FEATURE LEARNING FOR THE HIGH DIMENSIONAL STATIONARY SCHÖDINGER EQUATION WITH DEEP RITZ METHOD

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper investigates feature learning within the framework of the deep Ritz method for solving the stationary Schrödinger equation with Neumann boundary conditions. We first analyze the convergence of Riemannian gradient descent in an agnostic setting, where the hypothesis function is restricted to a single-index model while the PDE solution is arbitrary. We prove that gradient descent reaches an approximate global minimum: after $T = O(\log(1/\epsilon))$ iterations, the loss is within ϵ of a constant multiple of the optimal loss. We then examine the loss landscape when the source term of the PDE itself follows a single-index model, considering hypothesis functions given by either a single-index model or a two-neuron multi-index model. In the single-index case, we show that the minimum Ritz energy is attained at the feature vector aligned with that of the source term. In the two-neuron case, we study the landscape of regularized Ritz losses and characterize how a second feature emerges, given that the first feature is aligned with the source, as the regularization parameter varies. Finally, numerical experiments are presented to validate the feature emergence theory in the two-neuron setting.

1 INTRODUCTION

The past decade has witnessed remarkable progress in applying deep learning techniques to scientific computing, and in particular to the numerical solution of partial differential equations (PDEs). Classical numerical methods, such as finite difference, finite element, and spectral methods, often face challenges when addressing high-dimensional problems, irregular domains, or when solutions exhibit complex structures. Neural network–based approaches have emerged as promising alternatives, offering mesh-free approximations, strong expressive power, and flexibility in incorporating physical constraints. Several notable frameworks have been proposed for solving PDEs with neural networks. The deep Ritz method (E & Yu, 2018) formulates the variational problem associated with elliptic PDEs as an energy minimization task and leverages neural networks as trial functions. Physics-informed neural networks (PINNs) (Raissi et al., 2019) instead enforce PDE constraints through the residual of the governing equations, incorporating them into the training loss. The Deep Galerkin Method (Sirignano & Spiliopoulos, 2018) generalizes this perspective by introducing stochastic collocation strategies to enforce weak PDE formulations. More recently, Weak Adversarial Networks (Zang et al., 2020) have explored adversarial training principles to approximate PDE solutions in a weak sense.

Despite these advances, the theoretical foundations of deep learning for PDEs lag behind their empirical success. While substantial progress has been made in understanding approximation power and generalization error, much less is known about **how neural networks learn effective features or representations of PDE solutions**. *Feature learning* — the process of automatically discovering meaningful low-dimensional structure from high-dimensional data, is a key ingredient to the success of neural networks in a broad range of machine learning tasks. Many recent theoretical works have demonstrated that neural networks trained with gradient-based algorithms are able to learn certain high dimensional target functions with low-dimensional structures, e.g. single- or multi-index models, in regression settings. However, to the best of our knowledge, the emergence of feature learning has not been systematically studied in the context of neural network–based methods for PDEs.

In this work, we investigate the feature learning mechanism of simple neural network models applied to high dimensional stationary Schödinger equation. Specifically, we consider fitting the solution of the Schödinger equation in the framework of deep Ritz method where the hypothesis function is defined by either a single index model or a two-neuron network. Specifically, we study solutions within the framework of the deep Ritz method, where the hypothesis function is either a single-index model or a two-neuron network. Our overarching goal is to characterize the feature directions that minimize the Ritz loss and to identify which of these directions can be effectively captured by gradient descent.

1.1 Our Contributions

We highlight the major contributions of the paper as follows:

- We first investigate the convergence guarantees of gradient descent (GD) in the agnostic setting, where the hypothesis function is restricted to a single-index model while the PDE solution is generic. We show that GD achieves an approximate global minimum in the sense that the loss value after $T = O(\log 1/\epsilon)$ iterations of GD is within ϵ of a constant multiplier of the minimum loss.
- Next, we focus on the loss landscape when the source term of the PDE is itself a single-index model. We consider two cases for the hypothesis function: a single-index model or a multi-index model with two neurons. In the first case, we prove that the minimum Ritz energy is attained at the same feature vector as the source term. In the latter case, we analyze the landscape of a regularized Ritz loss, where the regularization is applied to the outer-layer weights, in the high-dimensional regime. We characterize the sets of local and global minimizers of the high-dimensional limit of the Ritz loss as the regularization parameter varies. Our main results demonstrate that, when one feature vector of the hypothesis function aligns with that of the source term, the behavior of the second feature vector depends on the strength of the regularization: the regularized Ritz loss may admit a local or global minimizer that deviates from the feature vector of the source term.
- Finally, we provide numerical experiments to validate the theory we established for the loss landscape of the two-neuron multi-index model.

1.2 Related Work

Analysis of neural networks for PDEs. Many recent theoretical studies have investigated the approximation power (Marwah et al., 2021; 2023; Grohs et al., 2022; 2023; Grohs & Herrmann, 2022; De Ryck & Mishra, 2024) of neural networks for representing solutions of PDEs. A number of works have also analyzed generalization error estimates (Mishra & Molinaro, 2023; De Ryck & Mishra, 2022; Shin et al., 2023; Lu et al., 2022b) within variational frameworks such as PINNs and the deep Ritz method. These analyses typically assume that PDE solutions lie in Sobolev or Hölder spaces, which leads to convergence rates that suffer from the curse of dimensionality: achieving an ϵ-accurate approximation of a d-dimensional solution requires $n = O(\epsilon^{-cd})$ network parameters and training samples. In contrast, another line of research (Chen et al., 2021; 2023; Feng & Lu, 2025; Weinan & Wojtowytsch, 2022) has established dimension-free approximation rates for certain high-dimensional PDEs by developing new regularity theory in Barron spaces. More recently, dimension-free generalization error bounds (Lu et al., 2021a; Lu & Lu, 2022) have also been derived for specific classes of elliptic PDEs. Also, there is a series of stochastic differential equation (SDE)-based neural PDE solvers for elliptic equations (Nüsken & Richter, 2021; Han et al., 2020; Nam et al., 2024).

Compared to approximation-theoretic and generalization analyses, the optimization aspect of neural networks for PDEs remains significantly more challenging and less well understood, primarily due to the highly non-convex nature of the associated loss landscapes. Several recent works (Luo & Yang, 2024; Bonfanti et al., 2024; Zhao & Luo, 2025; Xu et al., 2024; Gao et al., 2023; Jiao et al., 2024) have investigated the convergence of gradient-based algorithms for training highly over-parameterized neural networks within the Neural Tangent Kernel (NTK) framework (Jacot et al., 2018). However, NTK-based analyses are intrinsically tied to the lazy training regime and thus fail to capture feature learning. An alternative line of research considers the mean-field regime (Mei et al., 2018; Sirignano & Spiliopoulos, 2020; Rotskoff & Vanden-Eijnden, 2018), where over-parameterized neural networks are studied via their distributional dynamics, described by Wasserstein gradient flows. In this setting, Dus & Virginie (2024); Dus & Ehrlacher (2025) established convergence results for Wasserstein gradient flows of the Ritz energy with infinite-width two-layer neural networks, applied to problems such as the Poisson equation and the Schrödinger eigenvalue problem. Nonetheless, these mean-field results are limited to the infinite-width setting and do not directly extend to finite-width networks, where feature learning and optimization dynamics remain far less understood.

Feature learning of neural networks with gradient-based algorithms. Beyond the NTK regime, a growing body of work seeks to understand how neural networks trained with gradient descent (GD) can recover low-dimensional structures (feature directions) of relatively simple target functions in regression problems. Examples include polynomials Yehudai & Shamir (2019); Damian et al. (2022), single-index models Soltanolkotabi (2017); Dudeja & Hsu (2018); Damian et al. (2023); Bietti et al. (2022), multi-index models Ba et al. (2022); Dandi et al. (2024); Cui et al. (2024); Moniri et al. (2024); Bruna & Hsu (2025), and sparse Boolean functions Abbe et al. (2022; 2023), among others. A central challenge in extending these analyses from regression to PDE problems lies in the intrinsic mismatch between the PDE solutions and the neural network approximators. Although PDE solutions may exhibit low-dimensional structures, they are typically far more complex than simple neural network models such as multi-index functions. This places the learning problem into the so-called *agnostic* setting, where the hypothesis class does not perfectly capture the target function. While a line of research investigated the first-order methods to learn agnostically the single-index Frei et al. (2020); Wu (2022); Awasthi et al. (2023); Wang et al. (2023); Gollakota et al. (2023); Zarifis et al. (2024) and multi-index models Diakonikolas et al. (2024), these results do not directly generalize to PDEs because the loss functions in PDE learning involve complex differential operators. A rigorous understanding of feature learning in neural network–based PDE solvers remains largely open, even for simple architectures.

1.3 Notation

We use bold uppercase and lowercase letters to denote matrices and column vectors, respectively. For $\boldsymbol{x} \in \mathbb{R}^d$, we denote its p-norm by $|\boldsymbol{x}|_p$. When $p = 2$, we write $|\boldsymbol{x}| = |\boldsymbol{x}|_2$. Given a function f, we use $\|f\|_\infty$ to denote the sup norm of f. Let $\Omega = \mathcal{B}^d$ be the unit ball on $\mathbb{R}^d$ and $\partial\Omega$ be the boundary of Ω. Also, $\mathcal{S}^{d-1}$ denotes the sphere in d-dimension, $\mathcal{S}^{d-1} := \{\boldsymbol{x} \in \mathbb{R}^d : |\boldsymbol{x}| = 1\}$. We denote by $H^1(\Omega)$ the Sobolev space of square-integrable functions with square-integrable first derivatives. Given a unit vector $\boldsymbol{w} \in \mathcal{S}^{d-1}$, $\mathrm{P}_{\boldsymbol{w}^\perp} := I - \boldsymbol{w}\boldsymbol{w}^\top$ denotes the orthogonal projector onto the hyperplane perpendicular to $\boldsymbol{w}$, i.e., for $\boldsymbol{v} \in \mathbb{R}^d$, $\mathrm{P}_{\boldsymbol{w}^\perp}\boldsymbol{v} = \boldsymbol{v}^{\perp \boldsymbol{w}}$.

2 Set-Up and Main Results

We consider the following stationary Schrödinger equation

$$-\Delta u + u = f \text{ on } \Omega, \quad \frac{\partial u}{\partial \nu} = 0 \text{ on } \partial\Omega. \tag{1}$$

We assume that f is known and $f \in L^2(\Omega)$. Thanks to the standard well-posedness of PDEs, there exists a unique weak solution $u^* \in H^1(\Omega)$ to the equation (1) and

$$u^* = \underset{u \in H^1(\Omega)}{\arg\min}\, \mathcal{E}(u) := \underset{u \in H^1(\Omega)}{\arg\min}\, \frac{1}{2}\int_\Omega |\nabla u|^2 + u^2 - 2fudx. \tag{2}$$

It is also useful to note that for any $u \in H^1(\Omega)$,

$$\begin{aligned}\mathcal{E}(u) &= \frac{1}{2}\int_\Omega |\nabla u|^2 + u^2 - 2(-\Delta u^* + u^*)udx \\ &= \frac{1}{2}\int_\Omega |\nabla u - \nabla u^*|^2 + |u - u^*|^2 dx - \frac{1}{2}\int_\Omega |\nabla u^*|^2 + |u^*|^2 dx \\ &= \frac{1}{2}\int_\Omega |\nabla u - \nabla u^*|^2 + |u - u^*|^2 dx + \mathcal{E}(u^*),\end{aligned}$$

where we have used integration by parts in the second equality. As a consequence, minimizing the energy $\mathcal{E}$ is equivalent to minimizing the shifted loss function $\mathcal{L}$ defined as follows:

$$\begin{aligned}\mathcal{L}(u) :=&\mathcal{E}(u) - \mathcal{E}(u^*)\\ =&\frac{1}{2}\int_\Omega |\nabla u - \nabla u^*|^2 + |u - u^*|^2 dx\\ =&\frac{|\Omega|}{2}\mathbb{E}_{x\sim\mathcal{P}_\Omega}\left[|\nabla u(x) - \nabla u^*(x)|^2 + |u(x) - u^*(x)|^2\right],\end{aligned} \tag{3}$$

with $\mathcal{P}_\Omega$ denoting the uniform probability distributions on the domain Ω.

The deep Ritz method (DRM) E & Yu (2018) seeks an approximate solution of (2) by minimizing the energy $\mathcal{E}$ (or its empirical version) within a hypothesis function class $\mathcal{F}$ which is parameterized by neural networks. The resulting optimization problem over the neural network parameters is non-convex in general and it remains a challenging open problem whether standard gradient-based algorithms can find the globally optimal solution. The purpose of this paper is to approach this problem by understanding the feature learning mechanism of gradient descent in the minimization of the DRM loss. Specifically, we perform the analysis in three settings as discussed in what follows.

2.1 Single-Index Hypothesis in the fully Agnostic Setting

We first consider the case where the hypothesis function u is defined by a single-index model parameterized with a unit vector $\boldsymbol{w} \in \mathcal{S}^{d-1}$. More precisely, we assume that

$$u(\boldsymbol{x}) \equiv u_{\boldsymbol{w}}(\boldsymbol{x}) = \sigma^2(\boldsymbol{w}\cdot\boldsymbol{x}), \boldsymbol{x} \in \Omega \tag{4}$$

where the activation function $\sigma(\cdot)$ is the ReLU function (i.e., $\sigma(\cdot) = \max\{0, \cdot\}$). We consider the squared ReLU activation instead of ReLU its own as it yields a differentiable (Lipschitz continuous) loss function (see equation (5)) leading to a well-defined gradient descent algorithm. Also, it has been shown that shallow neural networks with squared ReLU or high-order power of ReLU activation enjoy quantitative approximation rate in Sobolev spaces (Lu et al., 2022a; Mao et al., 2024). Moreover, we make the following a-priori assumption on the right hand side f and the exact solution u^*. It is important to note that the ground-truth solution u^* still belongs to a large function class and does not necessarily admit the same single-index form (4) as u. This places the problem in the fully *agnostic learning* setting.

Assumption 2.1. *The following statements hold for the equation (1).*

1. *There exist a constant C_{u^*} such that $\|\nabla u^*\|_\infty \leqslant C_{u^*}$ and $\|u^*\|_\infty \leqslant C_{u^*}$.*
2. *There exists a constant C_f such that $\|f\|_\infty \leqslant C_f$.*

Ignoring the constant $\frac{|\Omega|}{2}$ and noting that $\sigma(\cdot)\sigma'(\cdot) = \sigma(\cdot)$ in the subgradient sense, one redefines the equivalent loss functions as follows

$$\begin{aligned}\mathcal{L}(\boldsymbol{w}) :=&\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}\left[|\nabla_{\boldsymbol{x}} u_{\boldsymbol{w}}(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} u^*(\boldsymbol{x})|^2 + |u_{\boldsymbol{w}}(\boldsymbol{x}) - u^*(\boldsymbol{x})|^2\right]\\ =&\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}\left[|2\sigma(\boldsymbol{w}\cdot\boldsymbol{x})\boldsymbol{w} - \nabla_{\boldsymbol{x}} u^*(\boldsymbol{x})|^2 + |\sigma^2(\boldsymbol{w}\cdot\boldsymbol{x}) - u^*(\boldsymbol{x})|^2\right]\end{aligned} \tag{5}$$

and

$$\begin{aligned}\mathcal{E}(\boldsymbol{w}) :=&\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}\left[|\nabla_{\boldsymbol{x}} u_{\boldsymbol{w}}(\boldsymbol{x})|^2 + |u_{\boldsymbol{w}}(\boldsymbol{x})|^2 - 2f(\boldsymbol{x})u_{\boldsymbol{w}}(\boldsymbol{x})\right]\\ =&\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}\left[|2\sigma(\boldsymbol{w}\cdot\boldsymbol{x})\boldsymbol{w}|^2 + |\sigma^2(\boldsymbol{w}\cdot\boldsymbol{x})|^2 - 2f(\boldsymbol{x})\sigma^2(\boldsymbol{w}\cdot\boldsymbol{x})\right].\end{aligned}$$

Moreover, since $\mathcal{L}(\boldsymbol{w})$ and $\mathcal{E}(\boldsymbol{w})$ only differs by a constant (see 3), one has that $\nabla\mathcal{L}(\boldsymbol{w}) = \nabla\mathcal{E}(\boldsymbol{w})$.

2.1.1 Optimization via Riemannian Gradient Descent

To optimize the loss $\mathcal{L}$ defined by 5, we adopt the classic Riemannian gradient descent (GD) method. First, observe that the gradient of the loss function $\mathcal{L}(\boldsymbol{w})$ with respect to $\boldsymbol{w}$ is given by

$$\begin{aligned}\nabla\mathcal{L}(\boldsymbol{w}) =&\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}\left[4\sigma'(\boldsymbol{w}\cdot\boldsymbol{x})\boldsymbol{x}\boldsymbol{w}^\top(2\sigma(\boldsymbol{w}\cdot\boldsymbol{x})\boldsymbol{w} - \nabla u^*) + 4(2\sigma(\boldsymbol{w}\cdot\boldsymbol{x})\boldsymbol{w} - \nabla u^*)\sigma(\boldsymbol{w}\cdot\boldsymbol{x})\right]\\ &+ \mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}\left[4(\sigma^2(\boldsymbol{w}\cdot\boldsymbol{x}) - u^*(\boldsymbol{x}))\sigma(\boldsymbol{w}\cdot\boldsymbol{x})\boldsymbol{x}\right].\end{aligned}$$

However, the expression above is intractable since the ground-truth u^* is unknown. Instead, noting that $\nabla\mathcal{L}(\boldsymbol{w}) = \nabla\mathcal{E}(\boldsymbol{w})$, we use the tractable gradient $\nabla\mathcal{E}(\boldsymbol{w})$ of the Ritz energy $\mathcal{E}$ given by

$$\nabla\mathcal{E}(\boldsymbol{w}) = 4\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}\left[2|\boldsymbol{w}|^2\sigma(\boldsymbol{w}\cdot\boldsymbol{x})\boldsymbol{x} + 2\sigma^2(\boldsymbol{w}\cdot\boldsymbol{x})\boldsymbol{w} + \sigma^3(\boldsymbol{w}\cdot\boldsymbol{x})\boldsymbol{x} - f(\boldsymbol{x})\sigma(\boldsymbol{w}\cdot\boldsymbol{x})\boldsymbol{x}\right].$$

Recall that $\mathrm{P}_{\boldsymbol{w}^\perp} := \boldsymbol{I} - \boldsymbol{w}\boldsymbol{w}^\top$. Then the Riemannian gradients of the loss function $\mathcal{L}(\boldsymbol{w})$ and the energy function $\mathcal{E}(\boldsymbol{w})$, denoted by $g_{\mathcal{L}}(\boldsymbol{w})$ and $g_{\mathcal{E}}(\boldsymbol{w})$ respectively are $g_{\mathcal{L}}(\boldsymbol{w}) := \mathrm{P}_{\boldsymbol{w}^\perp}\nabla\mathcal{L}(\boldsymbol{w})$ and $g_{\mathcal{E}}(\boldsymbol{w}) := \mathrm{P}_{\boldsymbol{w}^\perp}\nabla\mathcal{E}(\boldsymbol{w})$. Specifically, the gradient of the Ritz energy can be evaluated explicitly as

$$g_{\mathcal{E}}(\boldsymbol{w}) = 4\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}\left[2|\boldsymbol{w}|^2\sigma(\boldsymbol{w}\cdot\boldsymbol{x})\mathrm{P}_{\boldsymbol{w}^\perp}\boldsymbol{x} + \sigma^3(\boldsymbol{w}\cdot\boldsymbol{x})\mathrm{P}_{\boldsymbol{w}^\perp}\boldsymbol{x} - f(\boldsymbol{x})\sigma(\boldsymbol{w}\cdot\boldsymbol{x})\mathrm{P}_{\boldsymbol{w}^\perp}\boldsymbol{x}\right].$$

In practice, the expectation above is approximated by an empirical estimator $\hat{g}_{\mathcal{E}}$ computed with n iid uniform samples $\{\boldsymbol{x}_i\}_{i=1}^n$ on Ω by

$$\hat{g}_{\mathcal{E}}(\boldsymbol{w}^t) := \frac{1}{n}\sum_{i=1}^n 4\left(2|\boldsymbol{w}^t|^2\sigma(\boldsymbol{w}^t\cdot\boldsymbol{x}_i)\mathrm{P}_{(\boldsymbol{w}^t)^\perp}\boldsymbol{x}_i + \sigma^3(\boldsymbol{w}^t\cdot\boldsymbol{x}_i)\mathrm{P}_{(\boldsymbol{w}^t)^\perp}\boldsymbol{x}_i - f(\boldsymbol{x}_i)\sigma(\boldsymbol{w}^t\cdot\boldsymbol{x}_i)\mathrm{P}_{(\boldsymbol{w}^t)^\perp}\boldsymbol{x}_i\right).$$

We summarize the gradient descent procedure in Algorithm 1. It takes as input an initial guess $\boldsymbol{w}^0$ for the desired parameter, accuracy tolerance parameter ϵ (see Theorem 2.2), number of iterations T, step size η, and the sample distribution $\mathcal{P}_\Omega$. It outputs the estimated parameter $\boldsymbol{w}^T$ obtained by the Riemannian GD.

Algorithm 1 Riemannian GD

```
Input: $\boldsymbol{w}^0$, $\epsilon$, $T$, $\eta$; Sample distribution $\mathcal{P}_\Omega$.
Draw $n = \Theta(d(d+2)/\epsilon)$ samples $\{\boldsymbol{x}_i\}_{i=1}^n$ from $\mathcal{P}_\Omega$.
for $t = 0, \cdots, T-1$ do
    Compute the empirical estimate of $g_{\mathcal{E}}(\boldsymbol{w}^t)$:
        $\hat{g}_{\mathcal{E}}(\boldsymbol{w}^t) := \frac{1}{n}\sum_{i=1}^n 4\left(2|\boldsymbol{w}^t|^2\sigma(\boldsymbol{w}^t\cdot\boldsymbol{x}_i)\mathrm{P}_{(\boldsymbol{w}^t)^\perp}\boldsymbol{x}_i + \sigma^3(\boldsymbol{w}^t\cdot\boldsymbol{x}_i)\mathrm{P}_{(\boldsymbol{w}^t)^\perp}\boldsymbol{x}_i - f(\boldsymbol{x}_i)\sigma(\boldsymbol{w}^t\cdot\boldsymbol{x}_i)\mathrm{P}_{(\boldsymbol{w}^t)^\perp}\boldsymbol{x}_i\right).$
    Gradient descent and normalize: $\boldsymbol{w}^{t+1} = (\boldsymbol{w}^t - \eta\hat{g}_{\mathcal{E}}(\boldsymbol{w}^t))/|\boldsymbol{w}^t - \eta\hat{g}_{\mathcal{E}}(\boldsymbol{w}^t)|_2$.
end for
return $\boldsymbol{w}^T$
```

We measure the performance of $\boldsymbol{w}^T$ by comparing the loss value $\mathcal{L}(\boldsymbol{w}^T)$ with the minimum loss value OPT defined by

$$\mathrm{OPT} := \mathcal{L}(\boldsymbol{w}^*) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}\left[|\nabla_{\boldsymbol{x}}u_{\boldsymbol{w}^*}(\boldsymbol{x}) - \nabla_{\boldsymbol{x}}u^*(\boldsymbol{x})|^2 + |u_{w^*}(\boldsymbol{x}) - u^*(\boldsymbol{x})|^2\right],$$

where $\boldsymbol{w}^* := \arg\min_{\boldsymbol{w}\in\mathcal{S}^{d-1}}\mathcal{L}(\boldsymbol{w})$. Our first main theorem below shows that Algorithm 1 produces an approximate solution $u_{\boldsymbol{w}^T}$ to the problem (5) after $T = O(\log(1/\epsilon))$ iterations in the sense that the loss value $\mathcal{L}(\boldsymbol{w}^T)$ is within ϵ of a constant multiplier of the OPT.

Theorem 2.2. *Suppose that Assumption 2.1 holds. Consider Algorithm 1 with initial condition $\boldsymbol{w}^0$ satisfying $\angle(\boldsymbol{w}^0, \boldsymbol{w}^*) \in [0, \frac{\pi}{2}]$. Set the sample size $n = \Theta\left(\frac{d(d+2)}{\epsilon}\right)$ and the step size $\eta = \frac{d+2}{2048\pi}$. Then after $T = O(\log(1/\epsilon))$ iterations, with probability at least $1-\delta$, the output of Algorithm 1, $\boldsymbol{w}^T$, satisfies $\mathcal{L}(\boldsymbol{w}^T) < \gamma\cdot OPT + \frac{128}{d+2}\epsilon$, where the absolute constant $0 < \gamma < 2048\pi^2 + 2$.*

The proof of Theorem 2.2 can be found in Appendix A. We remark that the linear dependence of the step size η on the dimension d arises from the fact that the Riemannian gradient satisfies $g_{\mathcal{E}}(\boldsymbol{w}) = O(1/d)$. Consequently, scaling the step size with d ensures that the effective update of magnitude remains balanced and prevents vanishingly small progress in high dimensions.

2.2 Single-Index Hypothesis

In this subsection, we focus on the setting where f is given by a single-index model $f(\boldsymbol{x}) = \sigma^2(\boldsymbol{w}^*\cdot\boldsymbol{x})$ for some $\boldsymbol{w}^* \in \mathcal{S}^{d-1}$. We also consider the hypothesis function u is defined by a single index model $u(\boldsymbol{x}) = \sigma^2(\boldsymbol{w}\cdot\boldsymbol{x})$. Then the problem of minimizing the Ritz energy becomes

$$\min_{\boldsymbol{w}\in\mathcal{S}^{d-1}}\mathcal{E}(\boldsymbol{w}) := \frac{|\Omega|}{2}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}\left[|2\sigma(\boldsymbol{w}\cdot\boldsymbol{x})\boldsymbol{w}|^2 + |\sigma^2(\boldsymbol{w}\cdot\boldsymbol{x})|^2 - 2\sigma^2(\boldsymbol{w}^*\cdot\boldsymbol{x})\sigma^2(\boldsymbol{w}\cdot\boldsymbol{x})\right]. \quad (6)$$

Similar to the previous section, we remark that the exact solution u^* associated to the source term f is not a single index model. However, the proposition below shows that the minimum of the Ritz energy function $\mathcal{E}(\boldsymbol{w})$, defined in (6), attains its minimum in the same direction w^* as the source term f. We defer the proof of Proposition 2.3 to Appendix B.

Proposition 2.3. *The minimum value of the Ritz energy function $\mathcal{E}(\boldsymbol{w})$ is achieved when $\boldsymbol{w} = \boldsymbol{w}^*$.*

2.3 Multi-Index Hypothesis

In this subsection, we assume that f is given by a single-index model $f(\boldsymbol{x}) = \sigma^2(\boldsymbol{w}^* \cdot \boldsymbol{x})$, but consider a more complicated hypothesis model — two-neuron neural network, which is a special case of *multi-index model*. More concretely, we assume that

$$u(\boldsymbol{x}) = \sum_{i=1}^{2} a_i \sigma^2(\boldsymbol{w}_i \cdot \boldsymbol{x})$$

with $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathcal{S}^{d-1}$ and $a_1, a_2 \in \mathbb{R}$. Motivated by Proposition 2.3 for the single-index hypothesis function, we fix the first feature vector $\boldsymbol{w}_1 = \boldsymbol{w}^*$. In fact, we can also show that if the hypothesis function is $a_1\sigma^2(w_1 \cdot x) + a_2\sigma^2(w_2 \cdot x)$ with $a_i > 0$, then one of the features also aligns with $\boldsymbol{w}^*$. For simplicity, we investigate the mechanism by which the second feature $\boldsymbol{w}_2$ emerges under this constraint. Under this setting, the variational problem of minimizing the Ritz energy becomes

$$\begin{aligned}\mathcal{L}(\boldsymbol{w}, \boldsymbol{a}) &:= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}\left[|\nabla_{\boldsymbol{x}} u|^2 + |u|^2 - 2fu\right] \\ &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}\left[\left|2\sum_{i=1}^{2} a_i\sigma(\boldsymbol{w}_i\cdot x)\boldsymbol{w}_i\right|^2 + \left|\sum_{i=1}^{2} a_i\sigma^2(\boldsymbol{w}_i\cdot\boldsymbol{x})\right|^2 - 2\sigma^2(\boldsymbol{w}^*\cdot\boldsymbol{x})\sum_{i=1}^{2} a_i\sigma^2(\boldsymbol{w}_i\cdot\boldsymbol{x})\right].\end{aligned} \tag{7}$$

To write the loss function in a more compact form, we apply the similar techniques as in Cho & Saul (2009) and define the following kernels

$$\begin{aligned}K_1(\boldsymbol{w}_i, \boldsymbol{w}_j) &:= 4\boldsymbol{w}_i^\top \boldsymbol{w}_j \mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}[\sigma(\boldsymbol{w}_i\cdot\boldsymbol{x})\sigma(\boldsymbol{w}_j\cdot\boldsymbol{x})] = \frac{2}{\pi(d+2)}h_1(\theta_{ij}), \\ K_2(\boldsymbol{w}_i, \boldsymbol{w}_j) &:= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}[\sigma^2(\boldsymbol{w}_i\cdot\boldsymbol{x})\sigma^2(\boldsymbol{w}_j\cdot\boldsymbol{x})] = \frac{1}{2\pi(d+2)(d+4)}h_2(\theta_{ij}),\end{aligned}$$

where $\theta_{ij} = \angle(\boldsymbol{w}_i, \boldsymbol{w}_j)$ is the angle between w_i and w_j and

$$\begin{aligned}h_1(\theta_{ij}) &:= (\sin\theta_{ij}\cos\theta_{ij} + (\pi - \theta_{ij})\cos^2\theta_{ij}), \\ h_2(\theta_{ij}) &:= 3\sin\theta_{ij}\cos\theta_{ij} + (\pi - \theta_{ij})(1 + 2\cos^2\theta_{ij}),\end{aligned}$$

Note that we have also used $\boldsymbol{w}_i^\top \boldsymbol{w}_j = \cos\theta_{ij}$. Therefore, the loss function (7) can be written as

$$\mathcal{L}(\boldsymbol{w}, \boldsymbol{a}) = \boldsymbol{a}^\top \boldsymbol{K}_1 \boldsymbol{a} + \boldsymbol{a}^\top \boldsymbol{K}_2 \boldsymbol{a} - 2\boldsymbol{a}^\top \boldsymbol{K}_*, \tag{8}$$

where $\boldsymbol{K}_1 \in \mathbb{R}^{2\times 2}$, $\boldsymbol{K}_2 \in \mathbb{R}^{2\times 2}$ and $\boldsymbol{K}_* \in \mathbb{R}^{2\times 1}$ are defined as follows:

$$\boldsymbol{K}_1 = \begin{bmatrix} K_1(\boldsymbol{w}_1, \boldsymbol{w}_1) & K_1(\boldsymbol{w}_1, \boldsymbol{w}_2) \\ K_1(\boldsymbol{w}_2, \boldsymbol{w}_1) & K_1(\boldsymbol{w}_2, \boldsymbol{w}_2) \end{bmatrix}, \boldsymbol{K}_2 = \begin{bmatrix} K_2(\boldsymbol{w}_1, \boldsymbol{w}_1) & K_2(\boldsymbol{w}_1, \boldsymbol{w}_2) \\ K_2(\boldsymbol{w}_2, \boldsymbol{w}_1) & K_2(\boldsymbol{w}_2, \boldsymbol{w}_2) \end{bmatrix}, \boldsymbol{K}_* = \begin{bmatrix} K_2(\boldsymbol{w}^*, \boldsymbol{w}_1) \\ K_2(\boldsymbol{w}^*, \boldsymbol{w}_2) \end{bmatrix}.$$

In practice, an ℓ_2-regularization is commonly employed to promote weight decay. Here we consider two variants: one applied to $\boldsymbol{a} = (a_1, a_2)$ and another applied solely to a_2. In both regularization settings, our primary interest is to understand how the other feature $\boldsymbol{w}_2$ emerges in the high dimensional regime $d \to \infty$ given that the first feature is fixed at $\boldsymbol{w}_1 = \boldsymbol{w}^*$ as the regularization parameter λ varies.

2.3.1 Regularization applied to $\boldsymbol{a}$

We first consider the following regularized loss function:

$$\mathcal{L}_\lambda(\boldsymbol{w}, \boldsymbol{a}) = \boldsymbol{a}^\top \boldsymbol{K}_1 \boldsymbol{a} + \boldsymbol{a}^\top \boldsymbol{K}_2 \boldsymbol{a} - 2\boldsymbol{a}^\top \boldsymbol{K}_* + \lambda \boldsymbol{a}^\top \boldsymbol{a}, \tag{9}$$

where $\lambda > 0$ is the regularization parameter. To minimize the regularized loss function with respect to the parameters $\boldsymbol{w}$ and $\boldsymbol{a}$, noting first that for any fixed $\boldsymbol{w}$, the function $\boldsymbol{a} \mapsto \mathcal{L}_\lambda(\boldsymbol{w}, \boldsymbol{a})$ is quadratic and the minimizer $\boldsymbol{a}^*$ has a closed form solution and the regularized loss function can thus be reduced to a loss function depending only on $\boldsymbol{w}$. The details are presented in the following lemma.

Lemma 2.4. *For any fixed $\boldsymbol{w}$, the minimizer $\boldsymbol{a}^*$ of the regularized loss function (9) has a closed form $\boldsymbol{a}^* = (\boldsymbol{K}_1 + \boldsymbol{K}_2 + \lambda \boldsymbol{I}_2)^{-1}\boldsymbol{K}_*$ and $\mathcal{L}_\lambda(\boldsymbol{w}) = -\boldsymbol{K}_*^\top(\boldsymbol{K}_1 + \boldsymbol{K}_2 + \lambda \boldsymbol{I}_2)^{-1}\boldsymbol{K}_*$.*

Since by assumption $\boldsymbol{w}_1$ is aligned with $\boldsymbol{w}^*$, the loss function $\mathcal{L}_\lambda(\boldsymbol{w})$ can be rewritten as a loss of the angle $\theta := \angle(\boldsymbol{w}^*, \boldsymbol{w}_2) \in [0, \pi]$. More precisely, letting $c = \frac{2}{d+2}$ and defining $\lambda = \xi c$ with a rescaled regularization parameter ξ, it can be shown that the loss $\mathcal{L}_\lambda(\boldsymbol{w}) = \mathcal{L}_\xi(\theta)$ where

$$\mathcal{L}_\xi(\theta)$$
$$= -\frac{c^2}{16\pi^2(d+4)^2}\begin{bmatrix}h_2(0)\\ h_2(\theta)\end{bmatrix}^\top \begin{bmatrix}\frac{c}{\pi}h_1(0) + \frac{c}{4\pi(d+4)}h_2(0) + \xi c & \frac{c}{\pi}h_1(\theta) + \frac{c}{4\pi(d+4)}h_2(\theta)\\ \frac{c}{\pi}h_1(\theta) + \frac{c}{4\pi(d+4)}h_2(\theta) & \frac{c}{\pi}h_1(0) + \frac{c}{4\pi(d+4)}h_2(0) + \xi c\end{bmatrix}^{-1}\begin{bmatrix}h_2(0)\\ h_2(\theta)\end{bmatrix}.$$

Since our focus is on feature learning in the high-dimensional regime and noting that $h_1(0) = \pi$ and $h_2(0) = 3\pi$, we define the following limiting function as $d \to +\infty$:

$$\begin{aligned}\widetilde{\mathcal{L}}_\xi(\theta) &:= \lim_{d\to+\infty}\frac{16(d+4)^2}{c}\mathcal{L}_\xi(\theta)\\ &= -\frac{(9+9\xi) + \frac{1}{\pi^2}[(1+\xi)h_2^2(\theta) - 6h_1(\theta)h_2(\theta)]}{(1+\xi)^2 - \frac{1}{\pi^2}h_1^2(\theta)}.\end{aligned} \tag{10}$$

The theorem below characterizes the set of minimizers of the limiting function $\widetilde{\mathcal{L}}_\xi(\theta)$ as ξ varies.

Theorem 2.5. *Consider the minimization of the limiting loss function $\widetilde{\mathcal{L}}_\xi$ defined by* (10)*.*

1. *When $\xi \geqslant \frac{1}{2}$, $\widetilde{\mathcal{L}}_\xi(\theta)$ has a unique global minimizer at $\theta = 0$ for θ on $[0, \frac{5\pi}{6}]$.*
2. *When $\xi \leqslant \xi_0$ with some $\xi_0 < 1/2$, besides the local minimizer $\theta = 0$, there exists at least one additional local minimizer of $\widetilde{\mathcal{L}}_\xi(\theta)$ in the interval $(\frac{\pi}{4}, \frac{\pi}{2})$.*

We defer the proofs of Lemma 2.4 and Theorem 2.5 to Appendix C.

Remark. Although we can only rigorously prove the above theorem, our numerical experiments indicate a phase transition between a local minimizer (unique global minimizer) and two local minimizers, one at 0 and the other lying in $(\pi/4, \pi/2)$. This phase transition occurs approximately at the value $\xi_0 = 0.13$. More details can be found in Figure 1.

2.3.2 Regularization applied to a_2 solely

We now consider the regularized loss function in which only the coefficient a_2 is penalized:

$$\mathcal{L}_\lambda(\boldsymbol{w}, \boldsymbol{a}) = \boldsymbol{a}^\top\boldsymbol{K}_1\boldsymbol{a} + \boldsymbol{a}^\top\boldsymbol{K}_2\boldsymbol{a} - 2\boldsymbol{a}^\top\boldsymbol{K}_* + \lambda a_2^2. \tag{11}$$

To minimize the above regularized loss function with respect to parameters $\boldsymbol{w}$ and $\boldsymbol{a}$, we first note that for any fixed $\boldsymbol{w}$, the minimizer $\boldsymbol{a}^*$ has a closed form solution and the regularized loss function can be converted to a loss function that is only about $\boldsymbol{w}$. The details are presented in the following lemma.

Lemma 2.6. *For any fixed $\boldsymbol{w}$, the minimizer $\boldsymbol{a}^*$ of the regularized loss function (11) has a closed form $\boldsymbol{a}^* = \left(\boldsymbol{K}_1 + \boldsymbol{K}_2 + \begin{bmatrix}0 & 0\\ 0 & \lambda\end{bmatrix}\right)^{-1}\boldsymbol{K}_*$ and $\mathcal{L}_\lambda(\boldsymbol{w}) = -\boldsymbol{K}_*^\top\left(\boldsymbol{K}_1 + \boldsymbol{K}_2 + \begin{bmatrix}0 & 0\\ 0 & \lambda\end{bmatrix}\right)^{-1}\boldsymbol{K}_*$.*

Given that $\boldsymbol{w}_1$ is aligned with $\boldsymbol{w}^*$, the loss function $\mathcal{L}_\lambda(\boldsymbol{w})$ is equivalent to the following loss function $\mathcal{L}_\xi(\theta)$, upon substituting the definitions of $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$, where $\theta := \angle(\boldsymbol{w}^*, \boldsymbol{w}_2)$:

$$\mathcal{L}_\xi(\theta)$$
$$= -\frac{c^2}{16\pi^2(d+4)^2}\begin{bmatrix}h_2(0)\\ h_2(\theta)\end{bmatrix}^\top \begin{bmatrix}\frac{c}{\pi}h_1(0) + \frac{c}{4\pi(d+4)}h_2(0) & \frac{c}{\pi}h_1(\theta) + \frac{c}{4\pi(d+4)}h_2(\theta)\\ \frac{c}{\pi}h_1(\theta) + \frac{c}{4\pi(d+4)}h_2(\theta) & \frac{c}{\pi}h_1(0) + \frac{c}{4\pi(d+4)}h_2(0) + \xi c\end{bmatrix}^{-1}\begin{bmatrix}h_2(0)\\ h_2(\theta)\end{bmatrix},$$

where $c = \frac{2}{d+2}$ and $\lambda = \xi c$ for a reparameterized regularization parameter ξ. Similar to the previous section, we consider the following limiting function (as $d \to +\infty$):

$$\begin{aligned}\bar{\mathcal{L}}_\xi(\theta) &:= \lim_{d\to+\infty}\frac{16(d+4)^2}{c}\mathcal{L}_\xi(\theta)\\ &= -\frac{(9+9\xi) + \frac{1}{\pi^2}[h_2^2(\theta) - 6h_1(\theta)h_2(\theta)]}{1+\xi - \frac{1}{\pi^2}h_1^2(\theta)}.\end{aligned} \tag{12}$$

We are now ready to present the main result of this subsection and defer the proof of Lemma 2.6 and Theorem 2.7 to Appendix D.

Theorem 2.7. *Consider the minimization of the limiting loss function $\bar{\mathcal{L}}_\xi$ defined by* (12)*. For any $\xi > 0$, the function $\bar{\mathcal{L}}_\xi(\theta)$ has a **unique** global minimizer $\theta^* \in (\frac{\pi}{3}, \frac{\pi}{2})$.*

Comparing Theorem 2.7 with Theorem 2.5, we observe that different forms of regularization on $\boldsymbol{a}$ lead to distinct behaviors in feature emergence. In particular, penalizing only a_2 consistently yields the emergence of an additional feature, whereas strong penalization on both outer-layer weights results in feature collapse.

3 Numerical Experiments

In this section, we provide numerical results to validate our theory established in Section 2.3 for a multi-index model with two neurons.

First, we show the limiting function $\widetilde{\mathcal{L}}_\xi(\theta)$ in Figure 1 that corresponds to the regularized loss function with regularization applied on the whole vector $\boldsymbol{a}$ (Subsection 2.3.1). The left plot of Figure 1 shows the 3D surface defined by the function $(\xi, \theta) \mapsto \widetilde{\mathcal{L}}_\xi(\theta)$ with $\xi \in [0, 0.2]$ and $\theta \in [0, \pi]$. The right plot of the same figure shows the function $\widetilde{\mathcal{L}}_\xi(\theta)$ with some specifically chosen values of ξ. We also mark the global minimizers of $\widetilde{\mathcal{L}}_\xi(\theta)$ for $\xi = 0.01,\ 0.08,\ 0.09$. From the figure, we can observe the landscape of $\widetilde{\mathcal{L}}_\xi(\theta)$ exhibits the following phase transitions as ξ varies:

- When $0 < \xi \leqslant \xi_0 \approx 0.08$, $\widetilde{\mathcal{L}}_\xi(\theta)$ has two local minimizers, one at $\theta = 0$ and the other is in the interval $(\frac{\pi}{3}, \frac{\pi}{2})$. Moreover, the global minimizer is in the interval $(\frac{\pi}{3}, \frac{\pi}{2})$.
- When $\xi_0 \approx 0.08 < \xi \leqslant \xi_1 \approx 0.13$, $\widetilde{\mathcal{L}}_\xi(\theta)$ has two local minimizers, one at $\theta = 0$ and the other is in the interval $(\frac{\pi}{4}, \frac{\pi}{2})$. Moreover, $\theta = 0$ is the global minimizer.
- When $\xi > \xi_1 \approx 0.13$, $\widetilde{\mathcal{L}}_\xi(\theta)$ has only one local (and hence global) minimizer at $\theta = 0$.

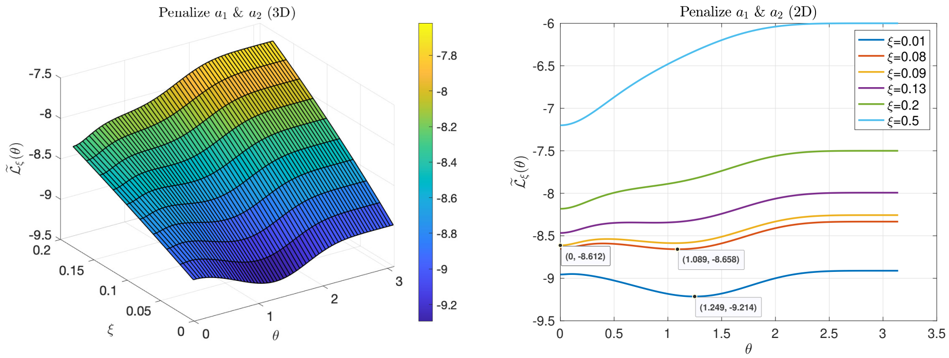

Figure 1: Graphs of the limiting function $\widetilde{\mathcal{L}}_\xi(\theta)$.

Next, we plot the limiting function $\bar{\mathcal{L}}_\xi(\theta)$ in Figure 2, which corresponds to the limiting regularized loss function with regularization applied only to a_2 defined by (12). In Figure 2, the left graph shows the 3D plot of the function $(\xi, \theta) \mapsto \bar{\mathcal{L}}_\xi(\theta)$ with $\xi \in [0.01, 0.5]$ and $\theta \in [0, \pi]$. The right graph shows plots of the function $\theta \mapsto \bar{\mathcal{L}}_\xi(\theta)$ with ξ taking a broader range of values. We also mark the global minimizers of $\bar{\mathcal{L}}_\xi(\theta)$ for $\xi = 0.01,\ 1,\ 5,\ 100$. Observe that the loss function $\bar{\mathcal{L}}_\xi(\theta)$ has a unique global minimizer in the interval $(\frac{\pi}{3}, \frac{\pi}{2})$ for $\xi > 0$.

Moreover, we plot the loss function $\mathcal{L}_\xi(\theta)$ with $d = 2$ when the regularization is applied to $\boldsymbol{a}$ and a_2 solely in Figure 3. We observe that the loss function has the same shape as the corresponding limiting functions.

We also perform numerical experiments on the landscapes for sigmoid and GELU activation functions with $d = 2$ and observe that they have different angle landscapes (Figure 4). For sigmoid

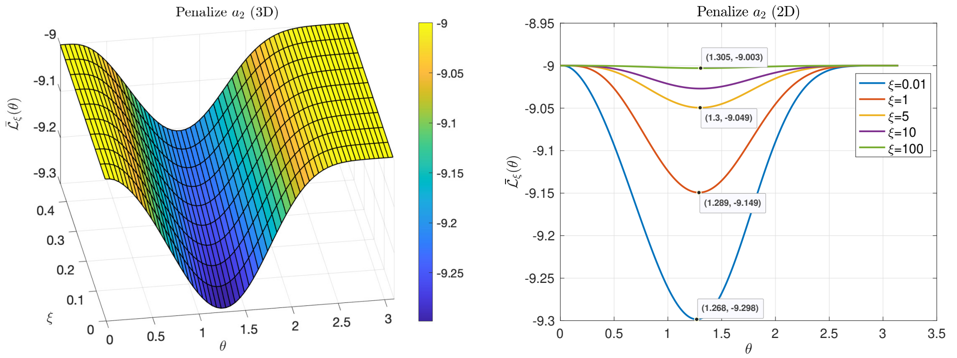

Figure 2: Graphs of the limiting function $\bar{\mathcal{L}}_\xi(\theta)$.

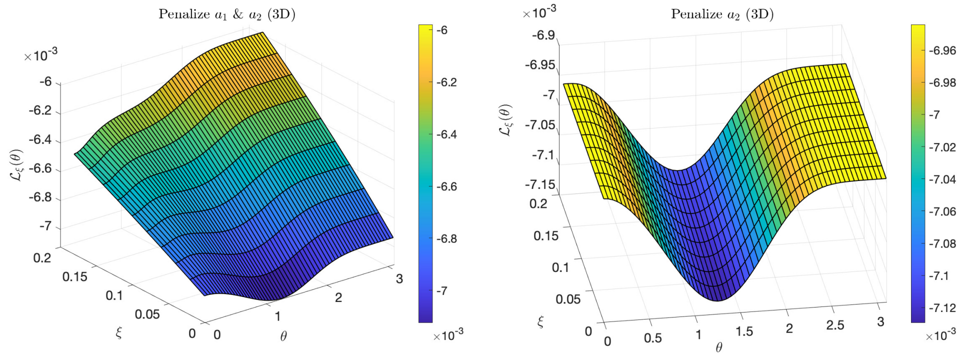

Figure 3: Graphs of the loss function $\mathcal{L}_\xi(\theta)$ with $d = 2$.

activation, the unique global minimum is achieved at $\theta = \pi$ when the regularization is applied to $\boldsymbol{a}$ and a_2 solely. For GELU, when the regularization is applied to $\boldsymbol{a}$, there are two local minimizers at $\theta = 0$ and $\theta = \pi$ and the global minimizer changes from $\theta = \pi$ to $\theta = 0$ as λ increases. When the regularization is applied to a_2 solely, there are two local minimizers, one is in $(1, 1.5)$ and the other is at $\theta = \pi$ and the global minimizer is at $\theta = \pi$.

4 Conclusion and Discussion

In this work, we present a systematic study of feature learning in the context of stationary Schrödinger equation using the deep Ritz method. Unlike prior analyses that rely on infinite-width limits or strong over-parameterization assumptions, our study focuses on the behavior of finite-width models and examines how low-dimensional features of PDE solutions emerge as a result of an optimization procedure. Specifically, in an agnostic setting where the PDE solution is generic, we established convergence guarantees for gradient descent under a single-index hypothesis class, showing that approximate global optimality can be achieved in logarithmic iteration complexity. Further, when the source term of the PDE follows a single-index structure, we characterize the loss landscape for both single- and two-neuron models. Our analysis of the regularized Ritz losses revealed how feature emergence depends critically on the regularization strength: penalizing a single outer-layer weight consistently produces an additional feature, whereas strong joint penalization can result in feature collapse. Collectively, these results provide a first step toward a rigorous theory of feature learning in finite-width neural networks for PDEs.

Our work opens several directions for future research. While our analysis focused on single-index and two-neuron models, it remains an open question to investigate multi-index models with more neurons and to determine whether feature emergence exhibits hierarchical or sequential patterns in such settings. Another important direction is to extend the analysis to broader classes of PDEs

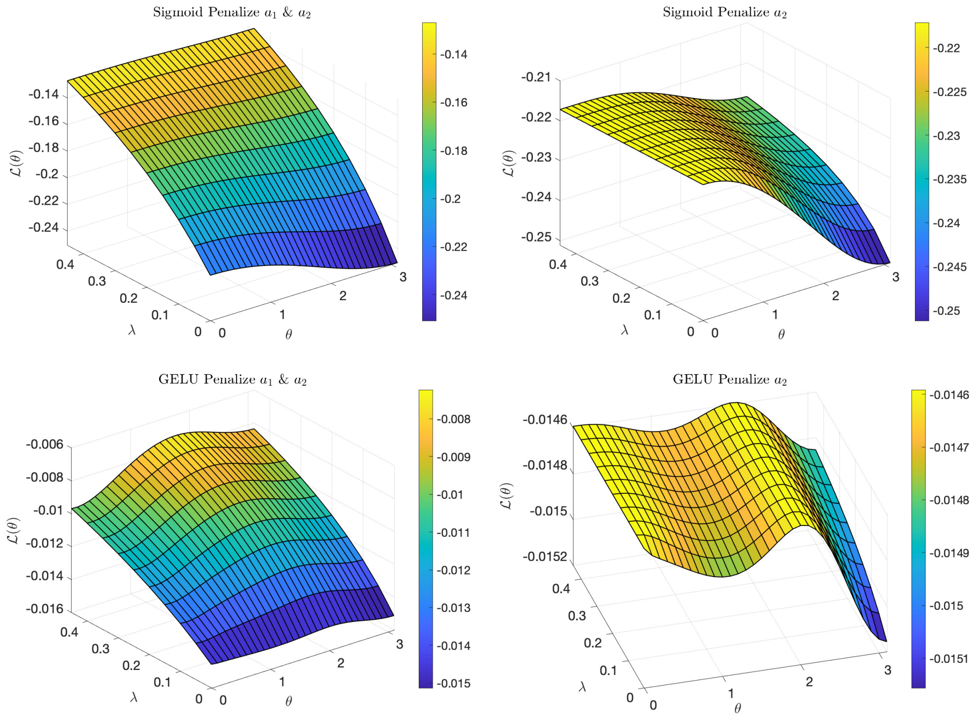

Figure 4: Graphs of the landscapes with $d = 2$ for sigmoid and GELU activation functions.

and boundary conditions, including higher-order or nonlinear equations, where feature structures may be more intricate. Beyond the deep Ritz method, it would also be valuable to explore whether optimizing losses defined by physics-informed neural networks exhibit analogous mechanisms of feature emergence. We plan to study questions along these directions in future work.

References

Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory*, pp. 4782–4887. PMLR, 2022.

Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 2552–2623. PMLR, 2023.

Pranjal Awasthi, Alex Tang, and Aravindan Vijayaraghavan. Agnostic learning of general relu activation using gradient descent. In *International Conference on Learning Representations (ICLR)*, 2023. poster.

Jimmy Ba, Murat A. Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems*, volume 35, 2022. NeurIPS 2022.

Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, 2022.

Andrea Bonfanti, Giuseppe Bruno, and Cristina Cipriani. The challenges of the nonlinear regime for physics-informed neural networks. *Advances in Neural Information Processing Systems*, 37: 41852–41881, 2024.

Joan Bruna and Daniel Hsu. Survey on algorithms for multi-index models. *arXiv preprint*, 2504.05426, 2025.

Ziang Chen, Jianfeng Lu, and Yulong Lu. On the representation of solutions to elliptic pdes in barron spaces. *Advances in neural information processing systems*, 34:6454–6465, 2021.

Ziang Chen, Jianfeng Lu, Yulong Lu, and Shengxuan Zhou. A regularity theory for static schrödinger equations on d in spectral barron spaces. *SIAM Journal on Mathematical Analysis*, 55(1):557–570, 2023.

Youngmin Cho and Lawrence K. Saul. Kernel methods for deep learning. In *Advances in Neural Information Processing Systems*, volume 22, pp. 342–350, 2009.

Hugo Cui, Luca Pesce, Yatin Dandi, Florent Krzakala, Yue Lu, Lenka Zdeborová, and Bruno Loureiro. Asymptotics of feature learning in two-layer networks after one gradient-step. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 9662–9695. PMLR, July 2024.

Alex Damian, Eshaan Nichani, Rong Ge, and Jason D. Lee. Smoothing the landscape boosts the signal for sgd: Optimal sample complexity for learning single index models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. Oral presentation at NeurIPS 2023.

Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, pp. 5413–5452. PMLR, 2022.

Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. How two-layer neural networks learn, one (giant) step at a time. *Journal of Machine Learning Research*, 25(349): 1–65, 2024.

Tim De Ryck and Siddhartha Mishra. Error analysis for physics-informed neural networks (pinns) approximating kolmogorov pdes. *Advances in Computational Mathematics*, 48(6):79, 2022.

Tim De Ryck and Siddhartha Mishra. Numerical analysis of physics-informed neural networks and related models in physics-informed machine learning. *Acta Numerica*, 33:633–713, 2024.

Ilias Diakonikolas, Daniel M Kane, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Agnostically learning multi-index models with queries. In *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 1931–1952. IEEE, 2024.

Rishabh Dudeja and Daniel Hsu. Learning single-index models in gaussian space. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet (eds.), *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1887–1930. PMLR, Jul 2018.

Mathias Dus and Virginie Ehrlacher. Numerical solution of poisson partial differential equation in high dimension using two-layer neural networks. *Mathematics of computation*, 94(351):159–208, 2025. ISSN 0025-5718.

Mathias Dus and Ehrlacher Virginie. Two-layers neural networks for schrödinger eigenvalue problems, 2024. URL `https://arxiv.org/abs/2409.01640`.

Weinan E and Bing Yu. The deep ritz method: A deep learning-based numerical algorithm for solving variational problems. *Communications in Mathematics and Statistics*, 6(1):1–12, Mar 2018. ISSN 2194-671X. doi: 10.1007/s40304-018-0127-z.

Ye Feng and Jianfeng Lu. Solution theory of hamilton-jacobi-bellman equations in spectral barron spaces. *arXiv preprint arXiv:2503.18656*, 2025.

Spencer Frei, Yuan Cao, and Quanquan Gu. Agnostic learning of a single neuron with gradient descent. *Advances in neural information processing systems*, 33:5417–5428, 2020.

Yihang Gao, Yiqi Gu, and Michael Ng. Gradient descent finds the global optima of two-layer physics-informed neural networks. In *International Conference on Machine Learning*, pp. 10676–10707. PMLR, 2023.

Aravind Gollakota, Parikshit Gopalan, Adam Klivans, and Konstantinos Stavropoulos. Agnostically learning single-index models using omnipredictors. In *Advances in Neural Information Processing Systems (NeurIPS) 36*, 2023. Poster Presentation.

Philipp Grohs and Lukas Herrmann. Deep neural network approximation for high-dimensional elliptic pdes with boundary conditions. *IMA Journal of Numerical Analysis*, 42(3):2055–2082, 2022.

Philipp Grohs, Arnulf Jentzen, and Diyora Salimova. Deep neural network approximations for solutions of pdes based on monte carlo algorithms. *Partial Differential Equations and Applications*, 3(4):45, 2022.

Philipp Grohs, Fabian Hornung, Arnulf Jentzen, and Philippe Von Wurstemberger. *A proof that artificial neural networks overcome the curse of dimensionality in the numerical approximation of Black–Scholes partial differential equations*, volume 284. American Mathematical Society, 2023.

Jihun Han, Mihai Nica, and Adam R. Stinchcombe. A derivative-free method for solving elliptic partial differential equations with deep neural networks. *Journal of Computational Physics*, 419: 109672:1–109672:18, 2020. doi: 10.1016/j.jcp.2020.109672.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Yuling Jiao, Ruoxuan Li, Peiying Wu, Jerry Zhijian Yang, and Pingwen Zhang. Drm revisited: A complete error analysis. 2024.

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

Yiping Lu, Haoxuan Chen, Jianfeng Lu, Lexing Ying, and Jose Blanchet. Machine learning for elliptic pdes: Fast rate generalization bound, neural scaling law and minimax optimality. In *Proceedings of the 10th International Conference on Learning Representations (ICLR 2022)*, 2022a. arXiv preprint arXiv:2110.06897 (2021), accepted ICLR 2022.

Yiping Lu, Haoxuan Chen, Jianfeng Lu, Lexing Ying, and Jose Blanchet. Machine learning for elliptic pdes: Fast rate generalization bound, neural scaling law and minimax optimality. In *10th International Conference on Learning Representations, ICLR 2022*, 2022b.

Yulong Lu, Jianfeng Lu, and Min Wang. A priori generalization analysis of the deep ritz method for solving high dimensional elliptic partial differential equations. In *Conference on learning theory*, pp. 3196–3241. PMLR, 2021a.

Yulong Lu, Jianfeng Lu, and Min Wang. A priori generalization analysis of the deep ritz method for solving high dimensional elliptic partial differential equations. In Mikhail Belkin and Samory Kpotufe (eds.), *Proceedings of the Thirty-Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 3196–3241. PMLR, August 15–19 2021b.

Tao Luo and Haizhao Yang. Two-layer neural networks for partial differential equations: optimization and generalization theory. In *Handbook of Numerical Analysis*, volume 25, pp. 515–554, 2024. ISBN 9780443239847.

Tong Mao, Jonathan W. Siegel, and Jinchao Xu. Approximation rates for shallow relu^k neural networks on sobolev spaces via the radon transform. *arXiv preprint*, 2024.

Tanya Marwah, Zachary Lipton, and Andrej Risteski. Parametric complexity bounds for approximating pdes with neural networks. *Advances in Neural Information Processing Systems*, 34: 15044–15055, 2021.

Tanya Marwah, Zachary Chase Lipton, Jianfeng Lu, and Andrej Risteski. Neural network approximations of pdes beyond linearity: A representational perspective. In *International Conference on Machine Learning*, pp. 24139–24172. PMLR, 2023.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

Siddhartha Mishra and Roberto Molinaro. Estimates on the generalization error of physics-informed neural networks for approximating pdes. *IMA Journal of Numerical Analysis*, 43(1):1–43, 2023.

Behrad Moniri, Donghwan Lee, Hamed Hassani, and Edgar Dobriban. A theory of non-linear feature learning with one gradient step in two-layer neural networks. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36106–36159. PMLR, July 2024.

Hong Chul Nam, Julius Berner, and Anima Anandkumar. Solving poisson equations using neural walk-on-spheres. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 37277–37292. PMLR, 2024.

Nikolas Nüsken and Lorenz Richter. Interpolating between BSDEs and PINNs: deep learning for elliptic and parabolic boundary value problems. *arXiv preprint arXiv:2112.03749*, 2021.

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. *Journal of Computational Physics*, 378:686–707, 2019. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2018.10.045.

Grant Rotskoff and Eric Vanden-Eijnden. Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks. *Advances in neural information processing systems*, 31, 2018.

Yeonjong Shin, Zhongqiang Zhang, and George Em Karniadakis. Error estimates of residual minimization using neural networks for linear pdes. *Journal of Machine Learning for Modeling and Computing*, 4(4), 2023.

Justin Sirignano and Konstantinos Spiliopoulos. Dgm: A deep learning algorithm for solving partial differential equations. *Journal of computational physics*, 375:1339–1364, 2018.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.

Mahdi Soltanolkotabi. Learning relus via gradient descent. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

Puqian Wang, Nikos Zarifis, Ilias Diakonikolas, and Jelena Diakonikolas. Robustly learning a single neuron via sharpness. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 36541–36577. PMLR, 2023.

E Weinan and Stephan Wojtowytsch. Some observations on high-dimensional partial differential equations with barron data. In *Mathematical and Scientific Machine Learning*, pp. 253–269. PMLR, 2022.

Michael M. Wolf. Mathematical foundations of supervised learning. URL `https://mediatum.ub.tum.de/doc/1723378/1723378.pdf`.

Lei Wu. Learning a single neuron for non-monotonic activation functions. In *International conference on artificial intelligence and statistics*, pp. 4178–4197. PMLR, 2022.

Xianliang Xu, Ting Du, Wang Kong, Bin Shan, Ye Li, and Zhongyi Huang. Convergence analysis of natural gradient descent for over-parameterized physics-informed neural networks. *arXiv preprint arXiv:2408.00573*, 2024.

Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. *Advances in neural information processing systems*, 32, 2019.

Yaohua Zang, Gang Bao, Xiaojing Ye, and Haomin Zhou. Weak adversarial networks for high-dimensional partial differential equations. *Journal of Computational Physics*, 411:109409, 2020.

Nikos Zarifis, Puqian Wang, Ilias Diakonikolas, and Jelena Diakonikolas. Robustly learning single-index models via alignment sharpness. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 58197–58243. PMLR, 2024.

Wei Zhao and Tao Luo. Convergence guarantees for gradient-based training of neural pde solvers: From linear to nonlinear pdes. *arXiv preprint arXiv:2505.14002*, 2025.

APPENDIX

A PROOF OF THEOREM 2.2

In this section, we provide the proof of Theorem 2.2. First recall that

$$\begin{aligned}\mathcal{L}(\boldsymbol{w}) :=&\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}\left[|\nabla_{\boldsymbol{x}} u_{\boldsymbol{w}}(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} u^*(\boldsymbol{x})|^2 + |u_{\boldsymbol{w}}(\boldsymbol{x}) - u^*(\boldsymbol{x})|^2\right]\\ =&\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}\left[|2\sigma(\boldsymbol{w}\cdot\boldsymbol{x})\boldsymbol{w} - \nabla_{\boldsymbol{x}} u^*(\boldsymbol{x})|^2 + |\sigma^2(\boldsymbol{w}\cdot\boldsymbol{x}) - u^*(\boldsymbol{x})|^2\right],\end{aligned}$$

$$\begin{aligned}g_{\mathcal{L}}(\boldsymbol{w}) :=&\mathrm{P}_{\boldsymbol{w}^\perp}\nabla\mathcal{L}(\boldsymbol{w})\\ =&4\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}\left[\sigma'(\boldsymbol{w}\cdot\boldsymbol{x})\boldsymbol{w}^\top(2\sigma(\boldsymbol{w}\cdot\boldsymbol{x})\boldsymbol{w} - \nabla u^*)\mathrm{P}_{\boldsymbol{w}^\perp}\boldsymbol{x} - \sigma(\boldsymbol{w}\cdot\boldsymbol{x})\mathrm{P}_{\boldsymbol{w}^\perp}\nabla u^* + (\sigma^2(\boldsymbol{w}\cdot\boldsymbol{x}) - u^*(\boldsymbol{x}))\sigma(\boldsymbol{w}\cdot\boldsymbol{x})\mathrm{P}_{\boldsymbol{w}^\perp}\boldsymbol{x}\right],\end{aligned}$$

and

$$\mathrm{OPT} := \mathcal{L}(\boldsymbol{w}^*) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}\left[|\nabla_{\boldsymbol{x}} u_{\boldsymbol{w}^*}(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} u^*(\boldsymbol{x})|^2 + |u_{\boldsymbol{w}^*}(\boldsymbol{x}) - u^*(\boldsymbol{x})|^2\right],$$

where $\boldsymbol{w}^* := \arg\min_{\boldsymbol{w}\in\mathcal{S}^{d-1}}\mathcal{L}(\boldsymbol{w})$. With the above definitions and by Young's inequality, we can get

$$\mathcal{L}(\boldsymbol{w}) \leqslant 2\mathcal{L}^*(\boldsymbol{w}) + 2\mathrm{OPT},$$

where $\mathcal{L}^*(\boldsymbol{w})$ represents the loss in realizable setting and is defined as

$$\begin{aligned}\mathcal{L}^*(\boldsymbol{w}) :=&\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}\left[|\nabla_{\boldsymbol{x}} u_{\boldsymbol{w}}(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} u_{\boldsymbol{w}^*}(\boldsymbol{x})|^2 + |u_{\boldsymbol{w}}(\boldsymbol{x}) - u_{\boldsymbol{w}^*}(\boldsymbol{x})|^2\right]\\ =&\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}\left[|2\sigma(\boldsymbol{w}\cdot\boldsymbol{x})\boldsymbol{w} - 2\sigma(\boldsymbol{w}^*\cdot\boldsymbol{x})\boldsymbol{w}^*|^2 + |\sigma^2(\boldsymbol{w}\cdot\boldsymbol{x}) - \sigma^2(\boldsymbol{w}^*\cdot\boldsymbol{x})|^2\right].\end{aligned}\tag{13}$$

The Riemannian gradient of the loss (13) is defined as

$$\begin{aligned}g^*(\boldsymbol{w}) :=&\mathrm{P}_{\boldsymbol{w}^\perp}\nabla\mathcal{L}^*(\boldsymbol{w})\\ =&\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}4\left[\sigma'(\boldsymbol{w}\cdot\boldsymbol{x})\boldsymbol{w}^\top(2\sigma(\boldsymbol{w}\cdot\boldsymbol{x})\boldsymbol{w} - 2\sigma(\boldsymbol{w}^*\cdot\boldsymbol{x})\boldsymbol{w}^*)\mathrm{P}_{\boldsymbol{w}^\perp}\boldsymbol{x} - 2\sigma(\boldsymbol{w}\cdot\boldsymbol{x})\sigma(\boldsymbol{w}^*\cdot\boldsymbol{x})\mathrm{P}_{\boldsymbol{w}^\perp}\boldsymbol{w}^*\right]\\ &+\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}\left[4(\sigma^2(\boldsymbol{w}\cdot\boldsymbol{x}) - \sigma^2(\boldsymbol{w}^*\cdot\boldsymbol{x}))\sigma(\boldsymbol{w}\cdot\boldsymbol{x})\mathrm{P}_{\boldsymbol{w}^\perp}\boldsymbol{x}\right].\end{aligned}$$

We denote the difference between the Riemannian gradient of the agnostic loss $g_{\mathcal{L}}(\boldsymbol{w})$ and the realizable loss $g^*(\boldsymbol{w})$ by $\xi(\boldsymbol{w})$, i.e.,

$$\begin{aligned}\xi(\boldsymbol{w}) :=&g_{\mathcal{L}}(\boldsymbol{w}) - g^*(\boldsymbol{w})\\ =&4\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}\left[\sigma'(\boldsymbol{w}\cdot\boldsymbol{x})\boldsymbol{w}^\top(2\sigma(\boldsymbol{w}^*\cdot\boldsymbol{x})\boldsymbol{w}^* - \nabla u^*)\mathrm{P}_{\boldsymbol{w}^\perp}\boldsymbol{x} + \sigma(\boldsymbol{w}\cdot\boldsymbol{x})(2\sigma(\boldsymbol{w}^*\cdot\boldsymbol{x})P_{\boldsymbol{w}^\perp}\boldsymbol{w}^* - \mathrm{P}_{\boldsymbol{w}^\perp}\nabla u^*)\right]\\ &+4\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}\left[(\sigma^2(\boldsymbol{w}^*\cdot\boldsymbol{x}) - u^*)\sigma(\boldsymbol{w}\cdot\boldsymbol{x})\mathrm{P}_{\boldsymbol{w}^\perp}\boldsymbol{x}\right].\end{aligned}\tag{14}$$

Throughout this section, we define a constant $C_d := \frac{8}{\sqrt{d+2}}$. We first show that the norm of $\xi(\boldsymbol{w})$ and the inner product of $\xi(\boldsymbol{w})$ and $\boldsymbol{w}^*$ are bounded.

Lemma A.1. *Let $\xi(\boldsymbol{w})$ as defined in (14). Then,*

$$|\xi(\boldsymbol{w})|_2 \leqslant C_d\sqrt{OPT} \quad \textit{and} \quad |\xi(\boldsymbol{w})\cdot\boldsymbol{w}^*| \leqslant C_d\sqrt{OPT}|(\boldsymbol{w}^*)^{\perp\boldsymbol{w}}|_2.$$

Proof. For simplicity, we use $\mathbb{E}$ to denote $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}$ throughout this proof. By the definition of $\xi(\boldsymbol{w})$ and the definition of the 2-norm, we have

$$\begin{aligned}&|\xi(\boldsymbol{w})|_2\\ =&\max_{\boldsymbol{v}\in\mathcal{S}^{d-1}}\mathbb{E}\left[4\sigma'(\boldsymbol{w}\cdot\boldsymbol{x})\boldsymbol{w}^\top(2\sigma(\boldsymbol{w}^*\cdot\boldsymbol{x})\boldsymbol{w}^* - \nabla_{\boldsymbol{x}} u^*)\mathrm{P}_{\boldsymbol{w}^\perp}\boldsymbol{x}\cdot\boldsymbol{v} + 4\sigma(\boldsymbol{w}\cdot\boldsymbol{x})(2\sigma(\boldsymbol{w}^*\cdot\boldsymbol{x})\mathrm{P}_{\boldsymbol{w}^\perp}\boldsymbol{w}^* - \mathrm{P}_{\boldsymbol{w}^\perp}\nabla_{\boldsymbol{x}} u^*)\cdot\boldsymbol{v}\right]\\ &+\mathbb{E}\left[4(\sigma^2(\boldsymbol{w}^*\cdot\boldsymbol{x}) - u^*)\sigma(\boldsymbol{w}\cdot\boldsymbol{x})\sigma'(\boldsymbol{w}^\top\boldsymbol{x})\mathrm{P}_{\boldsymbol{w}^\perp}\boldsymbol{x}\cdot\boldsymbol{v}\right]\\ =&\max_{\boldsymbol{v}\in\mathcal{S}^{d-1}}\mathbb{E}\left[4\sigma'(\boldsymbol{w}\cdot\boldsymbol{x})\boldsymbol{w}^\top(2\sigma(\boldsymbol{w}^*\cdot\boldsymbol{x})\boldsymbol{w}^* - \nabla_{\boldsymbol{x}} u^*)\boldsymbol{x}\cdot\mathrm{P}_{\boldsymbol{w}^\perp}\boldsymbol{v} + 4\sigma(\boldsymbol{w}\cdot\boldsymbol{x})(2\sigma(\boldsymbol{w}^*\cdot\boldsymbol{x})\mathrm{P}_{\boldsymbol{w}^\perp}\boldsymbol{w}^* - \mathrm{P}_{\boldsymbol{w}^\perp}\nabla_{\boldsymbol{x}} u^*)\cdot\boldsymbol{v}\right]\\ &+\mathbb{E}\left[4(\sigma^2(\boldsymbol{w}^*\cdot\boldsymbol{x}) - u^*)\sigma(\boldsymbol{w}\cdot\boldsymbol{x})\sigma'(\boldsymbol{w}^\top\boldsymbol{x})\boldsymbol{x}\cdot\mathrm{P}_{\boldsymbol{w}^\perp}\boldsymbol{v}\right]\\ \leqslant&\max_{\boldsymbol{v}\in\mathcal{S}^{d-1}}\sqrt{\mathbb{E}\left[(\boldsymbol{w}^\top(2\sigma(\boldsymbol{w}^*\cdot\boldsymbol{x})\boldsymbol{w}^* - \nabla_{\boldsymbol{x}} u^*))^2\right]\mathbb{E}\left[(4\sigma'(\boldsymbol{w}\cdot\boldsymbol{x})\boldsymbol{x}\cdot\mathrm{P}_{\boldsymbol{w}^\perp}\boldsymbol{v})^2\right]}\\ &+\sqrt{\mathbb{E}\left[((2\sigma(\boldsymbol{w}^*\cdot\boldsymbol{x})\boldsymbol{w}^* - \nabla_{\boldsymbol{x}} u^*)\cdot\mathrm{P}_{\boldsymbol{w}^\perp}\boldsymbol{v})^2\right]\mathbb{E}\left[(4\sigma(\boldsymbol{w}\cdot\boldsymbol{x}))^2\right]}\\ &+\sqrt{\mathbb{E}\left[(\sigma^2(\boldsymbol{w}^*\cdot\boldsymbol{x}) - u^*)^2\right]\mathbb{E}\left[(4\sigma(\boldsymbol{w}\cdot\boldsymbol{x})\sigma'(\boldsymbol{w}^\top\boldsymbol{x})\boldsymbol{x}\cdot\mathrm{P}_{\boldsymbol{w}^\perp}\boldsymbol{v})^2\right]}\\ \leqslant&\left(\frac{4}{\sqrt{2(d+2)}} + \frac{4}{\sqrt{2(d+2)}}\right)\sqrt{\mathbb{E}\left[|2\sigma(\boldsymbol{w}^*\cdot\boldsymbol{x})\boldsymbol{w}^* - \nabla_{\boldsymbol{x}} u^*|^2\right]} + \frac{4}{\sqrt{2(d+2)(d+4)}}\sqrt{\mathbb{E}\left[|\sigma^2(\boldsymbol{w}^*\cdot\boldsymbol{x}) - u^*|^2\right]}.\end{aligned}$$

Recall that

$$\begin{aligned}\mathrm{OPT} :=&\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}\left[|\nabla_{\boldsymbol{x}} u_{\boldsymbol{w}^*}(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} u^*(\boldsymbol{x})|^2 + |u_{\boldsymbol{w}^*}(\boldsymbol{x}) - u^*(\boldsymbol{x})|^2\right]\\ =&\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}\left[|2\sigma(\boldsymbol{w}^*\cdot\boldsymbol{x})\boldsymbol{w}^* - \nabla_{\boldsymbol{x}} u^*|^2 + |\sigma^2(\boldsymbol{w}^*\cdot\boldsymbol{x}) - u^*|^2\right].\end{aligned}$$

Hence,

$$|\xi(\boldsymbol{w})|_2 \leqslant \frac{8}{\sqrt{2(d+2)}}\sqrt{2\mathrm{OPT}} = C_d\sqrt{\mathrm{OPT}}.$$

Similarly, we can get

$$|\xi(\boldsymbol{w})\cdot\boldsymbol{w}^*| \leqslant C_d\sqrt{\mathrm{OPT}}|(\boldsymbol{w}^*)^{\perp\boldsymbol{w}}|_2.$$

□

Then, by the triangle inequality, we have the following Corollary.

Corollary A.2. *For any $\boldsymbol{w}\in\mathcal{S}^{d-1}$, $|g_{\mathcal{L}}(\boldsymbol{w})| \leqslant |\xi(\boldsymbol{w})| + |g^*(\boldsymbol{w})| \leqslant C_d\sqrt{OPT} + |g^*(\boldsymbol{w})|$.*

With the above results, we are ready to show the sharpness property of the Riemannian gradient $g_{\mathcal{L}}(\boldsymbol{w})$. Before that, we first note that the expectation in $\mathcal{L}^*(\boldsymbol{w})$ can be explicitly calculated by applying the similar techniques as in Cho & Saul (2009). We calculate the expectations as

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}[\sigma(\boldsymbol{w}\cdot\boldsymbol{x})\sigma(\boldsymbol{w}^*\cdot\boldsymbol{x})\boldsymbol{w}\cdot\boldsymbol{w}^*] &= \frac{1}{2\pi(d+2)}\left(\sin\theta\cos\theta + (\pi-\theta)\cos^2\theta\right),\\ \mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}[\sigma^2(\boldsymbol{w}\cdot\boldsymbol{x})\sigma^2(\boldsymbol{w}^*\cdot\boldsymbol{x})] &= \frac{1}{2\pi(d+2)(d+4)}(3\sin\theta\cos\theta + (\pi-\theta)(1+2\cos^2\theta)),\end{aligned}$$

where θ is the angle between $\boldsymbol{w}$ and $\boldsymbol{w}^*$ and $\theta\in[0,\pi]$. Then we have

$$\begin{aligned}\mathcal{L}^*(\boldsymbol{w}) = &\left(\frac{4}{d+2} - \frac{4}{\pi(d+2)}\left(\sin\theta\cos\theta + (\pi-\theta)\cos^2\theta\right)\right)\\ &+ \left(\frac{3}{(d+2)(d+4)} - \frac{1}{\pi(d+2)(d+4)}\left(3\sin\theta\cos\theta + (\pi-\theta)(1+2\cos^2\theta)\right)\right).\end{aligned}$$

The sharpness property of $g_{\mathcal{L}}(\boldsymbol{w})$ is shown as follows.

Lemma A.3 (Sharpness)**.** *Let $\theta := \angle(\boldsymbol{w},\boldsymbol{w}^*)$ (angle between $\boldsymbol{w}$ and $\boldsymbol{w}^*$). Assume $\theta\in[0,\frac{\pi}{2}]$ and $\sin\theta \geqslant \frac{32\pi\sqrt{OPT}}{C_d}$, then*

$$g_{\mathcal{L}}(\boldsymbol{w})\cdot\boldsymbol{w}^* \leqslant -\frac{1}{2}|g^*(\boldsymbol{w})|_2\sin\theta.$$

Proof. We take the Riemannian gradient of $\mathcal{L}^*(\boldsymbol{w})$ to get

$$g^*(\boldsymbol{w}) = \left(-\frac{4(\sin\theta + 2(\pi-\theta)\cos\theta)}{\pi(d+2)} - \frac{4\sin\theta + 4(\pi-\theta)\cos\theta)}{\pi(d+2)(d+4)}\right)\mathrm{P}_{\boldsymbol{w}^\perp}\boldsymbol{w}^*.$$

By noting that

$$\begin{aligned}g^*(\boldsymbol{w})\cdot\boldsymbol{w}^* =&\left(-\frac{4(\sin\theta + 2(\pi-\theta)\cos\theta)}{\pi(d+2)} - \frac{4\sin\theta + 4(\pi-\theta)\cos\theta)}{\pi(d+2)(d+4)}\right)\mathrm{P}_{\boldsymbol{w}^\perp}\boldsymbol{w}^*\cdot\boldsymbol{w}^*\\ =&-|g^*(\boldsymbol{w})|_2\sin\theta,\end{aligned}$$

and applying Lemma A.1, we can get

$$g_{\mathcal{L}}(\boldsymbol{w})\cdot\boldsymbol{w}^* = g^*(\boldsymbol{w})\cdot\boldsymbol{w}^* + \xi(\boldsymbol{w})\cdot\boldsymbol{w}^* \leqslant -\left(|g^*(\boldsymbol{w})|_2 - C_d\sqrt{\mathrm{OPT}}\right)\sin\theta.$$

Then we consider $|g^*(\boldsymbol{w})|$. We have that

$$|g^*(\boldsymbol{w})|_2 = \left(\frac{4(\sin\theta + 2(\pi-\theta)\cos\theta)}{\pi(d+2)} + \frac{4\sin\theta + 4(\pi-\theta)\cos\theta)}{\pi(d+2)(d+4)}\right)\sin\theta,$$

and note that when $\theta\in[0,\frac{\pi}{2}]$,

$$\underline{c}\sin\theta \leqslant |g^*(\boldsymbol{w})| \leqslant \bar{c}\sin\theta,$$

where $\bar{c} := \frac{8}{(d+2)} + \frac{4}{(d+2)(d+4)}$ and $\underline{c} := \frac{4}{\pi(d+2)} + \frac{4}{\pi(d+2)(d+4)}$. Therefore, by noting that $\underline{c} \geqslant \frac{C_d^2}{16\pi}$, we have that when $\sin\theta \geqslant \frac{2C_d\sqrt{\mathrm{OPT}}}{C_d^2/(16\pi)} \geqslant \frac{2C_d\sqrt{\mathrm{OPT}}}{\underline{c}}$,

$$|g^*(\boldsymbol{w})|_2 \geqslant 2C_d\sqrt{\mathrm{OPT}}, \quad \frac{1}{2}|g^*(\boldsymbol{w})|_2 \geqslant C_d\sqrt{\mathrm{OPT}}.$$

Hence, we can get

$$|g^*(\boldsymbol{w})|_2 - C_d\sqrt{\mathrm{OPT}} \geqslant |g^*(\boldsymbol{w})|_2 - \frac{1}{2}|g^*(\boldsymbol{w})|_2 = \frac{1}{2}|g^*(\boldsymbol{w})|_2,$$

and

$$g_{\mathcal{L}}(\boldsymbol{w}) \cdot \boldsymbol{w}^* \leqslant -\frac{1}{2}|g^*(\boldsymbol{w})|_2 \sin\theta.$$

□

If $\sin\theta < \frac{32\pi\sqrt{\mathrm{OPT}}}{C_d}$, we can get the approximate optimal solution immediately, which is shown in the following lemma.

Lemma A.4. *Let $\theta := \angle(\boldsymbol{w}, \boldsymbol{w}^*)$. If $\theta \in [0, \frac{\pi}{2}]$ and $\sin\theta < \frac{32\pi\sqrt{OPT}}{C_d}$, then*

$$\mathcal{L}(\boldsymbol{w}) < (256\pi^2 + 2)OPT.$$

Proof. By Young's inequality, we have

$$\begin{aligned}
\mathcal{L}(\boldsymbol{w}) \leqslant & 2\mathrm{OPT} + 2\mathcal{L}^*(\boldsymbol{w}) \\
= & 2\mathrm{OPT} + \frac{8}{d+2}\left(1 - \frac{1}{\pi}\sin\theta\cos\theta - (1 - \frac{\theta}{\pi})\cos^2\theta\right) \\
& + \frac{2}{(d+2)(d+4)}\left(3 - \frac{3}{\pi}\sin\theta\cos\theta - (1 - \frac{\theta}{\pi})(1 + 2\cos^2\theta)\right) \\
\leqslant & 2\mathrm{OPT} + \frac{8}{d+2}\left(\sin^2\theta + \frac{1}{\pi}\cos\theta(\theta\cos\theta - \sin\theta)\right) \\
& + \frac{2}{(d+2)(d+4)}\left(3\sin^2\theta + \frac{3}{\pi}\cos\theta(\theta\cos\theta - \sin\theta)\right) \\
\leqslant & 2\mathrm{OPT} + \frac{8}{d+2}\sin^2\theta + \frac{6}{(d+2)(d+4)}\sin^2\theta \\
\leqslant & \left(\left(\frac{8}{d+2} + \frac{6}{(d+2)(d+4)}\right)\frac{(32\pi)^2}{C_d^2} + 2\right)\mathrm{OPT},
\end{aligned}$$

where the second inequality comes from $\frac{\theta}{\pi} - 1 \leqslant 0$ and the third inequality comes from $\theta\cos\theta - \sin\theta \leqslant 0$. Since $\frac{8}{d+2} + \frac{6}{(d+2)(d+4)} < \frac{C_d^2}{4}$, we have

$$\mathcal{L}(\boldsymbol{w}) < (256\pi^2 + 2)\mathrm{OPT}.$$

□

Next, we provide the uniform upper bound on the number of samples required to approximate the Riemannian gradients $g_{\mathcal{E}}(\boldsymbol{w})$ in Lemma A.5 with the following definitions the empirical estimate of $g_{\mathcal{E}}(\boldsymbol{w})$:

$$\begin{aligned}
\hat{g}_{\mathcal{E}}(\boldsymbol{w}) &:= \frac{1}{n}\sum_{i=1}^{n} g_{\mathcal{E}}(\boldsymbol{w}; \boldsymbol{x}_i), \\
g_{\mathcal{E}}(\boldsymbol{w}; \boldsymbol{x}_i) &:= 4\left(2|\boldsymbol{w}|^2\sigma(\boldsymbol{w}\cdot\boldsymbol{x}_i)\mathrm{P}_{\boldsymbol{w}^\perp}\boldsymbol{x}_i + \sigma^3(\boldsymbol{w}\cdot\boldsymbol{x}_i)\mathrm{P}_{\boldsymbol{w}^\perp}\boldsymbol{x}_i - f(\boldsymbol{x}_i)\sigma(\boldsymbol{w}\cdot\boldsymbol{x}_i)\mathrm{P}_{\boldsymbol{w}^\perp}\boldsymbol{x}_i\right).
\end{aligned}$$

Lemma A.5. *Let $\boldsymbol{w}^*, \boldsymbol{w} \in \mathcal{S}^{d-1}$ and $\hat{g}_{\mathcal{E}}(\boldsymbol{w})$ be the empirical estimate of the Riemannian gradient $g_{\mathcal{E}}(\boldsymbol{w})$. Then, under Assumption 2.1, there exists a constant C_1 depending on δ, C_f such that with probability at least $1-\delta$,*

$$\sup_{\boldsymbol{w}\in\mathcal{S}^{d-1}} |\hat{g}_{\mathcal{E}}(\boldsymbol{w}) - g_{\mathcal{E}}(\boldsymbol{w})| \lesssim \sqrt{\frac{C_1 d}{n}},$$

$$\sup_{\boldsymbol{w}\in\mathcal{S}^{d-1}} (\hat{g}_{\mathcal{E}}(\boldsymbol{w}) - g_{\mathcal{E}}(\boldsymbol{w})) \cdot \boldsymbol{w}^* \lesssim \sqrt{\frac{C_1 d}{n}}.$$

Proof. Let $\mathcal{G} := \{g_{\mathcal{E}}(\boldsymbol{w}) : \boldsymbol{w} \in \mathcal{S}^{d-1}\}$, then by Assumption 2.1 and that $|\boldsymbol{x}| \leqslant 1$, we have that $|g_{\mathcal{E}}(\boldsymbol{w}, \boldsymbol{x})| \leqslant M$ with $M := 12 + 4C_f$. Thanks to Theorem 4.10 of Wainwright (2019), one has that with probability at least $1-\delta$,

$$\sup_{\boldsymbol{w}\in\mathcal{S}^{d-1}} |\hat{g}_{\mathcal{E}}(\boldsymbol{w}) - g_{\mathcal{E}}(\boldsymbol{w})| \leqslant 2\mathcal{R}_n(\mathcal{G}) + M\sqrt{\frac{2\log(1/\delta)}{n}}, \tag{15}$$

where $\mathcal{R}_n(\mathcal{G})$ is the Rademacher complexity of a function class $\mathcal{G}$, defined by

$$\mathcal{R}_n(\mathcal{G}) := \mathbb{E}_{\boldsymbol{x}_i,\varepsilon_i} \sup_{g\in\mathcal{G}} \Big|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i g(\boldsymbol{x}_i)\Big|,$$

where $\{\varepsilon_i\}_{i=1}^n$ is a sequence of i.i.d. Rademacher random variables. Moreover, under Assumption 2.1, one has that

$$\|g_{\mathcal{E}}(\boldsymbol{w}; \boldsymbol{x}) - g_{\mathcal{E}}(\boldsymbol{w}'; \boldsymbol{x})\|_\infty \leqslant L|\boldsymbol{w} - \boldsymbol{w}'|,$$

where $L := 64 + 16C_f$. This implies that

$$\mathcal{N}(\delta, \mathcal{G}, \|\cdot\|_\infty) \leqslant \mathcal{N}(\frac{\delta}{L}, \mathcal{S}^{d-1}, |\cdot|) \leqslant \left(\frac{3L}{\delta}\right)^d,$$

where $\mathcal{N}(\delta, \mathcal{G}, \|\cdot\|_\infty)$ denotes the δ-covering number of $\mathcal{G}$ w.r.t. the L_∞-norm and $\mathcal{N}(\frac{\delta}{L}, \mathcal{S}^{d-1}, |\cdot|)$ denotes the $\frac{\delta}{L}$-covering number of $\mathcal{S}^{d-1}$ w.r.t. the 2-norm. Applying Dudley's theorem (Wolf; Lu et al., 2021b) and noting that $\sup_{\boldsymbol{w}\in\mathcal{S}^{d-1}} \|g_{\mathcal{E}}(\boldsymbol{w})\|_\infty \leqslant M$, we can get

$$\begin{aligned}
\mathcal{R}_n(\mathcal{G}) \leqslant &\inf_{0\leqslant\delta\leqslant M} \left\{4\delta + \frac{12}{\sqrt{n}} \int_\delta^M \sqrt{\log\mathcal{N}(\tau, \mathcal{G}, \|\cdot\|_\infty)}\, d\tau\right\} \\
&\inf_{0\leqslant\delta\leqslant M} \left\{\leqslant 4\delta + \frac{12}{\sqrt{n}} \int_\delta^M \sqrt{d\log\left(\frac{3L}{\tau}\right)}\, d\tau\right\}.
\end{aligned}$$

By setting $\delta = 0$, we have

$$\begin{aligned}
\mathcal{R}_n(\mathcal{G}) &\leqslant 12\sqrt{\frac{d}{n}} \int_0^M \sqrt{\log(3L) + \log(1/\tau)}\, d\tau \\
&\lesssim 12\sqrt{\frac{d}{n}} \int_0^M \sqrt{\log(3L)} + \sqrt{\log(1/\tau)}\, d\tau \\
&\leqslant 12\sqrt{\frac{d}{n}} \left(M\sqrt{\log(3L)} + C\right) \\
&\leqslant \sqrt{\frac{\tilde{C}_1 d}{n}},
\end{aligned}$$

where $\tilde{C}_1$ depends at mostly polynomially on C_f. Plugging the bound above into (15), we can conclude that with probability at least $1-\delta$,

$$\sup_{\boldsymbol{w}\in\mathcal{S}^{d-1}} |\hat{g}_{\mathcal{E}}(\boldsymbol{w}) - g_{\mathcal{E}}(\boldsymbol{w})| \lesssim \sqrt{\frac{C_1 d}{n}}$$

and

$$\sup_{\boldsymbol{w}\in\mathcal{S}^{d-1}} (\hat{g}_{\mathcal{E}}(\boldsymbol{w}) - g_{\mathcal{E}}(\boldsymbol{w}))\cdot \boldsymbol{w}^* \leqslant \sup_{\boldsymbol{w}\in\mathcal{S}^{d-1}} |\hat{g}_{\mathcal{E}}(\boldsymbol{w}) - g_{\mathcal{E}}(\boldsymbol{w})| \\ \lesssim\sqrt{\frac{C_1 d}{n}}.$$

The constant C_1 depends on C_f and δ.

□

With above lemmas and corollary, we proceed to show the main result of this subsection.

Theorem A.6. *Suppose that Assumption 2.1 holds. Consider Algorithm 1 with initial condition $\boldsymbol{w}^0$ satisfying $\angle(\boldsymbol{w}^0, \boldsymbol{w}^*) \in [0, \frac{\pi}{2}]$. If we choose the sample size $n = \Theta\left(\frac{C_1 d}{C_d^2 \epsilon}\right)$ and the step size $\eta = \frac{1}{32\pi C_d^2}$, then after $T = O(\log(1/\epsilon))$ iterations, with probability at least $1-\delta$, the output of Algorithm 1 $\boldsymbol{w}^T$ satisfies $\mathcal{L}(\boldsymbol{w}^T) = O(OPT) + C_d^2\epsilon$.*

Proof. Since $|\boldsymbol{w}^t - \eta\hat{g}_{\mathcal{E}}(\boldsymbol{w}^t)|^2 = 1 + \eta^2|\hat{g}_{\mathcal{E}}(\boldsymbol{w}^t)|^2 \geqslant 1$, we have

$$\begin{aligned} |\boldsymbol{w}^{t+1} - \boldsymbol{w}^*|^2 \leqslant & |\boldsymbol{w}^t - \eta\hat{g}_{\mathcal{E}}(\boldsymbol{w}^t) - \boldsymbol{w}^*|^2 \\ = & |\boldsymbol{w}^t - \boldsymbol{w}^*|^2 + 2\eta\hat{g}_{\mathcal{E}}(\boldsymbol{w}^t)\cdot(\boldsymbol{w}^* - \boldsymbol{w}^t) + \eta^2|\hat{g}_{\mathcal{E}}(\boldsymbol{w}^t)|^2. \end{aligned} \tag{16}$$

Let us denote the angle between $\boldsymbol{w}^t$ and $\boldsymbol{w}^*$ by θ_t and assume that θ_t satisfies $\sin\theta_t \geqslant \frac{32\pi\sqrt{\mathrm{OPT}}}{C_d} + \sqrt{\epsilon}$, hence the condition for Lemma A.3 is satisfied. Note by definition $\hat{g}_{\mathcal{E}}(\boldsymbol{w}^t)\perp\boldsymbol{w}^t$, hence using Lemma A.3 and Lemma A.5, we have that with probability at least $1-\delta$,

$$\begin{aligned} \hat{g}_{\mathcal{E}}(\boldsymbol{w}^t)\cdot(\boldsymbol{w}^* - \boldsymbol{w}^t) = & \hat{g}_{\mathcal{E}}(\boldsymbol{w}^t)\cdot\boldsymbol{w}^* \\ = & (\hat{g}_{\mathcal{E}}(\boldsymbol{w}^t) - g_{\mathcal{L}}(\boldsymbol{w}^t))\cdot\boldsymbol{w}^* + g_{\mathcal{L}}(\boldsymbol{w}^t)\cdot\boldsymbol{w}^* \\ = & (\hat{g}_{\mathcal{E}}(\boldsymbol{w}^t) - g_{\mathcal{E}}(\boldsymbol{w}^t))\cdot\boldsymbol{w}^* + (g_{\mathcal{E}}(\boldsymbol{w}^t) - g_{\mathcal{L}}(\boldsymbol{w}^t))\cdot\boldsymbol{w}^* + g_{\mathcal{L}}(\boldsymbol{w}^t)\cdot\boldsymbol{w}^* \\ \leqslant & \sqrt{\frac{C_1 d}{n}} - \frac{1}{2}|g^*(\boldsymbol{w}^t)|_2\sin\theta_t, \end{aligned} \tag{17}$$

by noting that $g_{\mathcal{E}}(\boldsymbol{w}^t) = g_{\mathcal{L}}(\boldsymbol{w}^t)$. By Corollary A.2 and Lemma A.5, the squared norm term $|\hat{g}(\boldsymbol{w}^t)|^2$ in equation (16) can be bounded by

$$\begin{aligned} |\hat{g}_{\mathcal{E}}(\boldsymbol{w}^t)|^2 = & |\hat{g}_{\mathcal{E}}(\boldsymbol{w}^t) - g_{\mathcal{L}}(\boldsymbol{w}^t) + g_{\mathcal{L}}(\boldsymbol{w}^t)|^2 \\ \leqslant & 2|\hat{g}_{\mathcal{E}}(\boldsymbol{w}^t) - g_{\mathcal{L}}(\boldsymbol{w}^t)|^2 + 2|g_{\mathcal{L}}(\boldsymbol{w}^t)|^2 \\ \leqslant & 4|\hat{g}_{\mathcal{E}}(\boldsymbol{w}^t) - g_{\mathcal{E}}(\boldsymbol{w}^t)|^2 + 4|g_{\mathcal{E}}(\boldsymbol{w}^t) - g_{\mathcal{L}}(\boldsymbol{w}^t)|^2 + 2|g_{\mathcal{L}}(\boldsymbol{w}^t)|^2 \\ \leqslant & \frac{4C_1 d}{n} + 4C_d^2\mathrm{OPT} + 4|g^*(\boldsymbol{w}^t)|^2, \end{aligned} \tag{18}$$

by noting that $g_{\mathcal{E}}(\boldsymbol{w}^t) = g_{\mathcal{L}}(\boldsymbol{w}^t)$. Plugging (17) and (18) back into (16) and letting $\kappa_d := \sqrt{4C_1 d}$, we have that with probability at least $1-\delta$,

$$\begin{aligned} |\boldsymbol{w}^{t+1} - \boldsymbol{w}^*|^2 \leqslant & |\boldsymbol{w}^t - \boldsymbol{w}^*|^2 + 2\eta\left(\frac{\kappa_d}{\sqrt{n}} - \frac{1}{2}|g^*(\boldsymbol{w}^t)|\sin\theta_t\right) + \eta^2\left(\frac{\kappa_d^2}{n} + 4C_d^2\mathrm{OPT} + 4|g^*(\boldsymbol{w}^t)|^2\right) \\ = & |\boldsymbol{w}^t - \boldsymbol{w}^*|^2 + \frac{2\eta\kappa_d}{\sqrt{n}} - \eta|g^*(\boldsymbol{w}^t)|\sin\theta_t + \eta^2\left(\frac{\kappa_d^2}{n} + 4C_d^2\mathrm{OPT} + 4|g^*(\boldsymbol{w}^t)|^2\right). \end{aligned}$$

We first assume that $\theta_t \leqslant \theta_{t-1} \leqslant \cdots \leqslant \theta_0 \leqslant \delta = \frac{\pi}{2}$ and $\sin\theta_t \geqslant \frac{32\pi\sqrt{\mathrm{OPT}}}{C_d} + \sqrt{\epsilon}$. We will argue that $\theta_{t+1} \leqslant \theta_t$ in this case. Then, by an inductive argument, we immediately know that the assumption is valid and that θ_t is a decreasing sequence (as long as $\sin\theta_t \geqslant \frac{32\pi\sqrt{\mathrm{OPT}}}{C_d} + \sqrt{\epsilon}$). To prove $\theta_{t+1} \leqslant \theta_t$, we first recall that

$$|g^*(\boldsymbol{w}^t)|_2 = \left(\frac{4(\sin\theta_t + 2(\pi-\theta_t)\cos\theta_t)}{\pi(d+2)} + \frac{4\sin\theta_t + 4(\pi-\theta_t)\cos\theta_t)}{\pi(d+2)(d+4)}\right)\sin\theta_t,$$

and

$$\underline{c}\sin\theta_t \leqslant |g^*(\boldsymbol{w}^t)| \leqslant \bar{c}\sin\theta_t,$$

where $\bar{c} := \frac{8}{(d+2)} + \frac{4}{(d+2)(d+4)}$ and $\underline{c} := \frac{4}{\pi(d+2)} + \frac{4}{\pi(d+2)(d+4)}$. Noting that $C_d^2 < 8\bar{c}$, we have

$$\begin{aligned}|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*|^2 \leqslant& |\boldsymbol{w}^t - \boldsymbol{w}^*|^2 + \frac{2\eta\kappa_d}{\sqrt{n}} - \eta\underline{c}\sin^2\theta_t + \eta^2\left(\frac{\kappa_d^2}{n} + 4C_d^2\text{OPT} + 4\bar{c}^2\sin^2\theta_t\right)\\ <& |\boldsymbol{w}^t - \boldsymbol{w}^*|^2 + \frac{2\eta\kappa_d}{\sqrt{n}} - \eta\underline{c}\sin^2\theta_t + 4\eta^2\bar{c}^2\left(\frac{\kappa_d^2}{4n\bar{c}^2} + \frac{8}{\bar{c}}\text{OPT} + \sin^2\theta_t\right).\end{aligned}$$

Now choosing $n \gtrsim \kappa_d^2/(4\bar{c}^2\epsilon)$ and recalling that

$$\sin^2\theta_t \geqslant \left(\frac{32\pi\sqrt{\text{OPT}}}{C_d} + \sqrt{\epsilon}\right)^2 \geqslant \frac{(32\pi)^2\text{OPT}}{C_d^2} + \epsilon > \frac{8}{\bar{c}}\text{OPT} + \epsilon,$$

we can further bound $|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*|_2^2$ above as

$$\begin{aligned}|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*|^2 <& |\boldsymbol{w}^t - \boldsymbol{w}^*|^2 + \frac{2\eta\kappa_d}{\sqrt{n}} - \eta\underline{c}\sin^2\theta_t + 8\eta^2\bar{c}^2\sin^2\theta_t\\ <& |\boldsymbol{w}^t - \boldsymbol{w}^*|^2 - \eta\underline{c}\sin^2\theta_t + 8\eta^2\bar{c}^2\sin^2\theta_t.\end{aligned} \tag{19}$$

Recalling that

$$|\boldsymbol{w}^t - \boldsymbol{w}^*|^2 = 2 - 2\cos\theta_t = 4\sin^2(\theta_t/2),$$

and for any $\theta_t \leqslant \delta = \frac{\pi}{2}$, we have $\sqrt{2}\sin(\theta_t/2) \leqslant \sin\theta_t \leqslant 2\sin(\theta_t/2)$, hence

$$|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*|^2 < \left(1 - \frac{\eta\underline{c}}{2} + 8\eta^2\bar{c}^2\right)|\boldsymbol{w}^t - \boldsymbol{w}^*|^2.$$

Noting that $\bar{c} < \frac{C_d^2}{4}$, $\underline{c} > \frac{C_d^2}{16\pi}$ and choosing $\eta = \frac{1}{32\pi C_d^2}$ yields

$$\begin{aligned}4\sin^2(\theta_{t+1}/2) &= |\boldsymbol{w}^{t+1} - \boldsymbol{w}^*|^2\\ &< \left(1 - \frac{1}{2048\pi^2}\right)|\boldsymbol{w}^t - \boldsymbol{w}^*|^2\\ &= \left(1 - \frac{1}{2048\pi^2}\right)(4\sin^2(\theta_t/2)).\end{aligned} \tag{20}$$

This shows that $\theta_{t+1} \leqslant \theta_t$, hence completing the inductive argument. Furthermore, (20) implies that after at most $T = O(\log(1/\epsilon))$ iterations, it must hold that $\sin\theta_T \leqslant \frac{32\pi\sqrt{\text{OPT}}}{C_d} + \sqrt{\epsilon}$. Although (20) only holds when $\sin\theta_T \geqslant \frac{32\pi\sqrt{\text{OPT}}}{C_d} + \sqrt{\epsilon}$, we can further show that if after some iterations t^* we have $\sin\theta_{t^*} \leqslant \frac{32\pi\sqrt{\text{OPT}}}{C_d} + \sqrt{\epsilon}$, then $\sin\theta_{t^*+1}$ is still of order $\sqrt{\text{OPT}} + \sqrt{\epsilon}$. Concretely, if there exists some step $t^* \leqslant T$ such that $\sin\theta_{t^*} \leqslant \frac{32\pi\sqrt{\text{OPT}}}{C_d} + \sqrt{\epsilon}$, then at step $t^* + 1$ it must hold (by (19)):

$$\sin\theta_{t^*+1} \leqslant \sqrt{2 + 8\eta^2\bar{c}^2}\sin\theta_{t^*} \leqslant 2\sin\theta_{t^*} \leqslant 2\left(\frac{32\pi\sqrt{\text{OPT}}}{C_d} + \sqrt{\epsilon}\right)$$

In other words, for all steps $t^* \leqslant t \leqslant T$, it holds that $\sin\theta_t \leqslant 2\left(\frac{32\pi\sqrt{\mathrm{OPT}}}{C_d} + \sqrt{\epsilon}\right)$. Recall that by Young's inequality, we have

$$\begin{aligned}
\mathcal{L}(\boldsymbol{w}^t) \leqslant & 2\mathrm{OPT} + 2\mathcal{L}^*(\boldsymbol{w}^t) \\
= & 2\mathrm{OPT} + \frac{8}{d+2}\left(1 - \frac{1}{\pi}\sin\theta_t\cos\theta_t - (1-\frac{\theta_t}{\pi})\cos^2\theta_t\right) \\
& + \frac{2}{(d+2)(d+4)}\left(3 - \frac{3}{\pi}\sin\theta_t\cos\theta_t - (1-\frac{\theta_t}{\pi})(1+2\cos^2\theta_t)\right) \\
\leqslant & 2\mathrm{OPT} + \frac{8}{d+2}\left(\sin^2\theta_t + \frac{1}{\pi}\cos\theta_t(\theta_t\cos\theta_t - \sin\theta_t)\right) \\
& + \frac{2}{(d+2)(d+4)}\left(3\sin^2\theta_t + \frac{3}{\pi}\cos\theta_t(\theta_t\cos\theta_t - \sin\theta_t)\right) \\
\leqslant & 2\mathrm{OPT} + \frac{8}{d+2}\sin^2\theta_t + \frac{6}{(d+2)(d+4)}\sin^2\theta_t \\
\leqslant & \left(8\left(\frac{8}{d+2} + \frac{6}{(d+2)(d+4)}\right)\frac{(32\pi)^2}{C_d^2} + 2\right)\mathrm{OPT} + 8\left(\frac{8}{d+2} + \frac{6}{(d+2)(d+4)}\right)\epsilon.
\end{aligned}$$

Since $\frac{8}{d+2} + \frac{6}{(d+2)(d+4)} < \frac{C_d^2}{4}$, we have

$$\mathcal{L}(\boldsymbol{w}^t) < (2048\pi^2 + 2)\mathrm{OPT} + 2C_d^2\epsilon \lesssim \mathrm{OPT} + 2C_d^2\epsilon.$$

Thus, in summary, choosing $T = O(\log(1/\epsilon))$, we get that with probability at least $1-\delta$, $\sin\theta_T \lesssim \frac{\sqrt{\mathrm{OPT}}}{C_d} + \sqrt{\epsilon}$. Also, Riemannian GD (Algorithm 1) outputs $\boldsymbol{w}^T$ such that with probability at least $1-\delta$, $\mathcal{L}(\boldsymbol{w}^T) = O(\mathrm{OPT}) + 2C_d^2\epsilon$, with sample size $n = \Theta\left(\frac{C_1 d}{C_d^2\epsilon}\right)$. □

B Proof of Proposition 2.3

In this section, we present the proof of Proposition 2.3.

Proof. Applying similar arguments as in Cho & Saul (2009), one can obtain that

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}[\sigma(\boldsymbol{w}^*\cdot\boldsymbol{x})\sigma(\boldsymbol{w}\cdot\boldsymbol{x})] &= \frac{1}{2\pi(d+2)}(\sin\theta + (\pi-\theta)\cos\theta), \\
\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}[\sigma^2(\boldsymbol{w}^*\cdot\boldsymbol{x})\sigma^2(\boldsymbol{w}\cdot\boldsymbol{x})] &= \frac{1}{2\pi(d+2)(d+4)}(3\sin\theta\cos\theta + (\pi-\theta)(1+2\cos^2\theta)).
\end{aligned}$$

where $\theta := \angle(\boldsymbol{w}, \boldsymbol{w}^*)$. Hence, we have

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}[|2\sigma(\boldsymbol{w}\cdot\boldsymbol{x})\boldsymbol{w}|^2] = \frac{2}{d+2}, \quad \mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}[|\sigma^2(\boldsymbol{w}\cdot\boldsymbol{x})|^2] = \frac{3}{2(d+2)(d+4)}.$$

Therefore, $\min_{\boldsymbol{w}\in\mathcal{S}^{d-1}}\mathcal{E}(\boldsymbol{w})$ is equivalent to $\max_{\boldsymbol{w}\in\mathcal{S}^{d-1}}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\Omega}[\sigma^2(\boldsymbol{w}^*\cdot\boldsymbol{x})\sigma^2(\boldsymbol{w}\cdot\boldsymbol{x})]$, which is achieved when $\theta = 0$, i.e., $\boldsymbol{w} = \boldsymbol{w}^*$. □

C Proof of Lemma 2.4 and Theorem 2.5

In this section, we present the proof of Lemma 2.4 and Theorem 2.5. For completeness of the section, we include Lemma 2.4 and Theorem 2.5 again.

Lemma C.1. *For any fixed $\boldsymbol{w}$, the minimizer $\boldsymbol{a}^*$ of the regularized loss function (9) has a closed form $\boldsymbol{a}^* = (\boldsymbol{K}_1 + \boldsymbol{K}_2 + \lambda\boldsymbol{I}_2)^{-1}\boldsymbol{K}_*$ and $\mathcal{L}_\lambda(\boldsymbol{w}) = -\boldsymbol{K}_*^\top(\boldsymbol{K}_1 + \boldsymbol{K}_2 + \lambda\boldsymbol{I}_2)^{-1}\boldsymbol{K}_*$.*

Proof. Recall that

$$\mathcal{L}_\lambda(\boldsymbol{w}, \boldsymbol{a}) = \boldsymbol{a}^\top\boldsymbol{K}_1\boldsymbol{a} + \boldsymbol{a}^\top\boldsymbol{K}_2\boldsymbol{a} - 2\boldsymbol{a}^\top\boldsymbol{K}_* + \lambda\boldsymbol{a}^\top\boldsymbol{a}.$$

To find the minimizer $\boldsymbol{a}^*$, we take gradient of $\mathcal{L}_\lambda(\boldsymbol{w}, \boldsymbol{a})$ with respect to $\boldsymbol{a}$ and let the gradient be zero. The gradient of $\mathcal{L}_\lambda(\boldsymbol{w}, \boldsymbol{a})$ with respect to $\boldsymbol{a}$ is

$$\nabla_{\boldsymbol{a}} L_\lambda(\boldsymbol{w}, \boldsymbol{a}) = 2\boldsymbol{K}_1\boldsymbol{a} + 2\boldsymbol{K}_2\boldsymbol{a} - 2\boldsymbol{K}_* + 2\lambda \boldsymbol{I}_2 \boldsymbol{a},$$

with $\boldsymbol{I}_2$ denoting the 2×2 identity matrix. By letting $\nabla_{\boldsymbol{a}} L_\lambda(\boldsymbol{w}, \boldsymbol{a}) = 0$, we have

$$\boldsymbol{K}_1\boldsymbol{a} + \boldsymbol{K}_2\boldsymbol{a} = \boldsymbol{K}_* - \lambda \boldsymbol{I}_2 \boldsymbol{a},$$

and we solve for $\boldsymbol{a}^*$

$$\boldsymbol{a}^* = (\boldsymbol{K}_1 + \boldsymbol{K}_2 + \lambda \boldsymbol{I}_2)^{-1}\boldsymbol{K}_*.$$

Plugging $\boldsymbol{a}^*$ into the regularized loss function, we can get

$$\begin{aligned}
\mathcal{L}_\lambda(\boldsymbol{w}) &= \mathcal{L}_\lambda(\boldsymbol{w}, \boldsymbol{a}^*) \\
&= (\boldsymbol{a}^*)^\top \left(\boldsymbol{K}_* - \lambda \boldsymbol{I}_2 \boldsymbol{a}^*\right) - 2(\boldsymbol{a}^*)^\top \boldsymbol{K}_* + \lambda (\boldsymbol{a}^*)^\top \boldsymbol{a}^* \\
&= -(\boldsymbol{a}^*)^\top \boldsymbol{K}_* \\
&= -\boldsymbol{K}_*^\top (\boldsymbol{K}_1 + \boldsymbol{K}_2 + \lambda \boldsymbol{I}_2)^{-1} \boldsymbol{K}_*
\end{aligned}$$

□

Since by assumption $\boldsymbol{w}_1$ is aligned with $\boldsymbol{w}^*$, then the loss function $\mathcal{L}_\lambda(\boldsymbol{w})$ is equivalent to the following loss function $\mathcal{L}_\xi(\theta)$ by plugging in the definition of $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$, where $\theta := \angle(\boldsymbol{w}^*, \boldsymbol{w}_2)$.

$$\mathcal{L}_\xi(\theta) = -\frac{c^2}{16\pi^2(d+4)^2} \begin{bmatrix} h_2(0) \\ h_2(\theta) \end{bmatrix}^\top \begin{bmatrix} \frac{c}{\pi}h_1(0) + \frac{c}{4\pi(d+4)}h_2(0) + \xi c & \frac{c}{\pi}h_1(\theta) + \frac{c}{4\pi(d+4)}h_2(\theta) \\ \frac{c}{\pi}h_1(\theta) + \frac{c}{4\pi(d+4)}h_2(\theta) & \frac{c}{\pi}h_1(0) + \frac{c}{4\pi(d+4)}h_2(0) + \xi c \end{bmatrix}^{-1} \begin{bmatrix} h_2(0) \\ h_2(\theta) \end{bmatrix},$$

where $c = \frac{2}{d+2}$ and $\lambda = \xi c$ (ξ is a constant independent of d). When $d \to +\infty$, we have that

$$\mathcal{L}_\xi(\theta) \approx -\frac{c}{16\pi^2(d+4)^2} \frac{(\frac{1}{\pi}h_1(0) + \xi)(h_2^2(0) + h_2^2(\theta)) - 2h_2(0)h_2(\theta)\frac{1}{\pi}h_1(\theta)}{(\frac{1}{\pi}h_1(0) + \xi)^2 - (\frac{1}{\pi}h_1(\theta))^2}.$$

Noting that $h_1(0) = \pi$ and $h_2(0) = 3\pi$, we can get

$$\begin{aligned}
\mathcal{L}_\xi(\theta) &\approx -\frac{c}{16\pi^2(d+4)^2} \frac{(1+\xi)(9\pi^2 + h_2^2(\theta)) - 6h_1(\theta)h_2(\theta)}{(1+\xi)^2 - \frac{1}{\pi^2}h_1^2(\theta)} \\
&= -\frac{c}{16(d+4)^2} \frac{(9+9\xi) + \frac{1}{\pi^2}[(1+\xi)h_2^2(\theta) - 6h_1(\theta)h_2(\theta)]}{(1+\xi)^2 - \frac{1}{\pi^2}h_1^2(\theta)}.
\end{aligned}$$

Recall that we define

$$\begin{aligned}
\widetilde{\mathcal{L}}_\xi(\theta) &:= \lim_{d \to +\infty} \frac{16(d+4)^2}{c} \mathcal{L}_\xi(\theta) \\
&= -\frac{(9+9\xi) + \frac{1}{\pi^2}[(1+\xi)h_2^2(\theta) - 6h_1(\theta)h_2(\theta)]}{(1+\xi)^2 - \frac{1}{\pi^2}h_1^2(\theta)}.
\end{aligned} \tag{21}$$

Then we proceed to show the main result of $\widetilde{\mathcal{L}}_\xi(\theta)$.

Theorem C.2. *Consider the minimization of the limiting loss function $\widetilde{\mathcal{L}}_\xi$ defined by* (10).

1. *When $\xi \geqslant \frac{1}{2}$, $\widetilde{\mathcal{L}}_\xi(\theta)$ has a unique global minimizer at $\theta = 0$ for θ on $[0, \frac{5\pi}{6}]$.*
2. *When $\xi \leqslant \xi_0$ with some $\xi_0 < 1/2$, besides the local minimizer $\theta = 0$, there exists at least one additional local minimizer of $\widetilde{\mathcal{L}}_\xi(\theta)$ in the interval $(\frac{\pi}{4}, \frac{\pi}{2})$.*

Proof. Recall that

$$\begin{aligned}
\widetilde{\mathcal{L}}_\xi(\theta) &= -\frac{(9+9\xi) + \frac{1}{\pi^2}[(1+\xi)h_2^2(\theta) - 6h_1(\theta)h_2(\theta)]}{(1+\xi)^2 - \frac{1}{\pi^2}h_1^2(\theta)} \\
&= -\frac{9(1+\xi) + \frac{1}{\pi^2}[(1+\xi)B(\theta) - 6C(\theta)]}{(1+\xi)^2 - \frac{1}{\pi^2}A(\theta)},
\end{aligned}$$

where

$$A(\theta) := h_1^2(\theta), \quad B(\theta) := h_2^2(\theta), \quad C(\theta) := h_1(\theta)h_2(\theta).$$

We take the derivative of $\widetilde{\mathcal{L}}_\xi(\theta)$ with respect to θ to get

$$\mathcal{L}_\xi'(\theta) = -\frac{1}{\pi^2}\frac{\Gamma_\xi(\theta)}{[(1+\xi)^2 - \frac{1}{\pi^2}A(\theta)]^2},$$

where

$$\Gamma_\xi(\theta) := (1+\xi)^3 B'(\theta) - 6(1+\xi)^2 C'(\theta) + 9(1+\xi)A'(\theta) + \frac{1+\xi}{\pi^2}(A'(\theta)B(\theta) - A(\theta)B'(\theta)) \\ + \frac{6}{\pi^2}(A(\theta)C'(\theta) - A'(\theta)C(\theta)).$$

Recall that

$$\begin{aligned} h_1(\theta) &= \sin\theta\cos\theta + (\pi-\theta)\cos^2\theta, \\ h_1'(\theta) &= -\sin^2\theta - 2(\pi-\theta)\sin\theta\cos\theta, \\ h_2(\theta) &= 3\sin\theta\cos\theta + (\pi-\theta)(1+2\cos^2\theta), \\ h_2'(\theta) &= -4\sin^2\theta - 4(\pi-\theta)\sin\theta\cos\theta. \end{aligned}$$

By calculation, we have

$$\begin{aligned} A(\theta) &= \sin^2\theta\cos^2\theta + 2(\pi-\theta)\sin\theta\cos^3\theta + (\pi-\theta)^2\cos^4\theta \\ B(\theta) &= 9\sin^2\theta\cos^2\theta + 6(\pi-\theta)\sin\theta\cos\theta(1+2\cos^2\theta) + (\pi-\theta)^2(1+2\cos^2\theta)^2 \\ C(\theta) &= 3\sin^2\theta\cos^2\theta + (\pi-\theta)\sin\theta\cos\theta(1+5\cos^2\theta) + (\pi-\theta)^2(\cos^2\theta + 2\cos^4\theta) \end{aligned}$$

$$\begin{aligned} A'(\theta) &= -2\sin^3\theta\cos\theta - 6(\pi-\theta)\sin^2\theta\cos^2\theta - 4(\pi-\theta)^2\sin\theta\cos^3\theta \\ B'(\theta) &= -24\sin^3\theta\cos\theta - 8(\pi-\theta)\sin^2\theta(1+5\cos^2\theta) - 8(\pi-\theta)^2\sin\theta\cos\theta(1+2\cos^2\theta) \\ C'(\theta) &= -7\sin^3\theta\cos\theta - (\pi-\theta)\sin^2\theta(1+16\cos^2\theta) - 2(\pi-\theta)^2\sin\theta\cos\theta(1+4\cos^2\theta) \\ A'(\theta)B(\theta) - A(\theta)B'(\theta) &= 6\sin^5\theta\cos^3\theta + 2(\pi-\theta)\sin^4\theta\cos^2\theta(-2+5\cos^2\theta) \\ &\quad + 2(\pi-\theta)^2\sin^3\theta\cos\theta(-1-10\cos^2\theta+2\cos^4\theta) \\ &\quad + 2(\pi-\theta)^3\sin^2\theta\cos^2\theta(-3-12\cos^2\theta) \\ &\quad + 4(\pi-\theta)^4\sin\theta\cos^3\theta(-1-2\cos^2\theta) \\ A(\theta)C'(\theta) - A'(\theta)C(\theta) &= -\sin^5\theta\cos^3\theta + (\pi-\theta)\sin^4\theta\cos^2\theta(1-2\cos^2\theta) \\ &\quad + (\pi-\theta)^2\sin^3\theta\cos^3\theta(4-\cos^2\theta) \\ &\quad + 5(\pi-\theta)^3\sin^2\theta\cos^4\theta \\ &\quad + 2(\pi-\theta)^4\sin\theta\cos^5\theta \end{aligned}$$

Hence, by letting $\bar{\xi} := 1+\xi$, we have

$$\begin{aligned} \Gamma_{\bar{\xi}}(\theta) = (-24\bar{\xi}^3 + 42\bar{\xi}^2 - 18\bar{\xi})\sin^3\theta\cos\theta &+ (\pi-\theta)\sin^2\theta\left(6\bar{\xi}^2 - 8\bar{\xi}^3 + (-40\bar{\xi}^3 + 96\bar{\xi}^2 - 54\bar{\xi})\cos^2\theta\right) \\ &+ (\pi-\theta)^2\sin\theta\cos\theta\left(12\bar{\xi}^2 - 8\bar{\xi}^3 + (-16\bar{\xi}^3 + 48\bar{\xi}^2 - 36\bar{\xi})\cos^2\theta\right) \\ &+ \frac{1}{\pi^2}\Phi_{\bar{\xi}}(\theta), \end{aligned}$$

where

$$\begin{aligned} \Phi_{\bar{\xi}}(\theta) := 6(\bar{\xi}-1)\sin^5\theta\cos^3\theta &+ 2(\pi-\theta)\sin^4\theta\cos^2\theta(-2\bar{\xi} + 3 + (5\bar{\xi}-6)\cos^2\theta) \\ &+ 2(\pi-\theta)^2\sin^3\theta\cos\theta(-\bar{\xi} + (12-10\bar{\xi})\cos^2\theta + (2\bar{\xi}-3)\cos^4\theta) \\ &+ 2(\pi-\theta)^3\sin^2\theta\cos^2\theta(-3\bar{\xi} + (15-12\bar{\xi})\cos^2\theta) \\ &+ 4(\pi-\theta)^4\sin\theta\cos^3\theta(-\bar{\xi} + (3-2\bar{\xi})\cos^2\theta). \end{aligned}$$

We also note that $\Gamma_\xi(\theta)$ can be written as

$$\Gamma_\xi(\theta) = \gamma_0(\theta) + \xi\gamma_1(\theta) + \xi^2\gamma_2(\theta) + \xi^3\gamma_3(\theta),$$

where

$$\begin{aligned}\gamma_0(\theta) :=& -2(\pi-\theta)\sin^4\theta + 4(\pi-\theta)^2\sin^3\theta\cos\theta\\ &+\frac{1}{\pi^2}\left[2(\pi-\theta)\sin^6\theta\cos^2\theta - 2(\pi-\theta)^2\sin^7\theta\cos\theta - 6(\pi-\theta)^3\sin^4\theta\cos^2\theta - 4(\pi-\theta)^4\sin^3\theta\cos^3\theta\right],\end{aligned}$$

$$\gamma_1(\theta) := -6\sin^3\theta\cos\theta + (\pi-\theta)\sin^2\theta\left(-12+18\cos^2\theta\right) + 12(\pi-\theta)^2\sin\theta\cos^3\theta + \frac{1}{\pi^2}\tilde{\gamma}_1(\theta),$$

where

$$\begin{aligned}\tilde{\gamma}_1(\theta) := 6\sin^5\theta\cos^3\theta &+ 2(\pi-\theta)\sin^4\theta\cos^2\theta(-2+5\cos^2\theta)\\ &+ 2(\pi-\theta)^2\sin^3\theta\cos\theta(-1-10\cos^2\theta+2\cos^4\theta)\\ &- 6(\pi-\theta)^3\sin^2\theta\cos^2\theta(1+4\cos^2\theta)\\ &- 4(\pi-\theta)^4\sin\theta\cos^3\theta(1+2\cos^2\theta),\end{aligned}$$

$$\begin{aligned}\gamma_2(\theta) &:= -30\sin^3\theta\cos\theta - 6(\pi-\theta)\sin^2\theta\left(3+4\cos^2\theta\right) - 12(\pi-\theta)^2\sin\theta\cos\theta,\\ \gamma_3(\theta) &:= -24\sin^3\theta\cos\theta - 8(\pi-\theta)\sin^2\theta\left(1+5\cos^2\theta\right) - 8(\pi-\theta)^2\sin\theta\cos\theta(1+2\cos^2\theta).\end{aligned}$$

Claim 1: For any $\xi > 0$, $\widetilde{\mathcal{L}}_\xi(\theta)$ is increasing on $(\frac{\pi}{2}, \frac{5\pi}{6}]$.
We want to show that $\Gamma_\xi(\theta) < 0$ for $\theta \in (\frac{\pi}{2}, \frac{5\pi}{6}]$. To show that, we will show $\gamma_0(\theta) < 0$, $\gamma_1(\theta) < 0$, $\gamma_2(\theta) < 0$ and $\gamma_3(\theta) < 0$ respectively.
To show $\gamma_0(\theta) < 0$ on $(\frac{\pi}{2}, \pi)$, we observe that

$$\begin{aligned}\frac{2}{\pi^2}(\pi-\theta)\sin^6\theta\cos^2\theta &< \frac{2}{\pi^2}(\pi-\theta)\sin^4\theta,\\ -\frac{2}{\pi^2}(\pi-\theta)^2\sin^7\theta\cos\theta &< -\frac{2}{\pi^2}(\pi-\theta)^2\sin^3\theta\cos\theta,\\ -\frac{4}{\pi^2}(\pi-\theta)^4\sin^3\theta\cos^3\theta &< -\frac{4}{\pi^2}(\pi-\theta)^2\frac{\pi^2}{4}\sin^3\theta\cos\theta = -(\pi-\theta)^2\sin^3\theta\cos\theta,\end{aligned}$$

by using $\sin^2\theta < 1$, $\cos^2\theta < 1$ and $(\pi-\theta)^2 < (\frac{\pi}{2})^2$. Then we can get

$$\gamma_0(\theta) < -\left(2-\frac{2}{\pi^2}\right)(\pi-\theta)\sin^4\theta + \left(3-\frac{2}{\pi^2}\right)(\pi-\theta)^2\sin^3\theta\cos\theta < 0.$$

Next, we show that $\gamma_1(\theta) < 0$ on $(\frac{\pi}{2}, \frac{5\pi}{6}]$. First, we note that

$$\begin{aligned}6\sin^5\theta\cos^3\theta \leqslant 0 \text{ (the first term in } \tilde{\gamma}_1(\theta)),\\ 2(\pi-\theta)\sin^4\theta\cos^2\theta(-2+5cos^2\theta) \leqslant \frac{7}{2}(\pi-\theta)\sin^2\theta\cos^2\theta \text{ (the second term in } \tilde{\gamma}_1(\theta))\end{aligned}$$

by using $\cos^2\theta \leqslant \frac{3}{4}$ and $\sin^2\theta \leqslant 1$. Then we focus on the sum of the forth term and the fifth term in $\tilde{\gamma}_1(\theta)$ and let $t := \pi - \theta$. For $\theta \in (\frac{\pi}{2}, \pi)$, $t \in (0, \frac{\pi}{2})$, then we can get

$$\begin{aligned}H_1(t) &:= -6t^3\sin^2 t\cos^2 t(1+4\cos^2 t) + 4t^4\sin t\cos^3 t(1+2\cos^2 t)\\ &= 2t^3\sin t\cos^2 t\left(-3\sin t(1+4\cos^2 t) + 2t\cos t(1+2\cos^2 t)\right)\\ &\leqslant 2t^3\sin t\cos^2 t\left(-3\sin t(1+4\cos^2 t) + 2\sin t(1+2\cos^2 t)\right)\\ &= 2t^3\sin t\cos^2 t\left(-\sin t - 8\sin t\cos^2 t\right)\\ &< 0,\end{aligned}$$

where the first inequality comes from the standard inequality $\tan x \geqslant x$ on $(0, \frac{\pi}{2})$. Combining the above inequalities, we can get on $(\frac{\pi}{2}, \frac{5\pi}{6}]$,

$$\begin{aligned}\gamma_1(\theta) <& -6\sin^3\theta\cos\theta - 12(\pi-\theta)\sin^2\theta + \left(18+\frac{7}{2\pi^2}\right)(\pi-\theta)\sin^2\theta\cos^2\theta + 12(\pi-\theta)^2\sin\theta\cos^3\theta\\ &+\frac{1}{\pi^2}\left[2(\pi-\theta)^2\sin^3\theta\cos\theta(-1-10\cos^2\theta+2\cos^4\theta)\right]\\ <& -6\sin^3\theta\cos\theta - 12(\pi-\theta)\sin^2\theta + 19(\pi-\theta)\sin^2\theta\cos^2\theta + 12(\pi-\theta)^2\sin\theta\cos^3\theta\\ &+\frac{1}{\pi^2}\left[2(\pi-\theta)^2\sin^3\theta\cos\theta(-1-10\cos^2\theta+2\cos^4\theta)\right].\end{aligned}$$

By letting $t := \pi - \theta$ and using properties $\sin\theta = \sin t > 0$ and $\cos\theta = -\cos t < 0$ when $\theta \in (\frac{\pi}{2}, \frac{5\pi}{6}]$, $t \in [\frac{\pi}{6}, \frac{\pi}{2})$. The above inequality can be written as

$$\gamma_1(\theta) < \sin t J_1(t),$$

where

$$\begin{aligned} J_1(t) :=& 6\sin^2 t\cos t - 12t\sin t + 19t\sin t\cos^2 t - 12t^2\cos^3 t \\ &+ \frac{1}{\pi^2}\left[2t^2\sin^2 t\cos t(1 + 10\cos^2 t - 2\cos^4 t)\right]. \end{aligned}$$

Since $\cos^2 t \leqslant \frac{3}{4}$ and $\sin^2 t \geqslant \frac{1}{4}$ on $[\frac{\pi}{6}, \frac{\pi}{2})$, we have that

$$\begin{aligned} J_1(t) \leqslant& 6\sin^2 t\cos t - 12t\sin t + 19t\sin t\cos^2 t - 12t^2\cos^3 t + \frac{1}{\pi^2}\left[17t^2\sin^2 t\cos t - 4t^2\sin^2 t\cos^5 t\right] \\ \leqslant& 6\sin^2 t\cos t - 12t\sin t + 19t\sin t\cos^2 t - 12t^2\cos^3 t + \frac{1}{\pi^2}\left[17t^2\sin^2 t\cos t - t^2\cos^5 t\right] \\ <& 6\sin^2 t\cos t - 12t\sin t + 19t\sin t\cos^2 t - 12t^2\cos^3 t + \frac{17}{\pi^2}t^2\sin^2 t\cos t \\ <& 6\sin^2 t\cos t - 12t\sin t + 19t\sin t\cos^2 t - 12t^2\cos^3 t + 2t^2\sin^2 t\cos t. \end{aligned}$$

We divide the interval $[\frac{\pi}{6}, \frac{\pi}{2})$ into two sub-intervals $[\frac{\pi}{6}, 1.4]$ and $[1.4, \frac{\pi}{2})$. On $[1.4, \frac{\pi}{2})$, by using inequalities $\sin t \in [\sin(1.4), 1)$, $\cos t \in (0, \cos(1.4)]$ and $t \in [1.4, \frac{\pi}{2})$, we can get

$$J_1(t) < 6\cos(1.4) - 12 \cdot 1.4\sin(1.4) + 19\frac{\pi}{2}\cos^2(1.4) + 2\frac{\pi^2}{4}\cos(1.4) \approx -13.83 < 0$$

On $[\frac{\pi}{6}, 1.4]$, we take the first derivative of $J_1(t)$ to get

$$\begin{aligned} J_1'(t) &= 19\sin t - 37\sin^3 t - 46t\cos t + 29t\cos^3 t + 40t^2\sin t\cos^2 t - 2t^2\sin^3 t \\ &\leqslant 19\sin t - 37\sin^3 t - \frac{97}{4}t\cos t + 40t^2\sin t\cos^2 t - 2t^2\sin^3 t \\ &= 19\sin t + 40t^2\sin t - 37\sin^3 t - 42t^2\sin^3 t - \frac{97}{4}t\cos t, \end{aligned}$$

where the inequality comes from the fact that $\cos t \leqslant \frac{\sqrt{3}}{2}$ on $[\frac{\pi}{6}, 1.4]$. By using following standard Taylor polynomial bounds

$$\sin t \geqslant t - \frac{t^3}{6}, \quad \sin t \leqslant t - \frac{t^3}{6} + \frac{t^5}{120}, \quad \cos t \geqslant 1 - \frac{t^2}{2},$$

we can get

$$J_1'(t) \leqslant P_1(t) := \frac{7}{36}t^{11} - \frac{719}{216}t^9 + \frac{73}{4}t^7 - \frac{3601}{120}t^5 + \frac{287}{24}t^3 - \frac{21}{4}t.$$

We can check that $P_1(t) < 0$ on $[\frac{\pi}{6}, 1.4]$, hence, $J_1'(t) < 0$ on $[\frac{\pi}{6}, 1.4]$, which implies that $J_1(t)$ is decreasing on $[\frac{\pi}{6}, 1.4]$ and hence $J_1(t) \leqslant J_1(\frac{\pi}{6}) < 0$ on $[\frac{\pi}{6}, 1.4]$. Then we can conclude that $\gamma_1(\theta) < 0$ on $(\frac{\pi}{2}, \frac{5\pi}{6}]$.

To show $\gamma_2(\theta) < 0$, we let $t := \pi - \theta$, then $t \in (0, \frac{\pi}{2})$. We use properties $\sin\theta = \sin t > 0$ and $\cos\theta = -\cos t < 0$, then we have

$$\gamma_2(\theta) = -6\sin t J_2(t),$$

where $J_2(t) := t\sin t(3 + 4\cos^2 t) - \cos t(2t^2 + 5\sin^2 t)$. Next, we just need to show that $J_2(t) > 0$. Take the first derivative and second derivative of $J_2(t)$, we can get

$$\begin{aligned} J_2'(t) &= 12\sin t + 3t\cos t - 4\sin^3 t + 2t^2\sin t - 12t\sin^2 t\cos t - 15\sin t\cos^2 t, \\ J_2''(t) &= -23t\sin t + 36t\sin^3 t + 21\sin^2 t\cos t + 2t^2\cos t. \end{aligned}$$

By using following standard Taylor polynomial bounds on $[0, \frac{\pi}{2}]$ for $J_2''(t)$

$$\sin t \geqslant t - \frac{t^3}{6}, \quad \sin t \leqslant t - \frac{t^3}{6} + \frac{t^5}{120}, \quad \cos t \geqslant 1 - \frac{t^2}{2},$$

we can show that

$$J_2''(t) \geqslant P_2(t) := t^4 R_2(t^2),$$

where $R_2(x) = -\frac{1}{6}x^3 + \frac{65}{24}x^2 - \frac{1693}{120}x + \frac{64}{3}$. By analyzing $R_2'(x)$, it is easy to check that $R_2(x)$ is decreasing on $[0, (\frac{\pi}{2})^2]$. Hence, $R_2(t^2) \geqslant R_2((\frac{\pi}{2})^2) > 0$, which means $J_2''(t) > 0$ on $[0, \frac{\pi}{2}]$. Therefore, $J_2'(t) > J_2'(0) = 0$ on $(0, \frac{\pi}{2})$, which is equivalent to $J_2(t) > J_2(0) = 0$ on $(0, \frac{\pi}{2})$. At last, we can conclude that $\gamma_2(\theta) < 0$ on $(\frac{\pi}{2}, \pi)$.

To show $g_3(\theta) < 0$, similarly, we let $t := \pi - \theta$, then $t \in (0, \frac{\pi}{2})$. We use properties $\sin\theta = \sin t > 0$ and $\cos\theta = -\cos t < 0$, then we have

$$\gamma_3(\theta) = 8\sin t J_3(t),$$

where $J_3(t) := 3\sin^2 t\cos t - t\sin t(1 + 5\cos^2 t) + t^2\cos t(1 + 2\cos^2 t)$. Take the first derivative of $J_3(t)$, we can get

$$J_3'(t) = \sin t H_3(t),$$

where

$$H_3(t) := (6t^2 - 4)\sin^2 t - 7t^2 + 11t\sin t\cos t.$$

By using following standard Taylor polynomial bounds on $[0, \frac{\pi}{2}]$ for $H_3(t)$

$$\sin t \geqslant t - \frac{t^3}{6}, \quad \sin t \leqslant t - \frac{t^3}{6} + \frac{t^5}{120}, \quad \cos t \geqslant 1 - \frac{t^2}{2},$$

we can show that

$$H_3(t) \leqslant P_3(t) := t^6 R_3(t^2),$$

where $R_3(x) = \frac{1}{2400}x^3 - \frac{37}{2880}x^2 + \frac{13}{90}x - \frac{87}{135}$. By analyzing $R_3'(x)$, it is easy to check that $R_3(x)$ is increasing on $[0, (\frac{\pi}{2})^2]$. Hence, $R_3(t^2) \leqslant R_3((\frac{\pi}{2})^2) < 0$, which means that $H_3(t) < 0$ on $(0, \frac{\pi}{2})$ and further implies that $J_3(t)$ is decreasing on $(0, \frac{\pi}{2})$. Therefore, $J_3(t) < J_3(0) = 0$ on $(0, \frac{\pi}{2})$, which further implies that $\gamma_3(\theta) < 0$ on $(\frac{\pi}{2}, \pi)$.

In summary, we have shown that $\Gamma_\xi(\theta) < 0$ on $(\frac{\pi}{2}, \frac{5\pi}{6}]$ for any $\xi > 0$, which means that $\widetilde{\mathcal{L}}_\xi(\theta)$ is increasing on $(\frac{\pi}{2}, \frac{5\pi}{6}]$ for any $\xi > 0$.

Claim 2: $\theta = 0$ is a local minimizer for any $\xi > 0$.
First, we observe that $\widetilde{\mathcal{L}}_\xi'(0) = 0$. Then We take the second derivative of $\widetilde{\mathcal{L}}_\xi(\theta)$ with respect to θ to get

$$\widetilde{\mathcal{L}}_\xi''(\theta) = -\frac{1}{\pi^2}\frac{\Gamma_\xi'(\theta)[(1+\xi)^2 - \frac{1}{\pi^2}A(\theta)]^2 + \Gamma_\xi(\theta)\frac{2}{\pi^2}[(1+\xi)^2 - \frac{1}{\pi^2}A(\theta)]A'(\theta)}{[(1+\xi)^2 - \frac{1}{\pi^2}A(\theta)]^4},$$

where

$$\begin{aligned}\Gamma_\xi(\theta) &= (1+\xi)^3 B'(\theta) - 6(1+\xi)^2 C'(\theta) + 9(1+\xi)A'(\theta) + \frac{1+\xi}{\pi^2}(A'(\theta)B(\theta) - A(\theta)B'(\theta)) \\ &\quad + \frac{6}{\pi^2}(A(\theta)C'(\theta) - A'(\theta)C(\theta)), \\ \Gamma_\xi'(\theta) &= (1+\xi)^3 B''(\theta) - 6(1+\xi)^2 C''(\theta) + 9(1+\xi)A''(\theta) + \frac{1+\xi}{\pi^2}(A''(\theta)B(\theta) - A(\theta)B''(\theta)) \\ &\quad + \frac{6}{\pi^2}(A(\theta)C''(\theta) - A''(\theta)C(\theta)).\end{aligned}$$

By calculation, we have

$$\begin{aligned}A''(\theta) &= 2\sin^4\theta - 4(\pi-\theta)\sin\theta\cos\theta(4\cos^2\theta - 3) - 4(\pi-\theta)^2\cos^2\theta(4\cos^2\theta - 3) \\ B''(\theta) &= 32\sin^4\theta - 24\sin^2\theta\cos^2\theta - 16(\pi-\theta)\sin\theta\cos\theta(8\cos^2\theta - 5) - 8(\pi-\theta)^2(-1 - 4\cos^2\theta + 8\cos^4\theta) \\ C''(\theta) &= 8\sin^4\theta - 4\sin^2\theta\cos^2\theta - 2(\pi-\theta)\sin\theta\cos\theta(-17 + 24\cos^2\theta) - 2(\pi-\theta)^2(-1 - 10\cos^2\theta + 16\cos^4\theta)\end{aligned}$$

and

$$\widetilde{\mathcal{L}}_\xi''(0) = -\frac{-24\xi^3 - 12\xi^2 - 48\xi - 48}{(\xi^2 + 2\xi)^2} > 0,$$

for any $\xi > 0$. Therefore, $\theta = 0$ is a local minimizer for any $\xi > 0$.

Claim 3: When $\xi \geqslant \frac{1}{2}$, $\widetilde{\mathcal{L}}_\xi(\theta)$ is increasing on $(0, \frac{\pi}{2})$.
Recall that with $\bar{\xi} := \xi + 1$,

$$\begin{aligned}\Gamma_{\bar{\xi}}(\theta) = (-24\bar{\xi}^3 + 42\bar{\xi}^2 - 18\bar{\xi})\sin^3\theta\cos\theta &+ (\pi-\theta)\sin^2\theta\left(6\bar{\xi}^2 - 8\bar{\xi}^3 + (-40\bar{\xi}^3 + 96\bar{\xi}^2 - 54\bar{\xi})\cos^2\theta\right) \\ &+ (\pi-\theta)^2\sin\theta\cos\theta\left(12\bar{\xi}^2 - 8\bar{\xi}^3 + (-16\bar{\xi}^3 + 48\bar{\xi}^2 - 36\bar{\xi})\cos^2\theta\right) \\ &+ \frac{1}{\pi^2}\Phi_{\bar{\xi}}(\theta),\end{aligned}$$

where

$$\begin{aligned}\Phi_{\bar{\xi}}(\theta) = 6(\bar{\xi} - 1)\sin^5\theta\cos^3\theta &+ 2(\pi-\theta)\sin^4\theta\cos^2\theta(-2\bar{\xi} + 3 + (5\bar{\xi} - 6)\cos^2\theta) \\ &+ 2(\pi-\theta)^2\sin^3\theta\cos\theta(-\bar{\xi} + (12 - 10\bar{\xi})\cos^2\theta + (2\bar{\xi} - 3)\cos^4\theta) \\ &+ 2(\pi-\theta)^3\sin^2\theta\cos^2\theta(-3\bar{\xi} + (15 - 12\bar{\xi})\cos^2\theta) \\ &+ 4(\pi-\theta)^4\sin\theta\cos^3\theta(-\bar{\xi} + (3 - 2\bar{\xi})\cos^2\theta).\end{aligned}$$

Note that when $\xi \geqslant \frac{1}{2}, \bar{\xi} \geqslant \frac{3}{2}$ and we have following inequality

$$(-24\bar{\xi}^3 + 42\bar{\xi}^2 - 18\bar{\xi}) \sin^3\theta \cos\theta + \frac{6}{\pi^2}(\bar{\xi} - 1) \sin^5\theta \cos^3\theta < (-24\bar{\xi}^3 + 42\bar{\xi}^2 - 12\bar{\xi} - 6) \sin^3\theta \cos\theta < 0,$$
$$(\pi - \theta)^2 \sin\theta \cos\theta \left(12\bar{\xi}^2 - 8\bar{\xi}^3 + (-16\bar{\xi}^3 + 48\bar{\xi}^2 - 36\bar{\xi}) \cos^2\theta\right) < 0$$

Also, for other terms in $\Phi_{\bar{\xi}}(\theta)$, we have that

$$2(\pi - \theta) \sin^4\theta \cos^2\theta(-2\bar{\xi} + 3 + (5\bar{\xi} - 6) \cos^2\theta) < 2(\pi - \theta) \sin^2\theta(3\bar{\xi} - 3),$$
$$2(\pi - \theta)^2 \sin^3\theta \cos\theta(-\bar{\xi} + (12 - 10\bar{\xi}) \cos^2\theta + (2\bar{\xi} - 3) \cos^4\theta) < 2(\pi - \theta)^2 \sin^3\theta \cos\theta(-\bar{\xi} + (9 - 8\bar{\xi}) \cos^2\theta) < 0,$$
$$2(\pi - \theta)^3 \sin^2\theta \cos^2\theta(-3\bar{\xi} + (15 - 12\bar{\xi}) \cos^2\theta) < 0,$$
$$4(\pi - \theta)^4 \sin\theta \cos^3\theta(-\bar{\xi} + (3 - 2\bar{\xi}) \cos^2\theta) < 0.$$

Then we combine the second term of $\Gamma_{\bar{\xi}}(\theta)$ and the first inequality above to get

$$(\pi - \theta) \sin^2\theta \left(6\bar{\xi}^2 - 8\bar{\xi}^3 + (-40\bar{\xi}^3 + 96\bar{\xi}^2 - 54\bar{\xi}) \cos^2\theta\right) + \frac{2}{\pi^2}(\pi - \theta) \sin^4\theta \cos^2\theta(-2\bar{\xi} + 3 + (5\bar{\xi} - 6) \cos^2\theta)$$
$$<(\pi - \theta) \sin^2\theta \left(6\bar{\xi}^2 - 8\bar{\xi}^3 + 6\bar{\xi} - 6 + (-40\bar{\xi}^3 + 96\bar{\xi}^2 - 54\bar{\xi}) \cos^2\theta\right) < 0$$

Therefore, we can get $\Gamma_{\bar{\xi}}(\theta) < 0$ when $\bar{\xi} \geqslant \frac{3}{2}$ and $\theta \in (0, \frac{\pi}{2})$. Hence, $\widetilde{\mathcal{L}}'_\xi(\theta)$ is positive, which implies that $\widetilde{\mathcal{L}}_\xi(\theta)$ is increasing on $(0, \frac{\pi}{2})$ when $\bar{\xi} \geqslant \frac{3}{2}$.

Claim 4: When $\xi \leqslant 0.13$, there is at least one additional local minimizer on $(\frac{\pi}{4}, \frac{\pi}{2})$.
We first note that

$$\Gamma_{\bar{\xi}}(\frac{\pi}{2}) = \frac{\pi}{2}(6\bar{\xi}^2 - 8\bar{\xi}^3) < 0 \Rightarrow \mathcal{L}'_\xi(\frac{\pi}{2}) > 0.$$

and

$$\gamma_0(\frac{\pi}{4}) \approx 1.7777, \quad \gamma_1(\frac{\pi}{4}) \approx -2.0681, \quad \gamma_2(\frac{\pi}{4}) \approx -76.1528, \quad \gamma_3(\frac{\pi}{4}) \approx -58.8614,$$
$$\Gamma_{\bar{\xi}}(\frac{\pi}{4}) = \gamma_0(\frac{\pi}{4}) + \xi\gamma_1(\frac{\pi}{4}) + \xi^2\gamma_2(\frac{\pi}{4}) + \xi^3\gamma_3(\frac{\pi}{4}).$$

By calculation, we can get when $\xi \leqslant 0.13$,

$$\Gamma_{\bar{\xi}}(\theta)(\frac{\pi}{4}) \geqslant \gamma_0(\frac{\pi}{4}) + 0.13 \cdot \gamma_1(\frac{\pi}{4}) + 0.13^2 \cdot \gamma_2(\frac{\pi}{4}) + 0.13^3 \cdot \gamma_3(\frac{\pi}{4}) > 0 \Rightarrow \mathcal{L}'_\xi(\frac{\pi}{4}) < 0.$$

Hence, $\mathcal{L}_\xi(\theta)$ has at least one local minimizer in the interval $(\frac{\pi}{4}, \frac{\pi}{2})$.

Finally, by combining above claims, we complete the proof of the theorem.

□

D Proof of Lemma 2.6 and Theorem 2.7

In this section, we present the proof of Lemma 2.6 and Theorem 2.7. For completeness of the section, we include Lemma 2.6 and Theorem 2.7 again.

Lemma D.1. *For any fixed $\boldsymbol{w}$, the minimizer $\boldsymbol{a}^*$ of the regularized loss function (11) has a closed form* $\boldsymbol{a}^* = \left(\boldsymbol{K}_1 + \boldsymbol{K}_2 + \begin{bmatrix} 0 & 0 \\ 0 & \lambda \end{bmatrix}\right)^{-1} \boldsymbol{K}_*$ *and* $\mathcal{L}_\lambda(\boldsymbol{w}) = -\boldsymbol{K}_*^\top \left(\boldsymbol{K}_1 + \boldsymbol{K}_2 + \begin{bmatrix} 0 & 0 \\ 0 & \lambda \end{bmatrix}\right)^{-1} \boldsymbol{K}_*$.

Proof. Recall that

$$\mathcal{L}_\lambda(\boldsymbol{w}, \boldsymbol{a}) = \boldsymbol{a}^\top \boldsymbol{K}_1 \boldsymbol{a} + \boldsymbol{a}^\top \boldsymbol{K}_2 \boldsymbol{a} - 2\boldsymbol{a}^\top \boldsymbol{K}_* + \lambda a_2^2.$$

To find the minimizer $\boldsymbol{a}^*$, we take the gradient of $\mathcal{L}_\lambda(\boldsymbol{w}, \boldsymbol{a})$ with respect to $\boldsymbol{a}$ to get

$$\nabla_a L_\lambda(\boldsymbol{w}, \boldsymbol{a}) = 2\boldsymbol{K}_1\boldsymbol{a} + 2\boldsymbol{K}_2\boldsymbol{a} - 2\boldsymbol{K}_* + 2\begin{bmatrix} 0 \\ \lambda a_2 \end{bmatrix}.$$

By letting $\nabla_a \mathcal{L}_\lambda(\boldsymbol{w}, \boldsymbol{a}) = 0$, we have

$$\boldsymbol{K}_1\boldsymbol{a} + \boldsymbol{K}_2\boldsymbol{a} = \boldsymbol{K}_* - \begin{bmatrix} 0 \\ \lambda a_2 \end{bmatrix},$$

and we solve for $\boldsymbol{a}^*$

$$\boldsymbol{a}^* = \left(\boldsymbol{K}_1 + \boldsymbol{K}_2 + \begin{bmatrix} 0 & 0 \\ 0 & \lambda \end{bmatrix}\right)^{-1} \boldsymbol{K}_*.$$

Plugging $\boldsymbol{a}^*$ into the regularized loss function, we can get

$$\begin{aligned} \mathcal{L}_\lambda(\boldsymbol{w}) &= \mathcal{L}_\lambda(\boldsymbol{w}, \boldsymbol{a}^*) \\ &= (\boldsymbol{a}^*)^\top \left(\boldsymbol{K}_* - \begin{bmatrix} 0 \\ \lambda a_2^* \end{bmatrix}\right) - 2(\boldsymbol{a}^*)^\top \boldsymbol{K}_* + \lambda (a_2^*)^2 \\ &= -(\boldsymbol{a}^*)^\top \boldsymbol{K}_* \\ &= -\boldsymbol{K}_*^\top \left(\boldsymbol{K}_1 + \boldsymbol{K}_2 + \begin{bmatrix} 0 & 0 \\ 0 & \lambda \end{bmatrix}\right)^{-1} \boldsymbol{K}_* \end{aligned} \tag{22}$$

□

Since by assumption $\boldsymbol{w}_1$ is aligned with $\boldsymbol{w}^*$, then the loss function $\mathcal{L}_\lambda(\boldsymbol{w})$ is equivalent to the following loss function $\mathcal{L}_\xi(\theta)$ by plugging in the definition of $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$, where $\theta := \angle(\boldsymbol{w}^*, \boldsymbol{w}_2)$.

$$\mathcal{L}_\xi(\theta) = -\frac{c^2}{16\pi^2(d+4)^2} \begin{bmatrix} h_2(0) \\ h_2(\theta) \end{bmatrix}^\top \begin{bmatrix} \frac{c}{\pi} h_1(0) + \frac{c}{4\pi(d+4)} h_2(0) & \frac{c}{\pi} h_1(\theta) + \frac{c}{4\pi(d+4)} h_2(\theta) \\ \frac{c}{\pi} h_1(\theta) + \frac{c}{4\pi(d+4)} h_2(\theta) & \frac{c}{\pi} h_1(0) + \frac{c}{4\pi(d+4)} h_2(0) + \xi c \end{bmatrix}^{-1} \begin{bmatrix} h_2(0) \\ h_2(\theta) \end{bmatrix},$$

where $c = \frac{2}{d+2}$ and $\lambda = \xi c$ (ξ is a constant independent of d). When $d \to +\infty$, we have that

$$\mathcal{L}_\xi(\theta) \approx -\frac{c}{16\pi^2(d+4)^2} \frac{h_2^2(0)(\frac{1}{\pi}h_1(0) + \xi) + h_2^2(\theta)\frac{1}{\pi}h_1(0) - 2h_2(0)h_2(\theta)\frac{1}{\pi}h_1(\theta)}{\frac{1}{\pi}h_1(0)(\frac{1}{\pi}h_1(0) + \xi) - (\frac{1}{\pi}h_1(\theta))^2}.$$

Noting that $h_1(0) = \pi$ and $h_2(0) = 3\pi$, we can get

$$\mathcal{L}_\xi(\theta) \approx -\frac{c}{16(d+4)^2} \frac{(9 + 9\xi) + \frac{1}{\pi^2}[h_2^2(\theta) - 6h_1(\theta)h_2(\theta)]}{1 + \xi - \frac{1}{\pi^2}h_1^2(\theta)}.$$

Recall that we define

$$\begin{aligned} \bar{\mathcal{L}}_\xi(\theta) &:= \lim_{d \to +\infty} \frac{16(d+4)^2}{c} \mathcal{L}_\xi(\theta) \\ &= -\frac{(9 + 9\xi) + \frac{1}{\pi^2}[h_2^2(\theta) - 6h_1(\theta)h_2(\theta)]}{1 + \xi - \frac{1}{\pi^2}h_1^2(\theta)}. \end{aligned}$$

Then we proceed to show the main result of $\bar{\mathcal{L}}_\xi(\theta)$.

Theorem D.2. *Consider the minimization of the limiting loss function $\bar{\mathcal{L}}_\xi$ defined by* (12)*, there exists a* ***unique*** *global minimizer $\theta^* \in (\frac{\pi}{3}, \frac{\pi}{2})$ for $\bar{\mathcal{L}}_\xi(\theta)$ for any $\xi > 0$.*

Proof. Recall that

$$\bar{\mathcal{L}}_\xi(\theta) = -\frac{(9 + 9\xi) + \frac{1}{\pi^2}[h_2^2(\theta) - 6h_1(\theta)h_2(\theta)]}{1 + \xi - \frac{1}{\pi^2}h_1^2(\theta)}.$$

We denote

$$\begin{aligned} E(\theta) :=& h_2^2(\theta) - 6h_1(\theta)h_2(\theta) = h_2(\theta)(h_2(\theta) - 6h_1(\theta)), \\ =& [3\sin\theta\cos\theta + (\pi - \theta)(1 + 2\cos^2\theta)][-3\sin\theta\cos\theta + (\pi - \theta)(1 - 4\cos^2\theta)] \\ F(\theta) :=& h_1^2(\theta) = [\sin\theta\cos\theta + (\pi - \theta)\cos^2\theta]^2. \end{aligned}$$

Then we take the derivative of $\bar{\mathcal{L}}_\xi(\theta)$ to get

$$\bar{\mathcal{L}}'_\xi(\theta) = -\frac{1}{\pi^2} \frac{(1 + \xi)(E'(\theta) + 9F'(\theta)) + \frac{1}{\pi^2}(E(\theta)F'(\theta) - E'(\theta)F(\theta))}{[1 + \xi - \frac{1}{\pi^2}F(\theta)]^2}.$$

Recall that

$$\begin{aligned}
h_1(\theta) &= \sin\theta\cos\theta + (\pi-\theta)\cos^2\theta, \\
h_1'(\theta) &= -\sin^2\theta - 2(\pi-\theta)\sin\theta\cos\theta, \\
h_2(\theta) &= 3\sin\theta\cos\theta + (\pi-\theta)(1+2\cos^2\theta), \\
h_2'(\theta) &= -4\sin^2\theta - 4(\pi-\theta)\sin\theta\cos\theta.
\end{aligned}$$

Then by calculation, we have the following equations.

$$\begin{aligned}
F'(\theta) &= -2\sin^3\theta\cos\theta - 6(\pi-\theta)\sin^2\theta\cos^2\theta - 4(\pi-\theta)^2\sin\theta\cos^3\theta, \\
E'(\theta) &= 18\sin^3\theta\cos\theta + 54(\pi-\theta)\sin^2\theta\cos^2\theta - 2(\pi-\theta)\sin^4\theta + 4(\pi-\theta)^2\sin\theta\cos\theta(1+8\cos^2\theta), \\
E'(\theta)+9F'(\theta) &= -2(\pi-\theta)\sin^4\theta + 4(\pi-\theta)^2\sin^3\theta\cos\theta, \\
F(\theta) &= h_1^2(\theta) = \sin^2\theta\cos^2\theta + 2(\pi-\theta)\sin\theta\cos^3\theta + (\pi-\theta)^2\cos^4\theta, \\
E(\theta) &= -9\sin^2\theta\cos^2\theta - 18(\pi-\theta)\sin\theta\cos^3\theta + (\pi-\theta)^2(1-2\cos^2\theta-8\cos^4\theta),
\end{aligned}$$

$$\begin{aligned}
&E(\theta)F'(\theta) \\
=&18\sin^5\theta\cos^3\theta + 90(\pi-\theta)\sin^4\theta\cos^4\theta + 144(\pi-\theta)^2\sin^3\theta\cos^5\theta \\
&- 2(\pi-\theta)^2\sin^3\theta\cos\theta(1-2\cos^2\theta-8\cos^4\theta) + 72(\pi-\theta)^3\sin^2\theta\cos^6\theta \\
&- 6(\pi-\theta)^3\sin^2\theta\cos^2\theta(1-2\cos^2\theta-8\cos^4\theta) - 4(\pi-\theta)^4\sin\theta\cos^3\theta(1-2\cos^2\theta-8\cos^4\theta), \\
&E'(\theta)F(\theta) \\
=&18\sin^5\theta\cos^3\theta + 90(\pi-\theta)\sin^4\theta\cos^4\theta + 158(\pi-\theta)^2\sin^3\theta\cos^5\theta + 118(\pi-\theta)^3\sin^2\theta\cos^6\theta \\
&- 2(\pi-\theta)\sin^6\theta\cos^2\theta - 4(\pi-\theta)^2\sin^5\theta\cos^3\theta - 2(\pi-\theta)^3\sin^4\theta\cos^4\theta \\
&+ 4(\pi-\theta)^2\sin^3\theta\cos^3\theta + 8(\pi-\theta)^3\sin^2\theta\cos^4\theta + 4(\pi-\theta)^4\sin\theta\cos^5\theta + 32(\pi-\theta)^4\sin\theta\cos^7\theta,
\end{aligned}$$

$$E(\theta)F'(\theta) - E'(\theta)F(\theta) = 2(\pi-\theta)\sin^6\theta\cos^2\theta - 2(\pi-\theta)^2\sin^7\theta\cos\theta - 6(\pi-\theta)^3\sin^4\theta\cos^2\theta - 4(\pi-\theta)^4\sin^3\theta\cos^3\theta.$$

Then the numerator of $\bar{\mathcal{L}}_\xi'(\theta)$ can be written as

$$2(\pi-\theta)\sin^3\theta\Psi_\xi(\theta),$$

where

$$\begin{aligned}
\Psi_\xi(\theta) := (1+\xi)(-\sin\theta + 2(\pi-\theta)\cos\theta) + \frac{1}{\pi^2}[&\sin^3\theta\cos^2\theta - (\pi-\theta)\sin^4\theta\cos\theta \\
&- 3(\pi-\theta)^2\sin\theta\cos^2\theta - 2(\pi-\theta)^3\cos^3\theta].
\end{aligned}$$

We note that the denominator of $\bar{\mathcal{L}}_\xi'(\theta)$ is always positive and let $\bar{\mathcal{L}}_\xi'(\theta) = 0$. We can get two obvious roots $\theta = 0$ and $\theta = \pi$. Then we want to show that there is only another root in $(\frac{\pi}{3}, \frac{\pi}{2})$. To show that, we first note

$$\Psi_\xi(\frac{\pi}{2}) = -(1+\xi) < 0, \quad \Psi_\xi(\frac{\pi}{3}) > 0.$$

Therefore, there is at least one root in $(\frac{\pi}{3}, \frac{\pi}{2})$.

We first show that $\Psi_\xi(\theta) < 0$ when $\theta \in (\frac{\pi}{2}, \pi)$. To show that, we have the following inequalities when $\theta \in (\frac{\pi}{2}, \pi)$.

$$\begin{aligned}
0 < \frac{1}{\pi^2}\sin^3\theta\cos^2\theta &< \frac{1}{\pi^2}\sin\theta, \\
0 < -\frac{1}{\pi^2}(\pi-\theta)\sin^4\theta\cos\theta &< -\frac{1}{\pi^2}(\pi-\theta)\cos\theta, \\
-\frac{3}{\pi^2}(\pi-\theta)^2\sin\theta\cos^2\theta &< 0, \\
0 < -\frac{2}{\pi^2}(\pi-\theta)^3\cos^3\theta &< -\frac{2}{\pi^2}\frac{\pi^2}{4}(\pi-\theta)\cos\theta = -\frac{1}{2}(\pi-\theta)\cos\theta.
\end{aligned}$$

From the above inequalities, we have that when $\theta \in (\frac{\pi}{2}, \pi)$,

$$\begin{aligned}
\Psi_\xi(\theta) <&(1+\xi)(-\sin\theta + 2(\pi-\theta)\cos\theta) + \frac{1}{\pi^2}\sin\theta - (\frac{1}{\pi^2} + \frac{1}{2})(\pi-\theta)\cos\theta \\
=&-\left(1+\xi-\frac{1}{\pi^2}\right)\sin\theta + \left(\frac{3}{2} + 2\xi - \frac{1}{\pi^2}\right)(\pi-\theta)\cos\theta \\
<&0.
\end{aligned}$$

We next show that $\Psi_\xi(\theta) > 0$ when $\theta \in (0, \frac{\pi}{3})$. First, we note that $\Psi_\xi(\theta)$ can be decomposed as follows.

$$\Psi_\xi(\theta) = \psi_0(\theta) + \xi\psi_1(\theta),$$

where

$$\begin{aligned}
&\psi_0(\theta) := -\sin\theta + 2(\pi-\theta)\cos\theta + \frac{1}{\pi^2}[\sin^3\theta\cos^2\theta - (\pi-\theta)\sin^4\theta\cos\theta - 3(\pi-\theta)^2\sin\theta\cos^2\theta \\
&\qquad - 2(\pi-\theta)^3\cos^3\theta], \\
&\psi_1(\theta) := -\sin\theta + 2(\pi-\theta)\cos\theta.
\end{aligned}$$

For $\psi_1(\theta)$, we use the facts that $\sin\theta < \theta$ and $\cos\theta > \frac{1}{2}$ on $(0, \frac{\pi}{3})$ and get

$$\psi_1(\theta) > -\theta + 2(\pi-\theta)\cdot\frac{1}{2} = \pi - 2\theta > \frac{\pi}{3} > 0.$$

For $\psi_0(\theta)$, by using Taylor bounds on $(0, \frac{\pi}{3})$

$$\sin\theta > \theta - \frac{\theta^3}{6},\ \sin\theta < \theta,\ \cos\theta > 1 - \frac{\theta^2}{2},\ \cos\theta < 1 - \frac{\theta^2}{2} + \frac{\theta^4}{24},$$

we get

$$\psi_0(\theta) > \theta^2 R(\theta),$$

where

$$\begin{aligned}
R(\theta) :=& 2\pi - 5\,\theta + \Big(-\frac{7\pi}{4} + \frac{2}{\pi}\Big)\theta^2 + \Big(\frac{17}{4} - \frac{1}{2\pi^2}\Big)\theta^3 + \Big(-\frac{11}{4\pi} + \frac{\pi}{2}\Big)\theta^4 \\
&+ \Big(-\frac{11}{8} + \frac{13}{12\pi^2}\Big)\theta^5 + \Big(-\frac{7\pi}{96} + \frac{29}{24\pi}\Big)\theta^6 + \Big(\frac{41}{192} - \frac{59}{108\pi^2}\Big)\theta^7 \\
&+ \Big(-\frac{5}{24\pi} + \frac{\pi}{192}\Big)\theta^8 + \Big(-\frac{1}{64} + \frac{161}{1728\pi^2}\Big)\theta^9 + \Big(-\frac{\pi}{6912} + \frac{1}{64\pi}\Big)\theta^{10} \\
&+ \Big(\frac{1}{2304} - \frac{11}{1728\pi^2}\Big)\theta^{11} - \frac{1}{2304\pi}\theta^{12} + \frac{1}{6912\pi^2}\theta^{13}.
\end{aligned}$$

Note that

$$\begin{aligned}
&\Big(\frac{1}{2304} - \frac{11}{1728\pi^2}\Big)\theta^{11} > -2.21\times10^{-4}\,\theta^{10},\ -\frac{1}{2304\pi}\theta^{12} > -1.516\times10^{-4}\,\theta^{10}, \\
&\Big(-\frac{5}{24\pi} + \frac{\pi}{192}\Big)\theta^8 > -0.0524\,\theta^7,\ \Big(-\frac{1}{64} + \frac{161}{1728\pi^2}\Big)\,\theta^9 > -6.7824\times10^{-3}\,\theta^7, \\
&\Big(-\frac{11}{8} + \frac{13}{12\pi^2}\Big)\theta^5 > \Big(-\frac{11\pi}{24} + \frac{13}{36\pi}\Big)\theta^4,
\end{aligned}$$

by using $\theta < \frac{\pi}{3}$. Then we can get

$$\begin{aligned}
R(\theta) >& 2\pi - 5\,\theta + \Big(-\frac{7\pi}{4} + \frac{2}{\pi}\Big)\theta^2 + \Big(\frac{17}{4} - \frac{1}{2\pi^2}\Big)\theta^3 + \Big(-\frac{43}{18\pi} + \frac{\pi}{24}\Big)\theta^4 \\
&+ \Big(-\frac{7\pi}{96} + \frac{29}{24\pi}\Big)\theta^6 + 0.099\theta^7 + 4.14\times10^{-3}\theta^{10} + \frac{1}{6912\pi^2}\theta^{13}.
\end{aligned}$$

After using $\theta < \frac{\pi}{3}$ for θ^4-term, we can further get

$$\begin{aligned}
R(\theta) >& 2\pi - 5\theta + \Big(-\frac{7\pi}{4} + \frac{2}{\pi}\Big)\theta^2 + \Big(\frac{373}{108} - \frac{1}{2\pi^2} + \frac{\pi^2}{72}\Big)\theta^3 \\
&+ 0.1555\theta^6 + 0.099\theta^7 + 4.14\times10^{-3}\theta^{10} + \frac{1}{6912\pi^2}\theta^{13} \\
>& 2\pi - 5\theta - 4.8612\theta^2 + 3.54\theta^3 + 0.1555\theta^6 + 0.099\theta^7 + 4.14\times10^{-3}\theta^{10} + \frac{1}{6912\pi^2}\theta^{13}.
\end{aligned}$$

We note that the coefficients for θ^{10} and θ^{13} are positive, so we can focus on the sum of other terms and denote it by $S(\theta)$,

$$S(\theta) := 2\pi - 5\theta - 4.8612\theta^2 + 3.54\theta^3 + 0.1555\theta^6 + 0.099\theta^7.$$

We take the derivative of $S(\theta)$ and use $\theta < \frac{\pi}{3}$ to get

$$\begin{aligned} S'(\theta) &= -5 - 9.7224\theta + 10.62\theta^2 + 0.933\theta^5 + 0.693\theta^6 \\ &< -5 - 9.7224\theta + \theta^2(10.62 + 0.933\frac{\pi^3}{27} + 0.693\frac{\pi^4}{81}) \\ &< -5 - 9.7224\theta + 12.525\theta^2. \end{aligned}$$

For the quadratic function $y(\theta) := -5 - 9.7224\theta + 12.525\theta^2$, we can easily check that $y(\theta) < 0$ on $(0, \frac{\pi}{3})$. Hence, $S'(\theta) < 0$ on $(0, \frac{\pi}{3})$ and further

$$S(\theta) > S(\frac{\pi}{3}) > 0.$$

Therefore, we can conclude that $R(\theta) > 0$ and $\psi_0(\theta) > 0$ on $(0, \frac{\pi}{3})$. We have shown that $\Psi_\xi(\theta) > 0$ on $(0, \frac{\pi}{3})$ as desired.

Finally, we want to show that there is only one root in $(\frac{\pi}{3}, \frac{\pi}{2})$. To show that, we find the derivative of $\Psi_\xi(\theta)$.

$$\begin{aligned} \Psi'_\xi(\theta) =& (1+\xi)(-3\cos\theta - 2(\pi-\theta)\sin\theta) \\ &+ \frac{1}{\pi^2}[3\sin^2\theta\cos^3\theta - \sin^4\theta\cos\theta - 4(\pi-\theta)\sin^3\theta\cos^2\theta + (\pi-\theta)\sin^5\theta + 6(\pi-\theta)\sin\theta\cos^2\theta \\ &+ 3(\pi-\theta)^2\cos^3\theta + 6(\pi-\theta)^2\sin^2\theta\cos\theta + 6(\pi-\theta)^3\sin\theta\cos^2\theta]. \end{aligned}$$

Notice that

$$\begin{aligned} 3\sin^2\theta\cos^3\theta - \sin^4\theta\cos\theta &= \sin^2\theta\cos\theta(3\cos^2\theta - \sin^2\theta) \\ &< \sin^2\theta\cos\theta(3\frac{9}{16}\sin^2\theta - \sin^2\theta) \\ &= \frac{11}{16}\sin^4\theta\cos\theta < \frac{11}{16}\cos\theta, \end{aligned}$$

where the first inequality comes from the fact that $\cos\theta < \frac{3}{4}\sin\theta$ in $(\frac{\pi}{3}, \frac{\pi}{2})$. Furthermore, the following inequalities hold

$$\begin{aligned} -4(\pi-\theta)\sin^3\theta\cos^2\theta + (\pi-\theta)\sin^5\theta &= (\pi-\theta)\sin^3\theta(-4\cos^2\theta + \sin^2\theta) \\ &< (\pi-\theta)\sin^5\theta < (\pi-\theta)\sin\theta, \\ 3(\pi-\theta)^2\cos^3\theta + 6(\pi-\theta)^2\sin^2\theta\cos\theta &= 3(\pi-\theta)^2\cos\theta(\cos^2\theta + 2\sin^2\theta) \\ &= 3(\pi-\theta)^2\cos\theta(1+\sin^2\theta) \\ &< 6(\frac{2\pi}{3})^2\cos\theta = \frac{8}{3}\pi^2\cos\theta, \\ 6(\pi-\theta)\sin\theta\cos^2\theta + 6(\pi-\theta)^3\sin\theta\cos^2\theta &= 6(\pi-\theta)\sin\theta\cos^2\theta(1+(\pi-\theta)^2) \\ &< 6(\pi-\theta)\sin\theta\cos^2\theta(1+\frac{4}{9}\pi^2) \\ &< 6(\pi-\theta)\sin\theta\frac{1}{4}(1+\frac{4}{9}\pi^2) \\ &= (\pi-\theta)\sin\theta(\frac{3}{2} + \frac{2}{3}\pi^2). \end{aligned}$$

As a result, we have

$$\begin{aligned} \Psi'_\xi(\theta) &< -3\cos\theta - 2(\pi-\theta)\sin\theta + \frac{11}{16\pi^2}\cos\theta + \frac{1}{\pi^2}(\pi-\theta)\sin\theta + \frac{8}{3}\cos\theta + (\frac{3}{2\pi^2} + \frac{2}{3})(\pi-\theta)\sin\theta \\ &= -\left(\frac{1}{3} - \frac{11}{16\pi^2}\right)\cos\theta - \left(\frac{4}{3} - \frac{5}{2\pi^2}\right)(\pi-\theta)\sin\theta \\ &< 0, \end{aligned}$$

which means that $\Psi_\xi(\theta)$ is decreasing in $(\frac{\pi}{3}, \frac{\pi}{2})$. Hence, there is only one root in $(\frac{\pi}{3}, \frac{\pi}{2})$ and we can conclude that it is the global minimizer for any $\xi > 0$. □