Spectral Neural Networks: Approximation Theory and Optimization Landscape

Chenghui Li [*][1], Rishi Sonthalia [†][2], and Nicolás García Trillos [‡][1]

[1] Department of Statistics, University of Wisconsin Madison, Madison, Wisconsin 53706, USA.
[2] Department of Mathematics, Boston College, Boston, MA 02109, USA.

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 [3, 8, 64], clustering [56, 72], dimensionality reduction [7, 20], and contrastive learning [34]. 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

[*] cli539@stat.wisc.edu
[†] sonthal@bc.edu
[‡] Corresponding author. garciatrillo@wisc.edu

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:

$$\min_\theta \|f_\theta(\mathcal{X}_n) - \mathbf{v}\|^2, \tag{1.1}$$

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 positive semidefinite (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 [61] and spectral neural network [34] 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 contrastive loss function

$$\min_{\theta \in \Theta} L(\theta) \overset{\text{def}}{=} \ell(\mathbf{Y}_\theta), \quad \ell(\mathbf{Y}) \overset{\text{def}}{=} \left\|\mathbf{Y}\mathbf{Y}^\top - \mathcal{A}_\mathbf{n}\right\|_\mathrm{F}^2, \quad \mathbf{Y} \in \mathbb{R}^{n \times r} \tag{1.2}$$

through first-order optimization methods. Fig. 1.1 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
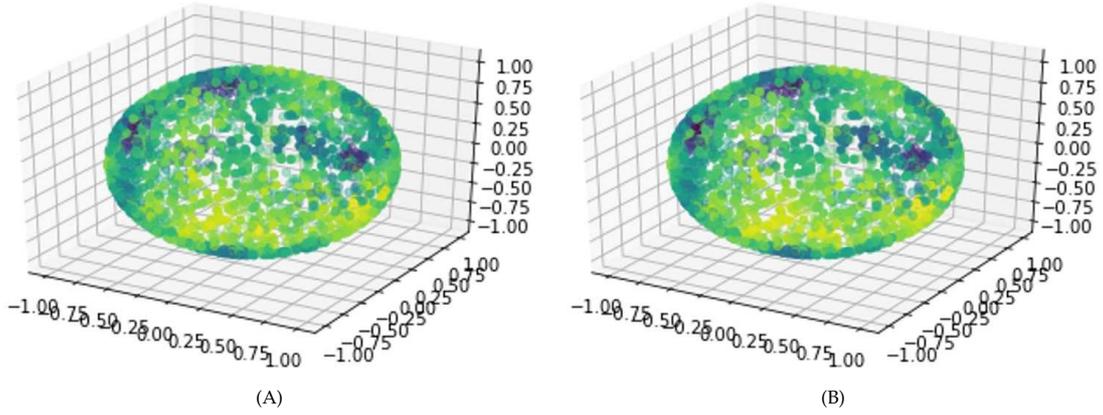

Figure 1.1: (B) shows the first eigenvector for the Laplacian of a proximity graph from data points sampled from S^2 obtained using an eigensolver. (A) shows the same eigenvector but obtained using SNN. The difference between the two figures is minor, showing that the neural network learns the eigenvector of the graph Laplacian well. See details in Appendix B.1.

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 1.1. 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 G^τ with entries

$$G_{ij}^\tau = \eta\left(\frac{\|x_i - x_j\|}{\tau}\right), \tag{1.3}$$

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 non-increasing, non-negative function such that $\eta(1) = 0, \eta(0) = 1$. In short, G^τ 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:

$$c_\eta \stackrel{\text{def}}{=} \int_{\mathbb{R}^m} |y_1|^2 \eta(\|y\|) dy, \tag{1.4}$$

where y_1 is the first coordinate of y. c_η is used to scale the weighted Laplace-Beltrami operator on $\mathcal{M}$. Notice that multiplication of a matrix by this factor does not change its eigenvectors.

From G^τ we define the adjacency matrix $\mathcal{A}_{\mathbf{n}}$, which we will use within Eq. (1.2), by

$$\mathcal{A}_{\mathbf{n}} \stackrel{\text{def}}{=} \mathbf{D}_G^{-\frac{1}{2}} G \mathbf{D}_G^{-\frac{1}{2}} + a\mathbf{I}, \tag{1.5}$$

where

$$(\mathbf{D}_G)_{ii} = \sum_{j=1}^n G_{ij}$$

is the degree matrix associated to G, and $a > 1$ is a fixed quantity that we introduce to guarantee that $\mathcal{A}_{\mathbf{n}}$ is positive definite. Here, we distance ourselves slightly from the choice made in the original SNN paper [34], where $\mathcal{A}_{\mathbf{n}}$ is taken to be G itself, and instead consider a normalized version. This is due to the following key properties satisfied by our choice of $\mathcal{A}_{\mathbf{n}}$.

Proposition 1.1.   *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 [72])

$$\Delta_n \overset{\text{def}}{=} \mathbf{I} - \mathbf{D}_G^{-\frac{1}{2}} \mathbf{G} \mathbf{D}_G^{-\frac{1}{2}}. \tag{1.6}$$

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

$$\Delta_\rho f \overset{\text{def}}{=} -\frac{1}{\rho^{3/2}} \text{div}\left(\rho^2 \nabla \left(\frac{f}{\sqrt{\rho}}\right)\right) \tag{1.7}$$

for all smooth enough $f : \mathcal{M} \to \mathbb{R}$; see [30, Section 1.4]. 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

$$0 = \lambda_1^{\mathcal{M}} \le \lambda_2^{\mathcal{M}} \le \cdots.$$

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

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 Eq. 1.2 to build an approximating neural network?

Q3 What can be said about the landscape of the objective function in Eq. (1.2)?

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 a 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 3.2. 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 Eq. (1.2) 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 Eq. (1.2) approximates well 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 7.

1.1   Related work

Other NN-based Eigensolvers. Other types of NN-based Eigensolvers have been considered in [22, 58]. [58] 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. [22] takes a similar approach as [58] 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 [12–14, 25, 27, 28, 48, 69, 75]. In this paper, we use adaptations of the results in [14] to infer that, with very high probability, the eigenvectors of the normalized graph Laplacian matrix Δ_n defined in Eq. (1.6) 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 [16–18, 36]. Theoretical properties of contrastive representation learning were first studied by [4, 41, 71] where they assumed conditional independence. [34] relaxes the conditional independence assumption by imposing the manifold assumption. With the spectral contrastive loss Eq. (1.2) crucially in use, [34] provides an error bound for downstream tasks. In this work, we analyze how a neural network can approximate and optimize the spectral loss function in Eq. (1.2), which is the pertaining step of [34].

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 [50, 57]. Specifically, [62] looks at the problem of Hölder continuous functions on the unit cube, [63, 76] studies the class of continuous functions on the unit cube, and [34, 57, 60] 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 [77]. 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. [59] introduced physics informed neural networks as a method for solving PDEs using neural networks. Specifically, [9, 59, 74] use neural networks to parameterize the solution of a PDE which is trained by optimiz-

ing a loss function that is designed to be minimized when the equation is satisfied. Other works such as [2, 9, 33, 78] 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 [40, 46, 47] 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 Eq. (1.2). 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 [5,11,32,55,67,70], and generalization for one layer networks [6,24,35,39,45,65]. Of relevance are also other works in the literature studying optimization problems very closely related to Eq. (2.1). For example, in [43, Section 3], there is a landscape analysis for Eq. (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. On the other hand, the case studied in [19, Section 3] 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. [51] describes the global optimization landscape of Eq. (2.1) 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 that 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. In Section 2.1, we discuss the rotation invariance of the loss function ℓ, which makes ℓ into a non-convex function, and reformulate the ambient optimization problem as one over a suitable quotient manifold. In Section 2.2, we establish some preliminary properties of ℓ that will allow us to characterize its first order stationary points. In Section 2.3, we discuss some of the spectral convergence results in the literature of graph Laplacians that we use in the sequel to guarantee that SNNs can recover the eigenfunctions of the Laplace-Beltrami operator defined in Eq. (1.7). In Section 2.4, we discuss some useful regularity estimates of the eigenvectors of Δ_n which turn out to be important for our main approximation results for SNNs in later sections.

2.1 Rotational invariance

Recall the ambient optimization problem defined in Eq. (1.2) as

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

Suppose $\mathbf{Y}$ is a stationary point of Eq. (2.1). Then $\mathbf{Y}O$ is also a stationary point for any $r \times r$ orthogonal matrix $O \in \mathbb{O}_r$. This implies that the loss function ℓ is non-convex in any neighborhood of a stationary point [44]. 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., $\mathbf{Y} \to \mathbf{Y}O$, by defining the equivalence classes $[\mathbf{Y}] = \{\mathbf{Y}O : O \in \mathbb{O}_r\}$. From [42], $\mathcal{N}_{r+}^n \stackrel{\text{def}}{=} \overline{\mathcal{N}}_{r+}^n / \mathbb{O}_r$ is a quotient manifold of $\overline{\mathcal{N}}_{r+}^n$; for a detailed introduction to Riemannian optimization see [10]. Since ℓ is invariant within each of the equivalence classes of $\overline{\mathcal{N}}_{r+}^n$, we can reformulate our optimization problem as one on the quotient manifold $\mathcal{N}_{r+}^n$

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

which can be approached via Riemannian first-order methods. In this case, however, the Riemannian gradient descent of H is equivalent to the standard gradient descent of ℓ in the Euclidean geometry[1]. 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 quotient space problem Eq. (2.2).

2.2 First order stationary points

Since $\mathcal{A}_{\mathbf{n}}$ is a PSD matrix, the Eckart-Young-Mirsky theorem (see [26]) implies that the global optimizers of Eq. (2.1) are the matrices $\mathbf{Y}$ of the form $\mathbf{Y} = \mathbf{Y}^*O$, where $O \in \mathbb{O}_r$ and

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

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:

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

---

[1]There are two interpretations to this statement. First, note that, as we will soon see, the expression for the gradient of the function H at a given $[Y]$ in Eq. (4.1) coincides with the expression for the standard Euclidean gradient of $\ell/2$ at Y. This implies that the continuous-time gradient flow of H with respect to the quotient geometry can be written as $[Y_t]$, where Y_t is the continuous time gradient flow of $\ell/2$ with respect to the Euclidean geometry. On the other hand, because geodesics on the quotient manifold are images, through the quotient map, of straight lines in the Euclidean space (see [54, Theorem 2.2]), it follows that the updates in the Riemannian gradient descent scheme, with retraction map the exponential map, can be written as $[Y_{k+1}] = [Y_k - \tau\nabla\ell(Y_k)/2]$.

In this way we guarantee that

$$\|u_l\|_{L^2(\mathcal{X}_n)}^2 \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \left(u_l(x_i)\right)^2 = 1, \tag{2.3}$$

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

$$\mathbf{Y}^* = \begin{bmatrix} | & & | \\ \sqrt{\dfrac{\lambda_1(\mathcal{A}_\mathbf{n})}{n}}u_1 & \cdots & \sqrt{\dfrac{\lambda_r(\mathcal{A}_\mathbf{n})}{n}}u_r \\ | & & | \end{bmatrix}. \tag{2.4}$$

Remark 2.1. As discussed in Remark 2.3 below, under Assumptions 1.1 we can assume that all the $\lambda_s(\mathcal{A}_\mathbf{n})$ appearing in the above expression are quantities of order 1.

Next, we need to understand the other first-order stationary points (FOSP) of Eq. (2.2). We use singular value decomposition (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 Eq. 2.2). *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 $[n]$ with cardinality r we have that $[\overline{\mathbf{U}}_S \mathbf{\Lambda}_S]$ is a Riemannian FOSP of Eq.* (2.2)*. Further, these are the only Riemannian FOSPs.*

Remark 2.2. Although our optimization of ℓ does not include an explicit orthogonality constraint, Theorem 2.1 states that all first order critical points of ℓ have orthogonal columns. Thus, if we optimize ℓ through a first order method, in practice we should expect as output a matrix whose columns are nearly orthogonal. For this reason, we expect, under the right assumptions on the size of the neural network and sufficient time for convergence, the SNN solution to be useful in downstream unsupervised learning tasks such as clustering and dimensionality reduction.

Theorem 2.1 shows that linear combinations of eigenvectors can be used to construct Riemannian first-order stationary points (FOSP) for the objective in Eq. (2.2). This theorem also shows that there are multiple FOSPs, which is quite different from the setting studied in [51]. In general, gradient descent is known to converge to a FOSP. However, as our careful landscape analysis in Section 4 shows, an FOSP for our problem is either a strict saddle or the global minimizer, and we thus expect that a first order method initialized randomly will converge toward the global minimizer.

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 [13, Theorem 2.4] – which considers the unnormalized graph Laplacian case – relates the eigenvalues of Δ_n and Δ_ρ.

Theorem 2.2 (Convergence of Eigenvalues of Graph Laplacian; Adapted from [13, Theorem 2.4]). *Let* $l \in \mathbb{N}$ *be fixed. Under Assumptions* 1.1, *with probability at least* $1 - Cn \times \exp(-cn\tau^{m+4})$ *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 factor that depends on the kernel* η *determining the graph weights (see Eq.* (1.4)*). We also recall that* m *denotes the intrinsic dimension of the manifold* $\mathcal{M}$.

Remark 2.3. From Theorem 2.2 and Eq. (C.1) 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 $\|\mathbf{Y} - \mathbf{Y}^*\|_F = O(\sqrt{r}\tau)$.

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 [14] for the case of unnormalized graph Laplacians. A straightforward adaptation of [14, Theorem 2.6] implies the following.

Theorem 2.3 (Almost $C^{0,1}$ Convergence of Graph Laplacian Eigenvectors; Adapted from [14, Theorem 2.6]). *Let* $r \in \mathbb{N}$ *be fixed and let* $u_1, \ldots, u_r$ *be normalized eigenvectors of* Δ_n *as in Eq.* (2.3). *Under Assumptions* 1.1, *with probability at least* $1 - C\tau^{-6m} \exp(-cn\tau^{m+4})$ *over the sampling of the* x_i, *we have*

$$\|f_s - u_s\|_{L^\infty(\mathcal{X}_n)} + [f_s - u_s]_{\tau,\mathcal{X}_n} \leq C_r \tau, \quad \forall s = 1, \ldots, r, \tag{2.5}$$

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*

$$\|v\|_{L^\infty(\mathcal{X}_n)} \overset{def}{=} \max_{x_i \in \mathcal{X}_n} |v(x_i)|,$$

$d_{\mathcal{M}}(\cdot, \cdot)$ *denotes the geodesic distance on* $\mathcal{M}$, *and* $[\cdot]_{\tau,\mathcal{X}_n}$ *is the seminorm*

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

Remark 2.4. In the statement of Theorem 2.3, the probabilistic bound is only due to the randomness in the data samples and has nothing to do with neural networks. Specifically, we need to have sufficient data to guarantee a certain regularity of eigenvectors of Δ_n that we later use to establish our approximation results. Note that when considering i.i.d. data on $\mathcal{M}$, to even guarantee that a τ-proximity graph over the data is connected requires the quantity $n\tau^m$ to be sufficiently large. We observe that this condition is not too different from the implicit requirement on n and τ that appears in Theorem 2.3.

2.4 Regularity of the eigenvectors of Δ_n

An essential corollary of Theorem 2.3 is the following set of regularity estimates satisfied by eigenvectors of the normalized graph Laplacian Δ_n.

Corollary 2.1. *Under the same setting, notation, and assumptions as in Theorem 2.3, the functions u_s satisfy*

$$|u_s(x_i) - u_s(x_j)| \leq L_s\big(d_{\mathcal{M}}(x_i, x_j) + \tau^2\big), \quad \forall\, x_i, x_j \in \mathcal{X}_n \tag{2.6}$$

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

Proof. From Eq. (2.5) we have

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

It follows from the triangle inequality that

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

In the above, the second inequality follows from inequality (2.5) 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'. $\qquad\square$

Remark 2.5. We observe that the τ^2 term on the right hand side of Eq. (2.6) is strictly better than the term that appears in the explicit regularity estimates in [14, Remark 2.4], 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 3.1. *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 $\omega > 0$, does there exist $f_\theta \in \mathcal{F}$ such that f_θ nearly interpolates the data*

$$\max_{i=1,\dots,n} \|y_i - f_\theta(x_i)\| = O(\omega^2),$$

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

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 $\|f - f_\theta\|_{L^\infty(\mathcal{K})} < \omega$. 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 [6]. 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 [53]. Building on this, researchers have developed an interest in understanding near interpolators [73]. These are networks that do not exactly fit the data but nearly fit the data.

In this context, Problem 3.1 can be interpreted as asking for the existence of a near interpolator that benignly overfits the data. Problem 3.1 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 the 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 multilayer ReLU neural networks. To be precise, our neural networks are parameterized functions $f : \mathbb{R}^d \to \mathbb{R}^r$ of the form

$$f(\mathbf{x}) = \mathbf{W}_L \cdot \text{ReLU}\left(\mathbf{W}_{L-1} \cdots \text{ReLU}(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1) \cdots + \mathbf{b}_{L-1}\right) + \mathbf{b}_L, \quad \mathbf{x} \in \mathbb{R}^d. \tag{3.1}$$

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

$$\mathcal{F}(r, \kappa, L, p, N) = \left\{ f \mid f(\mathbf{x}) \text{ has the form 3.1, where} \right.$$

$$\mathbf{W}_l \in \mathbb{R}^{p \times p}, \ \mathbf{b}_l \in \mathbb{R}^p \text{ for } l = 2, \ldots, L-1,$$

$$\mathbf{W}_1 \in \mathbb{R}^{p \times d}, \ \mathbf{b}_1 \in \mathbb{R}^p, \ \mathbf{W}_L \in \mathbb{R}^{r \times p}, \ \mathbf{b}_L \in \mathbb{R}^r,$$

$$\|\mathbf{W}_l\|_{\infty,\infty} \leq \kappa, \ \|\mathbf{b}_l\|_{\infty} \leq \kappa \text{ for } l = 1, \ldots, L,$$

$$\left. \sum_{l=1}^{L} \|W_l\|_0 + \|\mathbf{b}_l\|_0 \leq N \right\}, \tag{3.2}$$

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 conve-

nience, 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 Eq. (3.1).

3.2 Spectral approximation with multilayer ReLU NNs

In this section, we answer question Q1 by proving Theorem 3.1 and Corollary 3.2, which are stated below. Specifically, we provide upper bounds on the size of a multilayer ReLU neural network that guarantee that with high probability there exists a neural network that nearly interpolates the data while approximating the 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 of the neural network.

We start with the following lemma that we later combine with our results from Sections 2.3 and 2.4. The proof is presented in Appendix E.

Lemma 3.1. *Let $u : \mathcal{X}_n \to \mathbb{R}$ be a function satisfying*

$$|u(x) - u(\tilde{x})| \leq L\big(d_{\mathcal{M}}(x, \tilde{x}) + \tau^2\big), \quad \forall\, x, \tilde{x} \in \mathcal{X}_n \tag{3.3}$$

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

$$\|u - \tilde{g}\|_{L^\infty(\mathcal{X}_n)} \leq 5L\tau^2. \tag{3.4}$$

Remark 3.1. Lemma 3.1 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 3.1 and Theorem 2.3, 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 1.1, there are constants c, C that depend on $\mathcal{M}, \rho$, and the embedding dimension r such that with probability at least*

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

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 Eq. (3.2)) 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: $L \leq C(\log(1/\varepsilon) + \log d)$, and its width, p, satisfies $p \leq C(\varepsilon^{-m} + d)$.*

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 $\kappa \leq C/n^{1/(2L)}$.

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 does not decay too quickly (i.e., we need $\tau \gg n^{-1/(m+4)}$), 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 for the results to hold. Note that these rates do not depend on the connectivity parameter τ.

Remark 3.2. 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 3.1. *For f_θ as in Theorem* 3.1 *we have*

$$\sqrt{n}\left\|\mathbf{Y}_\theta\mathbf{Y}_\theta^\top - \mathbf{Y}^*\mathbf{Y}^{*\top}\right\|_{\infty,\infty} \leq C_r(\varepsilon + \tau^2), \tag{3.5}$$

and thus also

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

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 [15, Theorem 1] to approximate the eigenfunctions $f_1, \ldots, f_r$ and then the convergence result for Δ_n to Δ_ρ (such as [13, Theorem 2.4]), we would get that $\|\mathbf{Y}_\theta - \mathbf{Y}^*\|_{\mathrm{F}} = O(\sqrt{r}\tau)$. However, we know from Remark 2.3, 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 2.1 and Remark 2.5 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 error bounds for the approximation of 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 3.2. *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.

If $\varepsilon = \tau^2$, then with Theorem 3.1 and Corollary 3.2 we thus provide an answer to Problem 3.1. Specifically, letting $\omega = \tau$, we see that we can approximate eigenvectors of the graph Laplacian over our training data with an error in the order of $O(\omega^2)$ (Theorem 3.1) and we can approximate the eigenfunctions of the Laplace-Beltrami operator on the manifold with an error in the order of $O(\omega)$ (Corollary 3.2). We notice that, while one could use existing memorization results (e.g., [77, Theorem 3.1]) 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 3.2. 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.

We present an experiment in Section 6.1 that shows that a sufficiently large network does learn the true eigenfunctions of the underlying manifold. We observe that the empirical convergence rate (in n) is faster than the rate in Corollary 3.2. We believe that a more careful analysis using some ideas as in [29] and other recent literature can lead to a theoretical quadratic convergence rate in τ.

On the other hand, we see that the size of neural network is the same as the size for universal approximation. That is, once we can guarantee the existence of a network that approximates sufficiently well the eigenfunctions, we can then guarantee the existence of a network that simultaneously approximates the eigenvectors of the graph Laplacian associated to the given data and the eigenfunctions of the underlying Laplace-Beltrami operator. The size of τ does not affect the size of the network needed to guarantee that these estimates hold.

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 we measure the distance to optimal solutions via the convex function $\|\mathbf{Y}_\theta - \mathbf{Y}^*\|_F$. However, this is not the loss function SNNs use to train the network. Instead, the nonconvex loss function $\|\mathbf{Y}_\theta \mathbf{Y}_\theta^T - \mathcal{A}_n\|_F$ 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*

$$f_{\theta^*} \in \underset{f_\theta \in \mathcal{F}(r,\kappa,L,p,N)}{\arg\min} \left\| \mathbf{Y}_\theta \mathbf{Y}_\theta^\top - \mathcal{A}_\mathbf{n} \right\|_F^2. \tag{3.6}$$

Under Assumptions 1.1, there are constants c, C that depend on $\mathcal{M}, \rho$, and the embedding dimension r such that, with probability at least $1 - C\tau^{-6m} \exp(-cn\tau^{m+4})$, for every $\tilde{\delta} \in (0, c)$ (i.e., $\tilde{\delta}$ sufficiently small) and for

$$\kappa = \frac{C}{n^{1/(2L)}}, \quad L = C\left(\log\frac{1}{\tilde{\delta}\tau} + \log d\right), \quad p = C((\tilde{\delta}\tau)^{-m} + d),$$

we have

$$\min_{O \in \mathbb{O}_r} \|\mathbf{Y}_{\theta^*} - \mathbf{Y}^* O\|_F \le C\tau(\tilde{\delta} + \tau). \tag{3.7}$$

In the above theorem $\tilde{\delta}\tau$ can be understood as ε in Theorem 3.1 and Corollary 3.2 with an extra assumption $\tilde{\delta} \le c\tau$ that guarantees that the solution found by minimizing L is energetically better than any saddle point of the ambient loss function ℓ.

Remark 3.3. Eq. (3.7) says that $\mathbf{Y}_{\theta^*}$ approximates a minimizer of the ambient problem (2.1) 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, as 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 2.3 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., $\|\mathbf{Y}^*\mathbf{Y}^{*\top} - \mathcal{A}_\mathbf{n}\|_F^2 - \|\mathbf{Y}_\theta\mathbf{Y}_\theta^\top - \mathcal{A}_\mathbf{n}\|_F^2$) is of order $\tau^2 C_r(\tilde{\delta} + \tau)^2$ (Corollary F.2), 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 2.3. 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 as shown in Section 2.1). 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 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 2.1 under a suitable Riemannian quotient geometry [1,10].

To analyze the landscape for Eq. (2.2), we need expressions for the Riemannian gradient, the Riemannian Hessian, as well as the geodesic distance d on this quotient manifold. By [51, Lemma 2], we have that

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

and from [51, Lemma 3], we have that

$$\begin{aligned}
\overline{\operatorname{grad} H([\mathbf{Y}])} &= 2(\mathbf{Y}\mathbf{Y}^\top - \mathcal{A}_{\mathbf{n}})\mathbf{Y}, \\
\overline{\operatorname{Hess} H([\mathbf{Y}])}[\theta_{\mathbf{Y}}, \theta_{\mathbf{Y}}] &= \|\mathbf{Y}\theta_{\mathbf{Y}}^\top + \theta_{\mathbf{Y}}\mathbf{Y}^\top\|_{\mathrm{F}}^2 + 2\langle \mathbf{Y}\mathbf{Y}^\top - \mathcal{A}_{\mathbf{n}}, \theta_{\mathbf{Y}}\theta_{\mathbf{Y}}^\top \rangle.
\end{aligned} \tag{4.1}$$

Finally, by the classical theory on low-rank approximation (Eckart-Young-Mirsky theorem [26]), $[\mathbf{Y}^*]$ is the unique global minimizer of Eq. (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 4.1 (Eigengap). $\sigma_{r+1}(\mathcal{A}_{\mathbf{n}})$ is strictly smaller than $\sigma_r(\mathcal{A}_{\mathbf{n}})$.

See Remark G.1 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):

$$\mathcal{R}_1 \stackrel{\text{def}}{=} \left\{ \mathbf{Y} \in \mathbb{R}_*^{n \times r} \mid d([\mathbf{Y}], [\mathbf{Y}^*]) \leqslant \frac{\mu \sigma_r(\mathbf{Y}^*)}{\kappa^*} \right\}, \tag{4.2a}$$

$$\mathcal{R}_2 \stackrel{\text{def}}{=} \left\{ \mathbf{Y} \in \mathbb{R}_*^{n \times r} \mid d([\mathbf{Y}], [\mathbf{Y}^*]) > \frac{\mu \sigma_r(\mathbf{Y}^*)}{\kappa^*}, \ \|\overline{\operatorname{grad} H([\mathbf{Y}])}\|_{\mathrm{F}} \leqslant \frac{\alpha \mu \sigma_r^3(\mathbf{Y}^*)}{4\kappa^*}, \right.$$
$$\left. \|\mathbf{Y}\| \leqslant \beta \|\mathbf{Y}^*\|, \ \|\mathbf{Y}\mathbf{Y}^\top\|_{\mathrm{F}} \leqslant \gamma \|\mathbf{Y}^* \mathbf{Y}^{*\top}\|_{\mathrm{F}} \right\}, \tag{4.2b}$$

$$\mathcal{R}_3' \stackrel{\text{def}}{=} \left\{ \mathbf{Y} \in \mathbb{R}_*^{n \times r} \mid \|\overline{\operatorname{grad} H([\mathbf{Y}])}\|_{\mathrm{F}} > \frac{\alpha \mu \sigma_r^3(\mathbf{Y}^*)}{4\kappa^*}, \ \|\mathbf{Y}\| \leqslant \beta \|\mathbf{Y}^*\|, \right.$$
$$\left. \|\mathbf{Y}\mathbf{Y}^\top\|_{\mathrm{F}} \leqslant \gamma \|\mathbf{Y}^* \mathbf{Y}^{*\top}\|_{\mathrm{F}} \right\}, \tag{4.2c}$$

$$\mathcal{R}_3'' \stackrel{\text{def}}{=} \left\{ \mathbf{Y} \in \mathbb{R}_*^{n \times r} \mid \|\mathbf{Y}\| > \beta \|\mathbf{Y}^*\|, \ \|\mathbf{Y}\mathbf{Y}^\top\|_{\mathrm{F}} \leqslant \gamma \|\mathbf{Y}^* \mathbf{Y}^{*\top}\|_{\mathrm{F}} \right\}, \tag{4.2d}$$

$$\mathcal{R}_3''' \stackrel{\text{def}}{=} \left\{ \mathbf{Y} \in \mathbb{R}_*^{n \times r} \mid \|\mathbf{Y}\mathbf{Y}^\top\|_F > \gamma \|\mathbf{Y}^*\mathbf{Y}^{*\top}\|_F \right\}. \tag{4.2e}$$

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 Eq. (2.2)). *Suppose* $0 \leqslant \mu \leqslant \kappa^*/3$. *Given that Assumption* 4.1 *holds, for any* $\mathbf{Y} \in \mathcal{R}_1$,

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

$$\sigma_{\max}\left(\overline{\text{Hess}\,H([\mathbf{Y}])}\right) \leqslant 4\left(\sigma_1(\mathbf{Y}^*) + \frac{\mu\sigma_r(\mathbf{Y}^*)}{\kappa^*} \right)^2 + \frac{14}{3}\mu\sigma_r^2(\mathbf{Y}^*).$$

In particular, if μ is further chosen such that

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

we have $H([\mathbf{Y}])$ is geodesically strongly convex and smooth in $\mathcal{R}_1$.

Theorem 4.1 guarantees that the optimization problem Eq. (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 [10, Theorem 11.29].

By combining this result with Theorem 3.2, when 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 Eq. (2.1).

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 4.2 (Parameters Settings). Denote e_1, e_2 and e_3 to be some error terms as follows:

$$e_1 \stackrel{\text{def}}{=} \frac{\alpha\mu\sigma_r^3(\mathbf{Y}^*)}{2\sqrt{2}\kappa^*\sigma_{r+1}(\mathbf{\Lambda})}, \quad e_2 = \frac{e_1}{\sqrt{2}}, \quad e_3 = e_2 \cdot \sigma_{r+1}(\mathbf{\Lambda}).$$

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. $\sigma_r^2(\boldsymbol{\Lambda}) - 2e_1 - \sigma_{r+1}^2(\boldsymbol{\Lambda}) > 0.$

2. $\sigma_r^2(\boldsymbol{\Lambda}) \left(1 - \dfrac{e_1^2}{\left|\sigma_r^2(\boldsymbol{\Lambda}) - e_1 - \sigma_{r+1}^2(\boldsymbol{\Lambda})\right|^2}\right) - e_1 - \sigma_{r+1}^2(\boldsymbol{\Lambda}) > 0.$

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

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 $-2(\sqrt{2}-1)\sigma_r^2(\mathbf{Y}^*)$ which is negative. Hence, Assumption 4.2 is only related to the eigengap assumption $\sigma_r(\boldsymbol{\Lambda})$ and $\sigma_{r+1}(\boldsymbol{\Lambda})$ in Assumption 4.1. As soon as α is small enough, Assumption 4.2 is satisfied.

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

Theorem 4.2 (Region with Negative Eigenvalue in the Riemannian Hessian of Eq. 2.1). *Assume that Assumption 4.1 holds. Given any $\mathbf{Y} \in \mathbb{R}_*^{n\times r}$ such that $\mathbf{Y} \in \mathcal{R}_2$, let*

$$\theta_{\mathbf{Y}}^1 = [\mathbf{0}, \mathbf{0}, \dots, \mathbf{0}, \boldsymbol{a}, \mathbf{0}, \dots, \mathbf{0}]\mathbf{V}^\top,$$

where $\boldsymbol{a}$ is such that

$$\boldsymbol{a} = \arg\max_{\boldsymbol{a}:\mathbf{Y}^\top \boldsymbol{a}=0} \frac{\boldsymbol{a}^\top \mathcal{A}_{\mathbf{n}} \boldsymbol{a}}{\|\boldsymbol{a}\|^2}, \tag{4.3}$$

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

$$\tilde{i} \overset{def}{=} \arg\min_{j\in[r]} \mathbf{D}_{jj}. \tag{4.4}$$

Denote $\theta_{\mathbf{Y}}^2 = \mathbf{Y} - \mathbf{Y}^*\mathbf{Q}$, 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 $\theta_{\mathbf{Y}}^1$ or $\theta_{\mathbf{Y}}^2$. Then

$$\overline{\mathrm{Hess}\, H([\mathbf{Y}])}[\theta_{\mathbf{Y}}, \theta_{\mathbf{Y}}]$$

$$\leqslant \min\left\{-\frac{\sigma_{r+1}^2(\boldsymbol{\Lambda})}{2}\|\theta_{\mathbf{Y}}\|^2, -2\left(\sigma_r^2(\boldsymbol{\Lambda})\left(1-\frac{e_1^2}{\left|\sigma_r^2(\boldsymbol{\Lambda})-e_1-\sigma_{r+1}^2(\boldsymbol{\Lambda})\right|^2}\right)-e_1-\sigma_{r+1}^2(\boldsymbol{\Lambda})\right)\|\theta_{\mathbf{Y}}\|^2,\right.$$

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

In particular, if α and μ satisfies Assumption 4.2, we have $\overline{\mathrm{Hess}\, H([\mathbf{Y}])}$ has at least one negative eigenvalue and $\theta_{\mathbf{Y}}^1$ or $\theta_{\mathbf{Y}}^2$ is the escaping direction.

Remark 4.1. Theorem 4.2 suggests that if some singular values of $\mathbf{Y}$ are small, then the descent direction $\theta_{\mathbf{Y}}^1$ should increase the singular value of $\mathbf{Y}$. If all of the singular values of $\mathbf{Y}$ are large enough compared with $\sigma_r(\boldsymbol{\Lambda})$, then $\theta_{\mathbf{Y}}^2$ 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 $\sigma_{r+1}^2(\mathbf{\Lambda})/2$, then we have

$$\overline{\mathrm{Hess}\,H([\mathbf{Y}])}[\theta_\mathbf{Y}^1, \theta_\mathbf{Y}^1] \leq -\frac{\sigma_{r+1}^2(\mathbf{\Lambda})}{2}\|\theta_\mathbf{Y}^1\|^2.$$

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

$$\overline{\mathrm{Hess}\,H([\mathbf{Y}])}[\theta_\mathbf{Y}^1, \theta_\mathbf{Y}^1]$$
$$\leq -2\left(\sigma_r^2(\mathbf{\Lambda})\left(1 - \frac{e_1^2}{\left|\sigma_r^2(\mathbf{\Lambda}) - e_1 - \sigma_{r+1}^2(\mathbf{\Lambda})\right|^2}\right) - e_1 - \sigma_{r+1}^2(\mathbf{\Lambda})\right)\|\theta_\mathbf{Y}^1\|^2.$$

- If all of the eigenvalues of $\mathbf{Y}\mathbf{Y}^\top$ are larger than $e_1 + \sigma_{r+1}^2(\mathbf{\Lambda})$, then we have

$$\overline{\mathrm{Hess}\,H([\mathbf{Y}])}[\theta_\mathbf{Y}^2, \theta_\mathbf{Y}^2]$$
$$< \left((\alpha - 2(\sqrt{2}-1))\sigma_r^2(\mathbf{Y}^*) + 6\frac{\alpha^2\sigma_r^4(\mathbf{Y}^*)\,\sigma_{r+1}^2(\mathbf{\Lambda})/16}{\left|\sigma_r^2(\mathbf{\Lambda}) - e_2 - \sigma_{r+1}^2(\mathbf{\Lambda})\right|^2}\right)\|\theta_\mathbf{Y}^2\|_\mathrm{F}^2.$$

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 Eq. (2.1)).

1. $\left\|\overline{\mathrm{grad}\,H([\mathbf{Y}])}\right\|_\mathrm{F} > \alpha\mu\sigma_r^3(\mathbf{Y}^*)/(4\kappa^*), \ \forall \mathbf{Y} \in \mathcal{R}_3'$;

2. $\left\|\overline{\mathrm{grad}\,H([\mathbf{Y}])}\right\|_\mathrm{F} \geq 2(\|\mathbf{Y}\|^3 - \|\mathbf{Y}\|\|\mathbf{Y}^*\|^2) > 2(\beta^3 - \beta)\|\mathbf{Y}^*\|^3, \ \forall \mathbf{Y} \in \mathcal{R}_3''$;

3. $\left\langle\overline{\mathrm{grad}\,H([\mathbf{Y}])}, \mathbf{Y}\right\rangle > 2(1 - 1/\gamma)\|\mathbf{Y}\mathbf{Y}^\top\|_\mathrm{F}^2, \ \forall \mathbf{Y} \in \mathcal{R}_3'''$.

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 Figs. 5.1 and 5.2. See a discussion in Section 5.

Remark 4.2. These results can be seen as an under-parameterized generalization to the regression problem of [51, Section 5]. The proof in [51] 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, [70] proves that in the setting $r \geq n$, the gradient flow for Eq. (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 $r < n$.

Remark 4.3. The major difference between our landscape characterization in Theorems 4.1-4.3 and most of the results in the literature studying similar problems is that here we analyze the optimization landscape of the ambient space problem according to

a suitable quotient geometry and not with respect to the standard Euclidean geometry. It is important to observe that the expressions of Riemannian gradient and Hessian under this quotient geometry are exactly the same as those in the Euclidean space, but that a different arises, for example, when studying escape directions around saddle points, an essential analysis for characterizing strict saddles. In particular, under the quotient geometry, a escape direction at $\theta_{\mathbf{Y}}$ must lie in the horizontal space $\mathcal{H}_{\mathbf{Y}}\overline{\mathcal{M}}_{r+}^{q}$.

The quotient geometry considered here is also useful for providing a fully geometric description of the optimization landscape of the ambient space optimization problem. Indeed, apart from characterizing all saddle points as strict saddles, we prove that the objective function is strongly convex around the global minimizer and provide quantitative estimates on its curvature and radius of convexity. Combined with our other quantitative bounds, we expect that these results will be useful when analyzing the convergence to a global minimizer in the parameterized optimization problem under the neural tangent kernel and mean-field regimes.

Remark 4.4. In the specific case of $\mathcal{A}_{\mathbf{n}}$ as in Eq. (1.5), and under Assumptions 1.1, 4.1 should be interpreted as $\lambda_r^{\mathcal{M}} < \lambda_{r+1}^{\mathcal{M}}$, as suggested by Remark 2.3. 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 4.5. Theorems 4.1-4.3 guarantee the benigness of the ambient optimization problem, which is necessary condition of the benigness of parameterized optimization problem Eq. (1.2). Also, these landscape results imply that perturbed gradient descent is guaranteed to converge to the global minima in polynomial time for Eq. (2.1) [21, 32, 38, 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 7, 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 Fig. 5.1, 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 Fig. 5.2. We notice a similar qualitative behavior with the training dynamics of the NN, suggesting that the landscape of problem (1.2), if the NN is properly overparameterized, inherits properties of the landscape of ℓ.

(a) Initialized near optimal   (b) Initialized near saddle   (c) Initialized near saddle
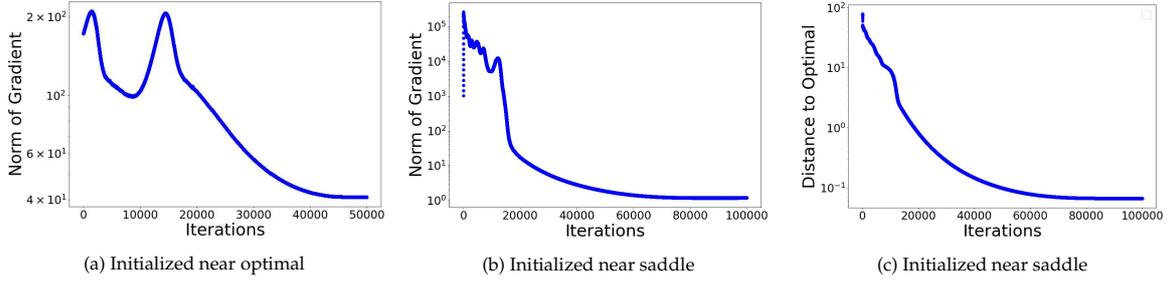
Figure 5.1: (a) and (b) Sum of the norms of the gradients for a two-layer ReLU neural network. In (a), the network is initialized near the global optimal solution and in (b) the network is initialized near a saddle point. (c) shows the distance between the current outputs of the neural network and the optimal solution for the case when it was initialized near a saddle point. More details are presented in Appendix B.2.

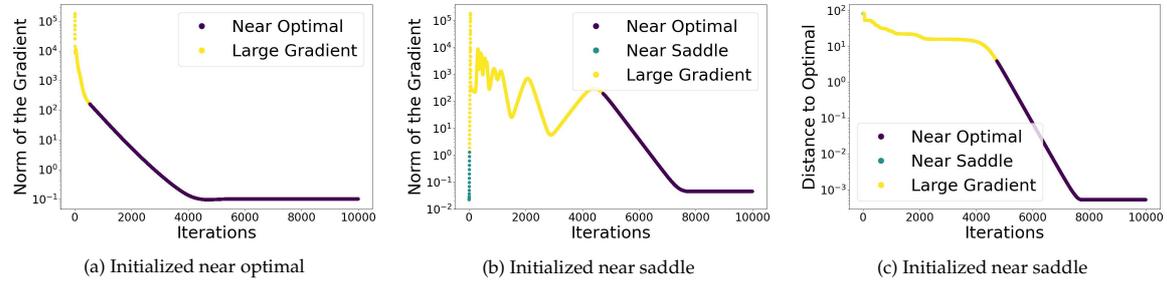(a) Initialized near optimal   (b) Initialized near saddle   (c) Initialized near saddle

Figure 5.2: 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.

6 Experiments

In this section, we evaluate the approximation capability of the SNN as described in Corollary 3.2, followed by testing its downstream clustering performance using the k-means algorithm on the obtained SNN embedding.

6.1 Laplace-Beltrami operator approximation

We generate $n = 8196$ sample points uniformly distributed on a 2-dimensional sphere and construct an ε-graph with $\varepsilon = n^{-1/(d+4)}$. In this setting, the goal of the SNN is to learn the first four eigenfunctions of the Laplace-Beltrami operator. We train a neural network with a single hidden layer containing $5 \times 2^{13} = 40960$ neurons to approximate the graph Laplacian of the constructed proximity graph. The training is performed on NVIDIA A100-SXM4-40 GB. The learning rate is set as 2×10^{-5}. After training, the network's output is evaluated at the points in $\mathcal{X}_n$ and compared to the sphere's eigenfunctions (the spherical harmonics) evaluated at $\mathcal{X}_n$.

The mean squared error (MSE) between the SNN-estimated eigenfunctions and the true Laplace-Beltrami eigenfunctions up to a QR decomposition is 0.0191. The first four func-
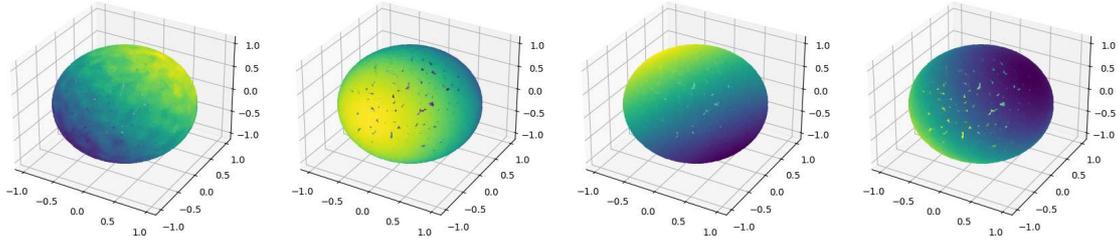
Figure 6.1: From left to right: SNN approximation of an orthonormal basis for the span of 1-st-4-th eigenfunctions of Laplace-Beltrami operator over a sphere. The SNN is trained by using Eq. (1.2), and then evaluated at $\mathcal{X}_n$.
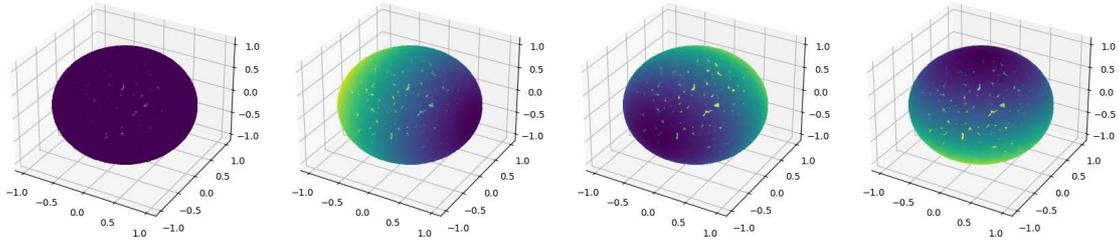

Figure 6.2: Plots of the first four spherical harmonics.

tions recovered by the SNN are illustrated in Fig. 6.1, while the first four spherical harmonics are depicted in Fig. 6.2. For benchmarking purposes, we also compute the eigenvectors obtained from a conventional eigendecomposition of the graph Laplacian, a method theoretically proven to be essentially minimax optimal for general approximation of eigenpairs of weighted Laplace-Beltrami operators (see [29]). This baseline approach achieves an MSE of 0.0295. Thus, the SNN demonstrates slightly improved performance.

Next, we test the out-of-sample generalization capability of the SNN. We randomly sample 8,196 points $\mathcal{Y}_n$ uniformly from the sphere and evaluate both the SNN and spherical harmonics on these points. The resulting MSE for the first four eigenfunctions is 0.0169, closely matching the training MSE and thus demonstrating strong generalization performance.

6.2 Detecting the bottleneck of a dumbbell

Since the SNN can effectively approximate eigenfunctions of the weighted Laplace-Beltrami operator of the underlying manifold, downstream distance-based clustering algorithms can exploit this capability to identify geometric properties of the manifold. A standard manifold learning task is to identify bottlenecks, or to construct smooth approximations of minimum cuts, which can be related to the task of clustering. To test SNN's performance in such downstream task, we consider 8,096 points uniformly sampled from a two-dimensional dumbbell. As discussed in the literature of manifold learning (see, e.g., [31]), an optimal two-way clustering for the dumbbell should cut through the bottleneck region. We use the same structure for the graph Laplacian and SNN described in Section 6.1. The sampled data points on the dumbbell manifold are visualized in Fig. 6.3.
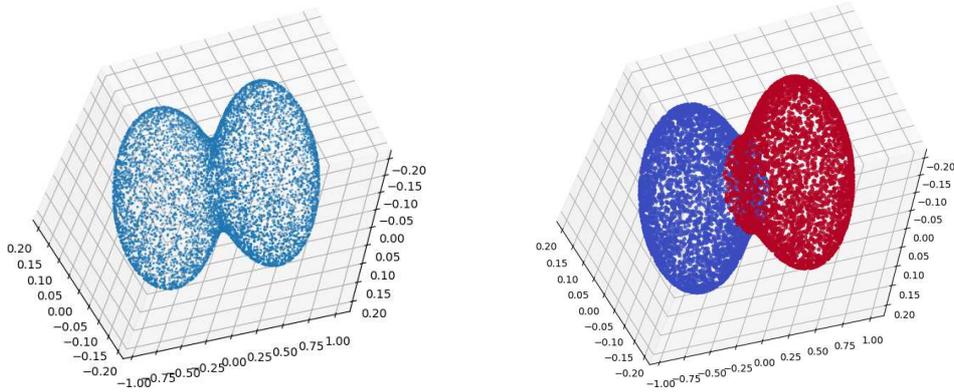
Figure 6.3: Left: 8192 data points uniformly sampled from a 2-dimensional dumbbell. Right: An application of k-means as a downstream clustering algorithm to the trained SNN embedding.

After training, we use the SNN to recover the first four eigenvectors of the graph Laplacian and apply the k-means algorithm to partition the data into two clusters. When the network is trained for $250,000$ epochs using a learning rate of 2×10^{-5}, the accuracy of this clustering task is 99.89%; the total training time is 4614 seconds. This aligns with our expectations given the proximity of the SNN embedding to the one induced by the eigenfunctions of Δ_ρ. If we decrease the number of epochs to $10,000$, the precision decreases to 98.82% with the total training time decreasing to 192 seconds. This presents a reasonable tradeoff between computation and accuracy of SNN. Furthermore, as suggested by our theoretical findings, the SNN embedding matrix is seen to be nearly orthogonal, with an average squared cosine similarity between different embedding columns of approximately 1.09×10^{-7}. Consequently, the SNN demonstrates strong potential in common machine learning applications, including clustering and PDE learning.

7 Conclusions

We have explored some theoretical aspects of spectral neural networks, 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 [34]. 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 [37, 49]. 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 (2.1). 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 or to a mean field limit. This difference is mainly due to the fact that the spectral contrastive loss ℓ (see (1.2)) 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.

Appendix 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

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

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,

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

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

$$\mathbf{Y}_\theta^\top \mathbf{Y}_\theta = n\mathbf{I}_r, \tag{A.3}$$

where $\mathbf{Y}_\theta = [f_\theta(x_1), \ldots, f_\theta(x_n)]$. This constraint is used to avoid a trivial solution. Note that Eq. (A.3) is a global constraint. [61] have established a stochastic coordinate descent fashion to efficiently train SpectralNets. However, the stochastic training process in [61] can only guarantee Eq. (A.3) holds approximately.

Conceptually, the SpectralNet loss Eq. (A.2) can also be written as

$$\mathcal{L}_{\text{SpectralNet}}(\theta) = \frac{2}{n^2} \text{trace}\left(\mathbf{Y}_\theta^\top (\mathbf{D}_G - G)\mathbf{Y}_\theta\right), \tag{A.4}$$

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

$$(\mathbf{D}_G)_{ii} = \sum_{j=1}^{n} G_{ij}. \tag{A.5}$$

The symmetric and positive semi-definite matrix $\mathbf{D}_G - G$ encodes the unnormalized graph Laplacian. Since $\mathbf{D}_G - G$ is positive semi-definite, the ambient problem of Eq. (A.4) is a constrained convex optimization problem. However, the parametrization and hard constraint (A.3) make understanding SpectralNet's training process from a theoretical perspective challenging.

Appendix 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. Fig. 1.1 shows a one-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 gram 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 Figs. 5.1 and 5.2 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 do not focus on the optimization problem of SNN, but use this to illustrate Theorems 4.1-4.3.

Appendix C   Properties of the matrix $\mathcal{A}_\mathbf{n}$ in Eq. (1.2)

C.1   Proof of Proposition 1.1

Proposition 1.1.   *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 [72])

$$\Delta_n \overset{def}{=} \mathbf{I} - \mathbf{D}_G^{-\frac{1}{2}} G \mathbf{D}_G^{-\frac{1}{2}}. \tag{1.6}$$

Proof of Proposition 1.1. Notice that

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

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 Eq. (1.5) and the fact that $a > 1$, it is sufficient to argue that all eigenvalues of $\mathbf{D}_G^{-1/2} G \mathbf{D}_G^{-1/2}$ have absolute value less than or equal to 1. This, however, follows from the following two facts: 1) the matrix $\mathbf{D}_G^{-1/2} G \mathbf{D}_G^{-1/2}$ is similar to the matrix $\mathbf{D}_G^{-1} G$, given that

$$\mathbf{D}_G^{\frac{1}{2}} (\mathbf{D}_G^{-1} G) \mathbf{D}_G^{-\frac{1}{2}} = \mathbf{D}_G^{-\frac{1}{2}} G \mathbf{D}_G^{-\frac{1}{2}},$$

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

Remark C.1. While one could set $\mathcal{A}_\mathbf{n}$ to be Δ_n itself (since Δ_n is PSD), solving the resulting problem (2.1) 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 (1.5) to guarantee that $\mathcal{A}_\mathbf{n}$ is always PSD, in this way simplifying the statements and proofs of our main results.

Appendix D   Auxiliary approximation results

D.1   Neural network approximation of Lipschitz functions on manifolds

[15] 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 does not grow exponentially with the ambient space dimension d. Precisely:

Theorem D.1 ([15, Theorem 1]). *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. $L \leq C_K \left(\log \dfrac{1}{\delta} + \log d \right)$, $p \leq C_K(\delta^{-m} + d)$,

2. $N \leq C_K \left(\delta^{-m} \log \dfrac{1}{\delta} + d \log \dfrac{1}{\delta} + d \log d \right)$, $\kappa \leq C_K$

such that there is a neural network $f_\theta \in \mathcal{F}(1, \kappa, L, p, N)$ (as defined in Eq. (3.2)), 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 [15] due to the fact that in their estimates the ambient space dimension d does not appear as an exponent.

Appendix E   Proofs of Theorem 3.1 and Corollary 3.2

We begin by proving an important lemma.

Lemma 3.1. *Let $u : \mathcal{X}_n \to \mathbb{R}$ be a function satisfying*

$$|u(x) - u(\tilde{x})| \leq L\big(d_{\mathcal{M}}(x, \tilde{x}) + \tau^2\big), \quad \forall x, \tilde{x} \in \mathcal{X}_n \tag{3.3}$$

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

$$\|u - \tilde{g}\|_{L^\infty(\mathcal{X}_n)} \leq 5L\tau^2. \tag{3.4}$$

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 $d_{\mathcal{M}}(x, \tilde{x}) \geq \tau^2/2$.

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

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

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 Eq. (3.4). 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

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

This completes the proof. $\qquad\square$

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 1.1, there are constants c, C that depend on $\mathcal{M}, \rho$, and the embedding dimension r such that with probability at least

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

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 Eq. (3.2)) such that

1. *$\sqrt{n}\|\mathbf{Y}_\theta - \mathbf{Y}^*\|_{\infty,\infty} \leq C(\varepsilon + \tau^2)$, and thus also $\|\mathbf{Y}_\theta - \mathbf{Y}^*\|_F \leq C\sqrt{r}(\varepsilon + \tau^2)$.*

2. *The depth of the network, L, satisfies: $L \leq C(\log(1/\varepsilon) + \log d)$, and its width, p, satisfies $p \leq C(\varepsilon^{-m} + d)$.*

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 $\kappa \leq C/n^{1/(2L)}$.

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 2.1, we know that, with very high probability, the function $u_s : \mathcal{X}_n \to \mathbb{R}$ satisfies

$$|u_s(x_i) - u_s(x_j)| \leq L_s\big(d_{\mathcal{M}}(x_i, x_j) + \tau^2\big), \quad \forall x_i, x_j \in \mathcal{X}_n \tag{E.1}$$

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 2.3) in combination with Lemma 3.1, we deduce the existence of a CL_s-Lipschitz function $g_s : \mathcal{M} \to \mathbb{R}$ satisfying

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

In turn, Theorem D.1 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

$$\|f_{\tilde{\theta}}(x) - g_s(x)\|_{L^\infty(\mathcal{M})} \leq \varepsilon. \tag{E.3}$$

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

$$\theta \overset{\text{def}}{=} \frac{1}{n^{1/(2L)}} \tilde{\theta},$$

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

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

and also, thanks to Eq. (E.2),

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

Stacking the scalar neural networks constructed above to approximate each of the functions u_s for $s = 1, \dots r$, and using Eq. (2.4), we obtain the desired vector valued neural network approximating $\mathbf{Y}^*$. $\square$

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

$$\sqrt{n} \left\| \mathbf{Y}_\theta \mathbf{Y}_\theta^\top - \mathbf{Y}^* \mathbf{Y}^{*\top} \right\|_{\infty,\infty} \leq C_r(\varepsilon + \tau^2), \tag{3.5}$$

and thus also

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

for some deterministic constant C_r.

Proof.

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

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

$$\|\mathbf{Y}^*\|_{\mathrm{F}}^2 = \sum_{s=1}^{r} \sigma_s(\mathcal{A}_{\mathbf{n}}) = \mathcal{O}(r).$$

The proof is complete. $\square$

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 3.2 for the convenience of the reader.

Corollary 3.2. *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*

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

In the above, $f_\theta^1,\dots,f_\theta^r$ are the coordinate functions of the vector-valued neural network f_θ, and the functions $f_1,\dots,f_r$ 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 Eq. (E.2) and recall that the scalar neural network f_θ constructed in the proof of Theorem 3.1 satisfies

$$\sqrt{n} \left\| f_\theta - \frac{1}{\sqrt{n}} g_s \right\|_{L^\infty(\mathcal{M})} \leq \delta. \tag{E.4}$$

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

$$\begin{aligned}
&\left| f_s(x) - \sqrt{\frac{n}{1+a}} f_\theta(x) \right| \\
&\leq |f_s(x) - f_s(x_i)| + |f_s(x_i) - u_s(x_i)| \\
&\quad + \left| u_s(x_i) - \frac{1}{\sqrt{\sigma_s(\mathcal{A}_{\mathbf{n}})}} g_s(x_i) \right| + \left| \frac{1}{\sqrt{\sigma_s(\mathcal{A}_{\mathbf{n}})}} g_s(x_i) - \frac{1}{\sqrt{1+a}} g_s(x_i) \right| \\
&\quad + \left| \frac{1}{\sqrt{1+a}} g_s(x_i) - \frac{1}{\sqrt{1+a}} g_s(x) \right| + \left| \frac{1}{\sqrt{1+a}} g_s(x) - \sqrt{\frac{n}{1+a}} f_\theta(x) \right| \\
&\leq C_s(\varepsilon + \tau), \tag{E.5}
\end{aligned}$$

where we have used the Lipschitz continuity of f_s and g_s, Theorem 2.3, Remark 2.3, and Eq. (E.4). The proof is complete. $\square$

Appendix F   Proof of Theorem 3.2

Recall that

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

We start our proof with a lemma from linear algebra.

Lemma F.1. *For any* $\mathbf{Y} \in \mathbb{R}^{n \times r}$ *we have*

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

Proof. A straightforward computation reveals that

$$
\begin{aligned}
&\left\| \mathbf{Y}\mathbf{Y}^\top - \mathcal{A}_\mathbf{n} \right\|_\mathrm{F}^2 - \left\| \mathbf{Y}^*\mathbf{Y}^{*\top} - \mathcal{A}_\mathbf{n} \right\|_\mathrm{F}^2 \\
&= \left\| (\mathbf{Y}\mathbf{Y}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top}) + (\mathbf{Y}^*\mathbf{Y}^{*\top} - \mathcal{A}_\mathbf{n}) \right\|_\mathrm{F}^2 - \left\| \mathbf{Y}^*\mathbf{Y}^{*\top} - \mathcal{A}_\mathbf{n} \right\|_\mathrm{F}^2 \\
&= \left\| \mathbf{Y}\mathbf{Y}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top} \right\|_\mathrm{F}^2 + 2\langle \mathbf{Y}\mathbf{Y}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top}, \mathbf{Y}^*\mathbf{Y}^{*\top} - \mathcal{A}_\mathbf{n} \rangle \\
&= \left\| \mathbf{Y}\mathbf{Y}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top} \right\|_\mathrm{F}^2 + 2\langle \mathbf{Y}\mathbf{Y}^\top, \mathbf{Y}^*\mathbf{Y}^{*\top} - \mathcal{A}_\mathbf{n} \rangle \\
&\le \left\| \mathbf{Y}\mathbf{Y}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top} \right\|_\mathrm{F}^2,
\end{aligned}
\tag{F.1}
$$

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

Corollary F.1. *With probability at least*

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

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*

$$\left\| \mathbf{Y}_\theta \mathbf{Y}_\theta^\top - \mathbf{Y}^*\mathbf{Y}^{*\top} \right\|_\mathrm{F} \le C_r \tau(\tilde{\varepsilon} + \tau). \tag{F.2}$$

Corollary F.2. *Let* f_θ *be as in Corollary F.1. Then*

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

Proof. Let θ be as in Corollary F.1. Then

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

where the second to last inequality follows from Lemma F.1.   □

In what follows we will write the SVD (eigendecomposition) of $\mathcal{A}_\mathbf{n}$ as $\overline{\mathbf{U}}\Sigma\overline{\mathbf{U}}^\top$. 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 F.1. τ and $\tilde{\varepsilon}$ in Corollary F.1 satisfy the following condition:

$$\tau^2 E < \sigma_r^2(\mathcal{A}_\mathbf{n}) - \sigma_{r+1}^2(\mathcal{A}_\mathbf{n}), \tag{F.3}$$

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

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

$$\left\|\mathbf{Y}^*\mathbf{Y}^{*\top} - \mathcal{A}_\mathbf{n}\right\|_\mathrm{F}^2 \leq \left\|\mathbf{Y}_{\theta*}\mathbf{Y}_{\theta*}^\top - \mathcal{A}_\mathbf{n}\right\|_\mathrm{F}^2 \leq \left\|\mathbf{Y}_\theta \mathbf{Y}_\theta^\top - \mathcal{A}_\mathbf{n}\right\|_\mathrm{F}^2. \tag{F.4}$$

Also,

$$\begin{aligned}
0 &\geq \left\|\mathbf{Y}_{\theta*}\mathbf{Y}_{\theta*}^\top - \mathcal{A}_\mathbf{n}\right\|_\mathrm{F}^2 - \left\|\mathbf{Y}_\theta \mathbf{Y}_\theta^\top - \mathcal{A}_\mathbf{n}\right\|_\mathrm{F}^2 \\
&= \left\|\left(\mathbf{Y}_{\theta*}\mathbf{Y}_{\theta*}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top}\right) + \left(\mathbf{Y}^*\mathbf{Y}^{*\top} - \mathcal{A}_\mathbf{n}\right)\right\|_\mathrm{F}^2 - \left\|\mathbf{Y}_\theta \mathbf{Y}_\theta^\top - \mathcal{A}_\mathbf{n}\right\|_\mathrm{F}^2 \\
&= \left\|\mathbf{Y}_{\theta*}\mathbf{Y}_{\theta*}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top}\right\|_\mathrm{F}^2 + \left\|\mathbf{Y}^*\mathbf{Y}^{*\top} - \mathcal{A}_\mathbf{n}\right\|_\mathrm{F}^2 \\
&\quad + 2\langle\mathbf{Y}_{\theta*}\mathbf{Y}_{\theta*}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top}, \mathbf{Y}^*\mathbf{Y}^{*\top} - \mathcal{A}_\mathbf{n}\rangle - \left\|\mathbf{Y}_\theta \mathbf{Y}_\theta^\top - \mathcal{A}_\mathbf{n}\right\|_\mathrm{F}^2 \\
&= \left\|\mathbf{Y}_{\theta*}\mathbf{Y}_{\theta*}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top}\right\|_\mathrm{F}^2 + \left\|\mathbf{Y}^*\mathbf{Y}^{*\top} - \mathcal{A}_\mathbf{n}\right\|_\mathrm{F}^2 \\
&\quad + 2\langle\mathbf{Y}_{\theta*}\mathbf{Y}_{\theta*}^\top, \mathbf{Y}^*\mathbf{Y}^{*\top} - \mathcal{A}_\mathbf{n}\rangle - \left\|\mathbf{Y}_\theta \mathbf{Y}_\theta^\top - \mathcal{A}_\mathbf{n}\right\|_\mathrm{F}^2,
\end{aligned} \tag{F.5}$$

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

$$\begin{aligned}
&\left\|\mathbf{Y}_{\theta*}\mathbf{Y}_{\theta*}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top}\right\|_\mathrm{F}^2 + 2\langle\mathbf{Y}_{\theta*}\mathbf{Y}_{\theta*}^\top, \mathbf{Y}^*\mathbf{Y}^{*\top} - \mathcal{A}_\mathbf{n}\rangle \\
&= \left\|\mathbf{Y}_{\theta*}\mathbf{Y}_{\theta*}^\top\right\|_\mathrm{F}^2 + \left\|\mathbf{Y}^*\mathbf{Y}^{*\top}\right\|_\mathrm{F}^2 - 2\langle\mathbf{Y}_{\theta*}\mathbf{Y}_{\theta*}^\top, \mathcal{A}_\mathbf{n}\rangle.
\end{aligned} \tag{F.6}$$

By combining Eq. (F.5), Lemma F.2 and Eq. (F.6), we have

$$\left\|\mathbf{Y}_{\theta*}\mathbf{Y}_{\theta*}^\top\right\|_\mathrm{F}^2 + \left\|\mathbf{Y}^*\mathbf{Y}^{*\top}\right\|_\mathrm{F}^2 - 2\langle\mathbf{Y}_{\theta*}\mathbf{Y}_{\theta*}^\top, \mathcal{A}_\mathbf{n}\rangle \leq \tau^2 E. \tag{F.7}$$

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

$$\begin{aligned}
\langle\mathbf{E}^1(\mathbf{E}^1)^\top, \mathbf{E}^2(\mathbf{E}^2)^\top\rangle &= 0, \\
\langle\mathbf{E}^1(\mathbf{E}^2)^\top, \mathbf{E}^2(\mathbf{E}^2)^\top\rangle &= 0, \\
\langle\mathbf{E}^1(\mathbf{E}^2)^\top, \mathbf{E}^1(\mathbf{E}^1)^\top\rangle &= 0, \\
\langle\mathbf{E}^2(\mathbf{E}^1)^\top, \mathbf{E}^1(\mathbf{E}^1)^\top\rangle &= 0, \\
\langle\mathbf{E}^2(\mathbf{E}^1)^\top, \mathbf{E}^2(\mathbf{E}^2)^\top\rangle &= 0.
\end{aligned} \tag{F.8}$$

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 Eq. (F.7), we deduce

$$
\begin{aligned}
\varepsilon^2 E \geq{}& \left\|\mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top\right\|_{\mathrm{F}}^2 + \left\|\mathbf{Y}^*\mathbf{Y}^{*\top}\right\|_{\mathrm{F}}^2 - 2\langle \mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top, \mathcal{A}_{\mathbf{n}}\rangle \\
={}& \left\|\overline{\mathbf{U}}(\mathbf{E}^1 + \mathbf{E}^2)(\mathbf{E}^1 + \mathbf{E}^2)^\top \overline{\mathbf{U}}^\top\right\|_{\mathrm{F}}^2 + \left\|\mathbf{Y}^*\mathbf{Y}^{*\top}\right\|_{\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 \\
={}& \left\|(\mathbf{E}^1 + \mathbf{E}^2)(\mathbf{E}^1 + \mathbf{E}^2)^\top\right\|_{\mathrm{F}}^2 + \left\|\mathbf{Y}^*\mathbf{Y}^{*\top}\right\|_{\mathrm{F}}^2 - 2\langle (\mathbf{E}^1 + \mathbf{E}^2)(\mathbf{E}^1 + \mathbf{E}^2)^\top, \mathbf{\Sigma}\rangle \\
\overset{\text{Eq. (F.8)}}{=}{}& \left\|\mathbf{E}^1(\mathbf{E}^1)^\top\right\|_{\mathrm{F}}^2 + \left\|\mathbf{E}^2(\mathbf{E}^2)^\top\right\|_{\mathrm{F}}^2 + 2\left\|\mathbf{E}^2(\mathbf{E}^1)^\top\right\|_{\mathrm{F}}^2 + 2\langle (\mathbf{E}^1)^\top \mathbf{E}^1, (\mathbf{E}^2)^\top \mathbf{E}^2\rangle \\
& + \left\|\mathbf{Y}^*\mathbf{Y}^{*\top}\right\|_{\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}}{=}{}& \left\|\mathbf{E}^1(\mathbf{E}^1)^\top\right\|_{\mathrm{F}}^2 + \left\|\mathbf{E}^2(\mathbf{E}^2)^\top\right\|_{\mathrm{F}}^2 + 2\left\|\mathbf{E}^2(\mathbf{E}^1)^\top\right\|_{\mathrm{F}}^2 + 2\langle (\mathbf{E}^1)^\top \mathbf{E}^1, (\mathbf{E}^2)^\top \mathbf{E}^2\rangle \\
& + \left\|\mathbf{Y}^*\mathbf{Y}^{*\top}\right\|_{\mathrm{F}}^2 - 2\langle \mathbf{E}^1(\mathbf{E}^1)^\top + \mathbf{E}^2(\mathbf{E}^2)^\top, \mathbf{\Sigma}\rangle \\
={}& \left\|\mathbf{E}^1(\mathbf{E}^1)^\top\right\|_{\mathrm{F}}^2 + \left\|\mathbf{E}^2(\mathbf{E}^2)^\top\right\|_{\mathrm{F}}^2 + 2\left\|\mathbf{E}^2(\mathbf{E}^1)^\top\right\|_{\mathrm{F}}^2 + 2\langle (\mathbf{E}^1)^\top \mathbf{E}^1, (\mathbf{E}^2)^\top \mathbf{E}^2\rangle \\
& + \left\|\mathbf{\Sigma}^1\right\|_{\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 \\
={}& \left\|\mathbf{E}^1(\mathbf{E}^1)^\top - \mathbf{\Sigma}^1\right\|_{\mathrm{F}}^2 + \left\|\mathbf{E}^2(\mathbf{E}^2)^\top\right\|_{\mathrm{F}}^2 \\
& + 2\left\|\mathbf{E}^2(\mathbf{E}^1)^\top\right\|_{\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{}& \left\|\mathbf{E}^1(\mathbf{E}^1)^\top - \mathbf{\Sigma}^1\right\|_{\mathrm{F}}^2 + \left\|\mathbf{E}^2(\mathbf{E}^2)^\top\right\|_{\mathrm{F}}^2 + 2\left\|\mathbf{E}^2(\mathbf{E}^1)^\top\right\|_{\mathrm{F}}^2 + 2\langle (\mathbf{E}^1)^\top \mathbf{E}^1, (\mathbf{E}^2)^\top \mathbf{E}^2\rangle \\
& - 2\left\|\mathbf{E}^2(\mathbf{E}^2)^\top\right\|_{\mathrm{F}} \cdot \sigma_{r+1}(\mathcal{A}_{\mathbf{n}}) \\
\geq{}& \left\|\mathbf{E}^1(\mathbf{E}^1)^\top - \mathbf{\Sigma}^1\right\|_{\mathrm{F}}^2 + \left\|\mathbf{E}^2(\mathbf{E}^2)^\top\right\|_{\mathrm{F}}^2 + \left(2\|\mathbf{E}^2\|_{\mathrm{F}}^2 + 2\left\|\mathbf{E}^2(\mathbf{E}^2)^\top\right\|_{\mathrm{F}}\right) \cdot \sigma_r^2(\mathbf{E}^1) \\
& - 2\left\|\mathbf{E}^2(\mathbf{E}^2)^\top\right\|_{\mathrm{F}} \cdot \sigma_{r+1}(\mathcal{A}_{\mathbf{n}}). \tag{F.9}
\end{aligned}
$$

On the other hand, we have

$$
\begin{aligned}
& \left\|\mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top}\right\|_{\mathrm{F}}^2 \\
={}& \left\|\mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top\right\|_{\mathrm{F}}^2 + \left\|\mathbf{Y}^*\mathbf{Y}^{*\top}\right\|_{\mathrm{F}}^2 - 2\langle \mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top, \mathcal{A}_{\mathbf{n}}\rangle + 2\langle \mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top, \mathcal{A}_{\mathbf{n}} - \mathbf{Y}^*\mathbf{Y}^{*\top}\rangle \\
={}& \left\|\mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top\right\|_{\mathrm{F}}^2 + \left\|\mathbf{Y}^*\mathbf{Y}^{*\top}\right\|_{\mathrm{F}}^2 - 2\langle \mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top, \mathcal{A}_{\mathbf{n}}\rangle + 2\langle \mathbf{E}^2(\mathbf{E}^2)^\top, \mathbf{\Sigma}^2\rangle \\
\leq{}& \tau^2 E + 2\left\|\mathbf{E}^2(\mathbf{E}^2)^\top\right\|_{\mathrm{F}} \cdot \sigma_{r+1}(\mathcal{A}_{\mathbf{n}}). \tag{F.10}
\end{aligned}
$$

It remains to show that $\left\|\mathbf{E}^2(\mathbf{E}^2)^\top\right\|_{\mathrm{F}}$ 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 $\left\|\mathbf{E}^2(\mathbf{E}^2)^\top\right\|_{\mathrm{F}}$ can be guaranteed to be small according to Eq. (F.9). Second, when $\sigma_r^2(\mathbf{E}^1)$ is small, we'll show that $\left\|\mathbf{E}^1(\mathbf{E}^1)^\top - \mathbf{\Sigma}^1\right\|_{\mathrm{F}}^2$ is large, which will contradict Eq. (F.9).

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

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 Eq. (F.9) and the fact that $\|AB\|_{\mathrm{F}} \leq \|A\|_{\mathrm{F}} \cdot \|B\|_{\mathrm{F}}$, we have

$$\|\mathbf{E}^1(\mathbf{E}^1)^\top - \mathbf{\Sigma}^1\|_{\mathrm{F}}^2 + \|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}}^2 + 2\|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}} \cdot \sigma_{r+1}(\mathcal{A}_\mathbf{n}) \leq \tau^2 E. \qquad \text{(F.11)}$$

This immediately implies

$$\|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}} \leq \frac{\tau^2 E}{\sigma_{r+1}(\mathcal{A}_\mathbf{n})}. \qquad \text{(F.12)}$$

Combining Eqs. (F.12) and (F.10), we obtain

$$\|\mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top}\|_{\mathrm{F}}^2 \leq \tau^2 E + \|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}} \cdot \sigma_{r+1}(\mathcal{A}_\mathbf{n}) \leq 2\tau^2 E. \qquad \text{(F.13)}$$

Case 2. If $0 \leq \sigma_r^2(\mathbf{E}^1) < 2\sigma_{r+1}(\mathcal{A}_\mathbf{n})/3$.

Invoking Eq. (F.9), we have

$$\begin{aligned}
\tau^2 E &\geq \|\mathbf{E}^1(\mathbf{E}^1)^\top - \mathbf{\Sigma}^1\|_{\mathrm{F}}^2 + \|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}}^2 + \left(2\|\mathbf{E}^2\|_{\mathrm{F}}^2 + 2\|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}}\right) \cdot \sigma_r^2(\mathbf{E}^1) \\
&\quad - 2\|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}} \cdot \sigma_{r+1}(\mathcal{A}_\mathbf{n}) \\
&\geq \left(\sigma_r^2(\mathbf{E}^1) - \sigma_r(\mathcal{A}_\mathbf{n})\right)^2 + \|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}}^2 + 4\|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}} \cdot \sigma_r^2(\mathbf{E}^1) \\
&\quad - 2\|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}} \cdot \sigma_{r+1}(\mathcal{A}_\mathbf{n}) \\
&= \left(\sigma_r^2(\mathbf{E}^1) - \sigma_r(\mathcal{A}_\mathbf{n})\right)^2 + \|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}}^2 - 2\|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}} \cdot \left(\sigma_{r+1}(\mathcal{A}_\mathbf{n}) - 2\sigma_r^2(\mathbf{E}^1)\right) \\
&= \left(\sigma_r^2(\mathbf{E}^1) - \sigma_r(\mathcal{A}_\mathbf{n})\right)^2 + \left(\|\mathbf{E}^2(\mathbf{E}^2)^\top\|_{\mathrm{F}} - \left(\sigma_{r+1}(\mathcal{A}_\mathbf{n}) - 2\sigma_r^2(\mathbf{E}^1)\right)\right)^2 \\
&\quad - \left(\sigma_{r+1}(\mathcal{A}_\mathbf{n}) - 2\sigma_r^2(\mathbf{E}^1)\right)^2 \\
&\geq \left(\sigma_r^2(\mathbf{E}^1) - \sigma_r(\mathcal{A}_\mathbf{n})\right)^2 - \left(\sigma_{r+1}(\mathcal{A}_\mathbf{n}) - 2\sigma_r^2(\mathbf{E}^1)\right)^2, \qquad \text{(F.14)}
\end{aligned}$$

where the second inequality follows from Weyl's inequality [66].

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 $0 \leq \sigma_r^2(\mathbf{E}^1) < 2\sigma_{r+1}(\mathcal{A}_\mathbf{n})/3$. 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 $(\sigma_r^2(\mathcal{A}_\mathbf{n}) - \sigma_{r+1}^2(\mathcal{A}_\mathbf{n}))/9$. However, the resulting inequality contradicts Assumption F.1. Case 2 is thus void.

By combining the aforementioned two cases, we conclude

$$\|\mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top}\|_{\mathrm{F}}^2 \leq 2E\tau^2. \qquad \text{(F.15)}$$

By using Eq. (G.3), we have

$$d^2\left([\mathbf{Y}_{\theta^*}], [\mathbf{Y}^*]\right) \leq \frac{1}{2(\sqrt{2}-1)\sigma_r^2(\mathbf{Y}^*)}\|\mathbf{Y}_{\theta^*}\mathbf{Y}_{\theta^*}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top}\|_{\mathrm{F}}^2 \leq \frac{\tau^2 E}{(\sqrt{2}-1)\sigma_r^2(\mathbf{Y}^*)}, \qquad \text{(F.16)}$$

where

$$d\left([\mathbf{Y}_{\theta^*}], [\mathbf{Y}^*]\right) = \min_{O \in \mathbb{O}_r}\|\mathbf{Y}_{\theta^*} - \mathbf{Y}^*O\|_{\mathrm{F}}.$$

This completes the proof. $\qquad \square$

Appendix 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., $\mathbf{Y} \to \mathbf{Y}\mathbf{O}$, by defining the equivalence classes $[\mathbf{Y}] = \{\mathbf{Y}\mathbf{O} : \mathbf{O} \in \mathbb{O}_r\}$. 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 $\overline{\mathcal{N}}_{r+}^n$ [42]. Moreover, we equip the tangent space $T_\mathbf{Y}\overline{\mathcal{N}}_{r+}^n = \mathbb{R}^{n \times r}$ with the metric $\bar{g}_\mathbf{Y}(\eta_\mathbf{Y}, \theta_\mathbf{Y}) = \text{tr}(\eta_\mathbf{Y}^\top \theta_\mathbf{Y})$.

For convenience, we recall the following:

$$
\begin{aligned}
\overline{\text{grad}\, H([\mathbf{Y}])} &= 2\big(\mathbf{Y}\mathbf{Y}^\top - \mathcal{A}_\mathbf{n}\big)\mathbf{Y}, \\
\overline{\text{Hess}\, H([\mathbf{Y}])}\big[\theta_\mathbf{Y}, \theta_\mathbf{Y}\big] &= \big\|\mathbf{Y}\theta_\mathbf{Y}^\top + \theta_\mathbf{Y}\mathbf{Y}^\top\big\|_\text{F}^2 + 2\big\langle \mathbf{Y}\mathbf{Y}^\top - \mathcal{A}_\mathbf{n}, \theta_\mathbf{Y}\theta_\mathbf{Y}^\top \big\rangle.
\end{aligned}
\tag{G.1}
$$

$$
\begin{aligned}
\mathcal{R}_1 &\overset{\text{def}}{=} \left\{ \mathbf{Y} \in \mathbb{R}_*^{n \times r} \mid d([\mathbf{Y}], [\mathbf{Y}^*]) \leqslant \frac{\mu \sigma_r(\mathbf{Y}^*)}{\kappa^*} \right\}, \\
\mathcal{R}_2 &\overset{\text{def}}{=} \left\{ \mathbf{Y} \in \mathbb{R}_*^{n \times r} \mid d([\mathbf{Y}], [\mathbf{Y}^*]) > \frac{\mu \sigma_r(\mathbf{Y}^*)}{\kappa^*}, \; \big\|\overline{\text{grad}\, H([\mathbf{Y}])}\big\|_\text{F} \leqslant \frac{\alpha \mu \sigma_r^3(\mathbf{Y}^*)}{4\kappa^*} \right. \\
&\qquad\qquad \left. \|\mathbf{Y}\| \leqslant \beta \|\mathbf{Y}^*\|, \; \|\mathbf{Y}\mathbf{Y}^\top\|_\text{F} \leqslant \gamma \|\mathbf{Y}^*\mathbf{Y}^{*\top}\|_\text{F} \right\}, \\
\mathcal{R}_3' &\overset{\text{def}}{=} \left\{ \mathbf{Y} \in \mathbb{R}_*^{n \times r} \mid \big\|\overline{\text{grad}\, H([\mathbf{Y}])}\big\|_\text{F} > \frac{\alpha \mu \sigma_r^3(\mathbf{Y}^*)}{4\kappa^*}, \; \|\mathbf{Y}\| \leqslant \beta \|\mathbf{Y}^*\|, \right. \\
&\qquad\qquad \left. \|\mathbf{Y}\mathbf{Y}^\top\|_\text{F} \leqslant \gamma \|\mathbf{Y}^*\mathbf{Y}^{*\top}\|_\text{F} \right\}, \\
\mathcal{R}_3'' &\overset{\text{def}}{=} \left\{ \mathbf{Y} \in \mathbb{R}_*^{n \times r} \mid \|\mathbf{Y}\| > \beta \|\mathbf{Y}^*\|, \; \|\mathbf{Y}\mathbf{Y}^\top\|_\text{F} \leqslant \gamma \|\mathbf{Y}^*\mathbf{Y}^{*\top}\|_\text{F} \right\}, \\
\mathcal{R}_3''' &\overset{\text{def}}{=} \left\{ \mathbf{Y} \in \mathbb{R}_*^{n \times r} \mid \|\mathbf{Y}\mathbf{Y}^\top\|_\text{F} > \gamma \|\mathbf{Y}^*\mathbf{Y}^{*\top}\|_\text{F} \right\},
\end{aligned}
\tag{G.2}
$$

Remark G.1. 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 [54] and [51].

Lemma G.1 ([51, Lemma 2]). *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.,* $\mathbf{Q}^* \in \arg \min_{\mathbf{Q} \in \mathcal{O}_r} \|\mathbf{Y}_2 \mathbf{Q} - \mathbf{Y}_1\|_F$ *and the geodesic distance between* $[\mathbf{Y}_1]$ *and* $[\mathbf{Y}_2]$ *is* $d([\mathbf{Y}_1], [\mathbf{Y}_2]) = \|\mathbf{Y}_2 \mathbf{Q}^* - \mathbf{Y}_1\|_F$.

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* $[\mathbf{Y}_1 + t(\mathbf{Y}_2 \mathbf{Q}^* - \mathbf{Y}_1)]$ *for* $t \in [0, 1]$.

Lemma G.2 ([51, Lemma 12]). *For any* $\mathbf{Y}_1, \mathbf{Y}_2 \in \mathbb{R}_*^{n \times r}$, *we have*

$$d^2([\mathbf{Y}_1], [\mathbf{Y}_2]) \leqslant \frac{1}{2(\sqrt{2} - 1)\sigma_r^2(\mathbf{Y}_2)} \|\mathbf{Y}_1 \mathbf{Y}_1^\top - \mathbf{Y}_2 \mathbf{Y}_2^\top\|_F^2, \tag{G.3}$$

$$\|(\mathbf{Y}_1 - \mathbf{Y}_2 \mathbf{Q})(\mathbf{Y}_1 - \mathbf{Y}_2 \mathbf{Q})^\top\|_F^2 \leqslant 2\|\mathbf{Y}_1 \mathbf{Y}_1^\top - \mathbf{Y}_2 \mathbf{Y}_2^\top\|_F^2, \tag{G.4}$$

where

$$\mathbf{Q} = \arg \min_{\mathbf{O} \in \mathcal{O}_r} \|\mathbf{Y}_1 - \mathbf{Y}_2 \mathbf{O}\|_F.$$

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

$$\|\mathbf{Y}_1 \mathbf{Y}_1^\top - \mathbf{Y}_2 \mathbf{Y}_2^\top\|_F \leqslant \frac{7}{3} \|\mathbf{Y}_2\| d([\mathbf{Y}_1], [\mathbf{Y}_2]). \tag{G.5}$$

Given any $\mathbf{Y} \in \mathbb{R}_*^{n \times r}$ and $x > 0$, let

$$B_x([\mathbf{Y}]) \overset{\text{def}}{=} \{[\mathbf{Y}_1] : d([\mathbf{Y}_1], [\mathbf{Y}]) < x\}$$

be the geodesic ball centered at $[\mathbf{Y}]$ with radius x. For any Riemannian manifold, there exists a convex geodesic ball at every point [23, Chapter 3.4]. The next result quantifies the convexity radius around a point $[\mathbf{Y}]$ in the manifold $\mathcal{N}_{r+}^n$.

Lemma G.3 ([51, Theorem 2]). *Given any* $\mathbf{Y} \in \mathbb{R}_*^{n \times r}$, *the geodesic ball centered at* $[\mathbf{Y}]$ *with radius* $x \leqslant r_{\mathbf{Y}} \overset{\text{def}}{=} \sigma_r(\mathbf{Y})/3$ *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 G.4 ([52, Proposition 2]). *Let* $\mathbf{Y} \in \mathbb{R}_*^{n \times r}$, *and let* $\mathbf{X} = \mathbf{Y}\mathbf{Y}^\top$. *Then*

$$2\sigma_r^2(\mathbf{Y}) \|\theta_{\mathbf{Y}}\|_F^2 \leqslant \|\mathbf{Y}\theta_{\mathbf{Y}}^\top + \theta_{\mathbf{Y}}\mathbf{Y}^\top\|_F^2 \leqslant 4\sigma_1^2(\mathbf{Y}) \|\theta_{\mathbf{Y}}\|_F^2$$

holds for all $\theta_{\mathbf{Y}} \in \mathcal{H}_{\mathbf{Y}} \overline{\mathcal{M}}_{r+}^q$.

Lemma G.5. *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*

$$\|\mathbf{A}\|_F \cdot \sigma_n(\mathbf{B}) \leq \|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_F \cdot \sigma_1(\mathbf{B}). \tag{G.6}$$

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

$$\|\mathbf{AB}\|_F^2 = \sum_{i=1}^m \|\mathbf{A}_i\mathbf{B}\|_F^2 \leq \sum_{i=1}^m \|\mathbf{A}_i\|_F \cdot \sigma_1(\mathbf{B}) = \|\mathbf{A}\|_F \cdot \sigma_1(\mathbf{B}).$$

Similarly,

$$\|\mathbf{AB}\|_F^2 = \sum_{i=1}^m \|\mathbf{A}_i\mathbf{B}\|_F^2 \geq \sum_{i=1}^m \|\mathbf{A}_i\|_F \cdot \sigma_n(\mathbf{B}) = \|\mathbf{A}\|_F \cdot \sigma_n(\mathbf{B}).$$

The proof is complete. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

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 Eq. (2.2)). *Suppose* $0 \leq \mu \leq \kappa^*/3$. *Given that Assumption 4.1 holds, for any* $\mathbf{Y} \in \mathcal{R}_1$,

$$\sigma_{\min}\left(\overline{\text{Hess}\, H([\mathbf{Y}])}\right) \geq \left(2\left(1 - \frac{\mu}{\kappa^*}\right)^2 - \left(\frac{14}{3}\right)\mu\right)\sigma_r(\mathcal{A}_{\mathbf{n}}) - 2\sigma_{r+1}(\mathcal{A}_{\mathbf{n}}),$$

$$\sigma_{\max}\left(\overline{\text{Hess}\, H([\mathbf{Y}])}\right) \leq 4\left(\sigma_1(\mathbf{Y}^*) + \frac{\mu\sigma_r(\mathbf{Y}^*)}{\kappa^*}\right)^2 + \frac{14}{3}\mu\sigma_r^2(\mathbf{Y}^*).$$

In particular, if μ *is further chosen such that*

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

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 Eq. (G.2), we have

$$\|\mathbf{Y} - \mathbf{Y}^*\mathbf{Q}\| \leq \|\mathbf{Y} - \mathbf{Y}^*\mathbf{Q}\|_F = d([\mathbf{Y}], [\mathbf{Y}^*]) \leq \frac{\mu\sigma_r(\mathbf{Y}^*)}{\kappa^*}. \qquad (G.7)$$

Thus

$$\begin{aligned}
\sigma_r(\mathbf{Y}) &= \sigma_r(\mathbf{Y} - \mathbf{Y}^*\mathbf{Q} + \mathbf{Y}^*\mathbf{Q}) \geq \sigma_r(\mathbf{Y}^*) - \|\mathbf{Y} - \mathbf{Y}^*\mathbf{Q}\| \overset{\text{Eq. (G.7)}}{\geq} \left(1 - \frac{\mu}{\kappa^*}\right)\sigma_r(\mathbf{Y}^*), \\
\sigma_1(\mathbf{Y}) &= \sigma_1(\mathbf{Y} - \mathbf{Y}^*\mathbf{Q} + \mathbf{Y}^*\mathbf{Q}) \leq \sigma_1(\mathbf{Y}^*) + \|\mathbf{Y} - \mathbf{Y}^*\mathbf{Q}\| \overset{\text{Eq. (G.7)}}{\leq} \sigma_1(\mathbf{Y}^*) + \frac{\mu\sigma_r(\mathbf{Y}^*)}{\kappa^*},
\end{aligned} \qquad (G.8)$$

where the first inequalities follow from Weyl's theorem [66]. Then,

$$\overline{\text{Hess}\, H([\mathbf{Y}])}[\theta_{\mathbf{Y}}, \theta_{\mathbf{Y}}]$$

$$
\begin{aligned}
&= \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{[Eq. (G.1)]} \\
&\geqslant 2\sigma_r^2(\mathbf{Y})\|\theta_\mathbf{Y}\|_\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 G.4]} \\
&= 2\sigma_r^2(\mathbf{Y})\|\theta_\mathbf{Y}\|_\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 && [\mathcal{A}_\mathbf{n} = \mathbf{Y}^*\mathbf{Y}^{*\top} + \mathbf{Z}\mathbf{Z}^\top] \\
&\geqslant 2\sigma_r^2(\mathbf{Y})\|\theta_\mathbf{Y}\|_\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\left\| \mathbf{Z}\mathbf{Z}^\top \right\|\left\| \theta_\mathbf{Y}\theta_\mathbf{Y}^\top \right\|_\mathrm{F} && [\langle A, B \rangle \leq \|A\|\|B\|_\mathrm{F}] \\
&\geqslant 2\sigma_r^2(\mathbf{Y})\|\theta_\mathbf{Y}\|_\mathrm{F}^2 - 2\left\| \mathbf{Y}\mathbf{Y}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top} \right\|\|\theta_\mathbf{Y}\|_\mathrm{F}^2 \\
&\quad - 2\left\| \mathbf{Z}\mathbf{Z}^\top \right\|\|\theta_\mathbf{Y}\|_\mathrm{F}^2 && [\|\theta_\mathbf{Y}\theta_\mathbf{Y}^\top\|_\mathrm{F} = \|\theta_\mathbf{Y}\|_\mathrm{F}^2] \\
&\geqslant 2\left(1 - \frac{\mu}{\kappa^*}\right)^2 \sigma_r^2(\mathbf{Y}^*)\|\theta_\mathbf{Y}\|_\mathrm{F}^2 - 2\left\| \mathbf{Z}\mathbf{Z}^\top \right\|\|\theta_\mathbf{Y}\|_\mathrm{F}^2 \\
&\quad - 2\left\| \mathbf{Y}\mathbf{Y}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top} \right\|\|\theta_\mathbf{Y}\|_\mathrm{F}^2 && \text{[Eq. (G.8)]} \\
&\geqslant 2\left(1 - \frac{\mu}{\kappa^*}\right)^2 \sigma_r^2(\mathbf{Y}^*)\|\theta_\mathbf{Y}\|_\mathrm{F}^2 - 2\left\| \mathbf{Z}\mathbf{Z}^\top \right\|\|\theta_\mathbf{Y}\|_\mathrm{F}^2 \\
&\quad - 2 \cdot \frac{7}{3}\|\mathbf{Y}^*\|\frac{\mu\sigma_r(\mathbf{Y}^*)}{\kappa^*}\|\theta_\mathbf{Y}\|_\mathrm{F}^2 && [\text{Lemma G.2}, \mathbf{Y} \in \mathcal{R}_1] \\
&= 2\left(1 - \frac{\mu}{\kappa^*}\right)^2 \sigma_r^2(\mathbf{Y}^*)\|\theta_\mathbf{Y}\|_\mathrm{F}^2 - 2 \cdot \frac{7}{3}\|\mathbf{Y}^*\|\frac{\mu\sigma_r(\mathbf{Y}^*)}{\kappa^*}\|\theta_\mathbf{Y}\|_\mathrm{F}^2 \\
&\quad - 2\sigma_{r+1}(\mathcal{A}_\mathbf{n})\|\theta_\mathbf{Y}\|_\mathrm{F}^2 && [\|\mathbf{Z}\mathbf{Z}^\top\| = \sigma_{r+1}(\mathcal{A}_\mathbf{n})] \\
&= \left(\left(2\left(1 - \frac{\mu}{\kappa^*}\right)^2 - \frac{14}{3}\mu\right)\sigma_r(\mathcal{A}_\mathbf{n}) - 2\sigma_{r+1}(\mathcal{A}_\mathbf{n})\right)\|\theta_\mathbf{Y}\|_\mathrm{F}^2. && \left[\kappa^* = \frac{\|\mathbf{Y}^*\|}{\sigma_r(\mathbf{Y}^*)}\right]
\end{aligned}
$$

Likewise,

$$
\begin{aligned}
&\overline{\mathrm{Hess}\, H([\mathbf{Y}])}[\theta_\mathbf{Y}, \theta_\mathbf{Y}] \\
&= \left\| \mathbf{Y}\theta_\mathbf{Y}^\top + \theta_\mathbf{Y}\mathbf{Y}^\top \right\|_\mathrm{F}^2 + 2\left\langle \mathbf{Y}\mathbf{Y}^\top - \mathcal{A}_\mathbf{n}, \theta_\mathbf{Y}\theta_\mathbf{Y}^\top \right\rangle && \text{[Eq. (G.1)]} \\
&\leq 4\sigma_1^2(\mathbf{Y})\|\theta_\mathbf{Y}\|_\mathrm{F}^2 + 2\left\langle \mathbf{Y}\mathbf{Y}^\top - \mathcal{A}_\mathbf{n}, \theta_\mathbf{Y}\theta_\mathbf{Y}^\top \right\rangle && \text{[Lemma G.4]} \\
&\leq 4\sigma_1^2(\mathbf{Y})\|\theta_\mathbf{Y}\|_\mathrm{F}^2 + 2\left\langle \mathbf{Y}\mathbf{Y}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top}, \theta_\mathbf{Y}\theta_\mathbf{Y}^\top \right\rangle && [\mathcal{A}_\mathbf{n} - \mathbf{Y}^*\mathbf{Y}^{*\top} \text{ is PSD}] \\
&\leqslant 4\sigma_1^2(\mathbf{Y})\|\theta_\mathbf{Y}\|_\mathrm{F}^2 + 2\left\| \mathbf{Y}\mathbf{Y}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top} \right\|\|\theta_\mathbf{Y}\|_\mathrm{F}^2 \\
&\leqslant 4\sigma_1^2(\mathbf{Y})\|\theta_\mathbf{Y}\|_\mathrm{F}^2 + 2\left\| \mathbf{Y}\mathbf{Y}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top} \right\|_\mathrm{F}\|\theta_\mathbf{Y}\|_\mathrm{F}^2 \\
&\leqslant 4\left(\sigma_1(\mathbf{Y}^*) + \frac{\mu\sigma_r(\mathbf{Y}^*)}{\kappa^*}\right)^2\|\theta_\mathbf{Y}\|_\mathrm{F}^2 + 2\left\| \mathbf{Y}\mathbf{Y}^\top - \mathbf{Y}^*\mathbf{Y}^{*\top} \right\|_\mathrm{F}\|\theta_\mathbf{Y}\|_\mathrm{F}^2 && \text{[Eq. (G.8)]} \\
&\leqslant \left(4\left(\sigma_1(\mathbf{Y}^*) + \frac{\mu\sigma_r(\mathbf{Y}^*)}{\kappa^*}\right)^2 + \frac{14}{3}\mu\sigma_r^2(\mathbf{Y}^*)\right)\|\theta_\mathbf{Y}\|_\mathrm{F}^2. && \text{[Lemma G.2]}
\end{aligned}
$$

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

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

we have $H([\mathbf{Y}])$ in Eq. (2.2) is geodesically strongly convex and smooth in $\mathcal{R}_1$ as $\mathcal{R}_1$ is a geodesically convex set by [51]. 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 $\sigma_r(\mathcal{A}_\mathbf{n}) > \sigma_{r+1}(\mathcal{A}_\mathbf{n})$. $\qquad\square$

Remark G.2. Compared with the bound in [51, Theorem 8], the smoothness and geodesically strongly convexity are as follows:

$$\sigma_{\min}\left(\overline{\operatorname{Hess} H([\mathbf{Y}])}\right) \geqslant \left(2\left(1 - \frac{\mu}{\kappa^*}\right)^2 - \frac{14}{3}\mu\right)\sigma_r^2(\mathbf{Y}^*),$$

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

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 Eq. 2.2). *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 $[n]$ with cardinality r we have that $[\overline{\mathbf{U}}_S\mathbf{\Lambda}_S]$ is a Riemannian FOSP of Eq. (2.2). Further, these are the only Riemannian FOSPs.*

Proof. From Eq. (G.1), the gradient can be written down as,

$$\overline{\operatorname{grad} H([\mathbf{Y}])} = 2(\mathbf{Y}\mathbf{Y}^\top - \mathcal{A}_\mathbf{n})\mathbf{Y} = 2(\mathbf{U}\mathbf{D}\mathbf{V}^\top(\mathbf{U}\mathbf{D}\mathbf{V}^\top)^\top - \mathcal{A}_\mathbf{n})\mathbf{U}\mathbf{D}\mathbf{V}^\top$$
$$= 2(\mathbf{U}\mathbf{D}^3\mathbf{V}^\top - \mathcal{A}_\mathbf{n}\mathbf{U}\mathbf{D}\mathbf{V}^\top).$$

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

$$\mathbf{U}\mathbf{D}^2 - \mathcal{A}_\mathbf{n}\mathbf{U} = \mathbf{0} \tag{G.9}$$

Since $\mathbf{D}^2$ is also a diagonal matrix, to satisfy Eq. (G.9), 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. $\qquad\square$

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 Eq. 2.1). *Assume that Assumption 4.1 holds. Given any $\mathbf{Y} \in \mathbb{R}_*^{n \times r}$ such that $\mathbf{Y} \in \mathcal{R}_2$, let*

$$\theta_\mathbf{Y}^1 = [\mathbf{0}, \mathbf{0}, \ldots, \mathbf{0}, a, \mathbf{0}, \ldots, \mathbf{0}]\mathbf{V}^\top,$$

where a is such that

$$a = \underset{a:\mathbf{Y}^\top a = 0}{\arg\max} \frac{a^\top \mathcal{A_n} a}{\|a\|^2}, \tag{4.3}$$

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

$$\tilde{i} \overset{def}{=} \underset{j \in [r]}{\arg\min} \, \mathbf{D}_{jj}. \tag{4.4}$$

Denote $\theta_\mathbf{Y}^2 = \mathbf{Y} - \mathbf{Y}^* \mathbf{Q}$, 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 $\theta_\mathbf{Y}^1$ or $\theta_\mathbf{Y}^2$. Then

$$\overline{\mathrm{Hess}\, H([\mathbf{Y}])}[\theta_\mathbf{Y}, \theta_\mathbf{Y}]$$

$$\leqslant \min\left\{ -\frac{\sigma_{r+1}^2(\mathbf{\Lambda})}{2}\|\theta_\mathbf{Y}\|^2, \; -2\left(\sigma_r^2(\mathbf{\Lambda}) \left(1 - \frac{e_1^2}{\left|\sigma_r^2(\mathbf{\Lambda}) - e_1 - \sigma_{r+1}^2(\mathbf{\Lambda})\right|^2} \right) - e_1 - \sigma_{r+1}^2(\mathbf{\Lambda}) \right) \|\theta_\mathbf{Y}\|^2, \right.$$

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

In particular, if α and μ satisfies Assumption 4.2, we have $\overline{\mathrm{Hess}\, H([\mathbf{Y}])}$ has at least one negative eigenvalue and $\theta_\mathbf{Y}^1$ or $\theta_\mathbf{Y}^2$ is the escaping direction.

Proof. By the definition of a, $a \in \mathrm{Span}\{\overline{\mathbf{U}}_{1,\ldots,r+1}\}$. 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, $\mathbf{U}^\top a = \mathbf{0}$. Then, by using Eq. (G.1), we have

$$\overline{\mathrm{Hess}\, H([\mathbf{Y}])}[\theta_\mathbf{Y}^1, \theta_\mathbf{Y}^1]$$

$$= \|\mathbf{Y}(\theta_\mathbf{Y}^1)^\top + \theta_\mathbf{Y}^1 \mathbf{Y}^\top\|_\mathrm{F}^2 + 2\langle \mathbf{YY}^\top - \mathcal{A_n}, \theta_\mathbf{Y}^1(\theta_\mathbf{Y}^1)^\top \rangle \qquad \text{[Eq. (G.1)]}$$

$$= \|\mathbf{Y}(\theta_\mathbf{Y}^1)^\top + \theta_\mathbf{Y}^1 \mathbf{Y}^\top\|_\mathrm{F}^2 - 2\langle \mathcal{A_n}, \theta_\mathbf{Y}^1(\theta_\mathbf{Y}^1)^\top \rangle \qquad [\mathbf{Y}^\top a = 0]$$

$$= 2\langle \mathbf{Y}^\top \mathbf{Y}, (\theta_\mathbf{Y}^1)^\top \theta_\mathbf{Y}^1 \rangle + 2\langle \mathbf{Y}(\theta_\mathbf{Y}^1)^\top, \theta_\mathbf{Y}^1 \mathbf{Y}^\top \rangle - 2\langle \mathcal{A_n}, \theta_\mathbf{Y}^1(\theta_\mathbf{Y}^1)^\top \rangle \qquad [\|A\|_\mathrm{F}^2 = \langle A, A\rangle]$$

$$= 2\langle \mathbf{Y}^\top \mathbf{Y}, (\theta_\mathbf{Y}^1)^\top \theta_\mathbf{Y}^1 \rangle - 2\langle \mathcal{A_n}, \theta_\mathbf{Y}^1(\theta_\mathbf{Y}^1)^\top \rangle \qquad [\mathbf{Y}^\top a = 0]$$

$$= 2\langle \mathbf{VD}^2\mathbf{V}^\top, (\theta_\mathbf{Y}^1)^\top \theta_\mathbf{Y}^1 \rangle - 2\langle \mathcal{A_n}, \theta_\mathbf{Y}^1(\theta_\mathbf{Y}^1)^\top \rangle$$

$$= 2\mathbf{D}_{\tilde{i}\tilde{i}}^2\|a\|^2 - 2a^\top \mathcal{A_n} a,$$

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

$$\overline{\mathrm{Hess}\, H([\mathbf{Y}])}[\theta_\mathbf{Y}^1, \theta_\mathbf{Y}^1] = 2\min_i \mathbf{D}_{ii}^2\|a\|^2 - 2a^\top \mathcal{A_n} a. \tag{G.10}$$

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

Case 1. When $\min_i \mathbf{D}_{ii}^2 < \sigma_{r+1}^2(\boldsymbol{\Lambda})/2$. For this case we must have that

$$\overline{\mathrm{Hess}\, H([\mathbf{Y}])}\, [\theta_{\mathbf{Y}}^1, \theta_{\mathbf{Y}}^1] \leq -\frac{\sigma_{r+1}^2(\boldsymbol{\Lambda})}{2} \|\theta_{\mathbf{Y}}^1\|^2.$$

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

Case 2. When $\min_i \mathbf{D}_{ii}^2 \geq \sigma_{r+1}^2(\boldsymbol{\Lambda})/2$. From the proof of Theorem 2.1, the gradient condition of $\mathcal{R}_2$ can be written as

$$
\begin{aligned}
\frac{\alpha \mu \sigma_r^3(\mathbf{Y}^*)}{4\kappa^*} &\geq \left\| \overline{\mathrm{grad}\, H([\mathbf{Y}])} \right\|_{\mathrm{F}} && [\mathbf{Y} \in \mathcal{R}_2] \\
&= \left\| 2(\mathbf{U}\mathbf{D}^3\mathbf{V}^\top - \mathcal{A}_{\mathbf{n}}\mathbf{U}\mathbf{D}\mathbf{V}^\top) \right\|_{\mathrm{F}} && [\text{Eq. (G.1)}] \\
&= \left\| 2(\mathbf{U}\mathbf{D}^2 - \mathcal{A}_{\mathbf{n}}\mathbf{U})\mathbf{D} \right\|_{\mathrm{F}}.
\end{aligned}
$$

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

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

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 $(C\mathbf{D}^2 - \boldsymbol{\Sigma}C)\mathbf{D}$ is $\mathbf{D}_{ii}^3 C_i - \mathbf{D}_{ii}\boldsymbol{\Sigma}C_i$. Therefore, the gradient condition of $\mathcal{R}_2$ can be written as

$$\sum_{i,j} \left(\mathbf{D}_{ii}^3 C_{ji} - \mathbf{D}_{ii}\boldsymbol{\Sigma}_{jj}C_{ji} \right)^2 \leq \frac{\alpha^2 \mu^2 \sigma_r^6(\mathbf{Y}^*)}{(4\kappa^*)^2}. \tag{G.11}$$

We fix i in the left-hand side of Eq. (G.11), we have

$$\sum_j \left(\mathbf{D}_{ii}^2 - \boldsymbol{\Sigma}_{jj} \right)^2 \mathbf{D}_{ii}^2 C_{ji}^2 \leq \frac{\alpha^2 \mu^2 \sigma_r^6(\mathbf{Y}^*)}{(4\kappa^*)^2}, \tag{G.12}$$

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

$$\min_j \left| \mathbf{D}_{ii}^2 - \boldsymbol{\Sigma}_{jj} \right|^2 \leq \sum_j \left(\mathbf{D}_{ii}^2 - \boldsymbol{\Sigma}_{jj} \right)^2 C_{ji}^2 \leq \frac{\alpha^2 \mu^2 \sigma_r^6(\mathbf{Y}^*)}{(4\kappa^*)^2 \sigma_{r+1}^2(\boldsymbol{\Lambda})/2}. \tag{G.13}$$

We use Eq. (G.12) for the second inequality. Equation G.13 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 $\alpha^2 \mu^2 \sigma_r^6(\mathbf{Y}^*)/((4\kappa^*)^2 \sigma_{r+1}^2(\boldsymbol{\Lambda})/2)$ is guaranteed small according to Assumption 4.2.

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

$$\|\xi^1\|^2 + \|\xi^2\|^2 = 1. \tag{G.14}$$

In the following, we divide all the cases into different regimes based on which of the eigenvalues of Λ is close to $D_{\bar{i}\bar{i}}$.

Case 2.1. When

$$\frac{\sigma^2_{r+1}(\Lambda)}{2} \leq D^2_{\bar{i}\bar{i}} \leq \frac{\alpha \mu \sigma^3_r(Y^*)}{2\sqrt{2}\kappa^* \sigma_{r+1}(\Lambda)} + \sigma^2_{r+1}(\Lambda).$$

Notice that the first assumption in Assumption 4.2 essentially guarantees a small

$$e_1 = \frac{\alpha \mu \sigma^3_r(Y^*)}{2\sqrt{2}\kappa^* \sigma_{r+1}(\Lambda)}.$$

Hence, we have

$$
\begin{aligned}
\frac{\alpha^2 \mu^2 \sigma^6_r(Y^*)}{(4\kappa^*)^2} &\geq \sum_j \left(D^2_{\bar{i}\bar{i}} - \Sigma_{jj}\right)^2 D^2_{\bar{i}\bar{i}} C^2_{j\bar{i}} &&\text{[Eq. (G.11)]} \\
&\geq \sum_{j \leq r} \left|\sigma^2_j(\Lambda) - D^2_{\bar{i}\bar{i}}\right|^2 \cdot D^2_{\bar{i}\bar{i}} \cdot C^2_{j\bar{i}} \\
&\geq \left|\sigma^2_r(\Lambda) - D^2_{\bar{i}\bar{i}}\right|^2 \cdot D^2_{\bar{i}\bar{i}} \cdot \|\xi^1\|^2 \\
&\geq \left|\sigma^2_r(\Lambda) - e_1 - \sigma^2_{r+1}(\Lambda)\right|^2 \cdot \frac{\sigma^2_{r+1}(\Lambda)}{2} \cdot \|\xi^1\|^2.
\end{aligned}
$$

Where in the last two inequalities, we use the condition $\sigma^2_{r+1}(\Lambda)/2 \leq D^2_{\bar{i}\bar{i}} \leq e_1 + \sigma^2_{r+1}(\Lambda)$ and that $e_1 < (\sigma^2_r(\Lambda) - \sigma^2_{r+1}(\Lambda))/2$ (follows from Assumption 4.2).

By reordering the inequality, we have

$$\|\xi^1\| \leq \frac{e_1}{|\sigma^2_r(\Lambda) - e_1 - \sigma^2_{r+1}(\Lambda)|}. \tag{G.15}$$

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

$$
\begin{aligned}
\max_{a:Y^\top a=0} \frac{a^\top \mathcal{A}_n a}{\|a\|^2} &= \max_{a:a^\top \overline{U} C=0} \frac{a^\top \mathcal{A}_n a}{\|a\|^2} \\
&= \max_{a:a^\top \overline{U} C=0} \frac{a^\top \overline{U} \Lambda \overline{U}^\top a}{\|a\|^2} &&[\mathcal{A}_n = \overline{U} \Lambda \overline{U}^\top] \\
&= \max_{b:b^\top C=0} \frac{b^\top \Lambda b}{\|b\|^2}. &&[\overline{U}^\top \overline{U} = I] \tag{G.16}
\end{aligned}
$$

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

Case 2.1.1. If $C_{1:r}$ is not full rank. In this case, there exists $\tilde{b}^1 \in \mathbb{R}^r$ such that $\|\tilde{b}^1\| > 0$ and $(\tilde{b}^1)^\top C_{1:r} = \mathbf{0}$. Therefore, by denoting $\bar{b}_{1:r} = t\tilde{b}^1 + b_{1:r}^1$, and $\bar{b}_{r+1:n} = b_{r+1:n}^2$. From the definition of $\bar{b}$ and the fact that $b^\top C = \mathbf{0}$, we have $\bar{b}^\top C = \mathbf{0}$. By letting $t \to \infty$, we have

$$\max_{b^\top C = \mathbf{0}} \frac{b^\top \Lambda b}{\|b\|^2} \geq \frac{\bar{b}^\top \Lambda \bar{b}}{\|\bar{b}\|^2} \geq \sigma_r^2(\Lambda). \tag{G.17}$$

Combining Eq. (G.17), Eq. (G.10) and the assumption that $\mathbf{D}_{\tilde{i}\tilde{i}}^2 \leq e_1 + \sigma_{r+1}^2(\Lambda)$, this implies

$$\overline{\text{Hess} \, H([\mathbf{Y}])}[\theta_\mathbf{Y}^1, \theta_\mathbf{Y}^1] \leqslant -\left(\sigma_r^2(\mathbf{Y}^*) - \sigma_{r+1}(\Lambda) - e_1\right)\|\theta_\mathbf{Y}^1\|_\text{F}^2. \tag{G.18}$$

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

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

$$b^1 \cdot \xi^1 + \|\xi^2\|^2 = 0. \tag{G.19}$$

By using the Cauchy inequality, this further implies that

$$\|b^1\| \geq \frac{\|\xi^2\|^2}{\|\xi^1\|}. \tag{G.20}$$

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

$$\begin{aligned}
\max_{b^\top C = \mathbf{0}} \frac{b^\top \Lambda b}{\|b\|^2} &\geq \frac{(b^1 + b^2)^\top \Lambda (b^1 + b^2)}{\|b^1 + b^2\|^2} \\
&= \frac{(b^1)^\top \Lambda b^1 + (b^2)^\top \Lambda b^2}{\|b^1\|^2 + \|b^2\|^2} \\
&\geq \frac{(b^1)^\top \Lambda b^1}{\|b^1\|^2 + \|\xi^2\|^2} \\
&\geq \frac{\|b^1\|^2 \cdot \sigma_r^2(\Lambda)}{\|b^1\|^2 + \|\xi^2\|^2}
\end{aligned}$$

$$\geq \frac{(\|\xi^2\|^4/\|\xi^1\|^2) \cdot \sigma_r^2(\mathbf{\Lambda})}{\|\xi^2\|^4/\|\xi^1\|^2 + \|\xi^2\|^2}$$

$$= \|\xi^2\|^2 \cdot \sigma_r^2(\mathbf{\Lambda}), \tag{G.21}$$

where the first equality follows from the definition of b^1 and b^2; the second inequality follows from the assumption that $\mathbf{\Lambda}$ is PSD, and $b^2 = \xi^2$; the third inequality follows from the fact that $b_i^1 = 0$ for $i > r$; the fourth inequality follows from Eq. (G.20); the last equality follows from Eq. (G.14). By using Eq. (G.16), this can be written as

$$\max_{a:\mathbf{Y}^\top a = 0} \frac{a^\top \mathcal{A}_\mathbf{n} a}{\|a\|^2} \geq \|\xi^2\|^2 \cdot \sigma_r^2(\mathbf{\Lambda}). \tag{G.22}$$

By the definition in Eqs. (4.3) and (G.10), we have

$$\overline{\mathrm{Hess}\,H([\mathbf{Y}])}\big[\theta_\mathbf{Y}^1, \theta_\mathbf{Y}^1\big]$$

$$= 2\min_i \mathbf{D}_{ii}^2 \cdot \|a\|^2 - 2a^\top \mathcal{A}_\mathbf{n} a \qquad\qquad [\text{Eq. (G.10)}]$$

$$\leq 2\mathbf{D}_{\tilde{i}\tilde{i}}^2 \cdot \|a\|^2 - 2\sigma_r^2(\mathbf{\Lambda})\|\xi^2\|^2 \cdot \|a\|^2 \qquad\qquad [\text{Eq. (G.22)}]$$

$$= 2\mathbf{D}_{\tilde{i}\tilde{i}}^2 \cdot \|a\|^2 - 2\sigma_r^2(\mathbf{\Lambda})(1 - \|\xi^1\|^2) \cdot \|a\|^2 \qquad\qquad [\|\xi^1\|^2 + \|\xi^2\|^2 = 1]$$

$$\leq \big(-2\sigma_r^2(\mathbf{\Lambda})(1 - \|\xi^1\|^2) + 2e_1 + 2\sigma_{r+1}^2(\mathbf{\Lambda})\big)\|\theta_\mathbf{Y}^1\|^2,$$

where the last inequality follows from $\mathbf{D}_{\tilde{i}\tilde{i}}^2 \leq e_1 + \sigma_{r+1}^2(\mathbf{\Lambda})$ and the fact that $\|\theta_\mathbf{Y}^1\| = \|a\|$. Finally, by applying Eq. (G.15) to control $\|\xi^1\|$, we conclude that

$$\overline{\mathrm{Hess}\,H([\mathbf{Y}])}\big[\theta_\mathbf{Y}^1, \theta_\mathbf{Y}^1\big]$$

$$\leq -2\left(\sigma_r^2(\mathbf{\Lambda})\left(1 - \frac{e_1^2}{\big|\sigma_r^2(\mathbf{\Lambda}) - e_1 - \sigma_{r+1}^2(\mathbf{\Lambda})\big|^2}\right) - e_1 - \sigma_{r+1}^2(\mathbf{\Lambda})\right)\|\theta_\mathbf{Y}^1\|^2. \tag{G.23}$$

According to the second assumption in Assumption 4.2, Eq. (G.23) guarantees an escape direction.

Case 2.2. When $\mathbf{D}_{\tilde{i}\tilde{i}}^2 > e_1 + \sigma_{r+1}^2(\mathbf{\Lambda})$. Recall the first assumption in Assumption 4.2, we have e_1 is small enough, which is viewed as an error term. In the following, we will show that $\theta_\mathbf{Y}^2$ is the escaping direction. We have

$$\min_j \mathbf{D}_{\tilde{i}\tilde{i}}^2 \big|\mathbf{D}_{\tilde{i}\tilde{i}}^2 - \mathbf{\Sigma}_{jj}\big|^2 \leq \mathbf{D}_{\tilde{i}\tilde{i}}^2 \sum_j \big(\mathbf{D}_{\tilde{i}\tilde{i}}^2 - \mathbf{\Sigma}_{jj}\big)^2 C_{j\tilde{i}}^2 \leq \frac{\alpha^2 \mu^2 \sigma_r^6(\mathbf{Y}^*)}{(4\kappa^*)^2}, \tag{G.24}$$

where we use Eq. (G.12) in the last inequality.

Recall that Assumption 4.2 guarantees small e_1 and e_2, by combining Eq. (G.24) and the assumption $\mathbf{D}_{\tilde{i}\tilde{i}}^2 > \sigma_{r+1}^2(\mathbf{\Lambda}) + e_1$, we must have

$$\mathbf{D}_{\tilde{i}\tilde{i}}^2 \geq \sigma_r^2(\mathbf{\Lambda}) - e_2, \tag{G.25}$$
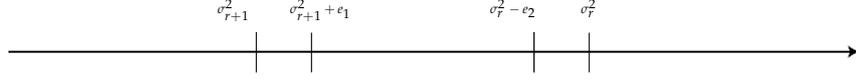
Figure G.1: The value of $\mathbf{D}_{\tilde{i}\tilde{i}}$ must be close to some $\sigma_i(\mathbf{\Lambda})$ according to Eq. (G.24). If $\mathbf{D}_{\tilde{i}\tilde{i}}^2 > \sigma_{r+1}^2 + e_1$, then we must have $\mathbf{D}_{\tilde{i}\tilde{i}}^2 \geq \sigma_r^2 - e_2$.

where e_2 is defined in Assumption 4.2. Otherwise, if $\sigma_{r+1}^2(\mathbf{\Lambda}) + e_1 < \mathbf{D}_{\tilde{i}\tilde{i}}^2 < \sigma_r^2(\mathbf{\Lambda}) - e_2$, this contradicts to Eq. (G.24), see an illustration of this fact in Fig. G.1.

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

$$\frac{\alpha^2\mu^2\sigma_r^6(\mathbf{Y}^*)}{(4\kappa^*)^2} \geq \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\mathbf{D}_{ii}^2 - \mathbf{\Sigma}_{jj}\right)^2\mathbf{D}_{ii}^2\mathbf{C}_{ji}^2 \qquad \text{[Eq. (G.11)]}$$

$$\geq \sum_{i=1}^{n}\sum_{j=r+1}^{n}\left|\sigma_j^2(\mathbf{\Lambda}) + e_2 - \sigma_r^2(\mathbf{\Lambda})\right|^2 \cdot \mathbf{D}_{ii}^2\mathbf{C}_{ji}^2,$$

where Eq. (G.25) and the first assumption in Assumption 4.2 guarantees the last inequality because e_2 is small with respect to $\sigma_r^2(\mathbf{\Lambda}) - \sigma_{r+1}^2(\mathbf{\Lambda})$. Therefore,

$$\sum_{i=1}^{n}\sum_{j=r+1}^{n}\mathbf{D}_{ii}^2\mathbf{C}_{ij}^2 \leq \frac{e_3^2}{\left|\sigma_r^2(\mathbf{\Lambda}) - e_2 - \sigma_{r+1}^2(\mathbf{\Lambda})\right|^2}, \qquad (G.26)$$

where e_3 is defined in Assumption 4.2. Recall that e_3 is small enough, guaranteed in Assumption 4.2. Also recall that $e_2 = e_1/\sqrt{2}$, which is guaranteed to be small enough as in the first assumption in Assumption 4.2, 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 $(r+1)$-th to n-th eigenvalues of $\mathbf{\Sigma}$, then we have

$$\left\langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \mathbf{Y}\mathbf{Y}^\top \right\rangle = \left\langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \mathbf{U}\mathbf{D}^2\mathbf{U}^\top \right\rangle = \left\langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \overline{\mathbf{U}}\mathbf{C}\mathbf{D}^2\mathbf{C}^\top\overline{\mathbf{U}}^\top \right\rangle$$

$$= \left\langle \mathbf{\Sigma}_{(r+1):n}, \mathbf{C}\mathbf{D}^2\mathbf{C}^\top \right\rangle \leq \sigma_{r+1}^2(\mathbf{\Lambda})\sum_{j=r+1}^{n}\sum_{i}\mathbf{C}_{ij}^2\mathbf{D}_{ii}^2$$

$$\leq \frac{e_3^2\sigma_{r+1}^2(\mathbf{\Lambda})}{\left|\sigma_r^2(\mathbf{\Lambda}) - e_2 - \sigma_{r+1}^2(\mathbf{\Lambda})\right|^2}, \qquad (G.27)$$

where the last inequality follows from Eq. (G.26). Eq. (G.27) directly implies,

$$\left\langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \theta_{\mathbf{Y}}^2\left(\theta_{\mathbf{Y}}^2\right)^\top \right\rangle = \left\langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \mathbf{Y}\mathbf{Y}^\top \right\rangle \leq \frac{e_3^2\sigma_{r+1}^2(\mathbf{\Lambda})}{\left|\sigma_r^2(\mathbf{\Lambda}) - e_2 - \sigma_{r+1}^2(\mathbf{\Lambda})\right|^2}, \qquad (G.28)$$

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

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

$$\mathbf{Y}(\theta_{\mathbf{Y}}^2)^{\top} - \mathbf{X}^* + \theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^{\top} = \mathbf{Y}(\theta_{\mathbf{Y}}^2)^{\top} + \theta_{\mathbf{Y}}^2\mathbf{Y}^{\top}, \tag{G.29}$$

and by using Eq. (G.1),

$$
\begin{aligned}
&\langle \overline{\operatorname{grad} H([\mathbf{Y}])}, \theta_{\mathbf{Y}}^2 \rangle \\
&= \langle 2(\mathbf{Y}\mathbf{Y}^{\top} - \mathcal{A}_{\mathbf{n}})\mathbf{Y}, \theta_{\mathbf{Y}}^2 \rangle \\
&= \langle 2(\mathbf{Y}\mathbf{Y}^{\top} - \mathcal{A}_{\mathbf{n}}), \theta_{\mathbf{Y}}^2\mathbf{Y}^{\top} \rangle \\
&= \langle \mathbf{Y}\mathbf{Y}^{\top} - \mathcal{A}_{\mathbf{n}}, \theta_{\mathbf{Y}}^2\mathbf{Y}^{\top} + \mathbf{Y}(\theta_{\mathbf{Y}}^2)^{\top} \rangle && \text{[first argument is symmetric]} \\
&= \langle \mathbf{Y}\mathbf{Y}^{\top} - \mathcal{A}_{\mathbf{n}}, \theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^{\top} + \mathbf{Y}\mathbf{Y}^{\top} - \mathbf{X}^* \rangle, \tag{G.30}
\end{aligned}
$$

where the last equality follows from Eq. (G.29)

$$
\begin{aligned}
&\overline{\operatorname{Hess} H([\mathbf{Y}])}[\theta_{\mathbf{Y}}^2, \theta_{\mathbf{Y}}^2] \\
&= \|\mathbf{Y}(\theta_{\mathbf{Y}}^2)^{\top} + \theta_{\mathbf{Y}}^2\mathbf{Y}^{\top}\|_{\mathrm{F}}^2 + 2\langle \mathbf{Y}\mathbf{Y}^{\top} - \mathcal{A}_{\mathbf{n}}, \theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^{\top} \rangle && \text{[Eq. (G.1)]} \\
&= \|\mathbf{Y}\mathbf{Y}^{\top} - \mathbf{X}^* + \theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^{\top}\|_{\mathrm{F}}^2 + 2\langle \mathbf{Y}\mathbf{Y}^{\top} - \mathcal{A}_{\mathbf{n}}, \theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^{\top} \rangle && \text{[Eq. (G.29)]} \\
&= \|\theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^{\top}\|_{\mathrm{F}}^2 + \|\mathbf{Y}\mathbf{Y}^{\top} - \mathbf{X}^*\|_{\mathrm{F}}^2 + 4\langle \mathbf{Y}\mathbf{Y}^{\top} - \mathbf{X}^*, \theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^{\top} \rangle \\
&\quad - 2\langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^{\top} \rangle \\
&= \|\theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^{\top}\|_{\mathrm{F}}^2 - 3\|\mathbf{Y}\mathbf{Y}^{\top} - \mathbf{X}^*\|_{\mathrm{F}}^2 + 4\langle \mathbf{Y}\mathbf{Y}^{\top} - \mathbf{X}^*, \mathbf{Y}\mathbf{Y}^{\top} - \mathbf{X}^* + \theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^{\top} \rangle \\
&\quad - 2\langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^{\top} \rangle \\
&= \|\theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^{\top}\|_{\mathrm{F}}^2 - 3\|\mathbf{Y}\mathbf{Y}^{\top} - \mathbf{X}^*\|_{\mathrm{F}}^2 + 4\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} \rangle \\
&\quad + 2\langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^{\top} \rangle + 4\langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \mathbf{Y}\mathbf{Y}^{\top} - \mathbf{X}^* \rangle \\
&= \|\theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^{\top}\|_{\mathrm{F}}^2 - 3\|\mathbf{Y}\mathbf{Y}^{\top} - \mathbf{X}^*\|_{\mathrm{F}}^2 + 4\langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \mathbf{Y}\mathbf{Y}^{\top} - \mathbf{X}^* \rangle \\
&\quad + 2\langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^{\top} \rangle + 4\langle \overline{\operatorname{grad} H([\mathbf{Y}])}, \theta_{\mathbf{Y}}^2 \rangle. && \text{[Eq. (G.30)]}
\end{aligned}
$$

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

First, for

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

we have

$$
\begin{aligned}
&\|\theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^{\top}\|_{\mathrm{F}}^2 - 3\|\mathbf{Y}\mathbf{Y}^{\top} - \mathbf{X}^*\|_{\mathrm{F}}^2 + 2\langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^{\top} \rangle \\
&\quad + 4\langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \mathbf{Y}\mathbf{Y}^{\top} - \mathbf{X}^* \rangle \\
&\leq -\|\mathbf{Y}\mathbf{Y}^{\top} - \mathbf{X}^*\|_{\mathrm{F}}^2 + 2\langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \theta_{\mathbf{Y}}^2(\theta_{\mathbf{Y}}^2)^{\top} \rangle \\
&\quad + 4\langle \mathcal{A}_{\mathbf{n}} - \mathbf{X}^*, \mathbf{Y}\mathbf{Y}^{\top} - \mathbf{X}^* \rangle && \text{[Eq. (G.4)]}
\end{aligned}
$$

$$= -\|\mathbf{Y}\mathbf{Y}^\top - \mathbf{X}^*\|_F^2 + 2\langle \mathcal{A}_\mathbf{n} - \mathbf{X}^*, \theta_\mathbf{Y}^2(\theta_\mathbf{Y}^2)^\top\rangle + 4\langle \mathcal{A}_\mathbf{n} - \mathbf{X}^*, \mathbf{Y}\mathbf{Y}^\top\rangle \qquad [\langle \mathcal{A}_\mathbf{n} - \mathbf{X}^*, \mathbf{X}^*\rangle = 0]$$

$$\leq -\|\mathbf{Y}\mathbf{Y}^\top - \mathbf{X}^*\|_F^2 + 6\frac{e_3^2\sigma_{r+1}^2(\boldsymbol{\Lambda})}{\left|\sigma_r^2(\boldsymbol{\Lambda}) - e_2 + \sigma_{r+1}^2(\boldsymbol{\Lambda})\right|^2} \qquad\qquad [\text{Eq. (G.27), Eq. (G.28)}]$$

$$\leq -2(\sqrt{2}-1)\sigma_r^2(\mathbf{Y}^*)\|\theta_\mathbf{Y}^2\|_F^2 + 6\frac{e_3^2\sigma_{r+1}^2(\boldsymbol{\Lambda})}{\left|\sigma_r^2(\boldsymbol{\Lambda}) - e_2 + \sigma_{r+1}^2(\boldsymbol{\Lambda})\right|^2}. \qquad [\text{Eq. (G.3)}]$$

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

$$\langle \overline{\mathrm{grad}\, H([\mathbf{Y}])}, \theta_\mathbf{Y}^2\rangle \leq \|\overline{\mathrm{grad}\, H([\mathbf{Y}])}\|_F \|\theta_\mathbf{Y}^2\|_F \leq \alpha\sigma_r^2(\mathbf{Y}^*)\|\theta_\mathbf{Y}^2\|_F^2,$$

where the last inequality is because

$$\|\overline{\mathrm{grad}\, H([\mathbf{Y}])}\|_F \leqslant \frac{\alpha\mu\sigma_r^3(\mathbf{Y}^*)}{4\kappa^*}.$$

According to the definition of $\mathcal{R}_2$ in Eq. (4.2), $\mathbf{Y} \in \mathcal{R}_2$ also implies

$$d([\mathbf{Y}], [\mathbf{Y}^*]) > \frac{\mu\sigma_r(\mathbf{Y}^*)}{\kappa^*},$$

then

$$\|\overline{\mathrm{grad}\, H([\mathbf{Y}])}\|_F \leqslant \frac{\alpha d([\mathbf{Y}], [\mathbf{Y}^*])\sigma_r^2(\mathbf{Y}^*)}{4} = \frac{\alpha\|\theta_\mathbf{Y}^2\|_F \sigma_r^2(\mathbf{Y}^*)}{4}.$$

By combining the above three inequalities, we have

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

where the last inequality follows from

$$\frac{\mu\sigma_r(\mathbf{Y}^*)}{\kappa^*} \leq d([\mathbf{Y}], [\mathbf{Y}^*]) = \|\theta_\mathbf{Y}\|_F,$$

and the definition of e_3 in Assumption 4.2.

Finally, according to the third assumption in Assumption 4.2, 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. $\qquad\square$

Remark G.3. 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. This theorem guarantees that when $\mathbf{Y} \in \mathcal{R}_3$, the magnitude of the Riemannian gradient is large. We restate it again for convenience.

Theorem 4.3 (Regions with Large Riemannian Gradient of Eq. (2.1))**.**

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

2. $\left\| \overline{\operatorname{grad} H([\mathbf{Y}])} \right\|_{\mathrm{F}} \geqslant 2(\|\mathbf{Y}\|^3 - \|\mathbf{Y}\| \|\mathbf{Y}^*\|^2) > 2(\beta^3 - \beta)\|\mathbf{Y}^*\|^3, \ \forall \mathbf{Y} \in \mathcal{R}_3''$;

3. $\left\langle \overline{\operatorname{grad} H([\mathbf{Y}])}, \mathbf{Y} \right\rangle > 2(1 - 1/\gamma)\|\mathbf{Y}\mathbf{Y}^\top\|_{\mathrm{F}}^2, \ \forall \mathbf{Y} \in \mathcal{R}_3'''$.

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 $\mathcal{R}_3'''$.

Proof. The proof of this result is the same as the proof of [51, Theorem 10] without any modifications. Indeed, notice that $\mathbf{Y} \in \mathcal{R}_3$ does not require Assumption 4.1 because $\mathcal{R}_3$ describes the case that $[\mathbf{Y}]$ is far away from the FOSP. $\qquad\square$

Acknowledgments

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, 2009.

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

[3] R. Ando and T. Zhang, Learning on graph with Laplacian regularization, in: *Advances in Neural Information Processing Systems*, Vol. 19, Curran Associates, Inc., 25–32, 2006.

[4] S. Arora, H. Khandeparkar, M. Khodak, O. Plevrakis, and N. Saunshi, A theoretical analysis of contrastive unsupervised representation learning, *arXiv:1902.09229*, 2019.

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

[6] P. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, Benign overfitting in linear regression, *Proc. Natl. Acad. Sci. USA*, 117(48):30063–30070, 2020.

[7] M. Belkin and P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, in: *Advances in Neural Information Processing Systems*, Vol. 14, MIT Press, 585–591, 2001.

[8] M. Belkin, P. Niyogi, and V. Sindhwani, Manifold regularization: A geometric framework for learning from labeled and unlabeled examples, *J. Mach. Learn. Res.*, 7(11):2399–2434, 2006.

[9] S. Bhatnagar, Y. Afshar, S. Pan, K. Duraisamy, and S. Kaushik, Prediction of aerodynamic flow fields using convolutional neural networks, *Comput. Mech.*, 64:525–545, 2019.

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

[11] P. Bréchet, K. Papagiannouli, J. An, and G. Montúfar, Critical points and convergence analysis of generative deep linear networks trained with Bures-Wasserstein loss, *Proc. Mach. Learn. Res.*, 202:3106–3147, 2023.

[12] D. Burago, S. Ivanov, and Y. Kurylev, A graph discretization of the Laplace-Beltrami operator, *J. Spectr. Theory*, 4(4):675–714, 2014.

[13] J. Calder and N. García Trillos, Improved spectral convergence rates for graph Laplacians on ε-graphs and k-NN graphs, *Appl. Comput. Harmon. Anal.*, 60:123–175, 2022.

[14] J. Calder, N. García Trillos, and M. Lewicka, Lipschitz regulariy of graph Laplacians on random data clouds, *SIAM J. Math. Anal.*, 54(1):1169–1222, 2022.

[15] M. Chen, H. Jiang, W. Liao, and T. Zhao, Nonparametric regression on low-dimensional manifolds using deep ReLU networks: Function approximation and statistical recovery, *Inf. Inference J. IMA*, 11(4):1203–1253, 2022.

[16] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, A simple framework for contrastive learning of visual representations, in: *Proceedings of the 37th International Conference on Machine Learning*, PMLR, 119:1597–1607, 2020.

[17] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, Big self-supervised models are strong semi-supervised learners, in: *Advances in Neural Information Processing Systems*, Vol. 33, Curran Associates, Inc., 22243–22255, 2020.

[18] X. Chen, H. Fan, R. Girshick, and K. He, Improved baselines with momentum contrastive learning, *arXiv:2003.04297*, 2020.

[19] Y. Chi, Y. M. Lu, and Y. Chen, Nonconvex optimization meets low-rank matrix factorization: An overview, *IEEE Trans. Signal Process.*, 67(20):5239–5269, 2019.

[20] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps, *Proc. Natl. Acad. Sci. USA*, 102(21):7426–7431, 2005.

[21] C. Criscitiello and N. Boumal, Efficiently escaping saddle points on manifolds, in: *Advances in Neural Information Processing Systems*, Vol. 32, Curran Associates, Inc., 5955–5965, 2019.

[22] Z. Deng, J. Shi, and J. Zhu, NeuralEF: Deconstructing kernels by deep neural networks, *Proc. Mach. Learn. Res.*, 162:4976–4992, 2022.

[23] M. P. Do Carmo and F. Flaherty, *Riemannian Geometry*, Springer, 1992.

[24] E. Dobriban and S. Wager, High-dimensional asymptotics of prediction: Ridge regression and classification, *Ann. Stat.*, 46(1):247–279, 2018.

[25] D. B. Dunson, H. T. Wu, and N. Wu, Spectral convergence of graph Laplacian and heat kernel reconstruction in l^∞ from random samples, *Appl. Comput. Harmon. Anal.*, 55:282–336, 2021.

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

[27] N. García Trillos, M. Gerlach, M. Hein, and D. Slepčev, Error estimates for spectral convergence of the graph Laplacian on random geometric graphs toward the Laplace-Beltrami operator, *Found. Comput. Math.*, 20(4):827–887, 2020.

[28] N. García Trillos, P. He, and C. Li, Large sample spectral analysis of graph-based multi-manifold clustering, *J. Mach. Learn. Res.*, 24(143):1–71, 2023.

[29] N. García Trillos, C. Li, and R. Venkatraman, Minimax rates for the estimation of eigenpairs of weighted Laplace-Beltrami operators on manifolds, *arXiv:2506.00171v1*, 2025.

[30] N. García Trillos and D. Slepcev, A variational approach to the consistency of spectral clustering, *Appl. Comput. Harmon. Anal.*, 45(2):239–281, 2018.

[31] N. García Trillos, D. Sleṕcev, J. Von Brecht, T. Laurent, and X. Bresson, Consistency of cheeger and ratio graph cuts, *J. Mach. Learn. Res.*, 17(181):1–46, 2016.

[32] R. Ge, F. Huang, C. Jin, and Y. Yuan, Escaping from saddle points – online stochastic gradient for tensor decomposition, in: *Proceedings of The 28th Conference on Learning Theory*, 40:797–842, PMLR , 2015.

[33] X. Guo, W. Li, and F. Iorio, Convolutional neural networks for steady flow approximation, in: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 481–490, 2016.

[34] J. Z. HaoChen, C. Wei, A. Gaidon, and T. Ma, Provable guarantees for self-supervised deep learning with spectral contrastive loss, in: *Advances in Neural Information Processing Systems*, Vol. 34, Curran Associates, Inc., 5000–5011, 2021.

[35] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, Surprises in high-dimensional ridgeless least squares interpolation, *The Ann. Stat.*, 50(2):949–986, 2022.

[36] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, Momentum contrast for unsupervised visual representation learning, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, 9729–9738, 2020.

[37] J. Hermann, Z. Schätzle, and F. Noé, Deep-neural-network solution of the electronic Schrödinger equation, *Nat. Chem.*, 12(10):891–897, 2020.

[38] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, How to escape saddle points efficiently, in: *Proceedings of the 34th International Conference on Machine Learning*, 70:1724-1732, PMLR, 2017.

[39] C. Kausik, K. Srivastava, and R. Sonthalia, Double descent and overfitting under noisy inputs and distribution shift for linear denoisers, *Trans. Mach. Learn. Res.*, 2024. `https://openreview.net/forum?id=HxfqTdLIRF`

[40] N. Kovachki, Z. Li, B. Liu, K. Azizzadenesheli, K. Bhattacharya, A. Stuart, and A. Anandkumar, Neural operator: Learning maps between function spaces, *arXiv:2108.08481*, 2021.

[41] J. D. Lee, Q. Lei, N. Saunshi, and J. Zhuo, Predicting what you already know helps: Provable self-supervised learning, in: *Advances in Neural Information Processing Systems*, Vol. 34, Curran Associates, Inc., 309–323, 2021.

[42] J. M. Lee, *Introduction to Riemannian Manifolds*, in: *Graduate Texts in Mathematics*, Vol. 176, Springer, 2018.

[43] Q. Li and G. Tang, The nonconvex geometry of low-rank matrix optimizations with general objective functions, in: *2017 IEEE Global Conference on Signal and Information Processing*, IEEE, 1235–1239, 2017.

[44] X. Li, J. Lu, R. Arora, J. Haupt, H. Liu, Z. Wang, and T. Zhao, Symmetry, saddle points, and global optimization landscape of nonconvex matrix factorization, *IEEE Trans. Inf. Theory*, 65(6):3489–3514, 2019.

[45] X. Li and R. Sonthalia, Least squares regression can exhibit under-parameterized double descent, in: *Advances in Neural Information Processing Systems*, Vol. 37, Curran Associates, Inc., 25510–25560, 2024.

[46] Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar, Fourier neural operator for parametric partial differential equations, *arXiv:2010.08895*, 2020.

[47] Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar, Neural operator: Graph kernel network for partial differential equations, *arXiv:2003.03485*, 2020.

[48] J. Lu, Graph approximations to the Laplacian spectra, *J. Topol. Anal.*, 14(01):111–145, 2022.

[49] J. Lu and Y. Lu, A priori generalization error analysis of two-layer neural networks for solving high dimensional Schrödinger eigenvalue problems, *Commun. Am. Math. Soc.*, 2(1):1–21, 2022.

[50] J. Lu, Z. Shen, H. Yang, and S. Zhang, Deep network approximation for smooth functions, *SIAM J. Math. Anal.*, 53(5):5465–5506, 2021.

[51] Y. Luo and N. García Trillos, Nonconvex matrix factorization is geodesically convex: Global landscape analysis for fixed-rank matrix optimization from a Riemannian perspective, *arXiv:2209.15130*, 2022.

[52] Y. Luo, X. Li, and A. R. Zhang, On geometric connections of embedded and quotient geometries in Riemannian fixed-rank matrix optimization, *arXiv:2110.12121*, 2021.

[53] N. R. Mallinar, J. B. Simon, A. Abedsoltan, P. Pandit, M. Belkin, and P. Nakkiran, Benign, tempered,

or catastrophic: Toward a refined taxonomy of overfitting, in: *Advances in Neural Information Processing Systems*, Vol. 35, Curran Associates, Inc., 1182–1195, 2022.

[54] E. Massart and P.-A. Absil, Quotient geometry with simple geodesics for the manifold of fixed-rank positive-semidefinite matrices, *SIAM J. Matrix Anal. Appl.*, 41(1):171–198, 2020.

[55] H. Min, S. Tarmoun, R. Vidal, and E. Mallada, On the explicit role of initialization on the convergence and implicit bias of overparametrized linear networks, *Proc. Mach. Learn. Res.*, 139:7760–7768, 2021.

[56] A. Ng, M. Jordan, and Y. Weiss, On spectral clustering: Analysis and an algorithm, in: *Advances in Neural Information Processing Systems*, Vol. 14, MIT Press, 849–856, 2001.

[57] P. C. Petersen, Neural Network Theory, *University of Vienna*, 2020. `http://pc-petersen.eu/Neural_Network_Theory.pdf`

[58] D. Pfau, S. Petersen, A. Agarwal, D. G. T. Barrett, and K. L. Stachenfeld, Spectral inference networks: Unifying deep and spectral learning, in: *7th International Conference on Learning Representations*, 2019. `https://openreview.net/forum?id=SJzqpj09YQ`

[59] M. Raissi, P. Perdikaris, and G. E. Karniadakis, Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, *J. Comput. Phys.*, 378:686–707, 2019.

[60] J. Schmidt-Hieber, Deep ReLU network approximation of functions on a manifold, *arXiv:1908.00695*, 2019.

[61] U. Shaham, K. Stanton, H. Li, R. Basri, B. Nadler, and Y. Kluger, SpectralNet: Spectral clustering using deep neural networks, in: *International Conference on Learning Representations*, 2018. `https://openreview.net/forum?id=HJ_aoCyRZ`

[62] Z. Shen, H. Yang, and S. Zhang, Nonlinear approximation via compositions, *Neural Networks*, 119:74–84, 2019.

[63] Z. Shen, H. Yang, and S. Zhang, Deep network approximation characterized by number of neurons, *Commun. Comput. Phys.*, 28(5):1768–1811, 2020.

[64] A. J. Smola and R. Kondor, Kernels and regularization on graphs, in: *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop*, Springer, 144–158, 2003.

[65] R. Sonthalia and R. R. Nadakuditi, Training data size induced double descent for denoising feedforward neural networks and the role of training noise, *Trans. Mach. Learn. Res.*, 2023. `https://openreview.net/forum?id=FdMWtpVT1I`

[66] G. W. Stewart, *Matrix Algorithms: Volume 1: Basic Decompositions*, SIAM, 1998.

[67] J. Sun, Q. Qu, and J. Wright, Complete dictionary recovery over the sphere I: Overview and the geometric picture, *IEEE Trans. Inf. Theory*, 63(2):853–884, 2017.

[68] Y. Sun, N. Flammarion, and M. Fazel, Escaping from saddle points on Riemannian manifolds, in: *Advances in Neural Information Processing Systems*, Vol. 32, Curran Associates, Inc., 7244–7254, 2019.

[69] W. Tao and Z. Shi, Convergence of Laplacian spectra from random samples, *J. Comput. Math.*, 38(6):952–984, 2020.

[70] S. Tarmoun, G. Franca, B. D. Haeffele, and R. Vidal, Understanding the dynamics of gradient flow in overparameterized linear models, in: *Proceedings of Machine Learning Research*, Vol. 139, PMLR, 10153–10161, 2021.

[71] C. Tosh, A. Krishnamurthy, and D. Hsu, Contrastive estimation reveals topic posterior information to linear models, *J. Mach. Learn. Res.*, 22(281):1–31, 2021.

[72] U. Von Luxburg, A tutorial on spectral clustering, *Stat. Comput.*, 17(4):395–416, 2007.

[73] Y. Wang, R. Sonthalia, and W. Hu, Near-interpolators: Rapid norm growth and the trade-off between interpolation and generalization, in: *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, PMLR, 238:4483–4491, 2024.

[74] E. Weinan and T. Yu, The Deep Ritz method: A deep learning-based numerical algorithm for solving variational problems, *Commun. Math. Stat.*, 6:1–12, 2017.

[75] C. L. Wormell and S. Reich, Spectral convergence of diffusion maps: Improved error bounds and an alternative normalization, *SIAM J. Numer. Anal.*, 59(3):1687–1734, 2021.

[76] D. Yarotsky, Optimal approximation of continuous functions by very deep ReLU networks, in: *Proceedings of Machine Learning Research*, PMLR, 75:639–649, 2018.

[77] C. Yun, S. Sra, and A. Jadbabaie, Small ReLU networks are powerful memorizers: A tight analysis of memorization capacity, in: *Advances in Neural Information Processing Systems*, Vol. 32, Curran Associates, Inc., 15479–15490, 2019.

[78] Y. Zhu and N. Zabaras, Bayesian deep convolutional encoder-decoder networks for surrogate modeling and uncertainty quantification, *J. Comput. Phys.*, 366:415–447, 2018.