically convex* in $\mathcal{R}_1$. $\mathcal{R}_2$ is then defined as the region outside of $\mathcal{R}_1$ such that the Riemannian gradient is small relative to μ. Hence this is the region in which we are close to the saddle points. We show that for this region there is *always an escape direction* (i.e., directions where the Hessian is strictly negative). $\mathcal{R}_3'$, $\mathcal{R}_3''$, and $\mathcal{R}_3'''$ are the remaining regions. We show that the *Riemannian gradient is large* (relative to μ) in these regions. Finally, it is easy to see that $\mathcal{R}_1 \bigcup \mathcal{R}_2 \bigcup \mathcal{R}_3' \cup \mathcal{R}_3'' \bigcup \mathcal{R}_3''' = \mathbb{R}_*^{n\times r}$.

We are now ready to state the first of our main results from this section.

Theorem 4.1 (Local Geodesic Strong Convexity and Smoothness of Equation 8)**.** *Suppose $0 \leqslant \mu \leqslant \kappa^*/3$. Given that Assumption 2 holds, for any $\mathbf{Y}\in\mathcal{R}_1$,*

$$\begin{aligned} \sigma_{\min}(\overline{\operatorname{Hess} H([\mathbf{Y}])}) &\geqslant \left(2\left(1-\mu/\kappa^*\right)^2 - (14/3)\mu\right)\sigma_r(\mathcal{A}_{\mathbf{n}}) - 2\sigma_{r+1}(\mathcal{A}_{\mathbf{n}}), \\ \sigma_{\max}(\overline{\operatorname{Hess} H([\mathbf{Y}])}) &\leqslant 4\left(\sigma_1(\mathbf{Y}^*) + \mu\sigma_r(\mathbf{Y}^*)/\kappa^*\right)^2 + 14\mu\sigma_r^2(\mathbf{Y}^*)/3 \end{aligned}$$

In particular, if μ is further chosen such that $\left(2\left(1-\mu/\kappa^*\right)^2-(14/3)\mu\right)\sigma_r\left(\mathcal{A}_{\mathbf{n}}\right)-2\sigma_{r+1}(\mathcal{A}_{\mathbf{n}})>0$*, we have $H([\mathbf{Y}])$ is geodesically strongly convex and smooth in $\mathcal{R}_1$.*

Theorem 4.1 guarantees that the optimization problem Equation 8 is geodesically strongly convex and smooth in a neighborhood of $[\mathbf{Y}^*]$. It also shows that if $\mathbf{Y}$ is close to the global minimizer, then Riemannian gradient descent stays in $\mathcal{R}_1$ and converges to the global minimizer of the quotient space linearly following the proof of 59, Theorem 11.29.

By combining this result with Theorem 3.2, when the number of neurons is large enough, $f_{\theta^*}(\mathcal{X}_n)\in\mathcal{R}_1$. Then, by applying gradient descent initiating at $f_{\theta^*}(\mathcal{X}_n)$, we gain a linear convergence rate to the eigenvector estimation of Equation 7.

In general, gradient descent is known to converge to a FOSP. Hence one might expect that if we initialized near one of the saddle points, then we might converge to that saddle point. However, our next main result of the section shows that even if we initialize near the saddle, there always exist escape directions. However, before we can prove this result, Theorem 4.2, we need to discuss specific assumptions to guarantee that α is sufficiently small.

Assumption 3 (Parameters Settings)**.** *Denote e_1, e_2 and e_3 to be some error terms as follows,*

$$e_1 \overset{def}{=} \frac{\alpha\mu\sigma_r^3\left(\mathbf{Y}^*\right)}{2\sqrt{2}\kappa^*\sigma_{r+1}(\mathbf{\Lambda})},\qquad e_2=\frac{e_1}{\sqrt{2}},\qquad \textit{and}\qquad e_3=e_2\cdot\sigma_{r+1}(\mathbf{\Lambda})$$

Note that $e_1, e_2, e_3\to 0$ as $\alpha\to 0$. Hence, pick α small enough such that the following conditions are true.

1. $\sigma_r^2(\mathbf{\Lambda})-2e_1-\sigma_{r+1}^2(\mathbf{\Lambda})>0$.
2. $\sigma_r^2(\mathbf{\Lambda})\left(1-\frac{e_1^2}{\left|\sigma_r^2(\mathbf{\Lambda})-e_1-\sigma_{r+1}^2(\mathbf{\Lambda})\right|^2}\right)-e_1-\sigma_{r+1}^2(\mathbf{\Lambda})>0$.
3. $(\alpha-2(\sqrt{2}-1))\sigma_r^2\left(\mathbf{Y}^*\right)+6\frac{\alpha^2\sigma_r^4(\mathbf{Y}^*)\sigma_{r+1}^2(\mathbf{\Lambda})/16}{\left|\sigma_r^2(\mathbf{\Lambda})-e_2-\sigma_{r+1}^2(\mathbf{\Lambda})\right|^2}<0$.

Note that given the eigengap assumption, the first two conditions can be satisfied as $\alpha\to 0$. For the last condition, we have that as $\alpha\to 0$, the left hand side converges to $-2(\sqrt{2}-1)\sigma_r^2\left(\mathbf{Y}^*\right)$ which is negative. Hence, Assumption 3 is only related to the eigengap assumption $\sigma_r(\mathbf{\Lambda})$ and $\sigma_{r+1}(\mathbf{\Lambda})$ in Assumption 2. As soon as α is small enough, Assumption 3 is satisfied.

We recall that the SVD decomposition of $\mathbf{Y}$ is $\mathbf{UDV}^\top$.

Theorem 4.2 (Region with Negative Eigenvalue in the Riemannian Hessian of Equation 7)**.** *Assume that Assumption 2 holds. Given any $\mathbf{Y}\in\mathbb{R}_*^{n\times r}$ such that $\mathbf{Y}\in\mathcal{R}_2$, let $\theta_{\mathbf{Y}}^1=[\mathbf{0},\mathbf{0},\ldots,\mathbf{0},\boldsymbol{a},\mathbf{0},\ldots,\mathbf{0}]\mathbf{V}^\top$ where $\boldsymbol{a}$ such that*

$$\boldsymbol{a}=\underset{\boldsymbol{a}:\mathbf{Y}^\top\boldsymbol{a}=\mathbf{0}}{\arg\max}\ \frac{\boldsymbol{a}^\top\mathcal{A}_{\mathbf{n}}\boldsymbol{a}}{\|\boldsymbol{a}\|^2} \tag{22}$$

and $[\mathbf{0}, \mathbf{0}, \ldots, \mathbf{0}, \boldsymbol{a}, \mathbf{0}, \ldots, \mathbf{0}] \in \mathbb{R}^{n\times r}$ *such that the* $\tilde{i}^{\text{th}}$ *columns is* $\boldsymbol{a}$ *and other columns are* $\mathbf{0}$ *where*

$$\tilde{i} \overset{def}{=} \arg\min_{j\in[r]} \mathbf{D}_{jj}. \tag{23}$$

Denote $\theta^2_{\mathbf{Y}} = \mathbf{Y} - \mathbf{Y}^*\mathbf{Q}$*, where* $\mathbf{Q} \in \mathbb{O}_r$ *is the best orthogonal matrix aligning* $\mathbf{Y}^*$ *and* $\mathbf{Y}$*. We choose* $\theta_{\mathbf{Y}}$ *to be either* $\theta^1_{\mathbf{Y}}$ *or* $\theta^2_{\mathbf{Y}}$*. Then*

$$\begin{aligned}\overline{\operatorname{Hess} H([\mathbf{Y}])}\,[\theta_{\mathbf{Y}}, \theta_{\mathbf{Y}}] \leqslant \min\Bigg\{ & -\frac{\sigma^2_{r+1}(\mathbf{\Lambda})}{2}\|\theta_{\mathbf{Y}}\|^2, \\ & -2\left(\sigma^2_r(\mathbf{\Lambda})\left(1 - \frac{e_1^2}{\left|\sigma^2_r(\mathbf{\Lambda}) - e_1 - \sigma^2_{r+1}(\mathbf{\Lambda})\right|^2}\right) - e_1 - \sigma^2_{r+1}(\mathbf{\Lambda})\right)\|\theta_{\mathbf{Y}}\|^2, \\ & \left((\alpha - 2(\sqrt{2}-1))\sigma^2_r(\mathbf{Y}^*) + 6\frac{\alpha^2\sigma^4_r(\mathbf{Y}^*)\,\sigma^2_{r+1}(\mathbf{\Lambda})/16}{\left|\sigma^2_r(\mathbf{\Lambda}) - e_2 - \sigma^2_{r+1}(\mathbf{\Lambda})\right|^2}\right)\|\theta_{\mathbf{Y}}\|^2_{\mathrm{F}}\Bigg\}\end{aligned}$$

In particular, if α *and* μ *satisfies Assumption 3, we have* $\overline{\operatorname{Hess} H([\mathbf{Y}])}$ *has at least one negative eigenvalue and* $\theta^1_{\mathbf{Y}}$ *or* $\theta^2_{\mathbf{Y}}$ *is the escaping direction.*

Remark 7. *Theorem 4.2 suggests that if some singular values of* $\mathbf{Y}$ *are small, then the descent direction* $\theta^1_{\mathbf{Y}}$ *should increase the singular value of* $\mathbf{Y}$*. If all of the singular values of* $\mathbf{Y}$ *are large enough compared with* $\sigma_r(\mathbf{\Lambda})$*, then* $\theta^2_{\mathbf{Y}}$ *should directly point* $[\mathbf{Y}]$ *to* $[\mathbf{Y}^*]$*. Thus, Theorem 4.2 fully characterizes the regime of* $\mathcal{R}_2$ *with respect to different minimum singular values of* $\mathbf{Y}$*.*

- *If any singular value of* $\mathbf{Y}\mathbf{Y}^\top$ *is smaller than* $\frac{\sigma^2_{r+1}(\mathbf{\Lambda})}{2}$*, then we have*

$$\overline{\operatorname{Hess} H([\mathbf{Y}])}\left[\theta^1_{\mathbf{Y}}, \theta^1_{\mathbf{Y}}\right] \leq -\frac{\sigma^2_{r+1}(\mathbf{\Lambda})}{2}\|\theta^1_{\mathbf{Y}}\|^2.$$

- *When the smallest singular value of* $\mathbf{Y}\mathbf{Y}^\top$ *is larger than* $\frac{\sigma^2_{r+1}(\mathbf{\Lambda})}{2}$ *and smaller than* $e_1 + \sigma^2_{r+1}(\mathbf{\Lambda})$*, then we have*

$$\overline{\operatorname{Hess} H([\mathbf{Y}])}\left[\theta^1_{\mathbf{Y}}, \theta^1_{\mathbf{Y}}\right] \leq -2\left(\sigma^2_r(\mathbf{\Lambda})\left(1 - \frac{e_1^2}{\left|\sigma^2_r(\mathbf{\Lambda}) - e_1 - \sigma^2_{r+1}(\mathbf{\Lambda})\right|^2}\right) - e_1 - \sigma^2_{r+1}(\mathbf{\Lambda})\right)\|\theta^1_{\mathbf{Y}}\|^2.$$

- *If all of the eigenvalues of* $\mathbf{Y}\mathbf{Y}^\top$ *are larger than* $e_1 + \sigma^2_{r+1}(\mathbf{\Lambda})$*, then we have* $\overline{\operatorname{Hess} H([\mathbf{Y}])}\left[\theta^2_{\mathbf{Y}}, \theta^2_{\mathbf{Y}}\right]$ *is smaller than*

$$\left((\alpha - 2(\sqrt{2}-1))\sigma^2_r(\mathbf{Y}^*) + 6\frac{\alpha^2\sigma^4_r(\mathbf{Y}^*)\,\sigma^2_{r+1}(\mathbf{\Lambda})/16}{\left|\sigma^2_r(\mathbf{\Lambda}) - e_2 - \sigma^2_{r+1}(\mathbf{\Lambda})\right|^2}\right)\left\|\theta^2_{\mathbf{Y}}\right\|^2_{\mathrm{F}}.$$

Finally, the next result says that if we are not close to a FOSP, then we have large gradients.

Theorem 4.3 ((Regions with Large Riemannian Gradient of Equation 7)**.**

1. $\|\overline{\operatorname{grad} H([\mathbf{Y}])}\|_{\mathrm{F}} > \alpha\mu\sigma_r^3(\mathbf{Y}^*)/(4\kappa^*), \forall \mathbf{Y} \in \mathcal{R}_3'$;
2. $\|\overline{\operatorname{grad} H([\mathbf{Y}])}\|_{\mathrm{F}} \geqslant 2\left(\|\mathbf{Y}\|^3 - \|\mathbf{Y}\|\,\|\mathbf{Y}^*\|^2\right) > 2\left(\beta^3 - \beta\right)\|\mathbf{Y}^*\|^3, \quad \forall \mathbf{Y} \in \mathcal{R}_3''$;
3. $\langle \overline{\operatorname{grad} H([\mathbf{Y}])}, \mathbf{Y}\rangle > 2(1-1/\gamma)\left\|\mathbf{Y}\mathbf{Y}^\top\right\|_{\mathrm{F}}^2, \quad \forall \mathbf{Y} \in \mathcal{R}_3'''$.

In particular, if $\beta > 1$ and $\gamma > 1$, we have the Riemannian gradient of $H([\mathbf{Y}])$ has large magnitude in all regions $\mathcal{R}_3', \mathcal{R}_3''$ and $\mathcal{R}_3'''$.

The behavior, implied by our theorems, of gradient descent as it goes through the regions $\mathcal{R}_1$, $\mathcal{R}_2$, $\mathcal{R}_3$ is illustrated in Figures 4 and 5. See a discussion in section 5.

Remark 8. *These results can be seen as an under-parameterized generalization to the regression problem of Section 5 in [56]. The proof in [56] is simpler because in their setting there are no saddle points or local minima that are not global in $\mathbb{R}_*^{n\times r}$. Conceptually, [46] proves that in the setting $r \geq n$, the gradient flow for Equation 7 converges to a global minimum linearly; in particular, in their setting there aren't any saddle points. We complement this result by studying the case $r < n$.*

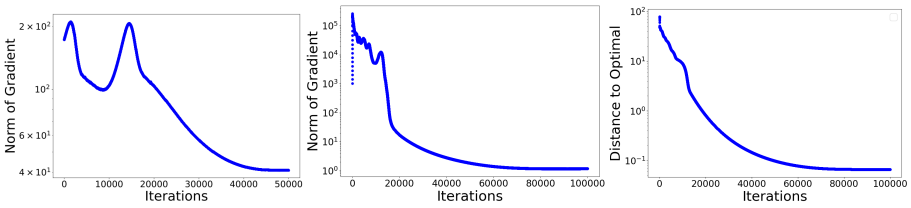
Remark 9. *In the specific case of $\mathcal{A}_{\mathbf{n}}$ as in Equation 5, and under Assumptions 1, Assumption 2 should be interpreted as $\lambda_r^{\mathcal{M}} < \lambda_{r+1}^{\mathcal{M}}$, as suggested by Remark 2. Also, μ must be taken to be in the order τ^2. The scale τ^2 is actually a natural scale for this problem, the energy gap between saddle points and the global minimizer $[\mathbf{Y}^*]$ is $O(\tau^2)$.*

Remark 10. *Theorem 4.1, 4.2 and 4.3 guarantee the benigness of the ambient optimization problem, which is necessary condition of the benigness of parameterized optimization problem Equation 2. Also, these landscape results imply that perturbed gradient descent is guaranteed to converge to the global minima in polynomial time for Equation 7 65, 66, 67, 68.*

5 Parameterized Loss Landscape

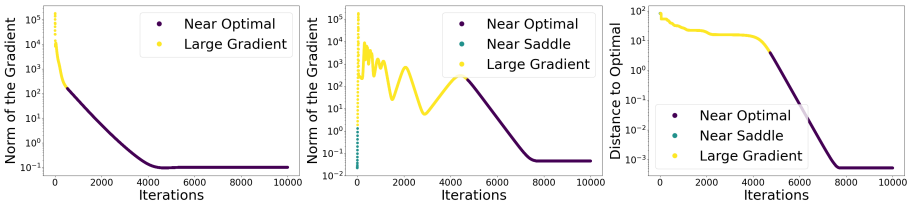
Finally, answering whether gradient descent converges to the global minimum for the parameterized problem (i.e., the NN training problem) is quite challenging. Hence, for this piece we explore the question experimentally. In Section 6, we explicitly state this as an interesting theoretical question that is worth exploring in the future.

Specifically, we present some numerical experiments where we consider different initializations for the training of SNN. Here we take 100 data points from MNIST and let $\mathcal{A}_n$ be the $n \times n$ gram matrix for the data points for simplicity. The detailed experimental design is provided in Appendix B.2. We remark that while we care about a $\mathcal{A}_n$ with a specific form for our approximation theory results, our analysis of the loss landscape described below holds for an arbitrary positive semi-definite matrix. In Figure 4, we plot the norm of the gradient during training when initialized in two different regions of parameter space.



(a) Initialized Near Optimal (b) Initialized Near Saddle (c) Initialized Near Saddle

Figure 4: (a) and (b) Sum of the norms of the gradients for a two-layer ReLU Neural Network. In (a), the network is initialized near the global optimal solution and in (b) the network is initialized near a saddle point. (c) shows the distance between the current outputs of the neural network and the optimal solution for the case when it was initialized near a saddle point. More details are presented in Appendix B.2.



(a) Initialized Near Optimal (b) Initialized Near Saddle (c) Initialized Near Saddle

Figure 5: Norms of the gradients for the ambient problem and the distance to the optimal solution. In (a), $\mathbf{Y}$ is initialized near the global optimal solution, and in (b) $\mathbf{Y}$ is initialized near a saddle point. c) shows the distance between $\mathbf{Y}$ and the optimal solution for the case when it was initialized near a saddle point.

Concretely, in a region of parameters for which $\mathbf{Y}_\theta$ is close to a solution $\mathbf{Y}^*$ to problem 7 and a region of parameters for which $\mathbf{Y}_\theta$ is close to a saddle point of the ambient loss ℓ. We compare these plots to the ones we produce from the gradient descent dynamics for the ambient problem 7, which are shown in Figure 5. We notice a similar qualitative behavior with the training dynamics of the NN, suggesting that the landscape of problem 2, if the NN is properly overparameterized, inherits properties of the landscape of ℓ.

6 Conclusions

We have explored some theoretical aspects of Spectral Neural Networks (SNN), a framework that substitutes the use of traditional eigensolvers with suitable neural network parameter optimization. Our emphasis has been on approximation theory, specifically identifying the minimum number of neurons of a multilayer NN required to capture spectral geometric properties in data, and investigating the optimization landscape of SNN, even in the face of its non-convex ambient loss function.

For our approximation theory results we have assumed a specific proximity graph structure over data points that are sampled from a distribution over a smooth low-dimensional manifold. A natural future direction worth of study is the generalization of these results to settings where data points, and their similarity graph, are sampled from other generative models, e.g., as in the application to contrastive learning in [8]. To carry out this generalization, an important first step is to study the regularity properties of eigenvectors of an adjacency matrix/graph Laplacian generated from other types of probabilistic models.

At a high level, our approximation theory results have sought to bridge the extensive body of research on graph-based learning methods, their ties to PDE theory on manifolds, and the approximation theory for neural networks. While our analysis has focused on eigenvalue problems, such as those involving graph Laplacians or Laplace Beltrami operators, we anticipate that this overarching objective can be extended to develop new provably consistent methods for solving a larger class of PDEs on manifolds with neural networks, such as Schrödinger equation as in [69, 70]. We believe this represents a significant and promising research avenue.

On the optimization front, we have focused on studying the landscape of the ambient space problem 7. This has been done anticipating the use of our estimates in a future analysis of the training dynamics of SNN. We reiterate that the setting of interest here is different from other settings in the literature that study the dynamics of neural network training in an appropriate scaling limit —leading to either a neural tangent kernel (NTK) or to a mean field limit. This difference is mainly due to the fact that the spectral contrastive loss ℓ (see 2) of SNN is non-convex, and even local strong convexity around a global minimizer does not hold in a standard sense and instead can only be guaranteed when considered under a suitable quotient geometry.

References

[1] Rie Ando and Tong Zhang. Learning on graph with laplacian regularization. *Advances in neural information processing systems*, 19, 2006.

[2] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(11), 2006.

[3] Alexander J Smola and Risi Kondor. Kernels and regularization on graphs. In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings*, pages 144–158. Springer, 2003.

[4] Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14, 2001.

[5] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[6] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in neural information processing systems*, 14, 2001.

[7] Ronald R Coifman, Stephane Lafon, Ann B Lee, Mauro Maggioni, Boaz Nadler, Frederick Warner, and Steven W Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the national academy of sciences*, 102(21):7426–7431, 2005.

[8] Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34, 2021.

[9] Uri Shaham, Kelly Stanton, Henry Li, Ronen Basri, Boaz Nadler, and Yuval Kluger. Spectralnet: Spectral clustering using deep neural networks. In *International Conference on Learning Representations*, 2018.

[10] Nicolás García Trillos and Dejan Slepvcev. A variational approach to the consistency of spectral clustering. *Applied and Computational Harmonic Analysis*, 45(2):239–281, 2018.

[11] David Pfau, Stig Petersen, Ashish Agarwal, David G. T. Barrett, and Kimberly L. Stachenfeld. Spectral inference networks: Unifying deep and spectral learning. In *International Conference on Learning Representations*, 2019.

[12] Zhijie Deng, Jiaxin Shi, and Jun Zhu. NeuralEF: Deconstructing kernels by deep neural networks. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 4976–4992. PMLR, 17–23 Jul 2022.

[13] Wenqi Tao and Zuoqiang Shi. Convergence of laplacian spectra from random samples. *Journal of Computational Mathematics*, 38(6):952–984, 2020.

[14] Dmitri Burago, Sergei Ivanov, and Yaroslav Kurylev. A graph discretization of the Laplace-Beltrami operator. *Journal of Spectral Theory*, 4(4):675–714, 2014.

[15] Nicolás García Trillos, Moritz Gerlach, Matthias Hein, and Dejan Slepćev. Error Estimates for Spectral Convergence of the Graph Laplacian on Random Geometric Graphs Toward the Laplace–Beltrami Operator. *Foundations of Computational Mathematics*, 20(4):827–887, August 2020.

[16] Jinpeng Lu. Graph approximations to the laplacian spectra. *Journal of Topology and Analysis*, 14(01):111–145, 2022.

[17] Jeff Calder and Nicolás García Trillos. Improved spectral convergence rates for graph laplacians on ε-graphs and k-nn graphs. *Applied and Computational Harmonic Analysis*, 60:123–175, 2022.

[18] Jeff Calder, Nicolás García Trillos, and Marta Lewicka. Lipschitz regularity of graph laplacians on random data clouds. *SIAM Journal on Mathematical Analysis*, 54(1):1169–1222, 2022.

[19] David B Dunson, Hau Tieng Wu, and Nan Wu. Spectral convergence of graph laplacian and heat kernel reconstruction in l^∞ from random samples. *Applied and Computational Harmonic Analysis*, 55:282–336, 2021.

[20] Caroline L Wormell and Sebastian Reich. Spectral convergence of diffusion maps: Improved error bounds and an alternative normalization. *SIAM Journal on Numerical Analysis*, 59(3):1687–1734, 2021.

[21] Nicolás García Trillos, Pengfei He, and Chenghui Li. Large sample spectral analysis of graph-based multi-manifold clustering. *Journal of Machine Learning Research*, 24(143):1–71, 2023.

[22] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[23] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020.

[24] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[25] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[26] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

[27] Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive estimation reveals topic posterior information to linear models. *Journal of Machine Learning Research*, 22(281):1–31, 2021.

[28] Jason D Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34:309–323, 2021.

[29] Philipp Christian Petersen. Neural network theory. *University of Vienna*, 2020.

[30] Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021.

[31] Zuowei Shen, Haizhao Yang, and Shijun Zhang. Nonlinear approximation via compositions. *Neural Networks*, 119:74–84, 2019.

[32] Dmitry Yarotsky. Optimal approximation of continuous functions by very deep relu networks. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, volume 75 of *Proceedings of Machine Learning Research*, pages 639–649. PMLR, 2018.

[33] Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation characterized by number of neurons. *Communications in Computational Physics*, 2020.

[34] Johannes Schmidt-Hieber. Deep relu network approximation of functions on a manifold. *arXiv preprint arXiv:1908.00695*, 2019.

[35] Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Small relu networks are powerful memorizers: a tight analysis of memorization capacity. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[36] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

[37] E Weinan and Ting Yu. The deep ritz method: A deep learning-based numerical algorithm for solving variational problems. *Communications in Mathematics and Statistics*, 6:1–12, 2017.

[38] Saakaar Bhatnagar, Yaser Afshar, Shaowu Pan, Karthik Duraisamy, and Shailendra Kaushik. Prediction of aerodynamic flow fields using convolutional neural networks. *Computational Mechanics*, 64:525–545, 2019.

[39] Xiaoxiao Guo, Wei Li, and Francesco Iorio. Convolutional neural networks for steady flow approximation. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, KDD '16, page 481–490. Association for Computing Machinery, 2016.

[40] Yinhao Zhu and Nicholas Zabaras. Bayesian deep convolutional encoder–decoder networks for surrogate modeling and uncertainty quantification. *Journal of Computational Physics*, 366:415–447, 2018.

[41] Jonas Adler and Ozan Öktem. Solving ill-posed inverse problems using iterative deep neural networks. *Inverse Problems*, 33(12):124007, 2017.

[42] Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces. *arXiv preprint arXiv:2108.08481*, 2021.

[43] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

[44] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Graph kernel network for partial differential equations. *arXiv preprint arXiv:2003.03485*, 2020.

[45] Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2(1):53–58, 1989.

[46] Salma Tarmoun, Guilherme Franca, Benjamin D Haeffele, and Rene Vidal. Understanding the dynamics of gradient flow in overparameterized linear models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10153–10161. PMLR, 18–24 Jul 2021.

[47] Hancheng Min, Salma Tarmoun, Rene Vidal, and Enrique Mallada. On the explicit role of initialization on the convergence and implicit bias of overparametrized linear networks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7760–7768. PMLR, 18–24 Jul 2021.

[48] Pierre Bréchet, Katerina Papagiannouli, Jing An, and Guido Montúfar. Critical points and convergence analysis of generative deep linear networks trained with bures-wasserstein loss. *arXiv preprint arXiv:2303.03027*, 2023.

[49] Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 2018.

[50] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in High-Dimensional Ridgeless Least Squares Interpolation. *The Annals of Statistics*, 2022.

[51] Peter Bartlett, Philip M. Long, Gabor Lugosi, and Alexander Tsigler. Benign Overfitting in Linear Regression. *Proceedings of the National Academy of Sciences*, 2020.

[52] Rishi Sonthalia and Raj Rao Nadakuditi. Training data size induced double descent for denoising feedforward neural networks and the role of training noise. *Transactions on Machine Learning Research*, 2023.

[53] Chinmaya Kausik, Kashvi Srivastava, and Rishi Sonthalia. Generalization error without independence: Denoising, linear regression, and transfer learning. *arXiv preprint arXiv:2305.17297*, 2023.

[54] Qiuwei Li and Gongguo Tang. The nonconvex geometry of low-rank matrix optimizations with general objective functions. In *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1235–1239, 2017.

[55] Yuejie Chi, Yue M. Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

[56] Yuetian Luo and Nicolás García Trillos. Nonconvex matrix factorization is geodesically convex: Global landscape analysis for fixed-rank matrix optimization from a riemannian perspective. *arXiv preprint arXiv:2209.15130*, 2022.

[57] Xingguo Li, Junwei Lu, Raman Arora, Jarvis Haupt, Han Liu, Zhaoran Wang, and Tuo Zhao. Symmetry, saddle points, and global optimization landscape of nonconvex matrix factorization. *IEEE Transactions on Information Theory*, 65(6):3489–3514, 2019.

[58] John M Lee. *Introduction to Riemannian manifolds*, volume 176. Springer, 2018.

[59] Nicolas Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023.

[60] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[61] Neil Rohit Mallinar, James B Simon, Amirhesam Abedsoltan, Parthe Pandit, Misha Belkin, and Preetum Nakkiran. Benign, tempered, or catastrophic: Toward a refined taxonomy of overfitting. In *Advances in Neural Information Processing Systems*, 2022.

[62] Yutong Wang, Rishi Sonthalia, and Wei Hu. Near-interpolators: Rapid norm growth and the trade-off between interpolation and generalization. In *AISTATS*, 2024.

[63] Minshuo Chen, Haoming Jiang, Wenjing Liao, and Tuo Zhao. Nonparametric regression on low-dimensional manifolds using deep relu networks: Function approximation and statistical recovery. *Information and Inference: A Journal of the IMA*, 11(4):1203–1253, 2022.

[64] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. Optimization algorithms on matrix manifolds. In *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

[65] Yue Sun, Nicolas Flammarion, and Maryam Fazel. Escaping from saddle points on riemannian manifolds. *Advances in Neural Information Processing Systems*, 32, 2019.

[66] Christopher Criscitiello and Nicolas Boumal. Efficiently escaping saddle points on manifolds. *Advances in Neural Information Processing Systems*, 32, 2019.

[67] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on learning theory*, pages 797–842. PMLR, 2015.

[68] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International conference on machine learning*, pages 1724–1732. PMLR, 2017.

[69] Jan Hermann, Zeno Schätzle, and Frank Noé. Deep-neural-network solution of the electronic schrödinger equation. *Nature Chemistry*, 12(10):891–897, 2020.

[70] Jianfeng Lu and Yulong Lu. A priori generalization error analysis of two-layer neural networks for solving high dimensional schrödinger eigenvalue problems. *Communications of the American Mathematical Society*, 2(1):1–21, 2022.

[71] Gilbert W Stewart. *Matrix algorithms: volume 1: basic decompositions*. SIAM, 1998.

[72] Estelle Massart and P.-A. Absil. Quotient geometry with simple geodesics for the manifold of fixed-rank positive-semidefinite matrices. *SIAM Journal on Matrix Analysis and Applications*, 41(1):171–198, 2020.

[73] Manfredo Perdigao Do Carmo and J Flaherty Francis. *Riemannian geometry*, volume 6. Springer, 1992.

[74] Yuetian Luo, Xudong Li, and Anru R Zhang. On geometric connections of embedded and quotient geometries in riemannian fixed-rank matrix optimization. *arXiv preprint arXiv:2110.12121*, 2021.

A Training of neural networks for spectral approximations

A.1 Training

Two of the main issues of standard eigensolvers are the need to store large matrices in memory and the need to redo computations from scratch if new data points are added. As mentioned, SNN can overcome this issue using mini-batch training. Specifically, the loss function $\ell(\mathbf{Y})$ can be written as,

$$\ell(\mathbf{Y}_\theta) = \sum_{i=1}^{n}\sum_{j=1}^{n}\left((\mathcal{A}_\mathbf{n})_{ij} - (\mathbf{Y}_\theta\mathbf{Y}_\theta^\top)_{ij})\right) = \sum_{i=1}^{n}\sum_{j=1}^{n}\Big((\mathcal{A}_\mathbf{n})_{ij} - \big\langle f_\theta(x_i), f_\theta(x_j)\big\rangle\Big)^2 \tag{24}$$

where $(\mathcal{A}_\mathbf{n})_{ij}$ represents the (i,j) entry of $\mathcal{A}_\mathbf{n}$ and f_θ is the neural network. Hence, in every iteration, one can randomly generate 1 index (i,j) from $[n]\times[n]$, compute the loss and gradient for that term in the summation, and then perform one iteration of gradient descent.

A.2 Other Training Approaches

SpectralNet: SpectralNet aims at minimizing the *SpectralNet loss*,

$$\mathcal{L}_{\text{SpectralNet}}(\theta) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\eta\left(\frac{|x_i - x_j|}{\varepsilon}\right)\|f_\theta(x_i) - f_\theta(x_j)\|^2 \tag{25}$$

where $f_\theta : \mathbb{R}^d \to \mathbb{R}^r$ encodes the spectral embedding of x_i while satisfying the constraint

$$\mathbf{Y}_\theta^\top\mathbf{Y}_\theta = n\mathbf{I}_r, \tag{26}$$

where $\mathbf{Y}_\theta = [f_\theta(x_1), \ldots, f_\theta(x_n)]$. This constraint is used to avoid a trivial solution. Note that Equation 26 is a global constraint. [9] have established a stochastic coordinate descent fashion to efficiently train SpectralNets. However, the stochastic training process in [9] can only guarantee Equation 26 holds approximately.

Conceptually, the SpectralNet loss Equation 25 can also be written as

$$\mathcal{L}_{\text{SpectralNet}}(\theta) = \frac{2}{n^2}\operatorname{trace}\left(\mathbf{Y}_\theta^\top\left(\mathbf{D}_{\boldsymbol{G}} - \boldsymbol{G}\right)\mathbf{Y}_\theta\right) \tag{27}$$

where $\boldsymbol{G} \in \mathbb{R}^{n\times n}$ such that $\boldsymbol{G}_{ij} = \eta\left(\frac{\|x_i - x_j\|}{\varepsilon}\right)$, and $\mathbf{D}_{\boldsymbol{G}}$ is a diagonal matrix where

$$(\mathbf{D}_{\boldsymbol{G}})_{ii} = \sum_{j=1}^{n}\boldsymbol{G}_{ij}. \tag{28}$$

The symmetric and positive semi-definite matrix $\mathbf{D}_{\boldsymbol{G}} - \boldsymbol{G}$ encodes the unnormalized graph Laplacian. Since $\mathbf{D}_{\boldsymbol{G}} - \boldsymbol{G}$ is positive semi-definite, the ambient problem of Equation 27 is a constrained convex optimization problem. However, the parametrization and hard constraint 26 make understanding SpectralNet's training process from a theoretical perspective challenging.

B Numerical Details

B.1 For Eigenvector Illustration

We sample 2000 data points x_i uniformly from a 2-dimensional sphere embedded in $\mathbb{R}^3$, and then construct a 30 nearest neighbor graph among these points. Figure 1 shows a 1-hidden layer neural network evaluated at x_i, with 10000 hidden neurons to learn the first eigenvector of the graph Laplacian. The Network is trained for 5000 epochs using the full batch *Adam* in *Pytorch* and a learning rate of $2 * 10^{-5}$.

B.2 Ambient vs Parameterized Problem

Data: We took 100 data points from MNIST. We normalized the pixel values to live in $[0, 1]$ and then computed $\mathcal{A}_n$ as the gran matrix.

Network Architecture: The neural network has one hidden layer with a width of 1000.

Initialization: To initialize the neural network near a saddle point, we randomly pick a saddle point and then pretrain the network to approach this saddle. We used full batch gradient descent with an initial learning rate of 3e-6. We trained the network for 10000 iterations and used Cosine annealing as the learning rate scheduler. When we initialized the network near the optimal solution, we followed the same procedure but pretrained the network for 1250 iterations.

Training Details: After pretraining the network, we trained the network with the true objective. We used full batch gradient descent with an initial learning rate of 3e-6. We trained the network for 10000 iterations and used Cosine annealing as the learning rate scheduler.

For the ambient problem, we used gradient descent with a learning rate 3e-6. We trained for 5000 iterations and again used Cosine annealing for the learning rate scheduler.

We remark that the sublinearity convergence rate in Figures 4 and 5 is due to the step size decaying in the optimizer. In $\mathcal{R}_1$, $H([\mathbf{Y}])$ has been shown to be strongly convex, so keeping the same step size should guarantee a linear rate. In this work, we don't focus on the optimization problem of SNN, but use this to illustrate Theorem 4.1, 4.2 and 4.3.

C Properties of the matrix $\mathcal{A}_\mathbf{n}$ in Equation 2

C.1 Proof of Proposition 1

The matrix $\mathcal{A}_\mathbf{n}$ defined in Equation 5 satisfies the following properties:

1. $\mathcal{A}_\mathbf{n}$ is symmetric positive definite.
2. $\mathcal{A}_\mathbf{n}$'s r *top* eigenvectors (the ones corresponding to the r largest eigenvalues) coincide with the eigenvectors of the r *smallest* eigenvalues of the symmetric normalized graph Laplacian matrix (see [5]):

$$\Delta_n \stackrel{\text{def}}{=} \mathbf{I} - \mathbf{D}_{\boldsymbol{G}}^{-1/2}\boldsymbol{G}\mathbf{D}_{\boldsymbol{G}}^{-1/2}. \tag{6}$$

Proof of Proposition 1. Notice that

$$\mathcal{A}_\mathbf{n} = -\Delta_n + (a+1)\mathbf{I}_n, \tag{29}$$

from where it follows that the eigenvectors of $\mathcal{A}_\mathbf{n}$ associated to its r largest eigenvalues coincide with the eigenvectors of Δ_n associated to its r smallest eigenvalues. Since $\mathcal{A}_\mathbf{n}$ is obviously symmetric, it remains to show that its eigenvalues are non-negative. In turn, from the definition of $\mathcal{A}_\mathbf{n}$ in Equation 5 and the fact that $a > 1$, it is sufficient to argue that all eigenvalues of $\mathbf{D}_{\boldsymbol{G}}^{-1/2}\boldsymbol{G}\mathbf{D}_{\boldsymbol{G}}^{-1/2}$ have absolute value less than or equal to 1. This, however, follows from the following two facts: 1) the matrix $\mathbf{D}_{\boldsymbol{G}}^{-1/2}\boldsymbol{G}\mathbf{D}_{\boldsymbol{G}}^{-1/2}$ is similar to the matrix $\mathbf{D}_{\boldsymbol{G}}^{-1}\boldsymbol{G}$, given that

$$\mathbf{D}_{\boldsymbol{G}}^{1/2}(\mathbf{D}_{\boldsymbol{G}}^{-1}\boldsymbol{G})\mathbf{D}_{\boldsymbol{G}}^{-1/2} = \mathbf{D}_{\boldsymbol{G}}^{-1/2}\boldsymbol{G}\mathbf{D}_{\boldsymbol{G}}^{-1/2},$$

implying that $\mathbf{D}_{\boldsymbol{G}}^{-1/2}\boldsymbol{G}\mathbf{D}_{\boldsymbol{G}}^{-1/2}$ and $\mathbf{D}_{\boldsymbol{G}}^{-1}\boldsymbol{G}$ have the same eigenvalues; and 2) all the eigenvalues of $\mathbf{D}_{\boldsymbol{G}}^{-1}\boldsymbol{G}$ have norm less than one, since $\mathbf{D}_{\boldsymbol{G}}^{-1}\boldsymbol{G}$ is a transition probability matrix. □

Remark 11. *While one could set $\mathcal{A}_\mathbf{n}$ to be Δ_n itself (since Δ_n is PSD), solving the resulting problem 7 would return the eigenvectors of Δ_n with the largest eigenvalues, which would not constitute a desirable output for data analysis, as the tail of the spectrum of Δ_n has little geometric information about the data set $\mathcal{X}_n$. It is interesting that we can still recover the relevant part of the spectrum of Δ_n indirectly, by studying the spectrum of the matrix $\mathcal{A}_\mathbf{n}$ that we use in this paper. Finally, it is worth mentioning that we add the term $a\mathbf{I}_n$ in the definition of $\mathcal{A}_\mathbf{n}$ in 5 to guarantee that $\mathcal{A}_\mathbf{n}$ is always PSD, in this way simplifying the statements and proofs of our main results.*

D Auxiliary Approximation Results

D.1 Neural Network Approximation of Lipschitz Functions on Manifolds

[63] shows that Lipschitz functions f defined over an m-dimensional smooth manifold $\mathcal{M}$ embedded in $\mathbb{R}^d$ can be approximated with a ReLU neural network

with a number of neurons that doesn't grow exponentially with the ambient space dimension d. Precisely:

Theorem 3 (Theorem 1 in [63]). *Let $f : \mathcal{M} \to \mathbb{R}$ be a Lipschitz function with Lipschitz constant less than K. Given any $\delta \in (0,1)$, there are κ, L, p, N satisfying:*

1. $L \leq C_K \left(\log \frac{1}{\delta} + \log d\right)$, *and* $p \leq C_K \left(\delta^{-m} + d\right)$,
2. $N \leq C_K \left(\delta^{-m} \log \frac{1}{\delta} + d \log \frac{1}{\delta} + d \log d\right)$, *and* $\kappa \leq C_K$,

such that there is a neural network $f_\theta \in \mathcal{F}(1, \kappa, L, p, N)$ (as defined in Equation 14), for which

$$\|f_\theta - f\|_{L^\infty(\mathcal{M})} \leq \delta.$$

In the above, C_K is a constant that depends on K and on the geometry of the manifold $\mathcal{M}$.

We remark that this result is not surprising because a Riemannian manifold locally behaves like a low dimensional Euclidean space. In this paper we utilize the results from [63] due to the fact that in their estimates the ambient space dimension d does not appear as an exponent.

E Proofs of Theorem 3.1 and Corollary 3

We begin by proving an important lemma.

Lemma 1. *Let $u : \mathcal{X}_n \to \mathbb{R}$ be a function satisfying*

$$|u(x) - u(\tilde{x})| \leq L(d_{\mathcal{M}}(x, \tilde{x}) + \tau^2), \quad \forall x, \tilde{x} \in \mathcal{X}_n \tag{15}$$

for some L and $\tau > 0$. Then there exists a $3L$-Lipschitz function $\tilde{g} : \mathcal{M} \to \mathbb{R}$ such that

$$\|u - \tilde{g}\|_{L^\infty(\mathcal{X}_n)} \leq 5L\tau^2. \tag{16}$$

Proof. We start by constructing a subset $\mathcal{X}_n'$ of $\mathcal{X}_n$ satisfying the following properties:

1. Any two points $x, \tilde{x} \in \mathcal{X}_n'$ (different from each other) satisfy $d_{\mathcal{M}}(x, \tilde{x}) \geq \frac{1}{2}\tau^2$.
2. For any $x \in \mathcal{X}_n$ there exists $\tilde{x} \in \mathcal{X}_n'$ such that $d_{\mathcal{M}}(x, \tilde{x}) \leq \tau^2$.

The set $\mathcal{X}_n'$ can be constructed inductively, as we explain next. First, we enumerate the points in $\mathcal{X}_n$ as $x_1, \ldots, x_n$. After having decided whether to include or not in $\mathcal{X}_n'$ the first s points in the list, we decide to include x_{s+1} as follows: if the ball of radius $\tau^2/2$ centered at x_{s+1} intersects any of the balls of radius $\tau^2/2$ centered around the points already included in $\mathcal{X}_n'$, then we do not include x_{s+1} in $\mathcal{X}_n'$, otherwise we include it. It is clear from this construction that the

resulting set $\mathcal{X}_n'$ satisfies the desired properties (property 2 follows from the triangle inequality).

Now, notice that the function $u : \mathcal{X}_n' \to \mathbb{R}$ (i.e., u restricted to $\mathcal{X}_n'$) is $3L$-Lipschitz, since

$$|u(x) - u(\tilde{x})| \leq L(d_{\mathcal{M}}(x, \tilde{x}) + \tau^2) \leq 3L d_{\mathcal{M}}(x, \tilde{x})$$

for any pair of points $x, \tilde{x}$ in $\mathcal{X}_n'$. Using McShane-Whitney theorem we can extend the function $u : \mathcal{X}_n' \to \mathbb{R}$ to a $3L$-Lipschitz function $\tilde{g} : \mathcal{M} \to \mathbb{R}$. It remains to prove Equation 16. To see this, let $x \in \mathcal{X}_n$ and let $\tilde{x} \in \mathcal{X}_n$ be as in property 2 of $\mathcal{X}_n'$. Then

$$\begin{aligned} |u(x) - \tilde{g}(x)| &\leq |u(x) - u(\tilde{x})| + |u(\tilde{x}) - g(x)| \\ &= |u(x) - u(\tilde{x})| + |g(\tilde{x}) - g(x)| \\ &\leq L(d_{\mathcal{M}}(x, \tilde{x}) + \tau^2) + 3L d_{\mathcal{M}}(x, \tilde{x}) \\ &\leq 5L\tau^2. \end{aligned}$$

This completes the proof. □

We are ready to prove Theorem 3.1, which here we restate for convenience.

Theorem 3.1 (Spectral approximation of normalized Laplacians with neural networks)**.** *Let $r \in \mathbb{N}$ be fixed. Under Assumptions 1, there are constants c, C that depend on $\mathcal{M}, \rho$, and the embedding dimension r, such that, with probability at least*

$$1 - C\tau^{-6m} \exp\left(-cn\tau^{m+4}\right),$$

for every $\varepsilon \in (0, 1)$ there are κ, L, p, N and a ReLU neural network $f_\theta \in \mathcal{F}(r, \kappa, L, p, N)$ (defined in Equation 14), such that:

1. *$\sqrt{n}\|\mathbf{Y}_\theta - \mathbf{Y}^*\|_{\infty,\infty} \leq C(\varepsilon + \tau^2)$, and thus also $\|\mathbf{Y}_\theta - \mathbf{Y}^*\|_{\mathrm{F}} \leq C\sqrt{r}(\varepsilon + \tau^2)$.*
2. *The depth of the network, L, satisfies: $L \leq C\left(\log \frac{1}{\varepsilon} + \log d\right)$, and its width, p, satisfies $p \leq C\left(\varepsilon^{-m} + d\right)$.*
3. *The number of neurons of the network, N, satisfies: $N \leq Cr\left(\varepsilon^{-m} \log \frac{1}{\varepsilon} + d \log \frac{1}{\varepsilon} + d \log d\right)$, and the range of weights, κ, satisfies $\kappa \leq \frac{C}{n^{1/(2L)}}$.*

Proof. Let $s \leq r$. As in the discussion we let u_s be a $\|\cdot\|_{L^2(\mathcal{X}_n)}$-normalized eigenvector of Δ_n corresponding to its s-th smallest eigenvalue. Thanks to Corollary 1, we know that, with very high probability, the function $u_s : \mathcal{X}_n \to \mathbb{R}$ satisfies

$$|u_s(x_i) - u_s(x_j)| \leq L_s(d_{\mathcal{M}}(x_i, x_j) + \tau^2), \quad \forall x_i, x_j \in \mathcal{X}_n, \tag{30}$$

for some deterministic constant L_s. Using the fact that $\sqrt{\sigma_s(\mathcal{A}_\mathbf{n})}$ is an order one quantity (according to Remark 2) in combination with Lemma 1, we deduce the existence of a CL_s-Lipschitz function $g_s : \mathcal{M} \to \mathbb{R}$ satisfying

$$\|g_s - \sqrt{\sigma_s(\mathcal{A}_\mathbf{n})} u_s\|_{L^\infty(\mathcal{X}_n)} \leq 5CL_s\tau^2. \tag{31}$$

In turn, Theorem 3 implies the existence of parameters κ, L, p, N as in the statement of the theorem and a (scalar-valued) neural network $f_{\tilde{\theta}}$ in the class $\mathcal{F}(1, \kappa, L, p, N)$ such that

$$\|f_{\tilde{\theta}}(x) - g_s(x)\|_{L^\infty(\mathcal{M})} \leq \varepsilon. \tag{32}$$

Using the fact that the ReLU is a homogeneous function of degree one, we can deduce that

$$\frac{1}{\sqrt{n}} f_{\tilde{\theta}} = f_\theta,$$

where $\theta \stackrel{\text{def}}{=} \frac{1}{n^{1/(2L)}}\tilde{\theta}$ and thus $f_\theta \in \mathcal{F}(1, \frac{\kappa}{n^{1/(2L)}}, L, p, N)$. It follows that the neural network f_θ satisfies

$$\sqrt{n}\|f_\theta - \frac{1}{\sqrt{n}} g_s\|_{L^\infty(\mathcal{M})} \leq \varepsilon,$$

and also, thanks to Equation 31,

$$\sqrt{n}\left\| f_\theta - \sqrt{\frac{\sigma_s(\mathcal{A}_\mathbf{n})}{n}} u_s \right\|_{L^\infty(\mathcal{X}_n)} \leq (5CL_s + 1)(\varepsilon + \tau^2).$$

Stacking the scalar neural networks constructed above to approximate each of the functions u_s for $s = 1, \dots r$, and using Equation 10, we obtain the desired vector valued neural network approximating $\mathbf{Y}^*$.

□

Corollary 2. *For f_θ as in Theorem 3.1 we have*

$$\sqrt{n}\|\mathbf{Y}_\theta \mathbf{Y}_\theta^\top - \mathbf{Y}^* \mathbf{Y}^{*\top}\|_{\infty,\infty} \leq C_r(\varepsilon + \tau^2), \tag{17}$$

and thus also

$$\|\mathbf{Y}_\theta \mathbf{Y}_\theta^\top - \mathbf{Y}^* \mathbf{Y}^{*\top}\|_{\mathrm{F}} \leq \sqrt{r} C_r(\varepsilon + \tau^2),$$

for some deterministic constant C_r.

Proof.

$$\begin{aligned}
\sqrt{n}\|\mathbf{Y}_\theta \mathbf{Y}_\theta^\top - \mathbf{Y}^* \mathbf{Y}^{*\top}\|_{\infty,\infty} &= \sqrt{n}\|\mathbf{Y}_\theta \left(\mathbf{Y}_\theta^\top - \mathbf{Y}^{*\top}\right) + \left(\mathbf{Y}_\theta - \mathbf{Y}^*\right)\mathbf{Y}^{*\top}\|_{\infty,\infty} \\
&\leq \sqrt{n}\|\mathbf{Y}_\theta \left(\mathbf{Y}_\theta^\top - \mathbf{Y}^{*\top}\right)\|_{\infty,\infty} + \sqrt{n}\|\left(\mathbf{Y}_\theta - \mathbf{Y}^*\right)\mathbf{Y}^{*\top}\|_{\infty,\infty} \\
&\leq \sqrt{nr}\|\mathbf{Y}_\theta\|_{\mathrm{F}}\|\mathbf{Y}_\theta^\top - \mathbf{Y}^{*\top}\|_{\infty,\infty} + \sqrt{nr}\|\mathbf{Y}_\theta - \mathbf{Y}^*\|_{\infty,\infty}\|\mathbf{Y}^{*\top}\|_{\mathrm{F}} \\
&\leq \sqrt{r}\big(C_r(\varepsilon + \tau^2) + 2\|\mathbf{Y}^*\|_{\mathrm{F}}\big) C_r(\varepsilon + \tau^2) \\
&\leq C_r(\varepsilon + \tau^2),
\end{aligned}$$

where the second to last inequality follows from our estimate for $\sqrt{n}\|\mathbf{Y}_\theta - \mathbf{Y}^*\|_{\infty,\infty} \leq C_r(\varepsilon + \tau^2)$ in Theorem 3.1, and the last inequality follows from the fact that $\|\mathbf{Y}^*\|_{\mathrm{F}}^2 = \sum_{s=1}^r \sigma_s(\mathcal{A}_\mathbf{n}) = \mathcal{O}(r)$. □

E.1 Eigenfunction approximation

The neural network f_θ constructed in the proof of Theorem 3.1 can be used to approximate eigenfunctions of Δ_ρ. We restate Corollary 3 for the convenience of the reader.

Corollary 3. *Under the same setting, notation, and assumptions as in Theorem 3.1, the neural network $f_\theta : \mathbb{R}^d \to \mathbb{R}^r$ can be chosen to satisfy*

$$\left\| \sqrt{\frac{n}{1+a}} f_\theta^i - f_i \right\|_{L^\infty(\mathcal{M})} \leq C(\varepsilon + \tau), \quad \forall i = 1, \ldots, r.$$

In the above, $f_\theta^1, \ldots, f_\theta^r$ are the coordinate functions of the vector-valued neural network f_θ, and the functions $f_1, \ldots, f_r$ are the normalized eigenfunctions of the Laplace-Beltrami operator Δ_ρ that are associated to Δ_ρ's r smallest eigenvalues.

Proof. Let $g_s : \mathcal{M} \to \mathbb{R}$ be the Lipschitz function appearing in Equation 31 and recall that the scalar neural network f_θ constructed in the proof of Theorem 3.1 satisfies

$$\sqrt{n} \| f_\theta - \frac{1}{\sqrt{n}} g_s \|_{L^\infty(\mathcal{M})} \leq \delta. \tag{33}$$

It can be shown that except on an event with probability less than $n \exp(-n\tau^m)$, for any $x \in \mathcal{M}$, there exists $x_i \in \mathcal{X}_n$ such that $d_{\mathcal{M}}(x_i, x) \leq \tau$. From the triangle inequality, it thus follows that

$$\begin{aligned}
|f_s(x) - \sqrt{n/(1+a)} f_\theta(x)| \leq & |f_s(x) - f_s(x_i)| + |f_s(x_i) - u_s(x_i)| \\
& + |u_s(x_i) - \frac{1}{\sqrt{\sigma_s(\mathcal{A}_{\mathbf{n}})}} g_s(x_i)| + |\frac{1}{\sqrt{\sigma_s(\mathcal{A}_{\mathbf{n}})}} g_s(x_i) - \frac{1}{\sqrt{1+a}} g_s(x_i)| \\
& + |\frac{1}{\sqrt{1+a}} g_s(x_i) - \frac{1}{\sqrt{1+a}} g_s(x)| \\
& + |\frac{1}{\sqrt{1+a}} g_s(x) - \sqrt{\frac{n}{1+a}} f_\theta(x)| \\
\leq & C_s(\varepsilon + \tau),
\end{aligned} \tag{34}$$

where we have used the Lipschitz continuity of f_s and g_s, Theorem 2, Remark 2, and Equation 33.

□

F Proof of Theorem 3.2

Recall that that $f_{\theta^*} \in \arg\min_{f_\theta \in \mathcal{F}(r,\kappa,L,p,N)} \| \mathbf{Y}_\theta \mathbf{Y}_\theta^\top - \mathcal{A}_{\mathbf{n}} \|_{\mathrm{F}}^2$. We start our proof with a lemma from linear algebra.

Lemma 2. *For any $\mathbf{Y} \in \mathbb{R}^{n\times r}$ we have*

$$\|\mathbf{Y}\mathbf{Y}^\top - \mathcal{A}_\mathbf{n}\|_\mathrm{F}^2 - \|\mathbf{Y}^*\mathbf{Y}^{*\top} - \mathcal{A}_\mathbf{n}\|_\mathrm{F}^2 \leq \|\mathbf{Y}\mathbf{Y}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top}\|_\mathrm{F}^2.$$

Proof. A straightforward computation reveals that

$$\begin{aligned}
&\|\mathbf{Y}\mathbf{Y}^\top - \mathcal{A}_\mathbf{n}\|_\mathrm{F}^2 - \|\mathbf{Y}^*\mathbf{Y}^{*\top} - \mathcal{A}_\mathbf{n}\|_\mathrm{F}^2 \\
&= \|(\mathbf{Y}\mathbf{Y}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top}) + (\mathbf{Y}^*\mathbf{Y}^{*\top} - \mathcal{A}_\mathbf{n})\|_\mathrm{F}^2 - \|\mathbf{Y}^*\mathbf{Y}^{*\top} - \mathcal{A}_\mathbf{n}\|_\mathrm{F}^2 \\
&= \|\mathbf{Y}\mathbf{Y}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top}\|_\mathrm{F}^2 + 2\langle \mathbf{Y}\mathbf{Y}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top}, \mathbf{Y}^*\mathbf{Y}^{*\top} - \mathcal{A}_\mathbf{n}\rangle \qquad (35)\\
&= \|\mathbf{Y}\mathbf{Y}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top}\|_\mathrm{F}^2 + 2\langle \mathbf{Y}\mathbf{Y}^\top, \mathbf{Y}^*\mathbf{Y}^{*\top} - \mathcal{A}_\mathbf{n}\rangle \\
&\leq \|\mathbf{Y}\mathbf{Y}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top}\|_\mathrm{F}^2,
\end{aligned}$$

where the last inequality follows thanks to the fact that $\mathbf{Y}\mathbf{Y}^\top$ is positive semi-definite and the fact that $\mathbf{Y}^*\mathbf{Y}^{*\top} - \mathcal{A}_\mathbf{n}$ is negative semi-definite, as can be easily deduced from the form of $\mathbf{Y}^*$. □

Invoking Corollary 2 with $\varepsilon = \tilde{\varepsilon}\tau$ we immediately obtain the following approximation estimate.

Corollary 4. *With probability at least*

$$1 - C\tau^{-6m}\exp\left(-cn\tau^{m+4}\right),$$

for every $\tilde{\varepsilon} \in (0,1)$ there is $f_\theta \in \mathcal{F}(r,\kappa,L,p,N)$ with κ, L, p, N as specified in Theorem 3.2 such that

$$\|\mathbf{Y}_\theta\mathbf{Y}_\theta^\top - \mathbf{Y}^*\mathbf{Y}^{*\top}\|_\mathrm{F} \leq C_r\tau(\tilde{\varepsilon} + \tau). \qquad (36)$$

Corollary 5. *Let f_θ be as in Corollary 4. Then*

$$\|\mathbf{Y}_\theta\mathbf{Y}_\theta^\top - \mathcal{A}_\mathbf{n}\|_\mathrm{F}^2 - \|\mathbf{Y}^*\mathbf{Y}^{*\top} - \mathcal{A}_\mathbf{n}\|_\mathrm{F}^2 \leq C_r\tau^2(\tilde{\varepsilon} + \tau)^2.$$

Proof. Let θ be as in Corollary 4. Then

$$\|\mathbf{Y}_\theta\mathbf{Y}_\theta^\top - \mathcal{A}_\mathbf{n}\|_\mathrm{F}^2 - \|\mathbf{Y}^*\mathbf{Y}^{*\top} - \mathcal{A}_\mathbf{n}\|_\mathrm{F}^2 \leq \|\mathbf{Y}_\theta\mathbf{Y}_\theta^\top - \mathbf{Y}^*\mathbf{Y}^{*\top}\|_\mathrm{F}^2 \leq C_r^2\tau^2(\tilde{\varepsilon} + \tau)^2,$$

where the second to last inequality follows from Lemma 2. □

In what follows we will write the SVD (eigendecomposition) of $\mathcal{A}_\mathbf{n}$ as $\overline{\mathbf{U}}\mathbf{\Sigma}\overline{\mathbf{U}}^\top$. Using the fact that $\overline{\mathbf{U}}$ is invertible (since it is an orthogonal matrix), we can easily see that $\mathbf{Y}_{\theta^*}$ can be written as $\mathbf{Y}_{\theta^*} = \overline{\mathbf{U}}(\mathbf{E}^1 + \mathbf{E}^2)$ where $\mathbf{E}^1, \mathbf{E}^2 \in \mathbb{R}^{n\times r}$ are such that the i^{th}row $\mathbf{E}_i^1 = \mathbf{0}$ for $i \geq r+1$, and i^{th} row $\mathbf{E}_i^2 = \mathbf{0}$ for $i \leq r$. Indeed, it suffices to select $\mathbf{E}_1$ and $\mathbf{E}_2$ so as to have $\mathbf{E}^1 + \mathbf{E}^2 = \overline{\mathbf{U}}^{-1}\mathbf{Y}_{\theta^*}$. We thus have $(\mathbf{E}^2)^\top\mathbf{E}^1 = \mathbf{0}$.

In what follows we will make the following assumption.

Assumption 4. *τ and $\tilde{\varepsilon}$ in Corollary 4 satisfy the following condition:*

$$\tau^2 E < \sigma_r^2(\mathcal{A}_{\mathbf{n}}) - \sigma_{r+1}^2(\mathcal{A}_{\mathbf{n}}), \tag{37}$$

where $E \overset{def}{=} C_r(\tilde{\delta} + \tau)^2$.

Proof of Theorem 3.2. Due to the definition of θ^*, we have

$$\|\mathbf{Y}^*\mathbf{Y}^{*\top} - \mathcal{A}_{\mathbf{n}}\|_{\mathrm{F}}^2 \leq \|\mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top - \mathcal{A}_{\mathbf{n}}\|_{\mathrm{F}}^2 \leq \|\mathbf{Y}_\theta\mathbf{Y}_\theta^\top - \mathcal{A}_{\mathbf{n}}\|_{\mathrm{F}}^2. \tag{38}$$

Also,

$$\begin{aligned}
0 &\geq \|\mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top - \mathcal{A}_{\mathbf{n}}\|_{\mathrm{F}}^2 - \|\mathbf{Y}_\theta\mathbf{Y}_\theta^\top - \mathcal{A}_{\mathbf{n}}\|_{\mathrm{F}}^2 \\
&= \|(\mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top}) + (\mathbf{Y}^*\mathbf{Y}^{*\top} - \mathcal{A}_{\mathbf{n}})\|_{\mathrm{F}}^2 - \|\mathbf{Y}_\theta\mathbf{Y}_\theta^\top - \mathcal{A}_{\mathbf{n}}\|_{\mathrm{F}}^2 \\
&= \|\mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top}\|_{\mathrm{F}}^2 + \|\mathbf{Y}^*\mathbf{Y}^{*\top} - \mathcal{A}_{\mathbf{n}}\|_{\mathrm{F}}^2 + 2\langle \mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top}, \mathbf{Y}^*\mathbf{Y}^{*\top} - \mathcal{A}_{\mathbf{n}}\rangle - \|\mathbf{Y}_\theta\mathbf{Y}_\theta^\top - \mathcal{A}_{\mathbf{n}}\|_{\mathrm{F}}^2 \\
&= \|\mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top}\|_{\mathrm{F}}^2 + \|\mathbf{Y}^*\mathbf{Y}^{*\top} - \mathcal{A}_{\mathbf{n}}\|_{\mathrm{F}}^2 + 2\langle \mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top, \mathbf{Y}^*\mathbf{Y}^{*\top} - \mathcal{A}_{\mathbf{n}}\rangle - \|\mathbf{Y}_\theta\mathbf{Y}_\theta^\top - \mathcal{A}_{\mathbf{n}}\|_{\mathrm{F}}^2
\end{aligned} \tag{39}$$

where the third equality follows from the fact that $\langle \mathbf{Y}^*\mathbf{Y}^{*\top}, \mathbf{Y}^*\mathbf{Y}^{*\top} - \mathcal{A}_{\mathbf{n}}\rangle = 0$. Notice that

$$\|\mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top}\|_{\mathrm{F}}^2 + 2\langle \mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top, \mathbf{Y}^*\mathbf{Y}^{*\top} - \mathcal{A}_{\mathbf{n}}\rangle = \|\mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top\|_{\mathrm{F}}^2 + \|\mathbf{Y}^*\mathbf{Y}^{*\top}\|_{\mathrm{F}}^2 - 2\langle \mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top, \mathcal{A}_{\mathbf{n}}\rangle \tag{40}$$

By combining Equation 39, Lemma 5 and Equation 40, we have

$$\|\mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top\|_{\mathrm{F}}^2 + \|\mathbf{Y}^*\mathbf{Y}^{*\top}\|_{\mathrm{F}}^2 - 2\langle \mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top, \mathcal{A}_{\mathbf{n}}\rangle \leq \tau^2 E \tag{41}$$

From $(\mathbf{E}^1)^\top\mathbf{E}^2 = \mathbf{0}$ and $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$, we have

$$\begin{aligned}
\langle \mathbf{E}^1(\mathbf{E}^1)^\top, \mathbf{E}^2(\mathbf{E}^2)^\top\rangle &= 0 \\
\langle \mathbf{E}^1(\mathbf{E}^2)^\top, \mathbf{E}^2(\mathbf{E}^2)^\top\rangle &= 0 \\
\langle \mathbf{E}^1(\mathbf{E}^2)^\top, \mathbf{E}^1(\mathbf{E}^1)^\top\rangle &= 0 \\
\langle \mathbf{E}^2(\mathbf{E}^1)^\top, \mathbf{E}^1(\mathbf{E}^1)^\top\rangle &= 0 \\
\langle \mathbf{E}^2(\mathbf{E}^1)^\top, \mathbf{E}^2(\mathbf{E}^2)^\top\rangle &= 0
\end{aligned} \tag{42}$$

Let $\mathbf{\Sigma}^1$ be the diagonal matrix such that $(\mathbf{\Sigma}^1)_{ii} = \mathbf{\Sigma}_{ii}$ for $i \leq r$, and $(\mathbf{\Sigma}^1)_{ii} = 0$ for $i > r$; let $\mathbf{\Sigma}^2$ be the diagonal matrix such that $(\mathbf{\Sigma}^2)_{ii} = 0$ for $i \leq r$, and $(\mathbf{\Sigma}^2)_{ii} = \mathbf{\Sigma}_{ii}$ for $i > r$. By plugging the decomposition of $\mathbf{Y}_{\theta^*}$ in Equation 41,

we deduce

$$
\begin{aligned}
\varepsilon^2 E \geq & \|\mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top\|_{\mathrm{F}}^2 + \|\mathbf{Y}^*\mathbf{Y}^{*\top}\|_{\mathrm{F}}^2 - 2\langle \mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top, \mathcal{A}_{\mathbf{n}}\rangle \\
= & \|\overline{\mathbf{U}}(\mathbf{E}^1+\mathbf{E}^2)(\mathbf{E}^1+\mathbf{E}^2)^\top\overline{\mathbf{U}}^\top\|_{\mathrm{F}}^2 + \|\mathbf{Y}^*\mathbf{Y}^{*\top}\|_{\mathrm{F}}^2 - 2\langle \overline{\mathbf{U}}(\mathbf{E}^1+\mathbf{E}^2)(\mathbf{E}^1+\mathbf{E}^2)^\top\overline{\mathbf{U}}^\top, \mathcal{A}_{\mathbf{n}}\rangle \\
= & \|(\mathbf{E}^1+\mathbf{E}^2)(\mathbf{E}^1+\mathbf{E}^2)^\top\|_{\mathrm{F}}^2 + \|\mathbf{Y}^*\mathbf{Y}^{*\top}\|_{\mathrm{F}}^2 - 2\langle (\mathbf{E}^1+\mathbf{E}^2)(\mathbf{E}^1+\mathbf{E}^2)^\top, \mathbf{\Sigma}\rangle \\
\overset{\text{Equation 42}}{=} & \|\mathbf{E}^1(\mathbf{E}^1)^\top\|_{\mathrm{F}}^2 + \|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}}^2 + 2\|\mathbf{E}^2(\mathbf{E}^1)^\top\|_{\mathrm{F}}^2 + 2\langle (\mathbf{E}^1)^\top\mathbf{E}^1, (\mathbf{E}^2)^\top\mathbf{E}^2\rangle \\
& + \|\mathbf{Y}^*\mathbf{Y}^{*\top}\|_{\mathrm{F}}^2 - 2\langle (\mathbf{E}^1+\mathbf{E}^2)(\mathbf{E}^1+\mathbf{E}^2)^\top, \mathbf{\Sigma}\rangle \\
\overset{(\mathbf{E}^1)^\top\mathbf{\Sigma}\mathbf{E}^2=\mathbf{0}}{=} & \|\mathbf{E}^1(\mathbf{E}^1)^\top\|_{\mathrm{F}}^2 + \|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}}^2 + 2\|\mathbf{E}^2(\mathbf{E}^1)^\top\|_{\mathrm{F}}^2 + 2\langle (\mathbf{E}^1)^\top\mathbf{E}^1, (\mathbf{E}^2)^\top\mathbf{E}^2\rangle \\
& + \|\mathbf{Y}^*\mathbf{Y}^{*\top}\|_{\mathrm{F}}^2 - 2\langle \mathbf{E}^1(\mathbf{E}^1)^\top + \mathbf{E}^2(\mathbf{E}^2)^\top, \mathbf{\Sigma}\rangle \\
= & \|\mathbf{E}^1(\mathbf{E}^1)^\top\|_{\mathrm{F}}^2 + \|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}}^2 + 2\|\mathbf{E}^2(\mathbf{E}^1)^\top\|_{\mathrm{F}}^2 + 2\langle (\mathbf{E}^1)^\top\mathbf{E}^1, (\mathbf{E}^2)^\top\mathbf{E}^2\rangle \\
& + \|\mathbf{\Sigma}^1\|_{\mathrm{F}}^2 - 2\langle \mathbf{E}^1(\mathbf{E}^1)^\top, \mathbf{\Sigma}^1\rangle - 2\langle \mathbf{E}^2(\mathbf{E}^2)^\top, \mathbf{\Sigma}^2\rangle \\
= & \|\mathbf{E}^1(\mathbf{E}^1)^\top - \mathbf{\Sigma}^1\|_{\mathrm{F}}^2 + \|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}}^2 + 2\|\mathbf{E}^2(\mathbf{E}^1)^\top\|_{\mathrm{F}}^2 + 2\langle (\mathbf{E}^1)^\top\mathbf{E}^1, (\mathbf{E}^2)^\top\mathbf{E}^2\rangle \\
& - 2\langle \mathbf{E}^2(\mathbf{E}^2)^\top, \mathbf{\Sigma}^2\rangle \\
\geq & \|\mathbf{E}^1(\mathbf{E}^1)^\top - \mathbf{\Sigma}^1\|_{\mathrm{F}}^2 + \|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}}^2 + 2\|\mathbf{E}^2(\mathbf{E}^1)^\top\|_{\mathrm{F}}^2 + 2\langle (\mathbf{E}^1)^\top\mathbf{E}^1, (\mathbf{E}^2)^\top\mathbf{E}^2\rangle \\
& - 2\|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}} \cdot \sigma_{r+1}(\mathcal{A}_{\mathbf{n}}) \\
\geq & \|\mathbf{E}^1(\mathbf{E}^1)^\top - \mathbf{\Sigma}^1\|_{\mathrm{F}}^2 + \|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}}^2 + (2\|\mathbf{E}^2\|_{\mathrm{F}}^2 + 2\|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}}) \cdot \sigma_r^2(\mathbf{E}^1) \\
& - 2\|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}} \cdot \sigma_{r+1}(\mathcal{A}_{\mathbf{n}}).
\end{aligned} \tag{43}
$$

On the other hand, we have

$$
\begin{aligned}
\|\mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top}\|_{\mathrm{F}}^2 = & \|\mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top\|_{\mathrm{F}}^2 + \|\mathbf{Y}^*\mathbf{Y}^{*\top}\|_{\mathrm{F}}^2 - 2\langle \mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top, \mathcal{A}_{\mathbf{n}}\rangle + 2\langle \mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top, \mathcal{A}_{\mathbf{n}} - \mathbf{Y}^*\mathbf{Y}^{*\top}\rangle \\
= & \|\mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top\|_{\mathrm{F}}^2 + \|\mathbf{Y}^*\mathbf{Y}^{*\top}\|_{\mathrm{F}}^2 - 2\langle \mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top, \mathcal{A}_{\mathbf{n}}\rangle + 2\langle \mathbf{E}^2(\mathbf{E}^2)^\top, \mathbf{\Sigma}^2\rangle \\
\leq & \tau^2 E + 2\|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}} \cdot \sigma_{r+1}(\mathcal{A}_{\mathbf{n}}).
\end{aligned} \tag{44}
$$

It remains to show that $\|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}}$ can be controlled by a term of the form $C\varepsilon^2 E$. We split the following discussion into two cases. First, we assume that $\sigma_r^2(\mathbf{E}^1)$ is large compared with $\sigma_{r+1}(\mathcal{A}_{\mathbf{n}})$. In this first case $\|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}}$ can be guaranteed to be small according to Equation 43. Second, when $\sigma_r^2(\mathbf{E}^1)$ is small, we'll show that $\|\mathbf{E}^1(\mathbf{E}^1)^\top - \mathbf{\Sigma}^1\|_{\mathrm{F}}^2$ is large, which will contradict Equation 43.

Case 1: If $\sigma_r^2(\mathbf{E}^1) \geq \frac{2}{3}\sigma_{r+1}(\mathcal{A}_{\mathbf{n}})$.

We have $3\|\mathbf{E}^2\|_{\mathrm{F}}^2 \cdot \sigma_r^2(\mathbf{E}^1) - 2\|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}} \cdot \sigma_{r+1}(\mathcal{A}_{\mathbf{n}}) \geq 0$. Then, from Equation 43 and the fact that $\|AB\|_{\mathrm{F}} \leq \|A\|_{\mathrm{F}} \cdot \|B\|_{\mathrm{F}}$, we have

$$
\|\mathbf{E}^1(\mathbf{E}^1)^\top - \mathbf{\Sigma}^1\|_{\mathrm{F}}^2 + \|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}}^2 + 2\|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}} \cdot \sigma_{r+1}(\mathcal{A}_{\mathbf{n}}) \leq \tau^2 E. \tag{45}
$$

This immediately implies

$$
\|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}} \leq \frac{\tau^2 E}{\sigma_{r+1}(\mathcal{A}_{\mathbf{n}})}. \tag{46}
$$

Combining Equation 46 and Equation 44, we obtain

$$\|\mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top}\|_{\mathrm{F}}^2 \leq \tau^2 E + \|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}} \cdot \sigma_{r+1}(\mathcal{A}_\mathbf{n}) \leq 2\tau^2 E. \tag{47}$$

Case 2: If $0 \leq \sigma_r^2(\mathbf{E}^1) < \frac{2}{3}\sigma_{r+1}(\mathcal{A}_\mathbf{n})$**.**
Invoking Equation 43, we have

$$\begin{aligned}
\tau^2 E &\geq \|\mathbf{E}^1(\mathbf{E}^1)^\top - \mathbf{\Sigma}^1\|_{\mathrm{F}}^2 + \|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}}^2 + (2\|\mathbf{E}^2\|_{\mathrm{F}}^2 + 2\|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}}) \cdot \sigma_r^2(\mathbf{E}^1) - 2\|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}} \cdot \sigma_{r+1}(\mathcal{A}_\mathbf{n}) \\
&\geq (\sigma_r^2(\mathbf{E}^1) - \sigma_r(\mathcal{A}_\mathbf{n}))^2 + \|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}}^2 + 4\|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}} \cdot \sigma_r^2(\mathbf{E}^1) - 2\|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}} \cdot \sigma_{r+1}(\mathcal{A}_\mathbf{n}) \\
&= (\sigma_r^2(\mathbf{E}^1) - \sigma_r(\mathcal{A}_\mathbf{n}))^2 + \|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}}^2 - 2\|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}} \cdot (\sigma_{r+1}(\mathcal{A}_\mathbf{n}) - 2\sigma_r^2(\mathbf{E}^1)) \\
&= (\sigma_r^2(\mathbf{E}^1) - \sigma_r(\mathcal{A}_\mathbf{n}))^2 + \big(\|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}} - (\sigma_{r+1}(\mathcal{A}_\mathbf{n}) - 2\sigma_r^2(\mathbf{E}^1))\big)^2 - (\sigma_{r+1}(\mathcal{A}_\mathbf{n}) - 2\sigma_r^2(\mathbf{E}^1))^2 \\
&\geq (\sigma_r^2(\mathbf{E}^1) - \sigma_r(\mathcal{A}_\mathbf{n}))^2 - (\sigma_{r+1}(\mathcal{A}_\mathbf{n}) - 2\sigma_r^2(\mathbf{E}^1))^2,
\end{aligned} \tag{48}$$

where the second inequality follows from Weyl's inequality [71].

It is straightforward to check that $(\sigma_r^2(\mathbf{E}^1) - \sigma_r(\mathcal{A}_\mathbf{n}))^2 - (\sigma_{r+1}(\mathcal{A}_\mathbf{n}) - 2\sigma_r^2(\mathbf{E}^1))^2$ is a decreasing function with respect to $\sigma_r^2(\mathbf{E}^1)$ in the range $0 \leq \sigma_r^2(\mathbf{E}^1) < \frac{2}{3}\sigma_{r+1}(\mathcal{A}_\mathbf{n})$. The smallest value of $(\sigma_r^2(\mathbf{E}^1) - \sigma_r(\mathcal{A}_\mathbf{n}))^2 - (\sigma_{r+1}(\mathcal{A}_\mathbf{n}) - 2\sigma_r^2(\mathbf{E}^1))^2$ in this range is thus larger than $\frac{1}{9}(\sigma_r^2(\mathcal{A}_\mathbf{n}) - \sigma_{r+1}^2(\mathcal{A}_\mathbf{n}))$. However, the resulting inequality contradicts Assumption 4. Case 2 is thus void.

By combining the aforementioned two cases, we conclude

$$\|\mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top}\|_{\mathrm{F}}^2 \leq 2E\tau^2. \tag{49}$$

By using Equation 53, we have

$$d^2([\mathbf{Y}_{\theta^*}], [\mathbf{Y}^*]) \leq \frac{1}{2(\sqrt{2}-1)\sigma_r^2(\mathbf{Y}^*)}\|\mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top}\|_{\mathrm{F}}^2 \leq \frac{\tau^2 E}{(\sqrt{2}-1)\sigma_r^2(\mathbf{Y}^*)}, \tag{50}$$

where $d([\mathbf{Y}_{\theta^*}], [\mathbf{Y}^*]) = \min_{\boldsymbol{O}\in\mathbb{O}_r}\|\mathbf{Y}_{\theta^*} - \mathbf{Y}^*\boldsymbol{O}\|_{\mathrm{F}}$. This completes the proof. □

G Ambient Optimization

This section contains the proof of the results from Section 4.

G.1 Setup from Main Text

Let us recall the quotient manifold that we are interested in. Let $\overline{\mathcal{N}}_{r+}^n$ be the space of $n \times r$ matrices with full column rank. To define the quotient manifold, we encode the invariance mapping, i.e., $\mathbf{Y} \to \mathbf{Y}\boldsymbol{O}$, by defining the equivalence classes $[\mathbf{Y}] = \{\mathbf{Y}\boldsymbol{O} : \boldsymbol{O} \in \mathbb{O}_r\}$. Since the invariance mapping is performed via the Lie group $\mathbb{O}_r$ smoothly, freely and properly, we have $\mathcal{N}_{r+}^n \overset{\text{def}}{=} \overline{\mathcal{N}}_{r+}^n/\mathbb{O}_r$

is a quotient manifold of $\overline{\mathcal{N}}_{r+}^n$ [58]. Moreover, we equip the tangent space $T_{\mathbf{Y}}\overline{\mathcal{N}}_{r_+}^n = \mathbb{R}^{n\times r}$ with the metric $\bar{g}_{\mathbf{Y}}\left(\eta_{\mathbf{Y}}, \theta_{\mathbf{Y}}\right) = \operatorname{tr}\left(\eta_{\mathbf{Y}}^\top \theta_{\mathbf{Y}}\right)$.

For convenience, we recall the following.

$$\begin{aligned} \overline{\operatorname{grad} H([\mathbf{Y}])} &= 2\left(\mathbf{Y}\mathbf{Y}^\top - \mathcal{A}_{\mathbf{n}}\right)\mathbf{Y}, \\ \overline{\operatorname{Hess} H([\mathbf{Y}])}\left[\theta_{\mathbf{Y}}, \theta_{\mathbf{Y}}\right] &= \left\|\mathbf{Y}\theta_{\mathbf{Y}}^\top + \theta_{\mathbf{Y}}\mathbf{Y}^\top\right\|_{\mathrm{F}}^2 + 2\left\langle \mathbf{Y}\mathbf{Y}^\top - \mathcal{A}_{\mathbf{n}}, \theta_{\mathbf{Y}}\theta_{\mathbf{Y}}^\top\right\rangle \end{aligned} \tag{51}$$

$$\begin{aligned} \mathcal{R}_1 &\overset{\text{def}}{=} \left\{\mathbf{Y} \in \mathbb{R}_*^{n\times r} \middle| d\left([\mathbf{Y}], [\mathbf{Y}^*]\right) \leqslant \mu\sigma_r\left(\mathbf{Y}^*\right)/\kappa^*\right\}, \\ \mathcal{R}_2 &\overset{\text{def}}{=} \left\{\mathbf{Y} \in \mathbb{R}_*^{n\times r} \middle| \begin{array}{l} d\left([\mathbf{Y}], [\mathbf{Y}^*]\right) > \mu\sigma_r\left(\mathbf{Y}^*\right)/\kappa^*, \|\overline{\operatorname{grad} H([\mathbf{Y}])}\|_{\mathrm{F}} \leqslant \alpha\mu\sigma_r^3\left(\mathbf{Y}^*\right)/\left(4\kappa^*\right) \\ \|\mathbf{Y}\| \leqslant \beta\left\|\mathbf{Y}^*\right\|, \left\|\mathbf{Y}\mathbf{Y}^\top\right\|_{\mathrm{F}} \leqslant \gamma\left\|\mathbf{Y}^*\mathbf{Y}^{*\top}\right\|_{\mathrm{F}} \end{array} \right\}, \\ \mathcal{R}_3' &\overset{\text{def}}{=} \left\{\mathbf{Y} \in \mathbb{R}_*^{n\times r} \middle| \|\overline{\operatorname{grad} H([\mathbf{Y}])}\|_{\mathrm{F}} > \alpha\mu\sigma_r^3\left(\mathbf{Y}^*\right)/\left(4\kappa^*\right), \|\mathbf{Y}\| \leqslant \beta\left\|\mathbf{Y}^*\right\|, \left\|\mathbf{Y}\mathbf{Y}^\top\right\|_{\mathrm{F}} \leqslant \gamma\left\|\mathbf{Y}^*\mathbf{Y}^{*\top}\right\|_{\mathrm{F}}\right\}, \\ \mathcal{R}_3'' &\overset{\text{def}}{=} \left\{\mathbf{Y} \in \mathbb{R}_*^{n\times r} \middle| \|\mathbf{Y}\| > \beta\|\mathbf{Y}^*\|, \|\mathbf{Y}\mathbf{Y}^\top\|_{\mathrm{F}} \leqslant \gamma\|\mathbf{Y}^*\mathbf{Y}^{*\top}\|_{\mathrm{F}}\right\}, \\ \mathcal{R}_3''' &\overset{\text{def}}{=} \left\{\mathbf{Y} \in \mathbb{R}_*^{n\times r} \middle| \|\mathbf{Y}\mathbf{Y}^\top\|_{\mathrm{F}} > \gamma\left\|\mathbf{Y}^*\mathbf{Y}^{*\top}\right|_{\mathrm{F}}\right\}, \end{aligned} \tag{52}$$

Remark 12. *To demonstrate strong geodesic convexity, the eigengap assumption is necessary as it prevents multiple global solutions. However, it is possible to relax this assumption and instead deduce a Polyak-Lojasiewicz condition, which would also imply a linear convergence rate for a first-order method.*

G.2 Some auxiliary inequalities

In this section, we collect results from prior work that will be useful for us. First, we provide the characterization of and results about the geodesic distance on $\mathcal{N}_{r+}^n$ from [72] and [56].

Lemma 3 (Lemma 2, [56])**.** *Let $\mathbf{Y}_1, \mathbf{Y}_2 \in \mathbb{R}_*^{n\times r}$, and $\mathbf{Q}_U\mathbf{\Sigma}\mathbf{Q}_V^\top$ be the SVD of $\mathbf{Y}_1^\top\mathbf{Y}_2$. Denote $\mathbf{Q}^* = \mathbf{Q}_V\mathbf{Q}_U^\top$. Then*

1. *$\mathbf{Y}_2\mathbf{Q}^* - \mathbf{Y}_1 \in \mathcal{H}_{\mathbf{Y}_1}\overline{\mathcal{N}}_{r+}^n, \mathbf{Q}^*$ is one of the best orthogonal matrices aligning $\mathbf{Y}_1$ and $\mathbf{Y}_2$, i.e., $\mathbf{Q}^* \in \arg\min_{\mathbf{Q}\in\mathbb{O}_r}\|\mathbf{Y}_2\mathbf{Q} - \mathbf{Y}_1\|_{\mathrm{F}}$ and the geodesic distance between $[\mathbf{Y}_1]$ and $[\mathbf{Y}_2]$ is $d\left([\mathbf{Y}_1], [\mathbf{Y}_2]\right) = \|\mathbf{Y}_2\mathbf{Q}^* - \mathbf{Y}_1\|_{\mathrm{F}}$;*
2. *if $\mathbf{Y}_1^\top\mathbf{Y}_2$ is nonsingular, then $\mathbf{Q}^*$ is unique and the Riemannian logarithm $\log_{[\mathbf{Y}_1]}[\mathbf{Y}_2]$ is uniquely defined and its horizontal lift at $\mathbf{Y}_1$ is given by $\overline{\log_{[\mathbf{Y}_1]}[\mathbf{Y}_2]} = \mathbf{Y}_2\mathbf{Q}^* - \mathbf{Y}_1$; moreover, the unique minimizing geodesic from $[\mathbf{Y}_1]$ to $[\mathbf{Y}_2]$ is $[\mathbf{Y}_1 + t\left(\mathbf{Y}_2\mathbf{Q}^* - \mathbf{Y}_1\right)]$ for $t \in [0, 1]$.*

Lemma 4 (Lemma 12 in [56])**.** *For any $\mathbf{Y}_1, \mathbf{Y}_2 \in \mathbb{R}_*^{n\times r}$, we have*

$$d^2\left([\mathbf{Y}_1], [\mathbf{Y}_2]\right) \leqslant \frac{1}{2(\sqrt{2}-1)\sigma_r^2\left(\mathbf{Y}_2\right)}\left\|\mathbf{Y}_1\mathbf{Y}_1^\top - \mathbf{Y}_2\mathbf{Y}_2^\top\right\|_{\mathrm{F}}^2 \tag{53}$$

and

$$\left\|\left(\mathbf{Y}_1-\mathbf{Y}_2\mathbf{Q}\right)\left(\mathbf{Y}_1-\mathbf{Y}_2\mathbf{Q}\right)^\top\right\|_{\mathrm{F}}^2 \leqslant 2\left\|\mathbf{Y}_1\mathbf{Y}_1^\top-\mathbf{Y}_2\mathbf{Y}_2^\top\right\|_{\mathrm{F}}^2, \tag{54}$$

where $\mathbf{Q}=\arg\min_{\mathbf{O}\in\mathbb{O}_r}\left\|\mathbf{Y}_1-\mathbf{Y}_2\mathbf{O}\right\|_{\mathrm{F}}$.

In addition, for any $\mathbf{Y}_1,\mathbf{Y}_2\in\mathbb{R}_*^{n\times r}$ *obeying* $d\left([\mathbf{Y}_1],[\mathbf{Y}_2]\right)\leqslant\frac{1}{3}\sigma_r\left(\mathbf{Y}_2\right)$*, we have*

$$\left\|\mathbf{Y}_1\mathbf{Y}_1^\top-\mathbf{Y}_2\mathbf{Y}_2^\top\right\|_{\mathrm{F}}\leqslant\frac{7}{3}\left\|\mathbf{Y}_2\right\|d\left([\mathbf{Y}_1],[\mathbf{Y}_2]\right) \tag{55}$$

Given any $\mathbf{Y}\in\mathbb{R}_*^{n\times r}$ and $x>0$, let $B_x([\mathbf{Y}])\overset{\text{def}}{=}\{[\mathbf{Y}_1]:d\left([\mathbf{Y}_1],[\mathbf{Y}]\right)<x\}$ be the geodesic ball centered at $[\mathbf{Y}]$ with radius x. For any Riemannian manifold, there exists a convex geodesic ball at every point (Chapter 3.4, [73]). The next result quantifies the convexity radius around a point $[\mathbf{Y}]$ in the manifold $\mathcal{N}_{r+}^n$.

Lemma 5 (Theorem 2, [56])**.** *Given any* $\mathbf{Y}\in\mathbb{R}_*^{n\times r}$*, the geodesic ball centered at* $[\mathbf{Y}]$ *with radius* $x\leqslant r_{\mathbf{Y}}\overset{def}{=}\sigma_r(\mathbf{Y})/3$ *is geodesically convex. In fact, for any two points* $[\mathbf{Y}_1],[\mathbf{Y}_2]\in B_x([\mathbf{Y}])$*, there is a unique shortest geodesic joining them, which is entirely contained in* $B_x([\mathbf{Y}])$*.*

Finally, we provide some useful inequalities.

Lemma 6 (Proposition 2 in [74])**.** *Let* $\mathbf{Y}\in\mathbb{R}_*^{n\times r}$*, and let* $\mathbf{X}=\mathbf{Y}\mathbf{Y}^\top$*. Then* $2\sigma_r^2(\mathbf{Y})\left\|\theta_{\mathbf{Y}}\right\|_{\mathrm{F}}^2\leqslant\left\|\mathbf{Y}\theta_{\mathbf{Y}}^\top+\theta_{\mathbf{Y}}\mathbf{Y}^\top\right\|_{\mathrm{F}}^2\leqslant 4\sigma_1^2(\mathbf{Y})\left\|\theta_{\mathbf{Y}}\right\|_{\mathrm{F}}^2$ *holds for all* $\theta_{\mathbf{Y}}\in\mathcal{H}_{\mathbf{Y}}\overline{\mathcal{M}}_{r+}^q$.

Lemma 7. *For* $\mathbf{A}\in\mathbb{R}^{m\times n}$*,* $\mathbf{B}\in\mathbb{R}^{n\times n}$ *where* $\mathbf{B}$ *is positive semi-definite, we have*

$$\|\mathbf{A}\|_{\mathrm{F}}\cdot\sigma_n(\mathbf{B})\leq\|\mathbf{A}\mathbf{B}\|_{\mathrm{F}}\leq\|\mathbf{A}\|_{\mathrm{F}}\cdot\sigma_1(\mathbf{B}) \tag{56}$$

Proof. When $m=1$, this statement is direct by the definition of the Frobenius norm. When $m>1$, we denote $\mathbf{A}_i$ to be the i^{th} row of $\mathbf{A}$, and then

$$\|\mathbf{A}\mathbf{B}\|_{\mathrm{F}}^2=\sum_{i=1}^m\|\mathbf{A}_i\mathbf{B}\|_{\mathrm{F}}^2\leq\sum_{i=1}^m\|\mathbf{A}_i\|_{\mathrm{F}}\cdot\sigma_1(\mathbf{B})=\|\mathbf{A}\|_{\mathrm{F}}\cdot\sigma_1(\mathbf{B})$$

Similarly,

$$\|\mathbf{A}\mathbf{B}\|_{\mathrm{F}}^2=\sum_{i=1}^m\|\mathbf{A}_i\mathbf{B}\|_{\mathrm{F}}^2\geq\sum_{i=1}^m\|\mathbf{A}_i\|_{\mathrm{F}}\cdot\sigma_n(\mathbf{B})=\|\mathbf{A}\|_{\mathrm{F}}\cdot\sigma_n(\mathbf{B})$$

□

G.3 Proof of Results

In this section, we provide the proofs for Theorems 4.1, 2.1, 4.2, and 4.3.

Theorem 4.1 (Local Geodesic Strong Convexity and Smoothness of Equation 8)**.** *Suppose* $0 \leqslant \mu \leqslant \kappa^*/3$*. Given that Assumption 2 holds, for any* $\mathbf{Y} \in \mathcal{R}_1$*,*

$$\sigma_{\min}(\overline{\operatorname{Hess} H([\mathbf{Y}])}) \geqslant \left(2\left(1-\mu/\kappa^*\right)^2 - (14/3)\mu\right)\sigma_r\left(\mathcal{A}_{\mathbf{n}}\right) - 2\sigma_{r+1}(\mathcal{A}_{\mathbf{n}}),$$
$$\sigma_{\max}(\overline{\operatorname{Hess} H([\mathbf{Y}])}) \leqslant 4\left(\sigma_1\left(\mathbf{Y}^*\right) + \mu\sigma_r\left(\mathbf{Y}^*\right)/\kappa^*\right)^2 + 14\mu\sigma_r^2\left(\mathbf{Y}^*\right)/3$$

In particular, if μ *is further chosen such that* $\left(2\left(1-\mu/\kappa^*\right)^2 - (14/3)\mu\right)\sigma_r\left(\mathcal{A}_{\mathbf{n}}\right) - 2\sigma_{r+1}(\mathcal{A}_{\mathbf{n}}) > 0$*, we have* $H([\mathbf{Y}])$ *is geodesically strongly convex and smooth in* $\mathcal{R}_1$*.*

Proof. Denote by $\mathbf{Q}$ the best orthogonal matrix that aligns $\mathbf{Y}$ and $\mathbf{Y}^*$. Then by the assumption on $\mathbf{Y} \in \mathcal{R}_1$ as defined in Equation 52, we have

$$\|\mathbf{Y} - \mathbf{Y}^*\mathbf{Q}\| \leqslant \|\mathbf{Y} - \mathbf{Y}^*\mathbf{Q}\|_{\mathrm{F}} = d\left([\mathbf{Y}], [\mathbf{Y}^*]\right) \leqslant \mu\sigma_r\left(\mathbf{Y}^*\right)/\kappa^*. \tag{57}$$

Thus

$$\begin{aligned}
\sigma_r(\mathbf{Y}) = \sigma_r\left(\mathbf{Y} - \mathbf{Y}^*\mathbf{Q} + \mathbf{Y}^*\mathbf{Q}\right) \geqslant \sigma_r\left(\mathbf{Y}^*\right) - \|\mathbf{Y} - \mathbf{Y}^*\mathbf{Q}\| &\overset{\text{Equation 57}}{\geqslant} \left(1-\mu/\kappa^*\right)\sigma_r\left(\mathbf{Y}^*\right) \\
\sigma_1(\mathbf{Y}) = \sigma_1\left(\mathbf{Y} - \mathbf{Y}^*\mathbf{Q} + \mathbf{Y}^*\mathbf{Q}\right) \leqslant \sigma_1\left(\mathbf{Y}^*\right) + \|\mathbf{Y} - \mathbf{Y}^*\mathbf{Q}\| &\overset{\text{Equation 57}}{\leqslant} \sigma_1\left(\mathbf{Y}^*\right) + \mu\sigma_r\left(\mathbf{Y}^*\right)/\kappa^*
\end{aligned} \tag{58}$$

where the first inequalities follow from Weyl's theorem [71]. Then,

$$
\begin{aligned}
\overline{\operatorname{Hess} H([\mathbf{Y}])}\left[\theta_{\mathbf{Y}}, \theta_{\mathbf{Y}}\right] &= \left\|\mathbf{Y}\theta_{\mathbf{Y}}^{\top} + \theta_{\mathbf{Y}}\mathbf{Y}^{\top}\right\|_{\mathrm{F}}^{2} + 2\left\langle \mathbf{Y}\mathbf{Y}^{\top} - \mathcal{A}_{\mathbf{n}}, \theta_{\mathbf{Y}}\theta_{\mathbf{Y}}^{\top}\right\rangle && \text{[Equation 51]} \\
&\geqslant 2\sigma_r^2(\mathbf{Y})\left\|\theta_{\mathbf{Y}}\right\|_{\mathrm{F}}^{2} + 2\left\langle \mathbf{Y}\mathbf{Y}^{\top} - \mathcal{A}_{\mathbf{n}}, \theta_{\mathbf{Y}}\theta_{\mathbf{Y}}^{\top}\right\rangle && \text{[Lemma 6]} \\
&= 2\sigma_r^2(\mathbf{Y})\left\|\theta_{\mathbf{Y}}\right\|_{\mathrm{F}}^{2} + 2\left\langle \mathbf{Y}\mathbf{Y}^{\top}, \theta_{\mathbf{Y}}\theta_{\mathbf{Y}}^{\top}\right\rangle - 2\left\langle \mathbf{Y}^{*}\mathbf{Y}^{*\top}, \theta_{\mathbf{Y}}\theta_{\mathbf{Y}}^{\top}\right\rangle \\
&\quad - 2\left\langle \mathbf{Z}\mathbf{Z}^{\top}, \theta_{\mathbf{Y}}\theta_{\mathbf{Y}}^{\top}\right\rangle && \left[\mathcal{A}_{\mathbf{n}} = \mathbf{Y}^{*}\mathbf{Y}^{*\top} + \mathbf{Z}\mathbf{Z}^{\top}\right] \\
&\geqslant 2\sigma_r^2(\mathbf{Y})\left\|\theta_{\mathbf{Y}}\right\|_{\mathrm{F}}^{2} - 2\left\|\mathbf{Y}\mathbf{Y}^{\top} - \mathbf{Y}^{*}\mathbf{Y}^{*\top}\right\|\left\|\theta_{\mathbf{Y}}\theta_{\mathbf{Y}}^{\top}\right\|_{\mathrm{F}} \\
&\quad - 2\|\mathbf{Z}\mathbf{Z}^{\top}\|\|\theta_{\mathbf{Y}}\theta_{\mathbf{Y}}^{\top}\|_{\mathrm{F}} && \left[\langle A, B\rangle \leq \|A\|\|B\|_{\mathrm{F}}\right] \\
&\geqslant 2\sigma_r^2(\mathbf{Y})\left\|\theta_{\mathbf{Y}}\right\|_{\mathrm{F}}^{2} - 2\left\|\mathbf{Y}\mathbf{Y}^{\top} - \mathbf{Y}^{*}\mathbf{Y}^{*\top}\right\|\left\|\theta_{\mathbf{Y}}\right\|_{\mathrm{F}}^{2} \\
&\quad - 2\|\mathbf{Z}\mathbf{Z}^{\top}\|\|\theta_{\mathbf{Y}}\|_{\mathrm{F}}^{2} && \left[\|\theta_{\mathbf{Y}}\theta_{\mathbf{Y}}^{\top}\|_{\mathrm{F}} = \|\theta_{\mathbf{Y}}\|_{\mathrm{F}}^{2}\right] \\
&\geqslant 2\left(1 - \frac{\mu}{\kappa^{*}}\right)^{2}\sigma_r^{2}\left(\mathbf{Y}^{*}\right)\left\|\theta_{\mathbf{Y}}\right\|_{\mathrm{F}}^{2} - 2\|\mathbf{Z}\mathbf{Z}^{\top}\|\|\theta_{\mathbf{Y}}\|_{\mathrm{F}}^{2} \\
&\quad - 2\left\|\mathbf{Y}\mathbf{Y}^{\top} - \mathbf{Y}^{*}\mathbf{Y}^{*\top}\right\|\left\|\theta_{\mathbf{Y}}\right\|_{\mathrm{F}}^{2} && \text{[Equation 58]} \\
&\geqslant 2\left(1 - \frac{\mu}{\kappa^{*}}\right)^{2}\sigma_r^{2}\left(\mathbf{Y}^{*}\right)\left\|\theta_{\mathbf{Y}}\right\|_{\mathrm{F}}^{2} - 2\|\mathbf{Z}\mathbf{Z}^{\top}\|\|\theta_{\mathbf{Y}}\|_{\mathrm{F}}^{2} \\
&\quad - 2\cdot\frac{7}{3}\left\|\mathbf{Y}^{*}\right\|\frac{\mu\sigma_r\left(\mathbf{Y}^{*}\right)}{\kappa^{*}}\left\|\theta_{\mathbf{Y}}\right\|_{\mathrm{F}}^{2} && \text{[Lemma 4, } \mathbf{Y} \in \mathcal{R}_1] \\
&= 2\left(1 - \frac{\mu}{\kappa^{*}}\right)^{2}\sigma_r^{2}\left(\mathbf{Y}^{*}\right)\left\|\theta_{\mathbf{Y}}\right\|_{\mathrm{F}}^{2} - 2\cdot\frac{7}{3}\left\|\mathbf{Y}^{*}\right\|\frac{\mu\sigma_r\left(\mathbf{Y}^{*}\right)}{\kappa^{*}}\left\|\theta_{\mathbf{Y}}\right\|_{\mathrm{F}}^{2} \\
&\quad - 2\sigma_{r+1}(\mathcal{A}_{\mathbf{n}})\|\theta_{\mathbf{Y}}\|_{\mathrm{F}}^{2} && \left[\|\mathbf{Z}\mathbf{Z}^{\top}\| = \sigma_{r+1}(\mathcal{A}_{\mathbf{n}})\right] \\
&= \left(\left(2\left(1 - \frac{\mu}{\kappa^{*}}\right)^{2} - \frac{14}{3}\mu\right)\sigma_r\left(\mathcal{A}_{\mathbf{n}}\right) - 2\sigma_{r+1}(\mathcal{A}_{\mathbf{n}})\right)\left\|\theta_{\mathbf{Y}}\right\|_{\mathrm{F}}^{2} && \left[\kappa^{*} = \frac{\|\mathbf{Y}^{*}\|}{\sigma_r(\mathbf{Y}^{*})}\right]
\end{aligned}
$$

Likewise,

$$
\begin{aligned}
\overline{\operatorname{Hess} H([\mathbf{Y}])}\left[\theta_{\mathbf{Y}}, \theta_{\mathbf{Y}}\right] &= \left\|\mathbf{Y}\theta_{\mathbf{Y}}^{\top} + \theta_{\mathbf{Y}}\mathbf{Y}^{\top}\right\|_{\mathrm{F}}^{2} + 2\left\langle \mathbf{Y}\mathbf{Y}^{\top} - \mathcal{A}_{\mathbf{n}}, \theta_{\mathbf{Y}}\theta_{\mathbf{Y}}^{\top}\right\rangle && \text{[Equation 5} \\
&\leq 4\sigma_1^2(\mathbf{Y})\left\|\theta_{\mathbf{Y}}\right\|_{\mathrm{F}}^{2} + 2\left\langle \mathbf{Y}\mathbf{Y}^{\top} - \mathcal{A}_{\mathbf{n}}, \theta_{\mathbf{Y}}\theta_{\mathbf{Y}}^{\top}\right\rangle && \text{[Lemma} \\
&\leqslant 4\sigma_1^2(\mathbf{Y})\left\|\theta_{\mathbf{Y}}\right\|_{\mathrm{F}}^{2} + 2\left\langle \mathbf{Y}\mathbf{Y}^{\top} - \mathbf{Y}^{*}\mathbf{Y}^{*\top}, \theta_{\mathbf{Y}}\theta_{\mathbf{Y}}^{\top}\right\rangle && [\mathcal{A}_{\mathbf{n}} - \mathbf{Y}^{*}\mathbf{Y}^{*\top} \text{ is PSI} \\
&\leqslant 4\sigma_1^2(\mathbf{Y})\left\|\theta_{\mathbf{Y}}\right\|_{\mathrm{F}}^{2} + 2\left\|\mathbf{Y}\mathbf{Y}^{\top} - \mathbf{Y}^{*}\mathbf{Y}^{*\top}\right\|\left\|\theta_{\mathbf{Y}}\right\|_{\mathrm{F}}^{2} \\
&\leqslant 4\sigma_1^2(\mathbf{Y})\left\|\theta_{\mathbf{Y}}\right\|_{\mathrm{F}}^{2} + 2\left\|\mathbf{Y}\mathbf{Y}^{\top} - \mathbf{Y}^{*}\mathbf{Y}^{*\top}\right\|_{\mathrm{F}}\left\|\theta_{\mathbf{Y}}\right\|_{\mathrm{F}}^{2} \\
&\leqslant 4\left(\sigma_1\left(\mathbf{Y}^{*}\right) + \frac{\mu\sigma_r\left(\mathbf{Y}^{*}\right)}{\kappa^{*}}\right)^{2}\left\|\theta_{\mathbf{Y}}\right\|_{\mathrm{F}}^{2} + 2\left\|\mathbf{Y}\mathbf{Y}^{\top} - \mathbf{Y}^{*}\mathbf{Y}^{*\top}\right\|_{\mathrm{F}}\left\|\theta_{\mathbf{Y}}\right\|_{\mathrm{F}}^{2} && \text{[Equation 5} \\
&\leqslant \left(4\left(\sigma_1\left(\mathbf{Y}^{*}\right) + \frac{\mu\sigma_r\left(\mathbf{Y}^{*}\right)}{\kappa^{*}}\right)^{2} + \frac{14}{3}\mu\sigma_r^{2}\left(\mathbf{Y}^{*}\right)\right)\left\|\theta_{\mathbf{Y}}\right\|_{\mathrm{F}}^{2} && \text{[Lemma}
\end{aligned}
$$

From the above we conclude that when μ is chosen such that

$$
\left(2\left(1 - \frac{\mu}{\kappa^{*}}\right)^{2} - \frac{14}{3}\mu\right)\sigma_r\left(\mathcal{A}_{\mathbf{n}}\right) - 2\sigma_{r+1}(\mathcal{A}_{\mathbf{n}}) > 0,
$$

we have $H([\mathbf{Y}])$ in Equation 8 is geodesically strongly convex and smooth in $\mathcal{R}_1$ as $\mathcal{R}_1$ is a geodesically convex set by [56]. Note that this is equivalent to

$$\left(\left(1-\frac{\mu}{\kappa^*}\right)^2-\frac{7}{3}\mu\right)>\frac{\sigma_{r+1}(\mathcal{A}_\mathbf{n})}{\sigma_r(\mathcal{A}_\mathbf{n})}.$$

Then note as $\mu\to 0$, the left hand side approaches 1 and the inequality becomes true as $\sigma_r(\mathcal{A}_\mathbf{n})>\sigma_{r+1}(\mathcal{A}_\mathbf{n})$. □

Remark 13. *Compared with the bound in Theorem 8 of [56], the smoothness and geodesically strongly convexity are as follows,*

$$\sigma_{\min}(\overline{\operatorname{Hess} H([\mathbf{Y}])})\geqslant\left(2\left(1-\mu/\kappa^*\right)^2-(14/3)\mu\right)\sigma_r^2\left(\mathbf{Y}^*\right),$$
$$\sigma_{\max}(\overline{\operatorname{Hess} H([\mathbf{Y}])})\leqslant 4\left(\sigma_1\left(\mathbf{Y}^*\right)+\mu\sigma_r\left(\mathbf{Y}^*\right)/\kappa^*\right)^2+14\mu\sigma_r^2\left(\mathbf{Y}^*\right)/3.$$

There is an extra term $-2\sigma_{r+1}(\mathcal{A}_\mathbf{n})$ in our lower bound of the strong convexity because even if $d([\mathbf{Y}],[\mathbf{Y}^*])$ is small, $\mathcal{A}_\mathbf{n}-\mathbf{Y}\mathbf{Y}^\top$ is not close to $\mathbf{0}$, which leads to the extra error term.

In the next three theorems, we show that for $\mathbf{Y}\notin\mathcal{R}_1$, either the Riemannian Hessian evaluated at $\mathbf{Y}$ has a large negative eigenvalue, or the norm of the Riemannian gradient is large. Let $\mathbf{Y}=\mathbf{U}\mathbf{D}\mathbf{V}^\top$, $\mathbf{Y}^*=\mathbf{U}^*\mathbf{\Sigma}^{*1/2}$.

Theorem 2.1 (FOSP of Equation 8)**.** *Let $\overline{\mathbf{U}}\mathbf{\Sigma}\overline{\mathbf{U}}^\top$ be $\mathcal{A}_\mathbf{n}$'s SVD factorization, and let $\mathbf{\Lambda}=\mathbf{\Sigma}^{1/2}$. Then for any S subset of $[n]$ we have that $\left[\overline{\mathbf{U}}_S\mathbf{\Lambda}_S\right]$ is a Riemannian FOSP of Equation 8. Further, these are the only Riemannian FOSPs.*

Proof. From Equation 51, the gradient can be written down as,

$$\begin{aligned}\overline{\operatorname{grad} H([\mathbf{Y}])}&=2\left(\mathbf{Y}\mathbf{Y}^\top-\mathcal{A}_\mathbf{n}\right)\mathbf{Y}=2\left(\mathbf{U}\mathbf{D}\mathbf{V}^\top(\mathbf{U}\mathbf{D}\mathbf{V}^\top)^\top-\mathcal{A}_\mathbf{n}\right)\mathbf{U}\mathbf{D}\mathbf{V}^\top\\&=2\left(\mathbf{U}\mathbf{D}^3\mathbf{V}^\top-\mathcal{A}_\mathbf{n}\mathbf{U}\mathbf{D}\mathbf{V}^\top\right)\end{aligned}$$

Therefore, whenever $\overline{\operatorname{grad} H([\mathbf{Y}])}=\mathbf{0}$, we have $\mathbf{U}\mathbf{D}^3\mathbf{V}^\top-\mathcal{A}_\mathbf{n}\mathbf{U}\mathbf{D}\mathbf{V}^\top=\mathbf{0}$. Since both $\mathbf{V}$ and $\mathbf{D}$ are of full rank, the condition is equivalent to

$$\mathbf{U}\mathbf{D}^2-\mathcal{A}_\mathbf{n}\mathbf{U}=\mathbf{0}\tag{59}$$

Since $\mathbf{D}^2$ is also a diagonal matrix, to satisfy Equation 59, the columns of $\mathbf{U}$ have to be the eigenvectors of $\mathcal{A}_\mathbf{n}$, and the diagonal of $\mathbf{D}^2$ has to be the eigenvalues of $\mathcal{A}_\mathbf{n}$. This completes the proof. □

Next, we prove Theorem 4.2. For the reader's convenience, we restate the theorem.

Theorem 4.2 (Region with Negative Eigenvalue in the Riemannian Hessian of Equation 7)**.** *Assume that Assumption 2 holds. Given any $\mathbf{Y}\in\mathbb{R}_*^{n\times r}$ such that $\mathbf{Y}\in\mathcal{R}_2$, let $\theta_\mathbf{Y}^1=[\mathbf{0},\mathbf{0},\ldots,\mathbf{0},\boldsymbol{a},\mathbf{0},\ldots,\mathbf{0}]\mathbf{V}^\top$ where $\boldsymbol{a}$ such that*

$$\boldsymbol{a}=\underset{\boldsymbol{a}:\mathbf{Y}^\top\boldsymbol{a}=\mathbf{0}}{\arg\max}\ \frac{\boldsymbol{a}^\top\mathcal{A}_\mathbf{n}\boldsymbol{a}}{\|\boldsymbol{a}\|^2}\tag{22}$$

and $[\mathbf{0}, \mathbf{0}, \ldots, \mathbf{0}, \boldsymbol{a}, \mathbf{0}, \ldots, \mathbf{0}] \in \mathbb{R}^{n\times r}$ *such that the* $\tilde{i}^{\text{th}}$ *columns is* $\boldsymbol{a}$ *and other columns are* $\mathbf{0}$ *where*

$$\tilde{i} \overset{def}{=} \arg\min_{j\in[r]} \mathbf{D}_{jj}. \tag{23}$$

Denote $\theta^2_{\mathbf{Y}} = \mathbf{Y} - \mathbf{Y}^*\mathbf{Q}$*, where* $\mathbf{Q} \in \mathbb{O}_r$ *is the best orthogonal matrix aligning* $\mathbf{Y}^*$ *and* $\mathbf{Y}$*. We choose* $\theta_{\mathbf{Y}}$ *to be either* $\theta^1_{\mathbf{Y}}$ *or* $\theta^2_{\mathbf{Y}}$*. Then*

$$\begin{aligned}\overline{\operatorname{Hess} H([\mathbf{Y}])}\left[\theta_{\mathbf{Y}}, \theta_{\mathbf{Y}}\right] \leqslant \min\Bigg\{ & -\frac{\sigma^2_{r+1}(\mathbf{\Lambda})}{2}\|\theta_{\mathbf{Y}}\|^2, \\ & -2\left(\sigma^2_r(\mathbf{\Lambda})\left(1 - \frac{e_1^2}{\left|\sigma^2_r(\mathbf{\Lambda}) - e_1 - \sigma^2_{r+1}(\mathbf{\Lambda})\right|^2}\right) - e_1 - \sigma^2_{r+1}(\mathbf{\Lambda})\right)\|\theta_{\mathbf{Y}}\|^2, \\ & \left((\alpha - 2(\sqrt{2}-1))\sigma^2_r(\mathbf{Y}^*) + 6\frac{\alpha^2\sigma^4_r(\mathbf{Y}^*)\,\sigma^2_{r+1}(\mathbf{\Lambda})/16}{\left|\sigma^2_r(\mathbf{\Lambda}) - e_2 - \sigma^2_{r+1}(\mathbf{\Lambda})\right|^2}\right)\|\theta_{\mathbf{Y}}\|^2_{\mathrm{F}}\Bigg\}\end{aligned}$$

In particular, if α *and* μ *satisfies Assumption 3, we have* $\overline{\operatorname{Hess} H([\mathbf{Y}])}$ *has at least one negative eigenvalue and* $\theta^1_{\mathbf{Y}}$ *or* $\theta^2_{\mathbf{Y}}$ *is the escaping direction.*

Proof. By the definition of $\boldsymbol{a}$, $\boldsymbol{a} \in \operatorname{Span}\{\overline{\mathbf{U}}_{1,\ldots,r+1}\}$. This is because the null space of $\mathbf{Y}$ has dimension $n - r$. Hence, its intersection with a dimension $r + 1$ space has a dimension of at least 1.

Using the SVD decomposition of $\mathbf{Y}$, we have, $\mathbf{U}^\top\boldsymbol{a} = \mathbf{0}$. Then, by using Equation 51, we have

$$\begin{aligned}\overline{\operatorname{Hess} H([\mathbf{Y}])}\left[\theta^1_{\mathbf{Y}}, \theta^1_{\mathbf{Y}}\right] &= \left\|\mathbf{Y}(\theta^1_{\mathbf{Y}})^\top + \theta^1_{\mathbf{Y}}\mathbf{Y}^\top\right\|^2_{\mathrm{F}} + 2\left\langle \mathbf{Y}\mathbf{Y}^\top - \mathcal{A}_{\mathbf{n}}, \theta^1_{\mathbf{Y}}(\theta^1_{\mathbf{Y}})^\top\right\rangle && \text{[Equation 51]} \\ &= \left\|\mathbf{Y}(\theta^1_{\mathbf{Y}})^\top + \theta^1_{\mathbf{Y}}\mathbf{Y}^\top\right\|^2_{\mathrm{F}} - 2\left\langle \mathcal{A}_{\mathbf{n}}, \theta^1_{\mathbf{Y}}(\theta^1_{\mathbf{Y}})^\top\right\rangle && [\mathbf{Y}^\top\boldsymbol{a} = 0] \\ &= 2\langle \mathbf{Y}^\top\mathbf{Y}, (\theta^1_{\mathbf{Y}})^\top\theta^1_{\mathbf{Y}}\rangle + 2\langle \mathbf{Y}(\theta^1_{\mathbf{Y}})^\top, \theta^1_{\mathbf{Y}}\mathbf{Y}^\top\rangle - 2\left\langle \mathcal{A}_{\mathbf{n}}, \theta^1_{\mathbf{Y}}(\theta^1_{\mathbf{Y}})^\top\right\rangle && [\|A\|^2_{\mathrm{F}} = \langle A, A\rangle] \\ &= 2\langle \mathbf{Y}^\top\mathbf{Y}, (\theta^1_{\mathbf{Y}})^\top\theta^1_{\mathbf{Y}}\rangle - 2\left\langle \mathcal{A}_{\mathbf{n}}, \theta^1_{\mathbf{Y}}(\theta^1_{\mathbf{Y}})^\top\right\rangle && [\mathbf{Y}^\top\boldsymbol{a} = 0] \\ &= 2\langle \mathbf{V}\mathbf{D}^2\mathbf{V}^\top, (\theta^1_{\mathbf{Y}})^\top\theta^1_{\mathbf{Y}}\rangle - 2\left\langle \mathcal{A}_{\mathbf{n}}, \theta^1_{\mathbf{Y}}(\theta^1_{\mathbf{Y}})^\top\right\rangle \\ &= 2\mathbf{D}^2_{\tilde{i}\tilde{i}}\|\boldsymbol{a}\|^2 - 2\boldsymbol{a}^\top\mathcal{A}_{\mathbf{n}}\boldsymbol{a}\end{aligned}$$

where the last equality comes from the definition of $\boldsymbol{a}$ and the fact that the $\mathbf{V}^\top\mathbf{V} = \mathbf{I}$ in $\theta^1_{\mathbf{Y}}(\theta^1_{\mathbf{Y}})^\top$. Recall $\tilde{i} = \arg\min \mathbf{D}_{ii}$, then

$$\overline{\operatorname{Hess} H([\mathbf{Y}])}\left[\theta^1_{\mathbf{Y}}, \theta^1_{\mathbf{Y}}\right] = 2\min_i \mathbf{D}^2_{ii}\|\boldsymbol{a}\|^2 - 2\boldsymbol{a}^\top\mathcal{A}_{\mathbf{n}}\boldsymbol{a} \tag{60}$$

In the following, we separate the proof into three regimes of $\min_i \mathbf{D}^2_{ii}$, corresponding to different escape directions.

Case 1: $\left(\textbf{When } \min_i \mathbf{D}^2_{ii} < \frac{\sigma^2_{r+1}(\mathbf{\Lambda})}{2}\right)$. For this case we must have that

$$\overline{\operatorname{Hess} H([\mathbf{Y}])}\left[\theta^1_{\mathbf{Y}}, \theta^1_{\mathbf{Y}}\right] \leq -\frac{\sigma^2_{r+1}(\mathbf{\Lambda})}{2}\|\theta^1_{\mathbf{Y}}\|^2.$$

This is because $\boldsymbol{a}^\top \mathcal{A}_{\mathbf{n}} \boldsymbol{a} \geq \sigma_{r+1}^2(\boldsymbol{\Lambda}) \|\boldsymbol{a}\|^2$ and $\|\boldsymbol{a}\| = \|\theta_{\mathbf{Y}}^1\|$.

Case 2: $\left(\textbf{When } \min_i \mathbf{D}_{ii}^2 \geq \frac{\sigma_{r+1}^2(\boldsymbol{\Lambda})}{2}\right)$.

From the proof of Theorem 2.1, the gradient condition of $\mathcal{R}_2$ can be written as

$$\begin{aligned}
\alpha\mu\sigma_r^3\left(\mathbf{Y}^*\right)/\left(4\kappa^*\right) &\geq \|\overline{\operatorname{grad} H([\mathbf{Y}])}\|_{\mathrm{F}} && [\mathbf{Y} \in \mathcal{R}_2] \\
&= \|2\left(\mathbf{U}\mathbf{D}^3\mathbf{V}^\top - \mathcal{A}_{\mathbf{n}}\mathbf{U}\mathbf{D}\mathbf{V}^\top\right)\|_{\mathrm{F}} && [\text{Equation 51}] \\
&= \|2\left(\mathbf{U}\mathbf{D}^2 - \mathcal{A}_{\mathbf{n}}\mathbf{U}\right)\mathbf{D}\|_{\mathrm{F}}
\end{aligned}$$

Assume $\mathbf{U} = \overline{\mathbf{U}}\boldsymbol{C}$ where $\boldsymbol{C} \in \mathbb{R}^{n\times r}$. Since $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_r$ and $\overline{\mathbf{U}}^\top\overline{\mathbf{U}} = \mathbf{I}_n$, we have $\boldsymbol{C}^\top\boldsymbol{C} = \mathbf{I}_r$. Furthermore,

$$\begin{aligned}
\|2\left(\mathbf{U}\mathbf{D}^2 - \mathcal{A}_{\mathbf{n}}\mathbf{U}\right)\mathbf{D}\|_{\mathrm{F}} &= \|2\left(\overline{\mathbf{U}}\boldsymbol{C}\mathbf{D}^2 - \mathcal{A}_{\mathbf{n}}\overline{\mathbf{U}}\boldsymbol{C}\right)\mathbf{D}\|_{\mathrm{F}} && [\mathbf{U} = \overline{\mathbf{U}}\boldsymbol{C}] \\
&= \|2\left(\overline{\mathbf{U}}\boldsymbol{C}\mathbf{D}^2 - \overline{\mathbf{U}}\boldsymbol{\Sigma}\boldsymbol{C}\right)\mathbf{D}\|_{\mathrm{F}} && [\mathcal{A}_{\mathbf{n}} = \overline{\mathbf{U}}\boldsymbol{\Sigma}\overline{\mathbf{U}}^\top] \\
&= 2\|\left(\boldsymbol{C}\mathbf{D}^2 - \boldsymbol{\Sigma}\boldsymbol{C}\right)\mathbf{D}\|_{\mathrm{F}}.
\end{aligned}$$

Here the third equality follows from $\overline{\mathbf{U}}^\top\overline{\mathbf{U}} = \mathbf{I}_n$. By a direct computation, the i^{th} column of $\left(\boldsymbol{C}\mathbf{D}^2 - \boldsymbol{\Sigma}\boldsymbol{C}\right)\mathbf{D}$ is $\mathbf{D}_{ii}^3\boldsymbol{C}_i - \mathbf{D}_{ii}\boldsymbol{\Sigma}\boldsymbol{C}_i$. Therefore, the gradient condition of $\mathcal{R}_2$ can be written as

$$\sum_{i,j}\left(\mathbf{D}_{ii}^3\boldsymbol{C}_{ji} - \mathbf{D}_{ii}\boldsymbol{\Sigma}_{jj}\boldsymbol{C}_{ji}\right)^2 \leq \alpha^2\mu^2\sigma_r^6\left(\mathbf{Y}^*\right)/\left(4\kappa^*\right)^2 \tag{61}$$

We fix i in the left hand side of Equation 61, we have

$$\sum_{j}\left(\mathbf{D}_{ii}^2 - \boldsymbol{\Sigma}_{jj}\right)^2\mathbf{D}_{ii}^2\boldsymbol{C}_{ji}^2 \leq \alpha^2\mu^2\sigma_r^6\left(\mathbf{Y}^*\right)/\left(4\kappa^*\right)^2 \tag{62}$$

where $\sum_j \boldsymbol{C}_{ji}^2 = 1$. From $\mathbf{D}_{ii}^2 \geq \frac{\sigma_{r+1}^2(\boldsymbol{\Lambda})}{2}$, we must have

$$\min_j |\mathbf{D}_{ii}^2 - \boldsymbol{\Sigma}_{jj}|^2 \leq \sum_j\left(\mathbf{D}_{ii}^2 - \boldsymbol{\Sigma}_{jj}\right)^2\boldsymbol{C}_{ji}^2 \leq \frac{\alpha^2\mu^2\sigma_r^6\left(\mathbf{Y}^*\right)}{\left(4\kappa^*\right)^2\frac{\sigma_{r+1}^2(\boldsymbol{\Lambda})}{2}}. \tag{63}$$

We use Equation 62 for the second inequality. Equation 63 is important in the proof because this essentially guarantees that $\mathbf{D}_{ii}^2$ must be close to some $\boldsymbol{\Sigma}_{jj}$. This is because $\frac{\alpha^2\mu^2\sigma_r^6(\mathbf{Y}^*)}{(4\kappa^*)^2\frac{\sigma_{r+1}^2(\boldsymbol{\Lambda})}{2}}$ is guaranteed small according to Assumption 3.

We decompose $\boldsymbol{C}_{\tilde{i}}$ into $\xi^1 + \xi^2$ where $\xi_j^1 = 0$ for all $j \geq r+1$ and $\xi_j^2 = 0$ for all $j \in [r]$. Since $\langle\xi^1, \xi^2\rangle = 0$ and $\boldsymbol{C}^\top\boldsymbol{C} = \mathbf{I}$,

$$\|\xi^1\|^2 + \|\xi^2\|^2 = 1 \tag{64}$$

In the following, we divide all the cases into different regimes based on which of the eigenvalues of $\boldsymbol{\Lambda}$ is close to $\mathbf{D}_{\tilde{i}\tilde{i}}$.

Case 2.1: $\left(\textbf{When } \frac{\sigma_{r+1}^2(\mathbf{\Lambda})}{2} \leq \mathbf{D}_{\tilde{i}\tilde{i}}^2 \leq \frac{\alpha\mu\sigma_r^3(\mathbf{Y}^*)}{2\sqrt{2}\kappa^*\sigma_{r+1}(\mathbf{\Lambda})} + \sigma_{r+1}^2(\mathbf{\Lambda})\right)$.

Notice that the first assumption in Assumption 3 essentially guarantees a small $e_1 = \frac{\alpha\mu\sigma_r^3(\mathbf{Y}^*)}{2\sqrt{2}\kappa^*\sigma_{r+1}(\mathbf{\Lambda})}$.

Hence, we have

$$\begin{aligned}
\alpha^2\mu^2\sigma_r^6\left(\mathbf{Y}^*\right)/\left(4\kappa^*\right)^2 &\geq \sum_j \left(\mathbf{D}_{\tilde{i}\tilde{i}}^2 - \mathbf{\Sigma}_{jj}\right)^2 \mathbf{D}_{\tilde{i}\tilde{i}}^2 \boldsymbol{C}_{j\tilde{i}}^2 && \text{[Equation 61]}\\
&\geq \sum_{j\leq r} \left|\sigma_j^2(\mathbf{\Lambda}) - \mathbf{D}_{\tilde{i}\tilde{i}}^2\right|^2 \cdot \mathbf{D}_{\tilde{i}\tilde{i}}^2 \cdot \boldsymbol{C}_{j\tilde{i}}^2\\
&\geq \left|\sigma_r^2(\mathbf{\Lambda}) - \mathbf{D}_{\tilde{i}\tilde{i}}^2\right|^2 \cdot \mathbf{D}_{\tilde{i}\tilde{i}}^2 \cdot \|\xi^1\|^2\\
&\geq \left|\sigma_r^2(\mathbf{\Lambda}) - e_1 - \sigma_{r+1}^2(\mathbf{\Lambda})\right|^2 \cdot \frac{\sigma_{r+1}^2(\mathbf{\Lambda})}{2} \cdot \|\xi^1\|^2.
\end{aligned}$$

Where in the last two inequalities, we use the condition $\frac{\sigma_{r+1}^2(\mathbf{\Lambda})}{2} \leq \mathbf{D}_{\tilde{i}\tilde{i}}^2 \leq e_1 + \sigma_{r+1}^2(\mathbf{\Lambda})$ and that $e_1 < (\sigma_r^2(\mathbf{\Lambda}) - \sigma_{r+1}^2(\mathbf{\Lambda}))/2$ (follows from Assumption 3).

By reordering the inequality, we have

$$\|\xi^1\| \leq \frac{e_1}{\left|\sigma_r^2(\mathbf{\Lambda}) - e_1 - \sigma_{r+1}^2(\mathbf{\Lambda})\right|} \tag{65}$$

Recall that $\mathbf{Y} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, then $\boldsymbol{a}^\top\mathbf{Y} = \mathbf{0}$ reduces to $\boldsymbol{a}^\top\mathbf{U}\mathbf{D}\mathbf{V}^\top = \mathbf{0}$. Since both $\mathbf{D}, \mathbf{V} \in \mathbb{R}^{r*r}$ are full rank, then we have $\boldsymbol{a}^\top\mathbf{U} = \mathbf{0}$, in turn $\boldsymbol{a}^\top\overline{\mathbf{U}}\boldsymbol{C} = \mathbf{0}$ because $\mathbf{U} = \overline{\mathbf{U}}\boldsymbol{C}$. Denote $\boldsymbol{b}^\top \overset{\text{def}}{=} \boldsymbol{a}^\top\overline{\mathbf{U}}$, then

$$\begin{aligned}
\max_{\boldsymbol{a}:\mathbf{Y}^\top\boldsymbol{a}=\mathbf{0}} \frac{\boldsymbol{a}^\top\mathcal{A}_{\mathbf{n}}\boldsymbol{a}}{\|\boldsymbol{a}\|^2} &= \max_{\boldsymbol{a}:\boldsymbol{a}^\top\overline{\mathbf{U}}\boldsymbol{C}=\mathbf{0}} \frac{\boldsymbol{a}^\top\mathcal{A}_{\mathbf{n}}\boldsymbol{a}}{\|\boldsymbol{a}\|^2}\\
&= \max_{\boldsymbol{a}:\boldsymbol{a}^\top\overline{\mathbf{U}}\boldsymbol{C}=\mathbf{0}} \frac{\boldsymbol{a}^\top\overline{\mathbf{U}}\mathbf{\Lambda}\overline{\mathbf{U}}^\top\boldsymbol{a}}{\|\boldsymbol{a}\|^2} \quad [\mathcal{A}_{\mathbf{n}} = \overline{\mathbf{U}}\mathbf{\Lambda}\overline{\mathbf{U}}^\top] \qquad (66)\\
&= \max_{\boldsymbol{b}:\boldsymbol{b}^\top\boldsymbol{C}=\mathbf{0}} \frac{\boldsymbol{b}^\top\mathbf{\Lambda}\boldsymbol{b}}{\|\boldsymbol{b}\|^2} \qquad [\overline{\mathbf{U}}^\top\overline{\mathbf{U}} = \mathbf{I}]
\end{aligned}$$

Since $\boldsymbol{a} \in \text{Span}\{\overline{\mathbf{U}}_{1,\ldots,r+1}\}$, we have $\boldsymbol{b}_j = 0$ for $j > r+1$. From $\boldsymbol{b}^\top\boldsymbol{C} = \mathbf{0}$, we have $\boldsymbol{b}^\top\boldsymbol{C}_{\tilde{i}} = 0$, which can be written as $\boldsymbol{b}^\top(\xi^1 + \xi^2) = 0$. Since there are in total r constraints in $\boldsymbol{b}^\top\boldsymbol{C} = \mathbf{0}$, there must exist a $\boldsymbol{b}$ satisfying the constraints $\boldsymbol{b}^\top\boldsymbol{C} = \mathbf{0}$, and the norm of $\boldsymbol{b}_{r+1:n}$ is relatively small compared with the norm of $\boldsymbol{b}_{1:r}$. Specifically, denote $\boldsymbol{C}_{1:r}$ to be the 1^{st} to r^{th} rows of $\boldsymbol{C}$. We consider $\boldsymbol{b}$ to be $\boldsymbol{b}^1 + \boldsymbol{b}^2$ such that $\boldsymbol{b}_i^1 = 0$ for $i > r$, and $\boldsymbol{b}_i^2 = 0$ for $i \in [r]$. We discuss two cases of $\boldsymbol{C}_{1:r} \in \mathbb{R}^{r*r}$ in the following.

Case 2.1.1: If $\boldsymbol{C}_{1:r}$ is not full rank.

In this case, there exists $\tilde{\boldsymbol{b}}^1 \in \mathbb{R}^r$ such that $\|\tilde{\boldsymbol{b}}^1\| > 0$ and $(\tilde{\boldsymbol{b}}^1)^\top\boldsymbol{C}_{1:r} = \mathbf{0}$. Therefore, by denoting $\bar{\boldsymbol{b}}_{1:r} = t\tilde{\boldsymbol{b}}^1 + \boldsymbol{b}_{1:r}^1$, and $\bar{\boldsymbol{b}}_{r+1:n} = \boldsymbol{b}_{r+1:n}^2$. From the definition of $\bar{\boldsymbol{b}}$ and the fact that $\boldsymbol{b}^\top\boldsymbol{C} = \mathbf{0}$, we have $\bar{\boldsymbol{b}}^\top\boldsymbol{C} = \mathbf{0}$. By letting $t \to \infty$, we have

$$\max_{\boldsymbol{b}^\top\boldsymbol{C}=\mathbf{0}} \frac{\boldsymbol{b}^\top\mathbf{\Lambda}\boldsymbol{b}}{\|\boldsymbol{b}\|^2} \geq \frac{\bar{\boldsymbol{b}}^\top\mathbf{\Lambda}\bar{\boldsymbol{b}}}{\|\bar{\boldsymbol{b}}\|^2} \geq \sigma_r^2(\mathbf{\Lambda}) \tag{67}$$

Combining Equation 67, Equation 60 and the Assumption that $\mathbf{D}^2_{\tilde{i}\tilde{i}} \leq e_1 + \sigma^2_{r+1}(\mathbf{\Lambda})$, this implies,

$$\overline{\text{Hess}\, H([\mathbf{Y}])}\left[\theta^1_{\mathbf{Y}}, \theta^1_{\mathbf{Y}}\right] \leqslant -(\sigma^2_r(\mathbf{Y}^*) - \sigma_{r+1}(\mathbf{\Lambda}) - e_1)\|\theta^1_{\mathbf{Y}}\|^2_{\text{F}} \tag{68}$$

According to Assumption 3, this satisfies the bound in Theorem 4.2 with $\theta^1_{\mathbf{Y}}$ being a negative escaping direction.

Case 2.1.2 : If $\boldsymbol{C}_{1:r}$ is full rank. In this case, we denote $\boldsymbol{b}^2 = \xi^2$. Since $\boldsymbol{C}_{1:r}$ is full rank, there exists $\boldsymbol{b}^1$ to have $(\boldsymbol{b}^1_{1:r})^\top \boldsymbol{C}_{1:r} = -(\boldsymbol{b}^2)^\top \boldsymbol{C}$; this is because $(\boldsymbol{b}^1_{1:r})^\top \boldsymbol{C}_{1:r} = -(\xi^2)^\top \boldsymbol{C}$ has in total r constraints, and there are in total r parameters in $\boldsymbol{b}^1_{1:r}$. Specifically, one can choose $\boldsymbol{b}^1$ to be $\boldsymbol{b}^1_{1:r} = -\xi^2 \boldsymbol{C}(\boldsymbol{C}_{1:r})^{-1}$ to satisfy $\boldsymbol{b}^\top \boldsymbol{C} = \mathbf{0}$. In addition, from the specific condition $\boldsymbol{b}^\top \boldsymbol{C}_{\tilde{i}} = \mathbf{0}$, we know that

$$\boldsymbol{b}^1 \cdot \xi^1 + \|\xi^2\|^2 = 0 \tag{69}$$

By using the Cauchy inequality, this further implies that

$$\|\boldsymbol{b}^1\| \geq \frac{\|\xi^2\|^2}{\|\xi^1\|} \tag{70}$$

Since we only choose a specific $\boldsymbol{b}$ such that $\boldsymbol{b}^\top \boldsymbol{C} = \mathbf{0}$ holds, we have

$$\begin{aligned}
\max_{\boldsymbol{b}^\top \boldsymbol{C} = \mathbf{0}} \frac{\boldsymbol{b}^\top \mathbf{\Lambda} \boldsymbol{b}}{\|\boldsymbol{b}\|^2} &\geq \frac{(\boldsymbol{b}^1 + \boldsymbol{b}^2)^\top \mathbf{\Lambda} (\boldsymbol{b}^1 + \boldsymbol{b}^2)}{\|\boldsymbol{b}^1 + \boldsymbol{b}^2\|^2} \\
&= \frac{(\boldsymbol{b}^1)^\top \mathbf{\Lambda} \boldsymbol{b}^1 + (\boldsymbol{b}^2)^\top \mathbf{\Lambda} \boldsymbol{b}^2}{\|\boldsymbol{b}^1\|^2 + \|\boldsymbol{b}^2\|^2} \\
&\geq \frac{(\boldsymbol{b}^1)^\top \mathbf{\Lambda} \boldsymbol{b}^1}{\|\boldsymbol{b}^1\|^2 + \|\xi^2\|^2} \\
&\geq \frac{\|\boldsymbol{b}^1\|^2 \cdot \sigma^2_r(\mathbf{\Lambda})}{\|\boldsymbol{b}^1\|^2 + \|\xi^2\|^2} \\
&\geq \frac{\frac{\|\xi^2\|^4}{\|\xi^1\|^2} \cdot \sigma^2_r(\mathbf{\Lambda})}{\frac{\|\xi^2\|^4}{\|\xi^1\|^2} + \|\xi^2\|^2} \\
&= \|\xi^2\|^2 \cdot \sigma^2_r(\mathbf{\Lambda})
\end{aligned} \tag{71}$$

where the first equality follows from the definition of $\boldsymbol{b}^1$ and $\boldsymbol{b}^2$; the second inequality follows from the assumption that $\mathbf{\Lambda}$ is PSD, and $\boldsymbol{b}^2 = \xi^2$; the third inequality follows from the fact that $\boldsymbol{b}^1_i = 0$ for $i > r$; the fourth inequality follows from Equation 70; the last equality follows from Equation 64. By using Equation 66, this can be written as

$$\max_{\boldsymbol{a}: \mathbf{Y}^\top \boldsymbol{a} = \mathbf{0}} \frac{\boldsymbol{a}^\top \mathcal{A}_{\mathbf{n}} \boldsymbol{a}}{\|\boldsymbol{a}\|^2} \geq \|\xi^2\|^2 \cdot \sigma^2_r(\mathbf{\Lambda}) \tag{72}$$

By the definition in Equation 22 and Equation 60, we have

$$\begin{aligned}
\overline{\operatorname{Hess} H([\mathbf{Y}])}\left[\theta^1_{\mathbf{Y}}, \theta^1_{\mathbf{Y}}\right] &= 2\min_i \mathbf{D}^2_{ii} \cdot \|\boldsymbol{a}\|^2 - 2\boldsymbol{a}^\top \mathcal{A}_{\mathbf{n}} \boldsymbol{a} && \text{[Equation 60]} \\
&\le 2\mathbf{D}^2_{\tilde{i}\tilde{i}} \cdot \|\boldsymbol{a}\|^2 - 2\sigma_r^2(\boldsymbol{\Lambda})\|\xi^2\|^2 \cdot \|\boldsymbol{a}\|^2 && \text{[Equation 72]} \\
&= 2\mathbf{D}^2_{\tilde{i}\tilde{i}} \cdot \|\boldsymbol{a}\|^2 - 2\sigma_r^2(\boldsymbol{\Lambda})(1 - \|\xi^1\|^2) \cdot \|\boldsymbol{a}\|^2 && [\|\xi^1\|^2 + \|\xi^2\|^2 = 1] \\
&\le \left(-2\sigma_r^2(\boldsymbol{\Lambda})(1 - \|\xi^1\|^2) + 2e_1 + 2\sigma^2_{r+1}(\boldsymbol{\Lambda})\right) \|\theta^1_{\mathbf{Y}}\|^2
\end{aligned}$$

where the last inequality follows from $\mathbf{D}^2_{\tilde{i}\tilde{i}} \le e_1 + \sigma^2_{r+1}(\boldsymbol{\Lambda})$ and the fact that $\|\theta^1_{\mathbf{Y}}\| = \|\boldsymbol{a}\|$. Finally, by applying Equation 65 to control $\|\xi^1\|$, we conclude that

$$\overline{\operatorname{Hess} H([\mathbf{Y}])}\left[\theta^1_{\mathbf{Y}}, \theta^1_{\mathbf{Y}}\right] \le -2\left(\sigma_r^2(\boldsymbol{\Lambda})\left(1 - \frac{e_1^2}{\left|\sigma_r^2(\boldsymbol{\Lambda}) - e_1 - \sigma^2_{r+1}(\boldsymbol{\Lambda})\right|^2}\right) - e_1 - \sigma^2_{r+1}(\boldsymbol{\Lambda})\right) \|\theta^1_{\mathbf{Y}}\|^2 \tag{73}$$

According to the second assumption in Assumption 3, Equation 73 guarantees an escape direction.

Case 2.2: $\left(\textbf{When } \mathbf{D}^2_{\tilde{i}\tilde{i}} > e_1 + \sigma^2_{r+1}(\boldsymbol{\Lambda})\right)$.

Recall the first assumption in Assumption 3, we have e_1 is small enough, which is viewed as an error term. In the following, we will show that $\theta^2_{\mathbf{Y}}$ is the escaping direction. We have

$$\min_j \mathbf{D}^2_{\tilde{i}\tilde{i}} |\mathbf{D}^2_{ii} - \boldsymbol{\Sigma}_{jj}|^2 \le \mathbf{D}^2_{\tilde{i}\tilde{i}} \sum_j \left(\mathbf{D}^2_{\tilde{i}\tilde{i}} - \boldsymbol{\Sigma}_{jj}\right)^2 \boldsymbol{C}^2_{j\tilde{i}} \le \frac{\alpha^2\mu^2\sigma_r^6\left(\mathbf{Y}^*\right)}{(4\kappa^*)^2} \tag{74}$$

where we use Equation 62 in the last inequality.

Recall that Assumption 3 guarantees small e_1 and e_2 , by combining Equation 74 and the assumption $\mathbf{D}^2_{\tilde{i}\tilde{i}} > \sigma^2_{r+1}(\boldsymbol{\Lambda}) + e_1$, we must have

$$\mathbf{D}^2_{\tilde{i}\tilde{i}} \ge \sigma_r^2(\boldsymbol{\Lambda}) - e_2 \tag{75}$$

where e_2 is defined in Assumption 3. Otherwise, if $\sigma^2_{r+1}(\boldsymbol{\Lambda}) + e_1 < \mathbf{D}^2_{\tilde{i}\tilde{i}} < \sigma_r^2(\boldsymbol{\Lambda}) - e_2$, this contradicts to Equation 74; see an illustration of this fact in Figure 6.

In this scenario, we consider the escaping direction $\theta^2_{\mathbf{Y}}$ to be $\mathbf{Y} - \mathbf{Y}^*\mathbf{Q}$. From the fact that $\mathbf{D}_{ii} \ge \mathbf{D}_{\tilde{i}\tilde{i}}$, we have

$$\begin{aligned}
\alpha^2\mu^2\sigma_r^6\left(\mathbf{Y}^*\right) / \left(4\kappa^*\right)^2 &\ge \sum_{i=1}^n \sum_{j=1}^n \left(\mathbf{D}^2_{ii} - \boldsymbol{\Sigma}_{jj}\right)^2 \mathbf{D}^2_{ii} \boldsymbol{C}^2_{ji} && \text{[Equation 61]} \\
&\ge \sum_{i=1}^n \sum_{j=r+1}^n \left|\sigma_j^2(\boldsymbol{\Lambda}) + e_2 - \sigma_r^2(\boldsymbol{\Lambda})\right|^2 \cdot \mathbf{D}^2_{ii} \boldsymbol{C}^2_{ji}
\end{aligned}$$

where Equation 75 and the first assumption in Assumption 3 guarantees the last inequality because e_2 is small with respect to $\sigma_r^2(\boldsymbol{\Lambda}) - \sigma^2_{r+1}(\boldsymbol{\Lambda})$. Therefore,

$$\sum_{i=1}^n \sum_{j=r+1}^n \mathbf{D}^2_{ii} \boldsymbol{C}^2_{ij} \le \frac{e_3^2}{\left|\sigma_r^2(\boldsymbol{\Lambda}) - e_2 - \sigma^2_{r+1}(\boldsymbol{\Lambda})\right|^2} \tag{76}$$

Figure 6: The value of $\mathbf{D}_{\tilde{i}\tilde{i}}$ must be close to some $\sigma_i(\mathbf{\Lambda})$ according to Equation 74. If $\mathbf{D}_{\tilde{i}\tilde{i}}^2 > \sigma_{r+1}^2 + e_1$, then we must have $\mathbf{D}_{\tilde{i}\tilde{i}}^2 \geq \sigma_r^2 - e_2$.

where e_3 is defined in Assumption 3. Recall that e_3 is small enough, guaranteed in Assumption 3. Also recall that $e_2 = \frac{e_1}{\sqrt{2}}$, which is guaranteed to be small enough as in the first assumption in Assumption 3, so $\sigma_r(\mathbf{\Lambda})^2 - e_3 - \sigma_{r+1}^2(\mathbf{\Lambda}) > 0$.

Denote $\mathbf{\Sigma}_{(r+1):n}$ to be a diagonal matrix with only $r+1^{\text{th}}$ to n^{th} eigenvalues of $\mathbf{\Sigma}$, then we have

$$\begin{aligned}
\langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \mathbf{Y}\mathbf{Y}^\top \rangle &= \langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \mathbf{U}\mathbf{D}^2\mathbf{U}^\top \rangle \\
&= \langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \overline{\mathbf{U}}\boldsymbol{C}\mathbf{D}^2\boldsymbol{C}^\top\overline{\mathbf{U}}^\top \rangle \\
&= \langle \mathbf{\Sigma}_{(r+1):n}, \boldsymbol{C}\mathbf{D}^2\boldsymbol{C}^\top \rangle \\
&\leq \sigma_{r+1}^2(\mathbf{\Lambda}) \sum_{j=r+1}^{n} \sum_i \boldsymbol{C}_{ij}^2 \mathbf{D}_{ii}^2 \\
&\leq \frac{e_3^2 \sigma_{r+1}^2(\mathbf{\Lambda})}{\left|\sigma_r^2(\mathbf{\Lambda}) - e_2 - \sigma_{r+1}^2(\mathbf{\Lambda})\right|^2}
\end{aligned} \tag{77}$$

where the last inequality follows from Equation 76. Equation 77 directly implies,

$$\begin{aligned}
\langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^\top \rangle &= \langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \mathbf{Y}\mathbf{Y}^\top \rangle \\
&\leq \frac{e_3^2 \sigma_{r+1}^2(\mathbf{\Lambda})}{\left|\sigma_r^2(\mathbf{\Lambda}) - e_2 - \sigma_{r+1}^2(\mathbf{\Lambda})\right|^2}
\end{aligned} \tag{78}$$

because $(\mathcal{A}_{\mathbf{n}} - \mathbf{X}^*)\mathbf{Y}^* = \mathbf{0}$ and $\theta_{\mathbf{Y}}^2 = \mathbf{Y} - \mathbf{Y}^*\boldsymbol{Q}$.

Recall $\mathbf{X}^* = \mathbf{Y}^*\mathbf{Y}^{*\top}$. A simple calculation yields

$$\mathbf{Y}(\theta_{\mathbf{Y}}^2)^\top - \mathbf{X}^* + \theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^\top = \mathbf{Y}(\theta_{\mathbf{Y}}^2)^\top + \theta_{\mathbf{Y}}^2\mathbf{Y}^\top \tag{79}$$

and by using Equation 51,

$$\begin{aligned}
\langle \overline{\operatorname{grad} H([\mathbf{Y}])}, \theta_{\mathbf{Y}}^2 \rangle &= \left\langle 2\left(\mathbf{Y}\mathbf{Y}^\top - \mathcal{A}_{\mathbf{n}}\right)\mathbf{Y}, \theta_{\mathbf{Y}}^2 \right\rangle \\
&= \left\langle 2(\mathbf{Y}\mathbf{Y}^\top - \mathcal{A}_{\mathbf{n}}), \theta_{\mathbf{Y}}^2\mathbf{Y}^\top \right\rangle \\
&= \left\langle \mathbf{Y}\mathbf{Y}^\top - \mathcal{A}_{\mathbf{n}}, \theta_{\mathbf{Y}}^2\mathbf{Y}^\top + \mathbf{Y}(\theta_{\mathbf{Y}}^2)^\top \right\rangle \qquad \text{[first argument is symmetric]} \\
&= \left\langle \mathbf{Y}\mathbf{Y}^\top - \mathcal{A}_{\mathbf{n}}, \theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^\top + \mathbf{Y}\mathbf{Y}^\top - \mathbf{X}^* \right\rangle.
\end{aligned} \tag{80}$$

where the last equality follows from Equation 79.

$$
\begin{aligned}
&\overline{\operatorname{Hess} H([\mathbf{Y}])}\left[\theta_{\mathbf{Y}}^2, \theta_{\mathbf{Y}}^2\right] = \left\|\mathbf{Y}(\theta_{\mathbf{Y}}^2)^\top + \theta_{\mathbf{Y}}^2\mathbf{Y}^\top\right\|_{\mathrm{F}}^2 + 2\left\langle \mathbf{Y}\mathbf{Y}^\top - \mathcal{A}_{\mathbf{n}}, \theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^\top\right\rangle && \text{[Equation 51]}\\
&= \left\|\mathbf{Y}\mathbf{Y}^\top - \mathbf{X}^* + \theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^\top\right\|_{\mathrm{F}}^2 + 2\left\langle \mathbf{Y}\mathbf{Y}^\top - \mathcal{A}_{\mathbf{n}}, \theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^\top\right\rangle && \text{[Equation 79]}\\
&= \left\|\theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^\top\right\|_{\mathrm{F}}^2 + \left\|\mathbf{Y}\mathbf{Y}^\top - \mathbf{X}^*\right\|_{\mathrm{F}}^2 + 4\left\langle \mathbf{Y}\mathbf{Y}^\top - \mathbf{X}^*, \theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^\top\right\rangle\\
&- 2\left\langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^\top\right\rangle\\
&= \left\|\theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^\top\right\|_{\mathrm{F}}^2 - 3\left\|\mathbf{Y}\mathbf{Y}^\top - \mathbf{X}^*\right\|_{\mathrm{F}}^2 + 4\left\langle \mathbf{Y}\mathbf{Y}^\top - \mathbf{X}^*, \mathbf{Y}\mathbf{Y}^\top - \mathbf{X}^* + \theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^\top\right\rangle\\
&- 2\left\langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^\top\right\rangle\\
&= \left\|\theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^\top\right\|_{\mathrm{F}}^2 - 3\left\|\mathbf{Y}\mathbf{Y}^\top - \mathbf{X}^*\right\|_{\mathrm{F}}^2 + 4\left\langle \mathbf{Y}\mathbf{Y}^\top - \mathcal{A}_{\mathbf{n}}, \mathbf{Y}\mathbf{Y}^\top - \mathbf{X}^* + \theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^\top\right\rangle\\
&+ 2\left\langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^\top\right\rangle + 4\left\langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \mathbf{Y}\mathbf{Y}^\top - \mathbf{X}^*\right\rangle\\
&= \left\|\theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^\top\right\|_{\mathrm{F}}^2 - 3\left\|\mathbf{Y}\mathbf{Y}^\top - \mathbf{X}^*\right\|_{\mathrm{F}}^2 + 4\left\langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \mathbf{Y}\mathbf{Y}^\top - \mathbf{X}^*\right\rangle\\
&+ 2\left\langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^\top\right\rangle + 4\left\langle \overline{\operatorname{grad} H([\mathbf{Y}])}, \theta_{\mathbf{Y}}^2\right\rangle && \text{[Equation 80]}
\end{aligned}
$$

This decomposes $\overline{H([\mathbf{Y}])}\left[\theta_{\mathbf{Y}}^2, \theta_{\mathbf{Y}}^2\right]$ into 2 parts, which will be bounded separately.

First, for $\left\|\theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^\top\right\|_{\mathrm{F}}^2 - 3\left\|\mathbf{Y}\mathbf{Y}^\top - \mathbf{X}^*\right\|_{\mathrm{F}}^2 + 2\left\langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^\top\right\rangle + 4\left\langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \mathbf{Y}\mathbf{Y}^\top - \mathbf{X}^*\right\rangle$, we have

$$
\begin{aligned}
&\left\|\theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^\top\right\|_{\mathrm{F}}^2 - 3\left\|\mathbf{Y}\mathbf{Y}^\top - \mathbf{X}^*\right\|_{\mathrm{F}}^2 + 2\left\langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^\top\right\rangle\\
&+ 4\left\langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \mathbf{Y}\mathbf{Y}^\top - \mathbf{X}^*\right\rangle\\
&\leq -\left\|\mathbf{Y}\mathbf{Y}^\top - \mathbf{X}^*\right\|_{\mathrm{F}}^2 + 2\left\langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^\top\right\rangle\\
&+ 4\left\langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \mathbf{Y}\mathbf{Y}^\top - \mathbf{X}^*\right\rangle && \text{[Equation 54]}\\
&= -\left\|\mathbf{Y}\mathbf{Y}^\top - \mathbf{X}^*\right\|_{\mathrm{F}}^2 + 2\left\langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^\top\right\rangle + 4\left\langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \mathbf{Y}\mathbf{Y}^\top\right\rangle && [\langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \mathbf{X}^*\rangle = 0]\\
&\leq -\left\|\mathbf{Y}\mathbf{Y}^\top - \mathbf{X}^*\right\|_{\mathrm{F}}^2 + 6\frac{e_3^2\sigma_{r+1}^2(\mathbf{\Lambda})}{\left|\sigma_r^2(\mathbf{\Lambda}) - e_2 + \sigma_{r+1}^2(\mathbf{\Lambda})\right|^2} && \text{[Equation 77,Equation 78]}\\
&\leq -2(\sqrt{2}-1)\sigma_r^2\left(\mathbf{Y}^*\right)\left\|\theta_{\mathbf{Y}}^2\right\|_{\mathrm{F}}^2 + 6\frac{e_3^2\sigma_{r+1}^2(\mathbf{\Lambda})}{\left|\sigma_r^2(\mathbf{\Lambda}) - e_2 + \sigma_{r+1}^2(\mathbf{\Lambda})\right|^2} && \text{[Equation 53]}
\end{aligned}
$$

Second, for $\left\langle \overline{\operatorname{grad} H([\mathbf{Y}])}, \theta_{\mathbf{Y}}^2\right\rangle$,

$$
\begin{aligned}
\left\langle \overline{\operatorname{grad} H([\mathbf{Y}])}, \theta_{\mathbf{Y}}^2\right\rangle \leq& \|\overline{\operatorname{grad} H([\mathbf{Y}])}\|_{\mathrm{F}}\left\|\theta_{\mathbf{Y}}^2\right\|_{\mathrm{F}}\\
\leq& \alpha\sigma_r^2\left(\mathbf{Y}^*\right)\left\|\theta_{\mathbf{Y}}^2\right\|_{\mathrm{F}}^2
\end{aligned}
$$

where the last inequality is because $\|\overline{\operatorname{grad} H([\mathbf{Y}])}\|_{\mathrm{F}} \leqslant \alpha\mu\sigma_r^3\left(\mathbf{Y}^*\right)/\left(4\kappa^*\right)$. According to the definition of $\mathcal{R}_2$ in Equation 21, $\mathbf{Y} \in \mathcal{R}_2$ also implies $d\left([\mathbf{Y}], [\mathbf{Y}^*]\right) > \mu\sigma_r\left(\mathbf{Y}^*\right)/\kappa^*$, then

$$
\|\overline{\operatorname{grad} H([\mathbf{Y}])}\|_{\mathrm{F}} \leqslant \alpha d\left([\mathbf{Y}], [\mathbf{Y}^*]\right)\sigma_r^2\left(\mathbf{Y}^*\right)/4 = \alpha\left\|\theta_{\mathbf{Y}}^2\right\|_{\mathrm{F}}\sigma_r^2\left(\mathbf{Y}^*\right)/4
$$

By combining the above three inequalities, we have

$$\begin{aligned}
&\overline{\operatorname{Hess} H([\mathbf{Y}])}\left[\theta_{\mathbf{Y}}^2, \theta_{\mathbf{Y}}^2\right] = \left\|\theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^\top\right\|_{\mathrm{F}}^2 - 3\left\|\mathbf{Y}\mathbf{Y}^\top - \mathbf{X}^*\right\|_{\mathrm{F}}^2 \\
&+ 4\left\langle\overline{\operatorname{grad} H([\mathbf{Y}])}, \theta_{\mathbf{Y}}^2\right\rangle + 2\left\langle\mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^\top\right\rangle + 4\left\langle\mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \mathbf{Y}\mathbf{Y}^\top - \mathbf{X}^*\right\rangle \\
&\leq (\alpha - 2(\sqrt{2}-1))\sigma_r^2\left(\mathbf{Y}^*\right)\left\|\theta_{\mathbf{Y}}^2\right\|_{\mathrm{F}}^2 + 6\frac{e_3^2\sigma_{r+1}^2(\mathbf{\Lambda})}{\left|\sigma_r^2(\mathbf{\Lambda}) - e_2 - \sigma_{r+1}^2(\mathbf{\Lambda})\right|^2} \\
&\leq \left((\alpha - 2(\sqrt{2}-1))\sigma_r^2\left(\mathbf{Y}^*\right) + 6\frac{\alpha^2\sigma_r^4\left(\mathbf{Y}^*\right)\sigma_{r+1}^2(\mathbf{\Lambda})/16}{\left|\sigma_r^2(\mathbf{\Lambda}) - e_2 - \sigma_{r+1}^2(\mathbf{\Lambda})\right|^2}\right)\left\|\theta_{\mathbf{Y}}^2\right\|_{\mathrm{F}}^2
\end{aligned}$$

where the last inequality follows from $\mu\sigma_r\left(\mathbf{Y}^*\right)/\kappa^* \leq d\left([\mathbf{Y}],[\mathbf{Y}^*]\right) = \|\theta_{\mathbf{Y}}\|_{\mathrm{F}}$ and the definition of e_3 in Assumption 3.

Finally, according to the third assumption in Assumption 3, one can guarantee the right-hand side of this bound is negative, which implies that $\theta_{\mathbf{Y}}^2$ is the escaping direction in this scenario.

Combining all the discussion, this finishes the proof of this theorem. □

Remark 14. *The eigengap assumption is crucial in discussing the three regions of the minimum singular value of* $\mathbf{Y}$ *for Theorem 4.2. Without this eigengap assumption and under the current quotient geometry, the third regime cannot lead to a negative eigenvalue of Hessian matrix because any span on the eigenspace are global solutions. To relax eigengap assumption, an alternative quotient geometry needs to be considered.*

Finally, we look at the last main result. Theorem 4.3 guarantees that when $\mathbf{Y} \in \mathcal{R}_3$, the magnitude of the Riemannian gradient descent is large. The proof of Theorem 4.3 directly follows from the proof of [56] without any modification. Hence, we do not repeat it here. Notice that $\mathbf{Y} \in \mathcal{R}_3$ does not require Assumption 2 because $\mathcal{R}_3$ describes the case that $[\mathbf{Y}]$ is far away from the FOSP.

Theorem 4.3 ((Regions with Large Riemannian Gradient of Equation 7)**.**

1. $\|\overline{\operatorname{grad} H([\mathbf{Y}])}\|_{\mathrm{F}} > \alpha\mu\sigma_r^3\left(\mathbf{Y}^*\right)/\left(4\kappa^*\right), \forall \mathbf{Y} \in \mathcal{R}_3'$;
2. $\|\overline{\operatorname{grad} H([\mathbf{Y}])}\|_{\mathrm{F}} \geqslant 2\left(\|\mathbf{Y}\|^3 - \|\mathbf{Y}\|\left\|\mathbf{Y}^*\right\|^2\right) > 2\left(\beta^3 - \beta\right)\left\|\mathbf{Y}^*\right\|^3, \quad \forall \mathbf{Y} \in \mathcal{R}_3''$;
3. $\langle\overline{\operatorname{grad} H([\mathbf{Y}])}, \mathbf{Y}\rangle > 2(1-1/\gamma)\left\|\mathbf{Y}\mathbf{Y}^\top\right\|_{\mathrm{F}}^2, \quad \forall \mathbf{Y} \in \mathcal{R}_3'''$.

In particular, if $\beta > 1$ *and* $\gamma > 1$*, we have the Riemannian gradient of* $H([\mathbf{Y}])$ *has large magnitude in all regions* $\mathcal{R}_3', \mathcal{R}_3''$ *and* $\mathcal{R}_3'''$.