Gradient Descent with Projection Finds Over-Parameterized Neural Networks for Learning Low-Degree Polynomials with Nearly Minimax Optimal Rate

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2026

Abstract

We study learning a low-degree spherical polynomial of degree $k_0 = \Theta(1)$ on the unit sphere in $\mathbb{R}^d$ using an over-parameterized two-layer neural network with augmented features. Our main result is an improved sample complexity: for any regression risk $\varepsilon \in (0, \Theta(d^{-k_0})]$, a network trained via Gradient Descent with Projection (GDP) achieves $n \asymp \Theta(\log(4/\delta) \cdot d^{k_0}/\varepsilon)$ with probability $1-\delta$, $\delta \in (0,1)$, improving over $\Theta(d^{k_0} \max\{\varepsilon^{-2}, \log d\})$. This rate is nearly optimal, yielding regression risk $\log(4/\delta) \cdot \Theta(d^{k_0}/n)$ with probability at least $1-\delta$, close to the minimax rate $\Theta(d^{k_0}/n)$ for kernels of rank $\Theta(d^{k_0})$. To our knowledge, this is the first sharp risk bound with algorithmic guarantees for over-parameterized networks on such tasks. Our approach goes beyond the NTK limit by learning a subspace of its eigenspace, using a projection operator to restrict the solution to a low-dimensional RKHS subspace, enabling the sharp bound.

1. Introduction

With the success of deep learning across machine learning [21], understanding neural network generalization is central. Prior work shows gradient-based methods (GD/SGD) achieve vanishing training loss in deep networks [1, 2, 13, 14, 30, 43]. Under over-parameterization, training dynamics are approximated by kernel methods such as the NTK [19], though infinite-width networks can still exhibit feature learning [36]. In this regime, weights stay near initialization, enabling first-order Taylor approximation and tractable generalization analysis [2, 7, 16].

Generalization can be studied via learning low-degree polynomials, motivated by spectral bias [8, 11, 25], where networks favor top eigenspaces of the NTK operator. For data on $\mathbb{S}^{d-1} \subset \mathbb{R}^d$, degree-$\ell$ polynomials admit representations via spherical harmonics up to degree ℓ, aligned with top NTK eigenvalues (see Section C and Theorem 11). Existing works study both NTK and feature learning beyond NTK for learning low-degree polynomials. QuadNTK [4] uses second-order expansion to learn sparse polynomials; Nichani et al. [24] combines NTK and QuadNTK for dense polynomials with sparse high-degree parts; other methods include two-stage optimization [12] and mean-field analyses [33]. However, sharp regression risk analysis for over-parameterized networks learning low-degree polynomials [4, 12, 16, 24, 33] is largely missing. For instance, Nichani et al. [24] obtain risk ε when $n \gtrsim d^{k_0} \max\left\{\varepsilon^{-2}, \log d\right\}$, while Ghorbani et al. [16] show vanishing NTK risk for $\tilde{\Theta}(d^{k_0}) \le n \le \Theta(d^{k_0+1-\delta})$ as $d \to \infty$ without rates or sharpness. Under fixed d settings common in sharp nonparametric regression [18, 23, 31, 38], even vanishing risk is not established.

Understanding sharp regression risk for learning low-degree polynomials remains important. In this paper, we assume f^* lies in the RKHS of an over-parameterized two-layer network with bounded norm, where f^* is a degree-k_0 polynomial on $\mathbb{S}^{d-1} \subset \mathbb{R}^d$, $k_0 \geq 1$. Our main result (Theorem 1) shows that training with Gradient Descent with Projection (GDP) and $n \geq \Theta(\log(4/\delta) \cdot d^{2k_0})$ yields risk $\log(4/\delta) \cdot \Theta(d^{k_0}/n)$ with probability $\geq 1-\delta$. Since the minimax risk for rank $r = \Theta(d^{k_0})$ kernels is $\Theta(r/n) = \Theta(d^{k_0}/n)$ [26, Theorem 2(a)], our rate is nearly optimal. This is, to our knowledge, the first nearly optimal risk bound with algorithmic guarantees for learning low-degree spherical polynomials via over-parameterized ReLU networks; unlike prior projected methods [35, 41], we design GDP tailored to neural networks achieving near-optimal rates.

We organize the paper as follows. Section 2 introduces the setup, Section 3 presents GDP and main results, and Section A of the appendix outlines proofs and techniques. Section E of the appendix provides the simulation results.

Notations. Bold letters denote matrices/vectors and regular letters denote scalars. $\mathbf{A}^{[i]}$ is the i-th column of a matrix $\mathbf{A}$, while subscripts indicate rows/elements; $\vec{\mathbf{x}}_i$ denotes the i-th feature. $\|\cdot\|_F$ and $\|\cdot\|_p$ denote Frobenius and ℓ^p/matrix p-norms. $[m\colon n]$ (or $[n]$) denotes integer ranges. $\mathrm{Var}[\cdot]$ is variance, $\mathbf{I}_n$ the identity matrix, and $\mathbb{1}_{\{E\}}$ an indicator function. A^c and $|A|$ denote complement and cardinality. $\mathrm{vec}(\cdot)$ and $\mathrm{tr}(\cdot)$ denote vectorization and trace. The unit sphere is $\mathbb{S}^{d-1} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 = 1\}$. Let $\mathcal{X}$ be the input space, and $L^p(\mathcal{X}, \mu)$ ($p \geq 1$) be the space of p-integrable functions with $\langle f, g\rangle_{L^p(\mu)} = \int fg \,\mathrm{d}\mu$ and $\|f\|^p_{L^p(\mu)} = \int |f|^p \,\mathrm{d}\mu < \infty$. $\mathbf{B}(\mathbf{x}; r)$ is the Euclidean ball. For $g : \mathcal{X} \to \mathbb{R}$, $\|g\|_\infty = \sup_{\mathbf{x}\in\mathcal{X}} |g(\mathbf{x})|$, and L^∞ is the bounded class. $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{H}}$ are Hilbert space inner product and norm. We use the following asymptotics: $a = \mathcal{O}(b)$ ($a \lesssim b$), $\tilde{\mathcal{O}}$ (refined constants), $a = o(b)$, $a = w(b)$, and $a \asymp b$ ($= \Theta(b)$). $\mathrm{Unif}\left(\mathbb{S}^{d-1}\right)$ is the uniform distribution. Constants may change line to line. $\mathbb{E}_P[\cdot]$ denotes expectation under P. $\mathbf{P}_{\mathcal{S}}$ is orthogonal projection, $\mathrm{Span}(\mathbf{A})$ the column span, and $\overline{A}$ the closure. Throughout this paper we set $\mathcal{X} = \mathbb{S}^{d-1}$.

2. Problem Setup

We introduce the problem setups for nonparametric regression with the target function as a low-degree spherical polynomial in this section.

2.1. Two-Layer Neural Network

We are given training data $\left\{(\vec{\mathbf{x}}_i, y_i)\right\}_{i=1}^n$ with $\vec{\mathbf{x}}_i \in \mathcal{X}$ and $y_i \in \mathbb{R}$, where $\vec{\mathbf{x}}_i \neq \vec{\mathbf{x}}_j$ for $i \neq j$. Let $\mathbf{S} = \left\{\vec{\mathbf{x}}_i\right\}_{i=1}^n$, P_n be the empirical distribution over $\mathbf{S}$, and $\mathbf{y} = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$. The responses satisfy $y_i = f^*(\vec{\mathbf{x}}_i) + w_i$, where $\{w_i\}_{i=1}^n$ are i.i.d. sub-Gaussian with mean 0 and variance proxy σ_0^2, i.e., $\mathbb{E}[\exp(\lambda w_i)] \leq \exp(\lambda^2\sigma_0^2/2)$ for all $\lambda \in \mathbb{R}$, and f^* is the target function. Define $\mathbf{w} = [w_1, \ldots, w_n]^\top$ and $f^*(\mathbf{S}) = \left[f^*(\vec{\mathbf{x}}_1), \ldots, f^*(\vec{\mathbf{x}}_n)\right]^\top$. The features are drawn i.i.d. from $P = \mathrm{Unif}\left(\mathbb{S}^{d-1}\right)$ with measure μ. We study a two-layer neural network (NN) with augmented feature:

$$f(\mathcal{W}, \mathbf{x}) = \frac{1}{\sqrt{m}} \sum_{r=1}^m a_r \sigma\left(\vec{\mathbf{w}}_r^{\top} \mathbf{x}\right) + \frac{1}{\sqrt{m}} \vec{\mathbf{w}}_{m+1}^{\top} \mathbf{F}(\mathbf{W}(0), \mathbf{x}), \tag{1}$$

where $\sigma(\cdot) = \max\{\cdot, 0\}$, $\mathcal{W} = \left\{\mathbf{W}, \vec{\mathbf{w}}_{m+1}\right\}$ denotes the weights of the network, $\mathbf{W} = \left\{\vec{\mathbf{w}}_r\right\}_{r=1}^m$, $\vec{\mathbf{w}}_{m+1}$ are weights with $\vec{\mathbf{w}}_r \in \mathbb{R}^d$, $\vec{\mathbf{w}}_{m+1} \in \mathbb{R}^m$, and m is the width. The augmented feature

$\mathbf{F}(\mathbf{W}(0),\mathbf{x}) \in \mathbb{R}^m$ satisfies $[\mathbf{F}(\mathbf{W}(0),\mathbf{x})]_r = \mathbb{1}_{\left\{\vec{\mathbf{w}}_r(0)^\top \mathbf{x} \geq 0\right\}}$, and $\boldsymbol{a} = [a_1,\ldots,a_m] \in \mathbb{R}^m$. We may write $\mathbf{W}, \vec{\mathbf{w}}_r$ as $\mathbf{W}_\mathbf{S}, \vec{\mathbf{w}}_{\mathbf{S},r}$ to indicate dependence on $\mathbf{S}$.

Novel Augmented Feature Compared to the Regular ReLU Network. Compared to the vanilla network $f^{(\text{vanilla})}(\mathbf{W},\mathbf{x}) = \frac{1}{\sqrt{m}}\sum_{r=1}^m a_r \sigma\left(\vec{\mathbf{w}}_r^\top \mathbf{x}\right)$, (1) includes the additional term $\frac{1}{\sqrt{m}}\vec{\mathbf{w}}_{m+1}^\top \mathbf{F}(\mathbf{W}(0),\mathbf{x})$, which ensures the associated NTK is a PSD kernel K (defined in (2)) with strictly positive eigenvalues (Theorem 31, Appendix D.6). In contrast, for $f^{(\text{vanilla})}$, the NTK eigenvalues $\left\{\tilde{\lambda}_j\right\}_{j\geq 0}$ satisfy $\tilde{\lambda}_{2t+1} = 0$ for $t \geq 1$ [6, Proposition 5], so the corresponding eigenspaces fail to span all spherical harmonics of order $2t+1$, limiting the learning of spherical polynomials with odd-degree (≥ 3) components.

2.2. Kernel and Kernel Regression for Nonparametric Regression

We define kernel functions for $\mathbf{u},\mathbf{v} \in \mathcal{X}$ by

$$K^{(0)}(\mathbf{u},\mathbf{v}) \coloneqq \frac{\pi - \arccos(\mathbf{u}^\top \mathbf{v})}{2\pi}, \quad K^{(1)}(\mathbf{u},\mathbf{v}) \coloneqq \mathbf{u}^\top\mathbf{v} K^{(0)}(\mathbf{u},\mathbf{v}), \quad K = K^{(0)} + K^{(1)}, \tag{2}$$

which corresponds to the NTK of the two-layer NN (1) with constant second-layer weights $\boldsymbol{a}$, and K is PSD. Let $\mathbf{K} \in \mathbb{R}^{n\times n}$ be the Gram matrix over $\mathbf{S}$, $\mathbf{K}_{ij} = K(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j)$, and $\mathbf{K}_n \coloneqq \mathbf{K}/n$ with analogous definitions for $\mathbf{K}^{(\alpha)}, \mathbf{K}_n^{(\alpha)}$, $\alpha = 0,1$. Let $\mathbf{K}_n = \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^\top$ with eigenvalues $\{\widehat{\lambda}_i\}_{i=1}^n$ in non-increasing order; $\mathbf{K}_n$ is non-singular [14] and $\widehat{\lambda}_1 \in (0,1)$ since $\sup_\mathbf{x} K(\mathbf{x},\mathbf{x}) = 1$. Let $\mathcal{H}_K$ be the RKHS of K. As K is continuous on compact $\mathcal{X}\times\mathcal{X}$, the operator $T_K f(\mathbf{x}) = \int_\mathcal{X} K(\mathbf{x},\mathbf{x}')f(\mathbf{x}')\mathrm{d}\mu(\mathbf{x}')$ is positive, self-adjoint, and compact. By the spectral theorem, there exist eigenpairs $\{(e_j,\lambda_j)\}_{j\geq 0}$ with $1 \geq \lambda_0 \geq \lambda_1 \geq \cdots > 0$ and $T_K e_j = \lambda_j e_j$. Let $\{\mu_\ell\}_{\ell\geq 0}$ be distinct eigenvalues with cumulative multiplicities m_ℓ. Then $\{v_j = \sqrt{\lambda_j} e_j\}_{j\geq 0}$ is an ONB of $\mathcal{H}_K$. For $\gamma_0 > 0$, define $\mathcal{H}_K(\gamma_0) = \{f \in \mathcal{H}_K : \|f\|_\mathcal{H} \leq \gamma_0\} = \{f = \sum_{j\geq 0}\beta_j e_j : \sum_{j\geq 0}\beta_j^2/\lambda_j \leq \gamma_0^2\}$. By Theorem 31, $\{e_j\}$ are spherical harmonics. Let $\mathcal{H}_\mathbf{S} = \overline{\{\sum_{i=1}^n K(\cdot,\vec{\mathbf{x}}_i)\alpha_i\}}$.

The task of nonparametric regression. We consider the target function $f^*(\mathbf{x}) = \sum_{\ell=0}^{k_0}\sum_{j=1}^{N(d,\ell)} a_{\ell,j} Y_{\ell,j}(\mathbf{x})$ where $\mathbf{x} \in \mathcal{X}$, $\{Y_{\ell,j}\}$ are degree-ℓ spherical harmonics forming an orthogonal basis of $\mathcal{H}_\ell$ with dimension $N(d,\ell)$. Background on $\mathbb{S}^{d-1}$ is in Section C, where it is also shown that $\{e_j\} = \{Y_{\ell,j}\}_{\ell\geq 0, j\in[N(d,\ell)]}$. We assume $f^* \in \mathcal{F}^*$ with

$$\mathcal{F}^* = \left\{f = \sum_{\ell=0}^{k_0}\sum_{j=1}^{N(d,\ell)} a_{\ell,j}Y_{\ell,j} : \sum_{\ell=0}^{k_0}\sum_{j=1}^{N(d,\ell)} a_{\ell,j}^2/\mu_\ell \leq \gamma_0^2\right\}. \tag{3}$$

By Theorem 11 in the appendix, $\mathcal{F}^*$ contains all degree-k_0 polynomials on $\mathcal{X}$ with finite $\mathcal{H}_K$-norm γ_0. The goal is to estimate $\widehat{f}$ from $\{(\vec{\mathbf{x}}_i, y_i)\}_{i=1}^n$ so that $\mathbb{E}_P\left[(\widehat{f} - f^*)^2\right]$ decays rapidly; we analyze the rate when $\widehat{f}$ is given by the over-parameterized NN (1) trained by GD.

Minimax Lower Risk Bound for Learning a Low-Degree Spherical Polynomial. From (3) and Theorem 31, $\mathcal{F}^* \subseteq \cup_{\ell=0}^{k_0}\mathcal{H}_\ell$ with $k_0 = \Theta(1)$. Define the finite-rank kernel with $r_0 \coloneqq m_{k_0} = \sum_{\ell'=0}^{k_0} N(d,\ell')$: $K^{(r_0)}(\mathbf{x},\mathbf{x}') \coloneqq \sum_{\ell=0}^{k_0}\sum_{j=1}^{N(d,\ell)} \mu_\ell Y_{\ell,j}(\mathbf{x})Y_{\ell,j}(\mathbf{x}')$ for all $\mathbf{x},\mathbf{x}' \in \mathcal{X}$, so that $\mathcal{F}^* \subseteq \mathcal{H}_{K^{(r_0)}}(\gamma_0)$. Lemma 12 shows $r_0 = \Theta(d^{k_0})$ for $k_0 = \Theta(1)$ and $d > \Theta(1)$. By [26, Theorem 2(a)], the minimax lower bound for the regression risk with $K^{(r_0)}$ is $\Theta(r_0/n) = \Theta(d^{k_0}/n)$.

3. Summary of Main Result

3.1. Training by Gradient Descent with Projection

In training the two-layer NN (1), $\boldsymbol{a}$ is randomly initialized to ± 1 with equal probability and fixed thereafter, and all other weight vectors are optimized. The quadratic loss $L(\mathbf{W}) = \frac{1}{2n}\sum_{i=1}^{n}(f(\mathcal{W}, \vec{\mathbf{x}}_i) - y_i)^2$ is minimized. At iteration $t+1$, GDP updates the network weights by

$$\begin{aligned} \mathrm{vec}\left(\mathbf{W}_{\mathbf{S}}(t+1)\right) - \mathrm{vec}\left(\mathbf{W}_{\mathbf{S}}(t)\right) &= -\frac{\eta}{n}\mathbf{Z}_{\mathbf{S}}(t)\mathbf{P}^{(r_0)}(\widehat{\mathbf{y}}(t) - \mathbf{y}), \\ \vec{\mathbf{w}}_{m+1}(t+1) - \vec{\mathbf{w}}_m(t) &= -\frac{\eta}{n\sqrt{m}}\mathbf{F}(\mathbf{W}(0), \mathbf{S})^\top \mathbf{P}^{(r_0)}(\widehat{\mathbf{y}}(t) - \mathbf{y}), \end{aligned} \tag{4}$$

where $\mathbf{F}(\mathbf{W}(0), \mathbf{S}) \in \mathbb{R}^{n\times m}$ with $[\mathbf{F}(\mathbf{W}(0), \mathbf{S})]_i = \mathbf{F}(\mathbf{W}(0), \vec{\mathbf{x}}_i)^\top$, $\widehat{\mathbf{y}}(t)_i = f(\mathcal{W}(t), \vec{\mathbf{x}}_i)$, and $f_t(\cdot) = f(\mathcal{W}(t), \cdot)$. The matrix $\mathbf{Z}_{\mathbf{S}}(t) \in \mathbb{R}^{md\times n}$ is specified by $[\mathbf{Z}_{\mathbf{S}}(t)]_{[(r-1)d+1:rd]i} = 1/\sqrt{m} \cdot \mathbb{1}_{\left\{\vec{\mathbf{w}}_r(t)^\top \vec{\mathbf{x}}_i \geq 0\right\}} \vec{\mathbf{x}}_i a_r$, and $\mathbf{P}^{(r_0)} = \mathbf{U}\boldsymbol{\Sigma}^{(r_0)}\mathbf{U}^\top$ projects onto an r_0-dimensional subspace of $\mathcal{H}_{\mathbf{S}}$ (absent in vanilla GD). Using symmetric initialization [10, 12] with even m, $\vec{\mathbf{w}}_{2r'}(0) \sim \mathcal{N}(\mathbf{0}, \kappa^2\mathbf{I}_d)$, $a_{2r'} \sim \mathrm{unif}(\{-1, 1\})$, and $\vec{\mathbf{w}}_{2r'-1}(0) = \vec{\mathbf{w}}_{2r'}(0)$, $a_{2r'-1} = -a_{2r'}$, ensures $\widehat{\mathbf{y}}(0) = \mathbf{0}$. We denote $\mathbf{W}(0) = \{\vec{\mathbf{w}}_r(0)\}_{r=1}^m$, and run Algorithm 1 deferred to Section A of the appendix for T steps.

3.2. Sharp Bound for Regression Risk

The main sharp risk bound is given in Theorem 1 (proved in Section D.2).

Theorem 1 *Suppose that $n \geq \Theta(\log(4/\delta) \cdot d^{2k_0})$, $\delta \in (0, 1)$, and $c_t \in (0, 1]$ is an arbitrary positive constant. Suppose the network width m satisfies*

$$m \gtrsim \left(\frac{n}{d^{k_0}}\right)^{\frac{25}{2}} d^{\frac{5}{2}}, \tag{5}$$

and the neural network $f(\mathbf{W}(t), \cdot)$ is trained by GDP using Algorithm 1 with the constant learning rate $\eta = \Theta(1) \in (0, 1)$, and $T \asymp n/d^{k_0}$. Then for every $t \in [c_t T \colon T]$, with probability at least $1 - \delta - \exp(-\Theta(n)) - 2\exp(-\Theta(r_0)) - 2/n$ over the random noise $\mathbf{w}$, the random training features $\mathbf{S}$ and the random initialization $\mathbf{W}(0)$, $f(\mathbf{W}(t), \cdot) = f_t$ satisfies

$$\mathbb{E}_P\left[(f_t - f^*)^2\right] \lesssim \log\frac{4}{\delta} \cdot \Theta\left(\frac{d^{k_0}}{n}\right). \tag{6}$$

Here $r_0 = m_{k_0} = \Theta(d^{k_0})$.

Theorem 1 shows that the neural network (1) trained by GDP via Algorithm 1 achieves a sharp regression risk rate $\Theta(\log(4/\delta) \cdot \Theta(d^{k_0}/n))$ for learning a degree-k_0 spherical polynomial. [42] establishes a minimax lower bound of order $\Theta(r_0/n)$ as described in Section 2.2. Compared to this lower bound, our result is nearly minimax optimal up to an additional logarithmic factor. It follows from (6) that the two-layer NN trained by GDP attains sample complexity $n \asymp \Theta(\log(4/\delta) \cdot d^{k_0}/\varepsilon)$ for any regression risk $\varepsilon \in (0, \Theta(d^{-k_0})]$, which is significantly smaller than $\Theta(d^{k_0} \max\{\varepsilon^{-2}, \log d\})$ in [24]. We further compare our result with competing approaches for learning low-degree spherical polynomials in Table 1, focusing on algorithmic guarantees, specifically, whether a finite-width neural network is trained—and the sharpness of the regression risk.

[24, Theorem 1] shows that achieving regression risk $\varepsilon > 0$ requires sample complexity $n \gtrsim d^{k_0} \max\{\varepsilon^{-2}, \log d\}$, implying convergence rate $\Theta(\sqrt{d^{k_0}/n})$ when the risk is below $1/\sqrt{\log d}$.

Table 1: Comparison between our result and existing works on learning low-degree polynomials on the spheres of $\mathbb{R}^d$ via training over-parameterized neural networks, with or without algorithmic guarantees. Most results adopt a common setup where $f^* \in \mathcal{H}_{\tilde{K}}$, with $\tilde{K}$ being the NTK of the specific neural network studied in each work, and the responses $\{y_i\}_{i=1}^n$ are corrupted by i.i.d. Gaussian or sub-Gaussian noise with zero mean; the only exception is [24], which assumes noise-free responses. It is noted that sample complexity can be directly derived from the regression risk. The regression risk in [12, Theorem 1] is characterized for risk below $1/\sqrt{\log d}$, where the meaning of r is given in Section 3.2, and $\tilde{\Theta}$ suppresses a logarithmic factor of $\log(mnd)$.

Existing Works and Our Result	**Finite-Width NN is Trained**	**Sharpness of the Regression Risk**
[16, Theorem 4]	No	Only matching the lower bound for pointwise kernel learning, not minimax optimal
[4, Theorem 7]	Yes	Not minimax optimal
[24, Theorem 1]	Yes	$\Theta(\sqrt{d^{k_0}/n})$, not minimax optimal
[12, Theorem 1]	Yes	L^1-norm regression risk $\tilde{\Theta}(\sqrt{dr^{k_0}/n} + \sqrt{r^p/m})$, not minimax optimal
Our Result (Theorem 1)	Yes	Nearly minimax optimal, $\log\frac{6}{\delta} \cdot \Theta\left(\frac{d^{k_0}}{n}\right)$

However, such a rate this is not minimax optimal and is looser than our bound. The two-stage feature learning method [12] assumes the target depends on $r \ll d$ directions, so GD confines the learned function to a rank-r RKHS subspace. Without this limitation (then $r = d$), its L^1-risk [12, Theorem 1] is at least $\tilde{\Theta}(\sqrt{d^{k_0+1}/n})$. Since L^p-norms increase with p, our L^2 bound (Theorem 1) yields a sharper L^1-risk $\Theta(\sqrt{d^{k_0}/n})$.

Apart from the feature learning methods in Table 1, the statistical learning literature has established rich results in the sharp convergence rates for the risk of nonparametric kernel regression [27, 29, 39, 40]. By training over-parameterized shallow [18, Theorem 5.2] or deep [31, Theorem 3.11] neural networks with traning features following spherical uniform distribution on the unit sphere, these results [18, 31] show that minimax optimal rate $\mathcal{O}(n^{-d/(2d-1)})$ is achieved for the regression risks when the target function is in $\mathcal{H}_{\tilde{K}}(\gamma_0)$ where $\tilde{K}$ is the NTK of a specific studied neural studied in each work. As discussed in Section 2.2, because the target function f^* as a degree-k_0 spherical polynomial lies in the union of the eigenspaces up to degree k_0, we need to learn the subspace $\cup_{\ell=0}^{k_0}\mathcal{H}_\ell$ of dimension $r_0 = m_{k_0}$ instead of the entire RKHS $\mathcal{H}_K(\gamma_0)$ for a sharp regression risk. This observation motivates our GDP algorithm, which fits f^* in an r_0-dimensional subspace $\mathcal{H}_{\mathbf{S},r_0}$ (with $r = r_0$). For $r \in [n-1]$, we write $\mathbf{U} = [\mathbf{U}^{(r)}\ \mathbf{U}^{(-r)}]$ and define $\mathcal{H}_{\mathbf{S},r} := \overline{\left\{\sum_{i=1}^n K(\cdot, \vec{\mathbf{x}}_i)\alpha_i : \boldsymbol{\alpha} \in \mathrm{Span}(\mathbf{U}^{(r)})\right\}}$, a subspace of $\mathcal{H}_{\mathbf{S}}$ of dimension r.

4. Conclusion

We study nonparametric regression by training an over-parameterized two-layer NN where the target function is in the RKHS associated with the NTK of the neural network and also a degree-k_0 spherical polynomial on the unit sphere in $\mathbb{R}^d$. We show that, if the neural network is trained by a novel Gradient Descent with Projection (GDP), a sharp and nearly minimax optimal rate of the order $\log(4/\delta) \cdot \Theta(d^{k_0}/n)$ can be obtained. A novel proof strategy is employed to achieve this result, and we compare our results to the current state-of-the-art with a detailed roadmap of our technical approach.

References

[1] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 242–252. PMLR, 2019.

[2] Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 322–332. PMLR, 2019.

[3] Francis R. Bach. Breaking the curse of dimensionality with convex neural networks. *J. Mach. Learn. Res.*, 18:19:1–19:53, 2017.

[4] Yu Bai and Jason D. Lee. Beyond linearization: On quadratic and higher-order approximation of wide neural networks. In *International Conference on Learning Representations*. OpenReview.net, 2020.

[5] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *Ann. Statist.*, 33(4):1497–1537, 08 2005.

[6] Alberto Bietti and Julien Mairal. On the inductive bias of neural tangent kernels. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems*, pages 12873–12884, 2019.

[7] Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems*, pages 10835–10845, 2019.

[8] Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan Gu. Towards understanding the spectral bias of deep learning. In Zhi-Hua Zhou, editor, *International Joint Conference on Artificial Intelligence*, pages 2205–2211. ijcai.org, 2021.

[9] T.S. Chihara. *An Introduction to Orthogonal Polynomials*. Dover Books on Mathematics. Dover Publications, 2011. ISBN 9780486479293.

[10] Lénaïc Chizat, Edouard Oyallon, and Francis Bach. *On lazy training in differentiable programming*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[11] Moulik Choraria, Leello Tadesse Dadi, Grigorios Chrysos, Julien Mairal, and Volkan Cevher. The spectral bias of polynomial neural networks. In *International Conference on Learning Representations*. OpenReview.net, 2022.

[12] Alexandru Damian, Jason D. Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In Po-Ling Loh and Maxim Raginsky, editors, *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pages 5413–5452. PMLR, 2022.

[13] Simon S. Du, Jason D. Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1675–1685. PMLR, 2019.

[14] Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019.

[15] Costas Efthimiou and Christopher Frye. *Spherical Harmonics in p Dimensions*. World Scientific Co., 2014. doi: 10.1142/9134.

[16] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *Ann. Statist.*, 49(2):1029 – 1054, 2021.

[17] R W Gosper. Decision procedure for indefinite hypergeometric summation. *Proc. Natl. Acad. Sci. U. S. A.*, 75(1):40–42, January 1978.

[18] Tianyang Hu, Wenjia Wang, Cong Lin, and Guang Cheng. Regularization matters: A nonparametric perspective on overparametrized neural network. In Arindam Banerjee and Kenji Fukumizu, editors, *International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 829–837. PMLR, 2021.

[19] Arthur Jacot, Clément Hongler, and Franck Gabriel. Neural tangent kernel: Convergence and generalization in neural networks. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems*, pages 8580–8589, 2018.

[20] Nicolai V. Krylov. Basics of harmonic polynomials and spherical functions. Technical report. URL https://www-users.cse.umn.edu/~nkrylov/Moscow_2019_Sphrcal.pdf.

[21] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–444, 2015.

[22] Michel. Ledoux. *Probability in Banach Spaces [electronic resource] : Isoperimetry and Processes / by Michel Ledoux, Michel Talagrand.* Classics in Mathematics. Springer Berlin Heidelberg, Berlin, Heidelberg, 1st ed. 1991. edition, 1991.

[23] Yicheng Li, Zixiong Yu, Guhan Chen, and Qian Lin. On the eigenvalue decay rates of a class of neural-network related kernel functions defined on general domains. *Journal of Machine Learning Research*, 25(82):1–47, 2024.

[24] Eshaan Nichani, Yu Bai, and Jason D. Lee. Identifying good directions to escape the NTK regime and efficiently learn low-degree plus sparse polynomials. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[25] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5301–5310. PMLR, 09–15 Jun 2019.

[26] Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Mach. Learn. Res.*, 13:389–427, 2012.

[27] Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *J. Mach. Learn. Res.*, 15(1):335–366, 2014.

[28] Lorenzo Rosasco, Mikhail Belkin, and Ernesto De Vito. On learning with integral operators. *J. Mach. Learn. Res.*, 11:905–934, 2010.

[29] Charles J. Stone. Additive Regression and Other Nonparametric Models. *Ann. Statist.*, 13(2): 689 – 705, 1985.

[30] Lili Su and Pengkun Yang. On learning over-parameterized neural networks: A functional approximation perspective. In *Advances in Neural Information Processing Systems*, pages 2637–2646, 2019.

[31] Namjoon Suh, Hyunouk Ko, and Xiaoming Huo. A non-parametric regression viewpoint : Generalization of overparametrized deep RELU network under noisy observations. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[32] G. Szegő. *Orthogonal Polynomials*. American Math. Soc: Colloquium publ. Amer. Math. Soc., 1975. ISBN 9780821810231.

[33] Shokichi Takakura and Taiji Suzuki. Mean-field analysis on two-layer neural networks from a kernel perspective. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

[34] F. T. Wright. A Bound on Tail Probabilities for Quadratic Forms in Independent Random Variables Whose Distributions are not Necessarily Symmetric. *Ann. Probab.*, 1(6):1068 – 1070, 1973.

[35] Xingyu Xu, Yandi Shen, Yuejie Chi, and Cong Ma. The power of preconditioning in overparameterized low-rank matrix sensing. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 38611–38654. PMLR, 23–29 Jul 2023.

[36] Greg Yang and Edward J. Hu. Tensor programs IV: feature learning in infinite-width neural networks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 11727–11737. PMLR, 2021.

[37] Yingzhen Yang. Sharp generalization for nonparametric regression by over-parameterized neural networks: A distribution-free analysis in spherical covariate. In *International Conference on Machine Learning (ICML)*, 2025.

[38] Yingzhen Yang and Ping Li. Gradient descent finds over-parameterized neural networks with sharp generalization for nonparametric regression. *arXiv preprint arXiv:2411.02904*, 2024. URL https://arxiv.org/abs/2411.02904.

[39] Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, 27(5):1564 – 1599, 1999.

[40] Ming Yuan and Ding-Xuan Zhou. Minimax optimal rates of estimation in high dimensional additive models. *Ann. Statist.*, 44(6):2564 – 2593, 2016.

[41] Gavin Zhang, Salar Fattahi, and Richard Y. Zhang. Preconditioned gradient descent for over-parameterized nonconvex burer–monteiro factorization with global optimality certification. *Journal of Machine Learning Research*, 24(163):1–55, 2023.

[42] Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates. *J. Mach. Learn. Res.*, 16(1): 3299–3340, January 2015. ISSN 1532-4435.

[43] Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems*, pages 2053–2062, 2019.

The appendix of this paper is organized as follows. We first describe the roadmap of proofs with our novel proof strategy in Section A. We then present the basic mathematical results employed in our proofs in Section B, and then introduce the detailed technical background about harmonic analysis on spheres in Section C. Detailed proofs are presented in Section D. In particular, more results about the eigenvalue decay rates are presented in Section D.6. The simulation results are presented in Section E.

Appendix A. Roadmap of Proofs

We present the roadmap of our theoretical results which lead to the main result, Theorem 1, in this section. We first present in Section A.1 the basic definitions used in our proofs, and then detail the roadmap and key technical results with our novel proof strategy for this work in Section A.2. The proof of Theorem 1 is presented in Section D.2. Section D.3 together with the remaining parts of the appendix present the proofs of the key results in Section A.2.

Algorithm 1 Training the Two-Layer NN by GDP

1: $\mathbf{W}(T) \leftarrow$ Training-by-GDP$(T, \mathbf{W}(0))$
2: **input:** $T, \mathbf{W}(0), \vec{\mathbf{w}}_{m+1} = \mathbf{0}, \eta$
3: **for** $t = 1, \ldots, T$ **do**
4: Perform the t-th step of GDP by (4)
5: **end for**
6: **return** $W(T), \vec{\mathbf{w}}_{m+1}(T)$

A.1. Basic Definitions

We introduce the following definitions for our analysis. We define

$$\mathbf{u}(t) \coloneqq \widehat{\mathbf{y}}(t) - \mathbf{y} \tag{7}$$

as the difference between the network output $\widehat{\mathbf{y}}(t)$ and the training response vector $\mathbf{y}$ right after the t-th step of GDP. Let $\tau \leq 1$ be a positive number. For $t \geq 0$ and $T \geq 1$ we define the following quantities: $c_{\mathbf{u}} \coloneqq \Theta(\gamma_0) + \sigma_0 + \tau + 1$,

$$R \coloneqq \frac{\eta c_{\mathbf{u}} T}{\sqrt{m}}, \tag{8}$$

$$\mathcal{V}_t \coloneqq \left\{ \mathbf{v} \in \mathbb{R}^n \colon \mathbf{v} = -\left(\mathbf{I}_n - \eta \mathbf{K}_n \mathbf{P}^{(r_0)}\right)^t f^*(\mathbf{S}) \right\}, \tag{9}$$

$$\mathcal{E}_{t,\tau} \coloneqq \left\{ \mathbf{e} \colon \mathbf{e} = \vec{\mathbf{e}}_1 + \vec{\mathbf{e}}_2 \in \mathbb{R}^n, \vec{\mathbf{e}}_1 = -\left(\mathbf{I}_n - \eta \mathbf{K}_n \mathbf{P}^{(r_0)}\right)^t \mathbf{w}, \left\| \vec{\mathbf{e}}_2 \right\|_2 \leq \sqrt{n}\tau \right\}. \tag{10}$$

In particular, Lemma 17 in the appendix shows that with high probability over the random noise $\mathbf{w}$, the distance of every weighting vector $\mathbf{w}_r(t)$ to its initialization $\mathbf{w}_r(0)$ is bounded by R. In addition, $\mathbf{u}(t)$ can be composed into two vectors, $\mathbf{u}(t) = \mathbf{v}(t) + \mathbf{e}(t)$ such that $\mathbf{v}(t) \in \mathcal{V}_t$ and $\mathbf{e}(t) \in \mathcal{E}_{t,\tau}$.

We then define the set of the neural network weights during the training by GDP using Algorithm 1 as follows:

$$\mathcal{W}(\mathbf{S},\mathbf{W}(0),T) \coloneqq \Bigg\{ \mathbf{W}\colon \exists t\in[T] \text{ s.t. } \mathrm{vec}\left(\mathbf{W}\right) = \mathrm{vec}\left(\mathbf{W}(0)\right) - \sum_{t'=0}^{t-1} \frac{\eta}{n} \mathbf{Z}_{\mathbf{S}}(t')\mathbf{P}^{(r_0)}\mathbf{u}(t'),$$
$$\mathbf{u}(t')\in\mathbb{R}^n, \mathbf{u}(t') = \mathbf{v}(t')+\mathbf{e}(t'), \mathbf{v}(t')\in\mathcal{V}_{t'}, \mathbf{e}(t')\in\mathcal{E}_{t',\tau}, \text{ for all } t'\in[0,t-1] \Bigg\}. \tag{11}$$

We will also show by Lemma 17 that with high probability over $\mathbf{w}$, $\mathcal{W}(\mathbf{S},\mathbf{W}(0),T)$ is the set of the weights of the two-layer NN (1) trained by GDP on the training features $\mathbf{S}$ with the random initialization $\mathbf{W}(0)$ and the number of steps of GDP not greater than T. The set of the functions represented by the neural network with weights in $\mathcal{W}(\mathbf{S},\mathbf{W}(0),T)$ is then defined as

$$\mathcal{F}_{\mathrm{NN}}(\mathbf{S},\mathbf{W}(0),T) \coloneqq \{f_t = f(\mathcal{W}(t),\cdot)\colon \exists\, t\in[T], \mathbf{W}(t)\in\mathcal{W}(\mathbf{S},\mathbf{W}(0),T)\}. \tag{12}$$

We also define the function class $\mathcal{F}(B,w)$ for any $B,w>0$ as

$$\mathcal{F}(B,w,\mathbf{S},r_0) \coloneqq \{f\colon f=h+e, h\in\mathcal{H}_K(B)\cap\mathcal{H}_{\mathbf{S},r_0}, \|e\|_\infty\le w\}. \tag{13}$$

We will show by Theorem 2 in the next subsection that with high probability over $\mathbf{w}$, $\mathcal{F}_{\mathrm{NN}}(\mathbf{S},\mathbf{W}(0),T)$ is a subset of $\mathcal{F}(B,w,\mathbf{S},r_0)$, where a smaller w requires a larger network width m, and $B_h>\gamma_0$ is an absolute positive constant defined by

$$B_h \coloneqq \gamma_0 + \Theta(1). \tag{14}$$

A.2. Detailed Roadmap and Key Results

The summary of the approaches and key technical results in the proofs are presented as follows. Our main result, Theorem 1, is built upon the following three significant technical results of independent interest.

First, using the novel GDP algorithm and the uniform convergence to the NTK (2) during the training process by GDP, we can have a nice decomposition of the neural network function at any step of GDP into a function in a r_0-dimensional subspace of the RKHS associated with the NTK (2), which is $\mathcal{H}_K(B_h)\cap\mathcal{H}_{\mathbf{S},r_0}$, and an error function with a small L^∞-norm. Formally, Theorem 2 states that with high probability over $\mathbf{w}$, $\mathcal{F}_{\mathrm{NN}}(\mathbf{S},\mathbf{W}(0),T)\subseteq\mathcal{F}(B_h,w,\mathbf{S},r_0)$.

Theorem 2 *Suppose $n\ge\Theta(\log(2/\delta)\cdot d^{2k_0})$, $\delta\in(0,1/2)$, $w\in(0,1)$, the network width m satisfies*

$$m \gtrsim \max\left\{T^{\frac{15}{2}} d^{\frac{5}{2}}/w^5, T^{\frac{25}{2}} d^{\frac{5}{2}}\right\}, \tag{15}$$

and the neural network $f_t = f(\mathcal{W}(t),\cdot)$ is trained by GDP using Algorithm 1 with the constant learning rate $\eta=\Theta(1)\in(0,1)$ and the random initialization $\mathbf{W}(0)\in\mathcal{W}_0$. Then for every $t\in[T]$ and every $\delta\in(0,1/2)$, with probability at least $1-2\delta-\exp\left(-\Theta(n)\right)-\exp\left(-\Theta(r_0)\right)$ over the random training features $\mathbf{S}$ the random noise $\mathbf{w}$, $f_t\in\mathcal{F}_{\mathrm{NN}}(\mathbf{S},\mathbf{W}(0),T)$, and f_t has the following decomposition on $\mathcal{X}$:

$$f_t = h_t + e_t, \tag{16}$$

where $h_t\in\mathcal{H}_K(B_h)\cap\mathcal{H}_{\mathbf{S},r_0}$ with B_h defined in (14), $e_t\in L^\infty$ with $\|e_t\|_\infty\le w$.

In particular, with the uniform convergence by Theorem 13 and the optimization results in Lemma 17 and Lemma 21 in the appendix, Theorem 2 shows that with high probability, the neural network function $f(\mathcal{W}(t),\cdot)$ right after the t-th step of GDP can be decomposed into two functions by $f(\mathcal{W}(t),\cdot) = f_t = h + e$, where $h \in \mathcal{H}_K(B_h) \cap \mathcal{H}_{\mathbf{S},r_0}$ is a function in a subspace of finite dimension r_0 of the RKHS associated with K with a bounded $\mathcal{H}_K$-norm. The error function e has a small L^∞-norm, that is, $\|e\|_\infty \le w$ with w being a small number controlled by the network width m, and larger m leads to smaller w.

Second, local Rademacher complexity is employed to tightly bound the risk of nonparametric regression in Theorem 3 below, which is based on the Rademacher complexity of a localized subset of the function class $\mathcal{F}(B_h, w, \mathbf{S}, r_0)$ in Lemma 27 deferred the appendix. We use Theorem 2, Lemma 27, and Lemma 28 deferred to the appendix to prove Theorem 3.

Theorem 3 *Suppose $n \ge \Theta(\log(2/\delta)\cdot d^{2k_0})$, $\delta \in (0,1/2)$, $w \in (0,1)$, m satisfies (15), and the neural network $f_t = f(\mathcal{W}(t),\cdot)$ is trained by GDP using Algorithm 1 with the constant learning rate $\eta = \Theta(1) \in (0,1)$ on the random initialization $\mathbf{W}(0) \in \mathcal{W}_0$. Then for every $t \in [T]$ and every $\delta \in (0,1/2)$, with probability at least $1 - 2\delta - \exp(-\Theta(n)) - 2\exp(-\Theta(r_0))$ over the random noise $\mathbf{w}$, the random training features $\mathbf{S}$ and the random initialization $\mathbf{W}(0)$,*

$$\mathbb{E}_P\left[(f_t - f^*)^2\right] - 2\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] \lesssim \sqrt{\log\frac{2}{\delta}}\cdot\frac{d^{k_0}}{n} + w. \tag{17}$$

Third, we have the following sharp upper bound for the training loss $\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right]$.

Theorem 4 *Suppose the neural network trained after the t-th step of GDP, $f_t = f(\mathcal{W}(t),\cdot)$, satisfies $\mathbf{u}(t) = f_t(\mathbf{S}) - \mathbf{y} = \mathbf{v}(t) + \mathbf{e}(t)$ with $\mathbf{v}(t) \in \mathcal{V}_t$, $\mathbf{e}(t) \in \mathcal{E}_{t,\tau}$. Let $n \ge \Theta(\log(2/\delta)\cdot d^{2k_0})$ and $\delta \in (0,1/2)$. If*

$$\eta \in (0,1), \quad \tau \le \sqrt{\frac{d^{k_0}}{n}}, \tag{18}$$

then for every $t \in [T]$, with probability at least $1 - 2\delta - \exp(-\Theta(r_0))$ over the random training features $\mathbf{S}$ and the random noise $\mathbf{w}$, we have

$$\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] \le \Theta\left(\frac{\gamma_0^2}{\eta t}\right) + \gamma_0^2\log\frac{2}{\delta}\cdot\Theta\left(\frac{d^{k_0}}{n}\right). \tag{19}$$

We then obtain Theorem 1 using the upper bound for the regression risk in (17) of Theorem 3 where w is set to d^{k_0}/n, with the empirical loss $\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right]$ bounded by $\Theta(\log(2/\delta)\cdot d^{k_0}/n)$ with high probability by Theorem 4.

A.3. Novel Proof Strategy

We remark that the proof strategy of our main result, Theorem 1, summarized above is significantly different from the existing works in training over-parameterized neural networks for nonparametric regression with minimax rates [18, 23, 31] and existing works about learning low-degree polynomials [4, 12, 16, 24].

First, GDP is carefully incorporated into the analysis about the uniform convergence results for NTK, leading to the crucial decomposition of the neural network function f_t in Theorem 2. It

is remarked that while existing works such as [23] also has uniform convergence results for over-parameterized neural network, our results about the uniform convergence (in Section D.1 of the appendix) do not depend on the Hölder continuity of the NTK.

Second, to the best of our knowledge, Theorem 3 is the first result about the sharp upper bound of the order $\Theta(\log(2/\delta) \cdot d^{k_0}/n)$ (with $w = d^{k_0}/n$) for the regression risk of the neural network function which has the decomposition in Theorem 2. We note that the RHS of this upper bound (17) is nearly $\Theta(d^{k_0}/n)$, which has the expected and the desired order since the target function is in a r_0-dimensional subspace of the RKHS $\mathcal{H}_K(\gamma_0)$ with $r_0 = \Theta(d^{k_0})$.

Third, a novel method based on the operator theory in RKHS has been developed to derive the sharp upper bound for the training loss in Theorem 4. As shown in Theorem 2, the network function f_t at every step t of GDP is approximately a function in the r_0-dimensional subspace, $\mathcal{H}_K(B_h) \cap \mathcal{H}_{\mathbf{S},r_0}$. We emphasize that while it is intuitive to only learn the r_0-dimensional subspace by projection, $\mathcal{H}_K(B_h) \cap \mathcal{H}_{\mathbf{S},r_0}$, since the target function lies in that function, it has been an open problem in the research community how to handle the incurred training loss by such projection. In particular, as pointed out by the existing work [24], learning in such a subspace leads to better alignment with the target function f^*, however, such alignment incurs additional training loss because the network function f_t only learns the information in such a subspace of dimension $r_0 < n$, and the information in the ground truth signal $f^*(\mathbf{S})$ not in the r_0-dimensional subspace is not learned by f_t. We manage to show that the information of $f^*(\mathbf{S})$ not in the r_0-dimensional subspace, which is $\mathbb{P}_{\mathbf{U}^{(-r_0)}}(f^*(\mathbf{S}))$ where $\mathbb{P}_{\mathbf{U}^{(-r_0)}} = \mathbb{P}_{\mathrm{Span}(\mathbf{U}^{(r_0)})^\perp}$, is sharply bounded in Lemma 24 of the appendix: $\left\|\mathbb{P}_{\mathbf{U}^{(-r_0)}}(f^*(\mathbf{S}))\right\|_2^2 \leq n\gamma_0^2 \log\frac{2}{\delta} \cdot \Theta\left(\frac{d^{k_0}}{n}\right)$. The proof of Lemma 24 relies on a novel result in operator theory developed in this work which is of independent interest in functional analysis. Let $\left\{\Phi^{(k)}\right\}_{k\geq 0}$ be an orthonormal basis of the RKHS $\mathcal{H}_K$ as an extension of the orthonormal basis of the RKHS $\mathcal{H}_{\mathbf{S}} \subseteq \mathcal{H}_K$, $\left\{\Phi^{(k)}\right\}_{k\in[0:n-1]}$. Using the bounded Hilbert-Schmidt norm of $P_{m_{k_0}}^{T_K} - P_{m_{k_0}}^{T_n}$, where the two operators are defined as $P_{m_{k_0}}^{T_K} h = \sum_{j=0}^{m_{k_0}-1} \langle h, v_j\rangle_{\mathcal{H}} v_j, P_{m_{k_0}}^{T_n} h = \sum_{j=0}^{m_{k_0}-1} \left\langle h, \Phi^{(j)}\right\rangle_{\mathcal{H}} \Phi^{(j)}$ for all $h \in \mathcal{H}_K$, we can prove the following theorem showing the bounded projection of f^* on the eigenfunctions $\left\{\Phi^{(q)}\right\}_{q\geq r_0}$:

Theorem 5 *With probability at least* $1-\delta$, $\sum_{q=r_0}^{\infty} \left\langle f^*, \Phi^{(q)}\right\rangle_{\mathcal{H}_K}^2 \leq \zeta_{n,\gamma_0,r_0,\delta} := \frac{32\gamma_0^2 \log\frac{2}{\delta}}{\left(\mu_{k_0}-\mu_{k_0+1}\right)^2 n}$.

Theorem 5 proves Lemma 24, which in turn proves Theorem 4.

A.4. Beyond the Regular NTK Limit

We remark that while an over-parameterized neural network is trained, our result goes beyond the regular NTK limit due to our new GDP algorithm. As shown in Theorem 2, the novel projection operator $\mathbf{P}^{(r_0)}$ in GDP ensures that the neural network function almost lies in a r_0-dimensional subspace of the RKHS $\mathcal{H}_K(\gamma_0)$ with $r_0 = \Theta(d^{k_0})$. Although such projection loses all the information of the ground truth signal $f^*(\mathbf{S})$ not lying in such a subspace, Theorem 4 shows that such information loss due to the projection is small enough to ensure a sharp regression risk bound. In contrast, the regular NTK-based analysis with vanilla GD must account for all eigenspaces associated with the NTK, and therefore cannot achieve such a sharp rate.

Simulation Results. Section E of the appendix provides the simulation results showing that a two-layer NN (1) trained by GDP always has lower test losses across different training data size than the vanilla gradient descent for learning a spherical polynomial with degree $k_0 \in \{1, 2, 3\}$.

Appendix B. Mathematical Tools

B.1. Concentration Inequalities for Supremum of Empirical Processes

The Rademacher complexity of a function class and its empirical version are defined below.

Definition 6 *Let $\boldsymbol{\sigma} = \{\sigma_i\}_{i=1}^n$ be n i.i.d. random variables such that $\Pr[\sigma_i = 1] = \Pr[\sigma_i = -1] = \frac{1}{2}$. The Rademacher complexity of a function class $\mathcal{F}$ is defined as*

$$\mathfrak{R}(\mathcal{F}) = \mathbb{E}_{\left\{\vec{\mathbf{x}}_i\right\}_{i=1}^n, \{\sigma_i\}_{i=1}^n} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(\vec{\mathbf{x}}_i) \right]. \tag{20}$$

The empirical Rademacher complexity is defined as

$$\widehat{\mathfrak{R}}(\mathcal{F}) = \mathbb{E}_{\{\sigma_i\}_{i=1}^n} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(\vec{\mathbf{x}}_i) \right], \tag{21}$$

For simplicity of notations, Rademacher complexity and empirical Rademacher complexity are also denoted by $\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(\vec{\mathbf{x}}_i)\right]$ and $\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(\vec{\mathbf{x}}_i)\right]$, respectively.

For data $\left\{\vec{\mathbf{x}}\right\}_{i=1}^n$ and a function class $\mathcal{F}$, we define the notation $R_n\mathcal{F}$ by $R_n\mathcal{F} \coloneqq \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(\vec{\mathbf{x}}_i)$. We have the contraction property for Rademacher complexity, which is due to Ledoux and Talagrand [22].

Theorem 7 *Let ϕ be a contraction,that is, $|\phi(x) - \phi(y)| \leq \mu |x - y|$ for $\mu > 0$. Then, for every function class $\mathcal{F}$,*

$$\mathbb{E}_{\{\sigma_i\}_{i=1}^n} \left[R_n \phi \circ \mathcal{F} \right] \leq \mu \mathbb{E}_{\{\sigma_i\}_{i=1}^n} \left[R_n \mathcal{F} \right], \tag{22}$$

where $\phi \circ \mathcal{F}$ is the function class defined by $\phi \circ \mathcal{F} = \{\phi \circ f \colon f \in \mathcal{F}\}$.

Definition 8 (Sub-root function,[5, Definition 3.1]) *A function $\psi \colon [0, \infty) \to [0, \infty)$ is sub-root if it is nonnegative, nondecreasing and if $\frac{\psi(r)}{\sqrt{r}}$ is nonincreasing for $r > 0$.*

Theorem 9 ([5, Theorem 3.3]) *Let $\mathcal{F}$ be a class of functions with ranges in $[a, b]$ and assume that there are some functional $T \colon \mathcal{F} \to \mathbb{R}+$ and some constant $\bar{B}$ such that for every $f \in \mathcal{F}$, $\operatorname{Var}[f] \leq T(f) \leq \bar{B}P(f)$. Let ψ be a sub-root function and let r^* be the fixed point of ψ. Assume that ψ satisfies that, for any $r \geq r^*$, $\psi(r) \geq \bar{B}\mathfrak{R}(\{f \in \mathcal{F} \colon T(f) \leq r\})$. Fix $x > 0$, then for any $K_0 > 1$, with probability at least $1 - e^{-x}$,*

$$\forall f \in \mathcal{F}, \quad \mathbb{E}_P[f] \leq \frac{K_0}{K_0 - 1} \mathbb{E}_{P_n}[f] + \frac{704 K_0}{\bar{B}} r^* + \frac{x\left(11(b-a) + 26\bar{B}K_0\right)}{n}.$$

Also, with probability at least $1 - e^{-x}$,

$$\forall f \in \mathcal{F}, \quad \mathbb{E}_{P_n}[f] \leq \frac{K_0 + 1}{K_0} \mathbb{E}_P[f] + \frac{704 K_0}{\bar{B}} r^* + \frac{x\left(11(b-a) + 26\bar{B}K_0\right)}{n}.$$

Appendix C. Detailed Technical Background about Harmonic Analysis on Spheres

In this section, we provide background materials on spherical harmonic analysis needed for our study of the RKHS. We refer the reader to [9, 15, 32] for further information on these topics. As mentioned above, expansions in spherical harmonics were used in the past in the statistics literature, such as [3, 6].

With $\ell \geq 0$, let $\mathcal{P}_\ell^{(\mathrm{hom})}$ denote the space of all the degree-ℓ homogeneous polynomials on $\mathcal{X} = \mathbb{S}^{d-1}$, and let $\mathcal{H}_\ell$ denote the space of degree-ℓ homogeneous harmonic polynomials on $\mathcal{X}$, or the degree-ℓ spherical harmonics. That is,

$$\mathcal{H}_\ell = \left\{ P \colon \mathcal{X} \to \mathbb{R} \colon P(\mathbf{x}) = \sum_{|\alpha|=\ell} c_\alpha \mathbf{x}^\alpha, \Delta P = 0 \right\}, \tag{23}$$

where $\alpha = [\alpha_1, \ldots, \alpha_d]$, $\mathbf{x}^\alpha = \prod_{i=1}^d \mathbf{x}_i^{\alpha_i}$, $|\alpha| = \sum_{i=1}^d \alpha_i$, and Δ is the Laplacian operator. For $\ell \neq \ell'$, the elements of $\mathcal{H}_\ell$ and $\mathcal{H}_{\ell'}$ are orthogonal to each other. All the functions in the following text of this section are assumed to be elements of $L^2(\mathcal{X}, v_{d-1})$, where v_{d-1} standards for the uniform distribution on the sphere $\mathcal{X} = \mathbb{S}^{d-1}$. We have $\langle f, g \rangle_{L^2} \coloneqq \int_{\mathcal{X}} f(x) g(x) \mathrm{d}v_{d-1}(x)$. We denote by $\{Y_{kj}\}_{j \in [N(d,k)]}$ the spherical harmonics of degree k which form an orthogonal basis of $\mathcal{H}_k$, where $N(d,k) = \frac{2k+d-2}{k}\binom{k+d-3}{d-2}$ is the dimension of $\mathcal{H}_k$. They form a orthonormal basis of $L^2(\mathcal{X}, v_{d-1})$. We have $\sum_{j=1}^{N(d,k)} Y_{kj}(\mathbf{x}) Y_{kj}(\mathbf{x}') = N(d,k) P_k(\langle \mathbf{x}, \mathbf{x}' \rangle)$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, where P_k is the k-th Legendre polynomial in dimension d, which is also known as Gegenbauer polynomials, given by the Rodrigues formula:

$$P_k(t) = (-\frac{1}{2})^k \frac{\Gamma\left(\frac{d-1}{2}\right)}{\Gamma\left(k + \frac{d-1}{2}\right)} \left(1 - t^2\right)^{(3-d)/2} \left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^k \left(1 - t^2\right)^{k+(d-3)/2}.$$

The polynomials $\{P_k\}$ are orthogonal in $L^2(\mathcal{X}, \mathrm{d}v_{d-1})$ where the measure $\mathrm{d}v_{d-1}$ is given by $\mathrm{d}v_{d-1}(t) = (1-t^2)^{(d-3)/2}\mathrm{d}t$, and we have

$$\int_{-1}^{1} P_k^2(t)(1-t^2)^{(d-3)/2}\mathrm{d}t = \frac{w_{d-1}}{w_{d-2}} \frac{1}{N(d,k)},$$

where $w_{d-1} \coloneqq \frac{2\pi^{d/2}}{\Gamma(d/2)}$ denotes the surface of the unit sphere $\mathbb{S}^{d-1}$. It follows from the orthogonality of spherical harmonics that

$$\int_{\mathcal{X}} P_j(\langle \mathbf{x}, \mathbf{w} \rangle) P_j(\langle \mathbf{x}', \mathbf{w} \rangle) \mathrm{d}v_{d-1}(\mathbf{w}) = \frac{\delta_{jk}}{N(d,k)} P_k(\langle \mathbf{x}, \mathbf{x}' \rangle),$$

where $\delta_{jk} = \mathbb{1}_{\{j=k\}}$. We have the following recurrence relation [15, Equation 4.36],

$$tP_k(t) = \frac{k}{2k+d-2} P_{k-1}(t) + \frac{k+d-2}{2k+d-2} P_{k+1}(t)$$

for all $k \geq 1$, and $tP_0(t) = P_1(t)$.

The Funk-Hecke formula is helpful for computing Fourier coefficients in the basis of spherical harmonics in terms of Legendre polynomials. For any $j \in [N(d,k)]$, we have

$$\int_{\mathcal{X}} f(\langle \mathbf{x}, \mathbf{x}' \rangle) Y_{kj}(\mathbf{x}') \mathrm{d}v_{d-1}(\mathbf{x}') = \frac{w_{d-2}}{w_{d-1}} Y_{kj}(\mathbf{x}) \int_{-1}^{1} f(t) P_k(t) (1-t^2)^{(d-3)/2} \mathrm{d}t.$$

For a positive-definite kernel $\tilde{K}(\mathbf{x}, \mathbf{x}') = \kappa(\langle \mathbf{x}, \mathbf{x}' \rangle)$ defined on $\mathcal{X}$, we have its Mercer decomposition as follows.

$$\tilde{K}(\mathbf{x}, \mathbf{x}') = \sum_{\ell \geq 0} \mu_\ell \sum_{j=1}^{N(d,\ell)} Y_{\ell,j}(\mathbf{x}) Y_{\ell,j}(\mathbf{x}') = \sum_{\ell \geq 0} \mu_\ell N(d,\ell) P_\ell(\langle \mathbf{x}, \mathbf{x}' \rangle),$$

where μ_ℓ is the eigenvalue of the integral operator $T_{\tilde{K}}$ associated with $\tilde{K}$ corresponding to $\mathcal{H}_\ell$. It follows that

$$\mu_\ell = \frac{w_{d-2}}{w_{d-1}} \int_{-1}^{1} \kappa(t) P_\ell(t) (1-t^2)^{(d-3)/2} \mathrm{d}t.$$

The above equation will be used to compute the eigenvalues of the PSD kernels define in (2) in Section D.6 of this appendix.

Proposition 10 ([20, Theorem 4.2]) *Let $p \in \mathcal{P}_\ell^{(\mathrm{hom})}$. Then there exists unique $h_{n-2i} \in \mathcal{H}_{n-2i}$ for $i \in \{0, 1, \ldots, \lfloor n/2 \rfloor\}$ such that*

$$p(\mathbf{x}) = h_n + h_{n-2} + \ldots + h_{n-2k}.$$

Theorem 11 *Every polynomial p defined on $\mathbb{S}^{d-1}$ of degree k for $k \geq 0$ can be represented as a linear combination of homogeneous harmonic polynomials up to degree k, that is,*

$$p = \sum_{i=0}^{k} c_i p_i,$$

where $p_i \in \mathcal{H}_i$ for $i \in \{0, 1, \ldots, k\}$.

Proof Every polynomial p defined on $\mathbb{S}^{d-1}$ of degree k can be represented as the sum of homogeneous polynomials on $\mathbb{S}^{d-1}$ by grouping the terms of p of the same degree together. It follows from Proposition 10 that every homogeneous polynomial is a linear combination of homogeneous harmonic polynomials up to degree k. As a result, the conclusion holds. ■

Lemma 12 (Estimation for $r_0 = m_{k_0}$) *For $k_0 = \Theta(1)$ and $d > \Theta(1)$, we have*

$$r_0 = \Theta(d^{k_0}). \tag{24}$$

Proof It follows from the direct calculation that $N(d,\ell) \asymp d^\ell$ under the given conditions, so that $r_0 = \sum_{\ell=0}^{k_0} N(d,\ell) \asymp d^{k_0}$. ■

Appendix D. Detailed Proofs

In Section D.4, we present the proofs of Theorem 4, and the lemmas required for the proofs in Section D.3.

D.1. Uniform Convergence to the NTK (2) and More

We define the following functions with $\mathbf{W} = \{\mathbf{w}_r\}_{r=1}^m$:

$$h(\mathbf{w}, \mathbf{u}, \mathbf{v}) := \mathbb{I}_{\{\mathbf{w}^\top \mathbf{u} \geq 0\}} \mathbb{I}_{\{\mathbf{w}^\top \mathbf{v} \geq 0\}}, \qquad \widehat{h}(\mathbf{W}, \mathbf{u}, \mathbf{v}) := \frac{1}{m} \sum_{r=1}^m h(\overrightarrow{\mathbf{w}}_r, \mathbf{u}, \mathbf{v}), \tag{25}$$

$$v_R(\mathbf{w}, \mathbf{u}) := \mathbb{I}_{\{|\mathbf{w}^\top \mathbf{u}| \leq R\}}, \qquad \widehat{v}_R(\mathbf{W}, \mathbf{u}) := \frac{1}{m} \sum_{r=1}^m v_R(\overrightarrow{\mathbf{w}}_r, \mathbf{u}), \tag{26}$$

where $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$. Then we have the following theorem stating the uniform convergence of $\widehat{h}(\mathbf{W}(0), \cdot, \cdot)$ to $K(\cdot, \cdot)$ and uniform convergence of $\widehat{v}_R(\mathbf{W}(0), \cdot)$ to $\frac{2R}{\sqrt{2\pi\kappa}}$ for a positive number $R \lesssim \eta T/\sqrt{m}$, and R is formally defined in (8). It is remarked that while existing works such as [23] also has uniform convergence results for over-parameterized neural network, our result does not depend on the Hölder continuity of the NTK.

Theorem 13 *The following results hold with $\eta \lesssim 1$, $m \gtrsim \max\left\{n^{2/d}, \Theta(T^{\frac{5}{3}})\right\}$, and $m/\log m \geq d$.*

(1) With probability at least $1 - 1/n$ over the random initialization $\mathbf{W}(0) = \left\{\overrightarrow{\mathbf{w}}_r(0)\right\}_{r=1}^m$,

$$\sup_{\mathbf{u} \in \mathcal{X}, \mathbf{v} \in \mathcal{X}} \left| K^{(\alpha)}(\mathbf{u}, \mathbf{v}) - \widehat{h}(\mathbf{W}(0), \mathbf{u}, \mathbf{v}) \right| \leq C_1(m/2, d, 1/n) \lesssim \sqrt{\frac{d \log m}{m}}, \alpha \in \{0, 1\}. \tag{27}$$

(2) With probability at least $1 - 1/n$ over the random initialization $\mathbf{W}(0) = \left\{\overrightarrow{\mathbf{w}}_r(0)\right\}_{r=1}^m$,

$$\sup_{\mathbf{u} \in \mathcal{X}} \widehat{v}_R(\mathbf{W}(0), \mathbf{u}) \leq \frac{2R}{\sqrt{2\pi\kappa}} + C_2(m/2, d, 1/n) \lesssim \sqrt{d} m^{-\frac{1}{5}} T^{\frac{1}{2}}, \tag{28}$$

where $C_1(m/2, d, 1/n), C_2(m/2, d, 1/n)$ are two positive numbers depending on (m, d, n), with their formal definitions deferred to (36) and (38) in Section D.3.

Proof This theorem follows from Theorem 15 and Theorem 16 in Section D.3. We note that

$$\widehat{h}(\mathbf{W}, \mathbf{u}, \mathbf{v}) = \frac{1}{m} \sum_{r=1}^m h(\overrightarrow{\mathbf{w}}_r, \mathbf{u}, \mathbf{v}) = \frac{1}{m/2} \sum_{r'=1}^{m/2} h(\overrightarrow{\mathbf{w}}_{2r'}(0), \mathbf{u}, \mathbf{v}),$$

then the first bound part (1) for $K^{(0)}$ directly follows from Theorem 15. Moreover, since $K^{(1)}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top \mathbf{v} K^{(0)}(\mathbf{u}, \mathbf{v})$, we have

$$\left| K^{(1)}(\mathbf{u}, \mathbf{v}) - \mathbf{u}^\top \mathbf{v} \cdot \widehat{h}(\mathbf{W}(0), \mathbf{u}, \mathbf{v}) \right| \leq \sup_{\mathbf{u} \in \mathcal{X}, \mathbf{v} \in \mathcal{X}} \left| K^{(0)}(\mathbf{u}, \mathbf{v}) - \widehat{h}(\mathbf{W}(0), \mathbf{u}, \mathbf{v}) \right|,$$

which leads to the second bound in part (1) for $K^{(1)}$. Part (2) directly follows from Theorem 16. ■

We define

$$\mathcal{W}_0 := \{\mathbf{W}(0) \colon (27), (28) \text{ hold}\} \tag{29}$$

as the set of all the good random initializations which satisfy (27) and (28) in Theorem 13. Theorem 13 shows that we have good random initialization with high probability, that is, $\Pr[\mathbf{W}(0) \in \mathcal{W}_0] \geq 1 - 2/n$. When $\mathbf{W}(0) \in \mathcal{W}_0$, the uniform convergence results, (27) and (28), hold with high probability, which is important for the analysis of the training dynamics of the two-layer NN (1) by GD.

D.2. Proofs for the Main Result, Theorem 1

Proof [**Proof of Theorem 1**] We use Theorem 3 and Theorem 4 to prove this theorem.

First of all, it follows by Theorem 4 that with probability at least $1 - 2\delta - \exp(-\Theta(r_0))$ over $\mathbf{S}$ and $\mathbf{w}$,

$$\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] \leq \Theta\left(\frac{\gamma_0^2}{\eta t}\right) + \gamma_0^2 \log\frac{2}{\delta} \cdot \Theta\left(\frac{d^{k_0}}{n}\right).$$

Plugging such bound for $\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right]$ in (17) of Theorem 3 leads to

$$\mathbb{E}_{P}\left[(f_t - f^*)^2\right] \lesssim \Theta\left(\frac{\gamma_0^2}{\eta t}\right) + \log\frac{2}{\delta} \cdot \frac{d^{k_0}}{n} + w. \tag{30}$$

Due to the definition of $T \asymp n/d^{k_0}$, we have

$$\frac{1}{\eta t} \asymp \frac{1}{\eta T} \asymp \frac{d^{k_0}}{n}. \tag{31}$$

We also have $\Pr[\mathcal{W}_0] \geq 1 - 2/n$. Let $w = d^{k_0}/n$, then that $w \in (0, 1)$ with $n > d^{k_0}$. (6) then follows from (30) with $w = d^{k_0}/n$, (31) and the union bound. We note that $c_{\mathbf{u}}$ is bounded by a positive constant, so that the condition on m in (15) in Theorem 2, together with $w = d^{k_0}/n$ and (31) leads to the condition on m in (5).

■

D.3. Proofs for Results in Section A.2

We present our key technical results regarding optimization and generalization of the two-layer NN (1) trained by GDP in this section. The following theorem,

Theorem 14 *Suppose K is a continuous and positive definite kernel on $\mathcal{X} \times \mathcal{X}$, and the target function $f^* \in \mathcal{H}_K(\gamma_0)$ is spanned by the orthogonal set $\{v_j\}_{j=0}^{r_0-1}$ in the first k_0 eigenspaces of T_K with $k_0 \geq 1$ and $r_0 = m_{k_0}$. That is,*

$$f^* = \sum_{j=0}^{r_0-1} \beta_j v_j, \quad \sum_{j=0}^{r_0-1} \beta_j^2 \leq \gamma_0^2. \tag{32}$$

Then with probability at least $1 - \delta$ over the random training features $\mathbf{S}$,

$$\sum_{q=r_0}^{\infty} \left\langle f^*, \Phi^{(q)} \right\rangle_{\mathcal{H}_K}^2 \leq \frac{32\gamma_0^2 \log\frac{2}{\delta}}{\left(\mu_{k_0} - \mu_{k_0+1}\right)^2 n} := \zeta_{n,\gamma_0,r_0,\delta}. \tag{33}$$

Similarly, for every $f \in \mathcal{F}(B_h, w, \mathbf{S}, r_0)$, with probability at least $1-\delta$ over the random training features $\mathbf{S}$,

$$\sum_{q=r_0}^{\infty} \langle f, v_q \rangle_{\mathcal{H}_K}^2 \leq \zeta_{n,B_h,r_0,\delta}. \tag{34}$$

D.3.1. RESULTS ABOUT UNIFORM CONVERGENCE

We have the following two theorems, Theorem 15 and Theorem 16, regarding the uniform convergence to the PSD kernel $K^{(0)}$ defined in (2) and the uniform convergence of $\widehat{v}_R$ to $\frac{2R}{\sqrt{2\pi\kappa}}$ on the unit sphere $\mathcal{X}$.

Theorem 15 (Adapted from [37, Theorem 6.1],[38, Theorem VI.7]) *Let $\mathbf{W}(0) = \left\{\overrightarrow{\mathbf{w}}_r(0)\right\}_{r=1}^m$, where each $\overrightarrow{\mathbf{w}}_r(0) \sim \mathcal{N}(\mathbf{0}, \kappa^2 \mathbf{I}_d)$ for $r \in [m]$. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$ over $\mathbf{W}(0)$,*

$$\sup_{\mathbf{u}\in\mathcal{X}, \mathbf{v}\in\mathcal{X}} \left| K^{(0)}(\mathbf{u}, \mathbf{v}) - \widehat{h}(\mathbf{W}(0), \mathbf{u}, \mathbf{v}) \right| \leq C_1(m, d, \delta), \tag{35}$$

where

$$C_1(m,d,\delta) \coloneqq \frac{1}{\sqrt{m}} \left(6(1+2B\sqrt{d}) + \sqrt{2\log\frac{(1+2m)^{2d}}{\delta}} \right) + \frac{7\log\frac{(1+2m)^{2d}}{\delta}}{3m}, \tag{36}$$

and B is an absolute positive constant. In addition, when $m \gtrsim n^{1/(2d)}$, $m/\log m \geq d$, and $\delta \asymp 1/n$, $C_1(m,d,\delta) \lesssim \sqrt{\frac{d\log m}{m}} + \frac{d\log m}{m} \lesssim \sqrt{\frac{d\log m}{m}}$.

Theorem 16 ([37, Theorem 6.1],[38, Theorem VI.8]) *Let $\mathbf{W}(0) = \left\{\overrightarrow{\mathbf{w}}_r(0)\right\}_{r=1}^m$, where each $\overrightarrow{\mathbf{w}}_r(0) \sim \mathcal{N}(\mathbf{0}, \kappa^2 \mathbf{I}_d)$ for $r \in [m]$. Suppose $\eta \lesssim 1$, $m \gtrsim 1$. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$ over $\mathbf{W}(0)$,*

$$\sup_{\mathbf{x}\in\mathcal{X}} \left| \widehat{v}_R(\mathbf{W}(0), \mathbf{x}) - \frac{2R}{\sqrt{2\pi\kappa}} \right| \leq C_2(m,d,\delta), \tag{37}$$

where

$$C_2(m,d,\delta) \coloneqq 3\sqrt{\frac{d}{\kappa}} m^{-\frac{1}{5}} T^{\frac{1}{2}} + \sqrt{\frac{2\log\frac{(1+2\sqrt{m})^d}{\delta}}{m}} + \frac{7\log\frac{(1+2\sqrt{m})^d}{\delta}}{3m}. \tag{38}$$

In addition, when $m \gtrsim n^{2/d}$, $m/\log m \geq d$, and $\delta \asymp 1/n$, $C_2(m,d,\delta) \lesssim \sqrt{d} m^{-\frac{1}{5}} T^{\frac{1}{2}}$.

D.3.2. PROOF OF THEOREM 2

We prove Theorem 2 in this subsection. The proof requires the following theorem, Lemma 17, about our main result about the optimization of the network (1). Lemma 17 states that with high probability over the random noise $\mathbf{w}$, the weights of the network $\mathbf{W}(t)$ obtained right after the t-th step of GD using Algorithm 1 belongs to $\mathcal{W}(\mathbf{S}, \mathbf{W}(0), T)$. Furthermore, every weighing vector $\mathbf{w}_r$ has bounded distance to the initialization $\mathbf{w}_r(0)$. The proof of Lemma 17 is based on Lemma 18, Lemma 19, and Lemma 20 deferred to Section D.4 of the appendix.

Lemma 17 *Suppose* $\delta \in (0, 1/2)$,

$$m \gtrsim T^{\frac{15}{2}} d^{\frac{5}{2}} / \tau^5, \tag{39}$$

the neural network $f(\mathcal{W}(t), \cdot)$ *trained by GDP using Algorithm 1 with the constant learning rate* $\eta = \Theta(1) \in (0, 1)$, *the random initialization* $\mathbf{W}(0) \in \mathcal{W}_0$. *Then for every* $\delta \in (0, 1/2)$ *with probability at least* $1 - 2\delta - \exp(-\Theta(n))$ *over the random training features* $\mathbf{S}$ *and the random noise* $\mathbf{w}$, $\mathbf{W}(t) \in \mathcal{W}(\mathbf{S}, \mathbf{W}(0), T)$ *for every* $t \in [T]$. *Moreover, for every* $t \in [0, T]$, $\mathbf{u}(t) = \mathbf{v}(t) + \mathbf{e}(t)$ *where* $\mathbf{u}(t) = \widehat{\mathbf{y}}(t) - \mathbf{y}$, $\mathbf{v}(t) \in \mathcal{V}_t$, $\mathbf{e}(t) \in \mathcal{E}_{t,\tau}$, $\|\mathbf{u}(t)\|_2 \le c_{\mathbf{u}}\sqrt{n}$, *and* $\left\|\vec{\mathbf{w}}_r(t) - \vec{\mathbf{w}}_r(0)\right\|_2 \le R$.

Proof [**Proof of Theorem 2**] In this proof we abbreviate f_t as f and $\mathbf{W}(t)$ as $\mathbf{W}$. It follows from Lemma 17 and its proof that conditioned on an event Ω with probability at least $1 - 2\delta - \exp(-\Theta(n))$, $f \in \mathcal{F}_{\mathrm{NN}}(\mathbf{S}, \mathbf{W}(0), T)$ with $\mathbf{W}(0) \in \mathcal{W}_0$. Moreover, $f = f(\mathcal{W}, \cdot)$ with $\mathbf{W} = \left\{\vec{\mathbf{w}}_r\right\}_{r=1}^m \in \mathcal{W}(\mathbf{S}, \mathbf{W}(0), T)$, and $\mathrm{vec}(\mathbf{W}) = \mathrm{vec}(\mathbf{W}_{\mathbf{S}}) = \mathrm{vec}(\mathbf{W}(0)) - \sum_{t'=0}^{t-1} \eta/n \cdot \mathbf{Z}_{\mathbf{S}}(t')\mathbf{u}(t')$ for some $t \in [T]$, where $\mathbf{u}(t') \in \mathbb{R}^n$, $\mathbf{u}(t') = \mathbf{v}(t') + \mathbf{e}(t')$ with $\mathbf{v}(t') \in \mathcal{V}_{t'}$ and $\mathbf{e}(t') \in \mathcal{E}_{t',\tau}$ for all $t' \in [0, t-1]$. It also follows from Lemma 17 that conditioned on Ω, $\left\|\vec{\mathbf{w}}_r(t) - \vec{\mathbf{w}}_r(0)\right\|_2 \le R$ for all $t \in [T]$.

$\vec{\mathbf{w}}_r$ is expressed as

$$\vec{\mathbf{w}}_r = \vec{\mathbf{w}}_{\mathbf{S},r}(t) = \vec{\mathbf{w}}_r(0) - \sum_{t'=0}^{t-1} \frac{\eta}{n} \left[\mathbf{Z}_{\mathbf{S}}(t')\right]_{[(r-1)d+1:rd]} \mathbf{P}^{(r_0)} \mathbf{u}(t'), \tag{40}$$

where the notation $\vec{\mathbf{w}}_{\mathbf{S},r}$ emphasizes that $\vec{\mathbf{w}}_r$ depends on the training features $\mathbf{S}$. We define the event

$$E_r(R) := \left\{\left|\vec{\mathbf{w}}_r(0)^\top \mathbf{x}\right| \le R\right\}, \quad \bar{E}_r(R) := \left\{\left|\vec{\mathbf{w}}_r(0)^\top \mathbf{x}\right| > R\right\}, \quad r \in [m].$$

We now approximate $f(\mathcal{W}, \mathbf{x})$ by $g(\mathbf{x}) := \frac{1}{\sqrt{m}} \sum_{r=1}^m a_r \mathbb{1}_{\left\{\vec{\mathbf{w}}_r(0)^\top \mathbf{x} \ge 0\right\}} \vec{\mathbf{w}}_r^\top \mathbf{x}$. We have

$$\begin{aligned}
|f(\mathcal{W}, \mathbf{x}) - g(\mathbf{x})| &= \frac{1}{\sqrt{m}} \left|\sum_{r=1}^m a_r \sigma\left(\vec{\mathbf{w}}_r^\top \mathbf{x}\right) - \sum_{r=1}^m a_r \mathbb{1}_{\left\{\vec{\mathbf{w}}_r(0)^\top \mathbf{x} \ge 0\right\}} \vec{\mathbf{w}}_r^\top \mathbf{x}\right| \\
&\le \frac{1}{\sqrt{m}} \sum_{r=1}^m \left|a_r \left(\mathbb{1}_{\{E_r(R)\}} + \mathbb{1}_{\{\bar{E}_r(R)\}}\right) \left(\sigma\left(\vec{\mathbf{w}}_r^\top \mathbf{x}\right) - \mathbb{1}_{\left\{\vec{\mathbf{w}}_r(0)^\top \mathbf{x} \ge 0\right\}} \vec{\mathbf{w}}_r^\top \mathbf{x}\right)\right| \\
&= \frac{1}{\sqrt{m}} \sum_{r=1}^m \mathbb{1}_{\{E_r(R)\}} \left|\sigma\left(\vec{\mathbf{w}}_r^\top \mathbf{x}\right) - \mathbb{1}_{\left\{\vec{\mathbf{w}}_r(0)^\top \mathbf{x} \ge 0\right\}} \vec{\mathbf{w}}_r^\top \mathbf{x}\right| \\
&= \frac{1}{\sqrt{m}} \sum_{r=1}^m \mathbb{1}_{\{E_r(R)\}} \left|\sigma\left(\vec{\mathbf{w}}_r^\top \mathbf{x}\right) - \sigma\left(\vec{\mathbf{w}}_r(0)^\top \mathbf{x}\right) - \mathbb{1}_{\left\{\vec{\mathbf{w}}_r(0)^\top \mathbf{x} \ge 0\right\}} (\vec{\mathbf{w}}_r - \vec{\mathbf{w}}_r(0))^\top \mathbf{x}\right| \\
&\le \frac{2R}{\sqrt{m}} \sum_{r=1}^m \mathbb{1}_{\{E_r(R)\}},
\end{aligned} \tag{41}$$

where first inequality follows from $\mathbb{1}_{\{\bar{E}_r(R)\}}\left(\sigma\left(\vec{\mathbf{w}}_r^\top\mathbf{x}\right)-\mathbb{1}_{\left\{\vec{\mathbf{w}}_r(0)^\top\mathbf{x}\geq 0\right\}}\vec{\mathbf{w}}_r^\top\mathbf{x}\right)=0$. Plugging $R=\frac{\eta c_{\mathbf{u}}T}{\sqrt{m}}$ in (41), since $\mathbf{W}(0)\in\mathcal{W}_0$, we have

$$\sup_{\mathbf{x}\in\mathcal{X}}|f(\mathcal{W},\mathbf{x})-g(\mathbf{x})|\leq 2\eta c_{\mathbf{u}}T\cdot\frac{1}{m}\sum_{r=1}^{m}\mathbb{1}_{\{E_r(R)\}}\leq 2\eta c_{\mathbf{u}}T\left(\frac{2R}{\sqrt{2\pi\kappa}}+C_2(m/2,d,1/n)\right). \tag{42}$$

Using (40), $g(\mathbf{x})$ is expressed as

$$\begin{aligned}
&g(\mathbf{x})\\
&=\frac{1}{\sqrt{m}}\sum_{r=1}^{m}a_r\sigma(\vec{\mathbf{w}}_r(0)^\top\mathbf{x})-\sum_{t'=0}^{t-1}\frac{1}{\sqrt{m}}\sum_{r=1}^{m}\mathbb{1}_{\left\{\vec{\mathbf{w}}_r(0)^\top\mathbf{x}\geq 0\right\}}\left(\frac{\eta}{n}\left[\mathbf{Z}_{\mathbf{S}}(t')\right]_{[(r-1)d+1:rd]}\mathbf{P}^{(r_0)}\mathbf{u}(t')\right)^\top\mathbf{x}\\
&\overset{①}{=}-\sum_{t'=0}^{t-1}\underbrace{\frac{\eta}{nm}\sum_{r=1}^{m}\mathbb{1}_{\left\{\vec{\mathbf{w}}_r(0)^\top\mathbf{x}\geq 0\right\}}\sum_{j=1}^{n}\mathbb{1}_{\left\{\vec{\mathbf{w}}_r(t')^\top\vec{\mathbf{x}}_j\geq 0\right\}}\left[\mathbf{P}^{(r_0)}\mathbf{u}(t')\right]_j\vec{\mathbf{x}}_j^\top\mathbf{x}}_{:=G_{t'}(\mathbf{x})},
\end{aligned}\tag{43}$$

where ① follows from the fact that $\frac{1}{\sqrt{m}}\sum_{r=1}^{m}a_r\sigma(\vec{\mathbf{w}}_r(0)^\top\mathbf{x})=f(\mathcal{W}(0),\mathbf{x})=0$ due to the particular initialization of the two-layer NN (1). For each $G_{t'}$ in the RHS of (43), we have

$$\begin{aligned}
G_{t'}(\mathbf{x})&\overset{②}{=}\frac{\eta}{nm}\sum_{r=1}^{m}\mathbb{1}_{\left\{\vec{\mathbf{w}}_r(0)^\top\mathbf{x}\geq 0\right\}}\sum_{j=1}^{n}\left(d_{t',r,j}+\mathbb{1}_{\left\{\vec{\mathbf{w}}_r(0)^\top\vec{\mathbf{x}}_j\geq 0\right\}}\right)\left[\mathbf{P}^{(r_0)}\mathbf{u}(t')\right]_j\vec{\mathbf{x}}_j^\top\mathbf{x}\\
&\overset{③}{=}\frac{\eta}{n}\sum_{j=1}^{n}K(\mathbf{x},\vec{\mathbf{x}}_j)\left[\mathbf{P}^{(r_0)}\mathbf{u}(t')\right]_j+\underbrace{\frac{\eta}{n}\sum_{j=1}^{n}q_j\left[\mathbf{P}^{(r_0)}\mathbf{u}(t')\right]_j}_{:=E_1(\mathbf{x})}\\
&\quad+\underbrace{\frac{\eta}{nm}\sum_{r=1}^{m}\mathbb{1}_{\left\{\vec{\mathbf{w}}_r(0)^\top\mathbf{x}\geq 0\right\}}\sum_{j=1}^{n}d_{t',r,j}\left[\mathbf{P}^{(r_0)}\mathbf{u}(t')\right]_j\vec{\mathbf{x}}_j^\top\mathbf{x}}_{:=E_2(\mathbf{x})}.
\end{aligned}\tag{44}$$

where $d_{t',r,j}:=\mathbb{1}_{\left\{\vec{\mathbf{w}}_r(t')^\top\vec{\mathbf{x}}_j\geq 0\right\}}-\mathbb{1}_{\left\{\vec{\mathbf{w}}_r(0)^\top\vec{\mathbf{x}}_j\geq 0\right\}}$ in ②, and $q_j:=\widehat{h}(\mathbf{W}(0),\vec{\mathbf{x}}_j,\mathbf{x})-K(\vec{\mathbf{x}}_j,\mathbf{x})$ for all $j\in[n]$ in ③. We now analyze each term on the RHS of (44). Let $h(\cdot,t')\colon\mathcal{X}\to\mathbb{R}$ be defined by $h(\mathbf{x},t'):=\frac{\eta}{n}\sum_{j=1}^{n}K(\mathbf{x},\vec{\mathbf{x}}_j)\left[\mathbf{P}^{(r_0)}\mathbf{u}(t')\right]_j$, then $h(\cdot,t')\in\mathcal{H}_{\mathbf{S},r_0}$ for each $t'\in[0,t-1]$. We further define

$$h_t(\cdot):=\sum_{t'=0}^{t-1}h(\cdot,t')\in\mathcal{H}_K,\tag{45}$$

Since $\mathbf{W}(0)\in\mathcal{W}_0$, $q_j\leq C_1(m/2,d,1/n)$ for all $j'\in[n]$ with $C_1(m/2,d,1/n)$ defined in (36). Moreover, $\mathbf{u}(t')\leq c_{\mathbf{u}}\sqrt{n}$ with high probability, so that we have

$$\|E_1\|_\infty=\left\|\frac{\eta}{n}\sum_{j=1}^{n}q_j\mathbf{u}_j(t')\right\|_\infty\leq\frac{\eta}{n}\|\mathbf{u}(t')\|_2\sqrt{n}C_1(m/2,d,1/n)\leq\eta c_{\mathbf{u}}C_1(m/2,d,1/n).\tag{46}$$

We now bound the last term on the RHS of (44). Define $\mathbf{X}' \in \mathbb{R}^{d\times n}$ with its j-column being $\mathbf{X}'^{[j]} = \frac{1}{m}\sum_{r=1}^{m} \mathbb{I}_{\left\{\vec{\mathbf{w}}_r(0)^\top \mathbf{x}\geq 0\right\}} d_{t',r,j}\vec{\mathbf{x}}_j$ for all $j \in [n]$, then $E_2(\mathbf{x}) = \frac{\eta}{n}\left(\mathbf{X}'\mathbf{P}^{(r_0)}\mathbf{u}(t')\right)^\top \mathbf{x}$.

We need to derive the upper bound for $\|\mathbf{X}'\|_2$. Because $\left\|\vec{\mathbf{w}}_r(t') - \vec{\mathbf{w}}_r(0)\right\|_2 \leq R$, it follows that $\mathbb{I}_{\left\{\vec{\mathbf{w}}_r(t')^\top \vec{\mathbf{x}}_j\geq 0\right\}} = \mathbb{I}_{\left\{\vec{\mathbf{w}}_r(0)^\top \vec{\mathbf{x}}_j\geq 0\right\}}$ when $\left|\vec{\mathbf{w}}_r(0)^\top \vec{\mathbf{x}}_j\right| > R$ for all $j \in [n]$. Therefore,

$$\left|d_{t',r,j'}\right| = \left|\mathbb{I}_{\left\{\vec{\mathbf{w}}_r(t')^\top \vec{\mathbf{x}}_j\geq 0\right\}} - \mathbb{I}_{\left\{\vec{\mathbf{w}}_r(0)^\top \vec{\mathbf{x}}_j\geq 0\right\}}\right| \leq \mathbb{I}_{\left\{\left|\vec{\mathbf{w}}_r(0)^\top \vec{\mathbf{x}}_j\right|\leq R\right\}},$$

and it follows that

$$\begin{aligned}\frac{\left|\sum_{r=1}^{m} \mathbb{I}_{\left\{\vec{\mathbf{w}}_r(0)^\top \vec{\mathbf{x}}_i\geq 0\right\}} d_{t',r,j}\right|}{m} &\leq \frac{\sum_{r=1}^{m}\left|d_{t',r,j}\right|}{m} \leq \frac{\sum_{r=1}^{m} \mathbb{I}_{\left\{\left|\vec{\mathbf{w}}_r(0)^\top \vec{\mathbf{x}}_j\right|\leq R\right\}}}{m} = \widehat{v}_R(\mathbf{W}(0), \vec{\mathbf{x}}_j) \\ &\leq \frac{2R}{\sqrt{2\pi\kappa}} + C_2(m/2, d, 1/n), \end{aligned} \tag{47}$$

where $\widehat{v}_R$ is defined by (26), and the last inequality follows from Theorem 16.

It follows from (47) that $\|\mathbf{X}'\|_2 \leq \sqrt{n}\left(\frac{2R}{\sqrt{2\pi\kappa}} + C_2(m/2, d, 1/n)\right)$, and we have

$$\|E_2(\mathbf{x})\|_\infty \leq \frac{\eta}{n}\|\mathbf{X}'\|_2\left\|\mathbf{P}^{(r_0)}\right\|_2\left\|\mathbf{u}(t')\right\|_2\|\mathbf{x}\|_2 \leq \eta c_{\mathbf{u}}\left(\frac{2R}{\sqrt{2\pi\kappa}} + C_2(m/2, d, 1/n)\right). \tag{48}$$

Combining (44), (46), and (48), for any $t' \in [0, t-1]$,

$$\begin{aligned}\sup_{\mathbf{x}\in\mathcal{X}}\left|G_{t'}(\mathbf{x}) - h(\mathbf{x}, t')\right| &\leq \|E_1\|_\infty + \|E_2\|_\infty \\ &\leq \eta c_{\mathbf{u}}\left(C_1(m/2, d, 1/n) + \frac{2R}{\sqrt{2\pi\kappa}} + C_2(m/2, d, 1/n)\right). \end{aligned} \tag{49}$$

Define $e_t(\mathbf{x}') = f(\mathcal{W}, \mathbf{x}') - h_t(\mathbf{x}')$ for $\mathbf{x}' \in \mathcal{X}$, and $e_t(\mathbf{x}') = 0$ for $\mathbf{x} \in \mathcal{X}\setminus\mathcal{X}$. It then follows from (42), (43), and (49) that

$$\begin{aligned}&\|e_t\|_\infty \leq \sup_{\mathbf{x}\in\mathcal{X}}|f(\mathcal{W}, \mathbf{x}) - g(\mathbf{x})| + \sup_{\mathbf{x}\in\mathcal{X}}|g(\mathbf{x}) - h_t(\mathbf{x})| \\ &\leq \sup_{\mathbf{x}\in\mathcal{X}}|f(\mathcal{W}, \mathbf{x}) - g(\mathbf{x})| + \sum_{t'=0}^{t-1}\sup_{\mathbf{x}\in\mathcal{X}}\left|G_{t'}(\mathbf{x}) - h(\mathbf{x}, t')\right| \\ &\overset{④}{\leq} 2\eta c_{\mathbf{u}} T\left(\frac{2R}{\sqrt{2\pi\kappa}} + C_2(m/2, d, 1/n)\right) + \eta c_{\mathbf{u}} T\left(C_1(m/2, d, 1/n) + \frac{2R}{\sqrt{2\pi\kappa}} + C_2(m/2, d, 1/n)\right) \\ &\leq \eta c_{\mathbf{u}} T\left(C_1(m/2, d, 1/n) + 3\left(\frac{2R}{\sqrt{2\pi\kappa}} + C_2(m/2, d, 1/n)\right)\right) := \Delta_{m,n,\eta,T}, \end{aligned} \tag{50}$$

where ④ follows from (42) and (49). We now give an estimate for $\Delta_{m,n,\eta,T}$. Since $\mathbf{W}(0) \in \mathcal{W}_0$, it follows from Theorem 13 that

$$\Delta_{m,n,\eta,T} \lesssim \sqrt{d} m^{-\frac{1}{5}} T^{\frac{3}{2}}.$$

It follows that, for any $w \in (0,1)$, when $m \gtrsim T^{\frac{15}{2}} d^{\frac{5}{2}}/w^5$, we have $\Delta_{m,n,\eta,T} \leq w$.

It follows from Lemma 21 that with probability at least $1-\delta-\exp(-\Theta(r_0))$ over $\mathbf{S}, \mathbf{w}$ $\|h_t\|_{\mathcal{H}_K} \leq B_h$, where B_h is defined in (14), and τ is required to satisfy $\tau \leq 1/(\eta T)$. Since $h(\mathbf{x}, t') \in \mathcal{H}_{\mathbf{S},r_0}$ for all $t' \in [0, t-1]$, $h_t \in \mathcal{H}_{\mathbf{S},r_0}$, so that $h_t \in \mathcal{H}_K(B_h) \cap \mathcal{H}_{\mathbf{S},r_0}$. Lemma 17 requires that $m \gtrsim T^{\frac{15}{2}} d^{\frac{5}{2}}/\tau^5$. As a result, we also have $m \gtrsim T^{\frac{25}{2}} d^{\frac{5}{2}}$. ■

D.3.3. PROOF OF THEOREM 3

We prove Theorem 3 using Theorem 2, Lemma 27, and Lemma 28.

Proof [Proof of Theorem 3]

It follows from Lemma 17 and Theorem 2 that for every $t \in [T]$, conditioned on an event Ω with probability at least $1-3\delta-\exp(-\Theta(n))-\exp(-\Theta(r_0))$ over $\mathbf{S}$ and $\mathbf{w}$, we have $\mathbf{W}(t) \in \mathcal{W}(\mathbf{S}, \mathbf{W}(0), T)$, and $f(\mathcal{W}(t), \cdot) = f_t \in \mathcal{F}_{\mathrm{NN}}(\mathbf{S}, \mathbf{W}(0), T)$. Moreover, conditioned on the event Ω, $f_t \in \mathcal{F}(B_h, w, \mathbf{S}, r_0)$, $f_t = h_t + e_t$ where $h_t \in \mathcal{H}_K(B_h) \cap \mathcal{H}_{\mathbf{S},r_0}$ and $e_t \in L^\infty$ with $\|e_t\|_\infty \leq w$. We then derive the sharp upper bound for $\mathbb{E}_P\left[(f_t - f^*)^2\right]$ by applying Theorem 9 to the function class $\mathcal{F} = \left\{F = (f-f^*)^2 : f \in \mathcal{F}(B_h, w, \mathbf{S}, r_0)\right\}$.

Since $B_0 \coloneqq (B_h + \gamma_0) + 1 \geq (B_h + \gamma_0) + w$, we have $\|F\|_\infty \leq B_0^2$ with $F \in \mathcal{F}$, so that $\mathbb{E}_P\left[F^2\right] \leq B_0^2 \mathbb{E}_P[F]$. Let $T(F) = B_0^2 \mathbb{E}_P[F]$ for $F \in \mathcal{F}$. Then $\mathrm{Var}[F] \leq \mathbb{E}_P\left[F^2\right] \leq T(F) = B_0^2 \mathbb{E}_P[F]$. We have

$$
\begin{aligned}
B_0^2 \mathfrak{R}\left(\{F \in \mathcal{F} \colon T(F) \leq r\}\right) &= B_0^2 \mathfrak{R}\left(\left\{(f-f^*)^2 \colon f \in \mathcal{F}(B_h, w, \mathbf{S}, r_0), \mathbb{E}_P\left[(f-f^*)^2\right] \leq \frac{r}{B_0^2}\right\}\right) \\
&\overset{①}{\leq} 2B_0^3 \mathfrak{R}\left(\left\{f - f^* \colon f \in \mathcal{F}(B_h, w, \mathbf{S}, r_0), \mathbb{E}_P\left[(f-f^*)^2\right] \leq \frac{r}{B_0^2}\right\}\right) \\
&\overset{②}{\leq} 2B_0^3 \sqrt{\log\frac{2}{\delta}} \cdot \Theta\left(\frac{d^{k_0}}{n}\right) + 2B_0^2\sqrt{\frac{r r_0}{n}} + 4B_0^3 w. \coloneqq \psi(r).
\end{aligned} \tag{51}
$$

Here ① is due to the contraction property of Rademacher complexity in Theorem 7. ② holds with probability at least $1-\delta$ over $\mathbf{S}$, following from Lemma 27. ψ is a sub-root function since it is nonnegative, nondecreasing and $\psi(r)/\sqrt{r}$ is nonincreasing. Let r^* be the fixed point of ψ, and r be any nonnegative number such that $0 \leq r \leq r^*$. It follows from [5, Lemma 3.2] that $0 \leq r \leq \psi(r)$. Therefore, by the definition of ψ in (51), we have

$$
r \lesssim \frac{r_0}{n} + \sqrt{\log\frac{2}{\delta}} \cdot \frac{d^{k_0}}{n} + w \lesssim \sqrt{\log\frac{2}{\delta}} \cdot \frac{d^{k_0}}{n} + w \tag{52}
$$

since $r_0 = \Theta(d^{k_0})$ and $B_0 = \Theta(1)$. It then follows from Theorem 9 that with probability at least $1-\exp(-x)$ over the random training features $\mathbf{S}$,

$$
\mathbb{E}_P\left[(f_t - f^*)^2\right] - \frac{K_0}{K_0 - 1}\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] - \frac{x\left(11B_0^2 + 26B_0^2 K_0\right)}{n} \leq \frac{704 K_0}{B_0^2} \cdot B_0^4 r^*, \tag{53}
$$

or

$$
\mathbb{E}_P\left[(f_t - f^*)^2\right] - 2\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] \lesssim r^* + \frac{x}{n}, \tag{54}
$$

with $K_0 = 2$ in (53). It follows from (52) and (54) that

$$\mathbb{E}_P\left[(f_t - f^*)^2\right] - 2\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] \lesssim \sqrt{\log\frac{2}{\delta}} \cdot \frac{d^{k_0}}{n} + w + \frac{x}{n},$$

which proves (17) with $x = d^{k_0}$. ■

Proof [Proof of Theorem 4]

We have

$$f_t(\mathbf{S}) = f^*(\mathbf{S}) + \mathbf{w} + \mathbf{v}(t) + \mathbf{e}(t), \tag{55}$$

where $\mathbf{v}(t) \in \mathcal{V}_t$, $\mathbf{e}(t) \in \mathcal{E}_{t,\tau}$, $\mathbf{e}(t) = \vec{\mathbf{e}}_1(t) + \vec{\mathbf{e}}_2(t)$ with $\mathbf{v}(t) = -\left(\mathbf{I}_n - \eta\mathbf{K}_n\mathbf{P}^{(r_0)}\right)^t f^*(\mathbf{S})$, $\vec{\mathbf{e}}_1(t) = -\left(\mathbf{I}_n - \eta\mathbf{K}_n\mathbf{P}^{(r_0)}\right)^t \mathbf{w}$ and $\left\|\vec{\mathbf{e}}_2(t)\right\|_2 \le \sqrt{n}\tau$. Since $\widehat{\lambda}_1 \in (0,1)$, we have $\eta\widehat{\lambda}_1 \in (0,1)$ if $\eta \in (0,1)$. We use the simplified notations $\mathbb{P}_{\mathbf{U}^{(r_0)}} = \mathbb{P}_{\mathrm{Span}(\mathbf{U}^{(r_0)})}$ and $\mathbb{P}_{\mathbf{U}^{(-r_0)}} = \mathbb{P}_{\mathrm{Span}(\mathbf{U}^{(r_0)})^\perp}$, we then have

$$\begin{aligned} f_t(\mathbf{S}) - \mathbf{y} &= f_t(\mathbf{S}) - \mathbb{P}_{\mathbf{U}^{(r_0)}}\left(f^*(\mathbf{S}) + \mathbf{w}\right) - \mathbb{P}_{\mathbf{U}^{(-r_0)}}\left(f^*(\mathbf{S}) + \mathbf{w}\right) \\ &= \mathbb{P}_{\mathbf{U}^{(r_0)}}\left(\mathbf{v}(t) + \mathbf{e}(t)\right) + \mathbb{P}_{\mathbf{U}^{(-r_0)}}\left(\mathbf{v}(t) + \mathbf{e}(t)\right) \\ &= \mathbb{P}_{\mathbf{U}^{(r_0)}}\left(\mathbf{v}(t) + \mathbf{e}(t)\right) - \mathbb{P}_{\mathbf{U}^{(-r_0)}}\left(f^*(\mathbf{S})\right) + \mathbf{w}\right) + \mathbb{P}_{\mathbf{U}^{(-r_0)}}(\vec{\mathbf{e}}_2(t)). \end{aligned} \tag{56}$$

It follows from (56) that $f_t(\mathbf{S}) - \mathbb{P}_{\mathbf{U}^{(r_0)}}\left(f^*(\mathbf{S}) + \mathbf{w}\right) = \mathbb{P}_{\mathbf{U}^{(r_0)}}\left(\mathbf{v}(t) + \mathbf{e}(t)\right) + \mathbb{P}_{\mathbf{U}^{(-r_0)}}(\vec{\mathbf{e}}_2(t))$, or equivalently,

$$f_t(\mathbf{S}) = \mathbb{P}_{\mathbf{U}^{(r_0)}}(f_t(\mathbf{S})) + \mathbb{P}_{\mathrm{Span}(\mathbf{U}^{(r_0)})^\perp}(f_t(\mathbf{S})) \tag{57}$$

with

$$\begin{aligned} \mathbb{P}_{\mathbf{U}^{(r_0)}}(f_t(\mathbf{S})) &= \mathbb{P}_{\mathbf{U}^{(r_0)}}\left(f^*(\mathbf{S}) + \mathbf{v}(t) + \mathbf{w} + \mathbf{e}(t)\right), \\ \mathbb{P}_{\mathrm{Span}(\mathbf{U}^{(r_0)})^\perp}(f_t(\mathbf{S})) &= \mathbb{P}_{\mathbf{U}^{(-r_0)}}(\vec{\mathbf{e}}_2(t)). \end{aligned}$$

It follows from (57) that

$$\begin{aligned} &\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] = \frac{1}{n}\|f_t(\mathbf{S}) - f^*(\mathbf{S})\|_2^2 = \frac{1}{n}\|f_t(\mathbf{S}) - \mathbb{P}_{\mathbf{U}^{(r_0)}}(f^*(\mathbf{S})) - \mathbb{P}_{\mathbf{U}^{(-r_0)}}(f^*(\mathbf{S}))\|_2^2 \\ &= \frac{1}{n}\|\mathbb{P}_{\mathbf{U}^{(r_0)}}(\mathbf{v}(t) + \mathbf{w} + \mathbf{e}(t))\|_2^2 + \frac{1}{n}\left\|\mathbb{P}_{\mathbf{U}^{(-r_0)}}(\vec{\mathbf{e}}_2(t) - f^*(\mathbf{S}))\right\|_2^2 \\ &\overset{①}{\le} \frac{3}{n}\sum_{i=1}^{r_0}\left(1 - \eta\widehat{\lambda}_i\right)^{2t}\left[\mathbf{U}^{(r_0)\top} f^*(\mathbf{S})\right]_i^2 + \frac{3}{n}\sum_{i=1}^{r_0}\left(1 - \left(1 - \eta\widehat{\lambda}_i\right)^t\right)^2\left[\mathbf{U}^{(r_0)\top}\mathbf{w}\right]_i^2 \\ &\quad + \frac{2}{n}\|\mathbb{P}_{\mathbf{U}^{(-r_0)}}(f^*(\mathbf{S}))\|_2^2 + 5\tau^2 \\ &\overset{②}{\le} \frac{3\gamma_0^2}{2e\eta t} + \frac{3}{n}\left\|\mathbf{U}^{(r_0)\top}\mathbf{w}\right\|_2^2 + \gamma_0^2\log\frac{2}{\delta}\cdot\Theta\left(\frac{d^{k_0}}{n}\right) + 5\tau^2 \\ &\overset{③}{\le} \Theta\left(\frac{\gamma_0^2}{\eta t}\right) + \frac{3r_0(\sigma_0^2+1)}{n} + \gamma_0^2\log\frac{2}{\delta}\cdot\Theta\left(\frac{d^{k_0}}{n}\right), \end{aligned}$$

which completes the proof. Here ① follows by Cauchy-Scharz inequality and $\left\|\vec{\mathbf{e}}_2(t)\right\|_2 \leq \sqrt{n}\tau$, and ② follows from Lemma 22 since $f^* \in \mathcal{F}^* \subseteq \mathcal{H}_{K^{(r_0)}}(\gamma_0) \subseteq \mathcal{H}_K(\gamma_0)$, Lemma 23, and (100) in Lemma 24 which holds with probability at least $1-2\delta$. It follows from the concentration inequality about quadratic forms of sub-Gaussian random variables in [34] that

$$\Pr\left[\left\|\mathbf{U}^{(r_0)\top}\mathbf{w}\right\|_2^2 - \mathbb{E}\left[\left\|\mathbf{U}^{(r_0)\top}\mathbf{w}\right\|_2^2\right] > r_0\right] \leq \exp\left(-\Theta(r_0)\right),$$

so that with probability at least $1-\exp\left(-\Theta(r_0)\right)$,

$$\left\|\mathbf{U}^{(r_0)\top}\mathbf{w}\right\|_2^2 \leq \mathbb{E}\left[\left\|\mathbf{U}^{(r_0)\top}\mathbf{w}\right\|_2^2\right] + r_0 \leq \sigma_0^2 \mathrm{tr}\left(\mathbf{U}^{(r_0)}\mathbf{U}^{(r_0)\top}\right) + r_0 = r_0(\sigma_0^2+1), \tag{58}$$

which leads to ③ with $\tau = \sqrt{d^{k_0}/n}$.

■

D.4. Proof of the Lemmas Required for the Proofs in Section D.3

Proof [Proof of Lemma 17]

First, when $m \gtrsim T^{\frac{15}{2}} d^{\frac{5}{2}}/\tau^5$ with a proper constant, it can be verified that $\mathbf{E}_{m,n,\eta,R} \leq \tau\sqrt{n}/T$ where $\mathbf{E}_{m,n,\eta,R}$ is defined by (67) of Lemma 19. Also, Theorem 15 and Theorem 16 hold when (39) holds. We then use mathematical induction to prove this theorem. We will first prove that $\mathbf{u}(t) = \mathbf{v}(t) + \mathbf{e}(t)$ where $\mathbf{v}(t) \in \mathcal{V}_t$, $\mathbf{e}(t) \in \mathcal{E}_{t,\tau}$, and $\|\mathbf{u}(t)\|_2 \leq c_{\mathbf{u}}\sqrt{n}$ for for all $t \in [0,T]$.

When $t=0$, we have

$$\mathbf{u}(0) = -\mathbf{y} = \mathbf{v}(0) + \mathbf{e}(0), \tag{59}$$

where $\mathbf{v}(0) \coloneqq -f^*(\mathbf{S}) = -\left(\mathbf{I} - \eta\mathbf{K}_n\mathbf{P}^{(r_0)}\right)^0 f^*(\mathbf{S})$, $\mathbf{e}(0) = -\mathbf{w} = \vec{\mathbf{e}}_1(0) + \vec{\mathbf{e}}_2(0)$ with $\vec{\mathbf{e}}_1(0) = -\left(\mathbf{I} - \eta\mathbf{K}_n\mathbf{P}^{(r_0)}\right)^0 \mathbf{w}$ and $\vec{\mathbf{e}}_2(0) = \mathbf{0}$. Therefore, $\mathbf{v}(0) \in \mathcal{V}_0$ and $\mathbf{e}(0) \in \mathcal{E}_{0,\tau}$. Also, it follows from the proof of Lemma 18 that $\|\mathbf{u}(0)\|_2 \leq c_{\mathbf{u}}\sqrt{n}$ with probability at least $1-\exp\left(-\Theta(n)\right)$ over the random noise $\mathbf{w}$.

Suppose that for all $t_1 \in [0,t]$ with $t \in [0, T-1]$, $\mathbf{u}(t_1) = \mathbf{v}(t_1) + \mathbf{e}(t_1)$ where $\mathbf{v}(t_1) \in \mathcal{V}_{t_1}$, and $\mathbf{e}(t_1) = \vec{\mathbf{e}}_1(t_1) + \vec{\mathbf{e}}_2(t_1)$ with $\mathbf{v}(t_1) \in \mathcal{V}_{t_1}$ and $\mathbf{e}(t_1) \in \mathcal{E}_{t_1,\tau}$ for all $t_1 \in [0,t]$. Then it follows from Lemma 19 that the recursion $\mathbf{u}(t'+1) = \left(\mathbf{I} - \eta\mathbf{K}_n\mathbf{P}^{(r_0)}\right)\mathbf{u}(t') + \mathbf{E}(t'+1)$ holds for all $t' \in [0,t]$. As a result, we have

$$\begin{aligned}
\mathbf{u}(t+1) &= \left(\mathbf{I} - \eta\mathbf{K}_n\mathbf{P}^{(r_0)}\right)\mathbf{u}(t) + \mathbf{E}(t+1) \\
&= -\left(\mathbf{I} - \eta\mathbf{K}_n\mathbf{P}^{(r_0)}\right)^{t+1} f^*(\mathbf{S}) - \left(\mathbf{I} - \eta\mathbf{K}_n\right)^{t+1}\mathbf{w} + \sum_{t'=1}^{t+1}\left(\mathbf{I} - \eta\mathbf{K}_n\mathbf{P}^{(r_0)}\right)^{t+1-t'}\mathbf{E}(t') \\
&= \mathbf{v}(t+1) + \mathbf{e}(t+1),
\end{aligned} \tag{60}$$

where $\mathbf{v}(t+1)$ and $\mathbf{e}(t+1)$ are defined as

$$\mathbf{v}(t+1) \coloneqq -\left(\mathbf{I} - \eta\mathbf{K}_n\mathbf{P}^{(r_0)}\right)^{t+1} f^*(\mathbf{S}) \in \mathcal{V}_{t+1}, \tag{61}$$

$$\mathbf{e}(t+1) \coloneqq \underbrace{-\left(\mathbf{I}-\eta\mathbf{K}_n\mathbf{P}^{(r_0)}\right)^{t+1}\mathbf{w}}_{\vec{\mathbf{e}}_1(t+1)} + \underbrace{\sum_{t'=1}^{t+1}\left(\mathbf{I}-\eta\mathbf{K}_n\mathbf{P}^{(r_0)}\right)^{t+1-t'}\mathbf{E}(t')}_{\vec{\mathbf{e}}_2(t+1)}. \tag{62}$$

We now prove the upper bound for $\vec{\mathbf{e}}_2(t+1)$. With $\eta \in (0,2)$, we have $\left\|\mathbf{I}-\eta\mathbf{K}_n\mathbf{P}^{(r_0)}\right\|_2 \in (0,1)$. It follows that

$$\left\|\vec{\mathbf{e}}_2(t+1)\right\|_2 \le \sum_{t'=1}^{t+1}\left\|\mathbf{I}-\eta\mathbf{K}_n\mathbf{P}^{(r_0)}\right\|_2^{t+1-t'}\left\|\mathbf{E}(t')\right\|_2 \le \tau\sqrt{n}, \tag{63}$$

where the last inequality follows from the fact that $\|\mathbf{E}(t)\|_2 \le \mathbf{E}_{m,n,\eta,R} \le \tau\sqrt{n}/T$ for all $t \in [T]$. It follows that $\mathbf{e}(t+1) \in \mathcal{E}_{t+1,\tau}$. Also, it follows from Lemma 18 that with probability at least $1-2\delta-\exp(-\Theta(n))$ over $\mathbf{S}$ and $\mathbf{w}$,

$$\begin{aligned}\|\mathbf{u}(t+1)\|_2 &\le \|\mathbf{v}(t+1)\|_2 + \left\|\vec{\mathbf{e}}_1(t+1)\right\|_2 + \left\|\vec{\mathbf{e}}_2(t+1)\right\|_2 \\ &\le \left(\frac{\gamma_0}{\sqrt{2e\eta}} + \sigma_0 + \tau + 1\right)\sqrt{n} \le c_{\mathbf{u}}\sqrt{n}.\end{aligned}$$

The above inequality completes the induction step, which also completes the proof. It is noted that $\left\|\vec{\mathbf{w}}_r(t) - \vec{\mathbf{w}}_r(0)\right\|_2 \le R$ holds for all $t \in [T]$ by Lemma 20.

■

Lemma 18 *Let $0 \le t \le T$, $\mathbf{v} = -\left(\mathbf{I}-\eta\mathbf{K}_n\mathbf{P}^{(r_0)}\right)^t f^*(\mathbf{S})$, $\mathbf{e} = -\left(\mathbf{I}-\eta\mathbf{K}_n\mathbf{P}^{(r_0)}\right)^t\mathbf{w}$, and $\eta \in (0,1)$. Suppose $\delta \in (0,1/2)$, then with probability at least $1-2\delta-\exp(-\Theta(n))$ over the random training features $\mathbf{S}$ and the random noise $\mathbf{w}$,*

$$\|\mathbf{v}\|_2 + \|\mathbf{e}\|_2 \le (\Theta(\gamma_0) + \sigma_0 + 1)\cdot\sqrt{n}. \tag{64}$$

Proof When $t \in [T]$, we have

$$\begin{aligned}\|\mathbf{v}\|_2^2 &= \sum_{i=1}^{n}\left(1-\eta\widehat{\lambda}_i\right)^{2t}\left[\mathbf{U}^\top f^*(\mathbf{S})\right]_i^2 \\ &= \sum_{i=1}^{r_0}\left(1-\eta\widehat{\lambda}_i\right)^{2t}\left[\mathbf{U}^\top f^*(\mathbf{S})\right]_i^2 + \sum_{i=r_0+1}^{n}\left(1-\eta\widehat{\lambda}_i\right)^{2t}\left[\mathbf{U}^\top f^*(\mathbf{S})\right]_i^2 \\ &\le \sum_{i=1}^{n}\left(1-\eta\widehat{\lambda}_i\right)^{2t}\left[\mathbf{U}^\top f^*(\mathbf{S})\right]_i^2 + \left\|\mathbb{P}_{\mathbf{U}^{(-r_0)}}(f^*(\mathbf{S}))\right\|_2^2 \\ &\overset{①}{\le} \sum_{i=1}^{n}\frac{1}{2e\eta\widehat{\lambda}_i t}\left[\mathbf{U}^\top f^*(\mathbf{S})\right]_i^2 + n\gamma_0^2\log\frac{2}{\delta}\cdot\Theta\left(\frac{d^{k_0}}{n}\right) \\ &\overset{②}{\le} \frac{n\gamma_0^2}{2e\eta t} + n\gamma_0^2\log\frac{2}{\delta}\cdot\Theta\left(\frac{d^{k_0}}{n}\right) \le \frac{\gamma_0^2}{2e\eta}\cdot n.\end{aligned} \tag{65}$$

Here ① follows from Lemma 23 and (100) in Lemma 24 which holds with probability at least $1-2\delta$, ② follows by Lemma 22 since $f^* \in \mathcal{F}^* \subseteq \mathcal{H}_{K^{(r_0)}}(\gamma_0) \subseteq \mathcal{H}_K(\gamma_0)$. Moreover, it follows from the concentration inequality about quadratic forms of sub-Gaussian random variables in [34] that $\Pr\{\|\mathbf{w}\|_2^2 - \mathbb{E}\left[\|\mathbf{w}\|_2^2\right] > n\} \leq \exp\left(-\Theta(n)\right)$, so that $\|\mathbf{e}\|_2 \leq \|\mathbf{w}\|_2 \leq \sqrt{\mathbb{E}\left[\|\mathbf{w}\|_2^2\right]} + \sqrt{n} = \sqrt{n}(\sigma_0+1)$ with probability at least $1-\exp\left(-\Theta(n)\right)$. As a result, (64) follows from this inequality and (65) for $t \geq 1$. When $t=0$, $\|\mathbf{v}\|_2 \leq \gamma_0\sqrt{n}$, so that (64) still holds.

■

Lemma 19 *Let $0<\eta<1$, $0 \leq t \leq T-1$ for $T \geq 1$, and suppose that $\|\widehat{\mathbf{y}}(t') - \mathbf{y}\|_2 \leq c_{\mathbf{u}}\sqrt{n}$ holds for all $0 \leq t' \leq t$ and the random initialization $\mathbf{W}(0) \in \mathcal{W}_0$. Then*

$$\widehat{\mathbf{y}}(t+1) - \mathbf{y} = (\mathbf{I} - \eta\mathbf{K}_n)\left(\widehat{\mathbf{y}}(t) - \mathbf{y}\right) + \mathbf{E}(t+1), \tag{66}$$

where $\|\mathbf{E}(t+1)\|_2 \leq \mathbf{E}_{m,n,\eta,R}$, and $\mathbf{E}_{m,n,\eta,R}$ is defined by

$$\mathbf{E}_{m,n,\eta,R} \coloneqq \eta c_{\mathbf{u}}\sqrt{n}\left(4\left(\frac{2R}{\sqrt{2\pi\kappa}} + C_2(m/2,d,1/n)\right) + 2C_1(m/2,d,1/n)\right) \lesssim \sqrt{dn}m^{-\frac{1}{5}}T^{\frac{1}{2}}. \tag{67}$$

Proof

Because $\|\widehat{\mathbf{y}}(t') - \mathbf{y}\|_2 \leq \sqrt{n}c_{\mathbf{u}}$ holds for all $t' \in [0,t]$, by Lemma 20, we have

$$\left\|\vec{\mathbf{w}}_r(t') - \vec{\mathbf{w}}_r(0)\right\|_2 \leq R, \quad \forall\, 0 \leq t' \leq t+1. \tag{68}$$

We define $\mathbf{H}^{(0)} \coloneqq \mathbf{F}(\mathbf{W}(0),\mathbf{S})\mathbf{F}(\mathbf{W}(0),\mathbf{S})^\top/m \in \mathbb{R}^{n\times n}$. We also define two sets of indices

$$E_{i,R} \coloneqq \left\{r \in [m]\colon \left|\mathbf{w}_r(0)^\top\vec{\mathbf{x}}_i\right| > R\right\}, \quad \bar{E}_{i,R} \coloneqq [m] \setminus E_{i,R},$$

then we have

$$\begin{aligned}
\widehat{\mathbf{y}}_i(t+1) - \widehat{\mathbf{y}}_i(t) &= \frac{1}{\sqrt{m}}\sum_{r=1}^{m} a_r\left(\sigma\left(\vec{\mathbf{w}}_{\mathbf{S},r}^\top(t+1)\vec{\mathbf{x}}_i\right) - \sigma\left(\vec{\mathbf{w}}_{\mathbf{S},r}^\top(t)\vec{\mathbf{x}}_i\right)\right) \\
&\quad + \frac{1}{\sqrt{m}}\left(\vec{\mathbf{w}}_{m+1}(t+1) - \vec{\mathbf{w}}_{m+1}(t)\right)^\top \boldsymbol{\sigma}(\mathbf{W}(0),\vec{\mathbf{x}}_i) \\
&= \underbrace{\frac{1}{\sqrt{m}}\sum_{r\in E_{i,R}} a_r\left(\sigma\left(\vec{\mathbf{w}}_{\mathbf{S},r}^\top(t+1)\vec{\mathbf{x}}_i\right) - \sigma\left(\vec{\mathbf{w}}_{\mathbf{S},r}^\top(t)\vec{\mathbf{x}}_i\right)\right)}_{\coloneqq \mathbf{D}_i^{(1)}} \\
&\quad + \underbrace{\frac{1}{\sqrt{m}}\sum_{r\in \bar{E}_{i,R}} a_r\left(\sigma\left(\vec{\mathbf{w}}_{\mathbf{S},r}^\top(t+1)\vec{\mathbf{x}}_i\right) - \sigma\left(\vec{\mathbf{w}}_{\mathbf{S},r}^\top(t)\vec{\mathbf{x}}_i\right)\right)}_{\coloneqq \mathbf{E}_i^{(1)}} \\
&\quad - \frac{\eta}{nm}\mathbf{F}(\mathbf{W}(0),\vec{\mathbf{x}}_i)^\top\mathbf{F}(\mathbf{W}(0),\mathbf{S})^\top\mathbf{P}^{(r_0)}(\widehat{\mathbf{y}}(t) - \mathbf{y})
\end{aligned}$$

$$= \mathbf{D}_i^{(1)} + \mathbf{E}_i^{(1)} - \frac{\eta}{n}\left[\mathbf{H}^{(0)}\right]_i \mathbf{P}^{(r_0)}\left(\widehat{\mathbf{y}}(t) - \mathbf{y}\right), \tag{69}$$

and $\mathbf{D}^{(1)}, \mathbf{E}^{(1)} \in \mathbb{R}^n$ are vectors with their i-th element being $\mathbf{D}_i^{(1)}$ and $\mathbf{E}_i^{(1)}$ defined on the RHS of (69). Now we derive the upper bound for $\mathbf{E}_i^{(1)}$. For all $i \in [n]$ we have

$$\begin{aligned}
\left|\mathbf{E}_i^{(1)}\right| &= \left|\frac{1}{\sqrt{m}} \sum_{r \in \bar{E}_{i,R}} a_r \left(\sigma\left(\vec{\mathbf{w}}_{\mathbf{S},r}(t+1)^\top \vec{\mathbf{x}}_i\right) - \sigma\left(\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top \vec{\mathbf{x}}_i\right)\right)\right| \\
&\leq \frac{1}{\sqrt{m}} \sum_{r \in \bar{E}_{i,R}} \left|\vec{\mathbf{w}}_{\mathbf{S},r}(t+1)^\top \vec{\mathbf{x}}_i - \vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top \vec{\mathbf{x}}_i\right| \leq \frac{1}{\sqrt{m}} \sum_{r \in \bar{E}_{i,R}} \left\|\vec{\mathbf{w}}_{\mathbf{S},r}(t+1) - \vec{\mathbf{w}}_{\mathbf{S},r}(t)\right\|_2 \\
&\overset{①}{=} \frac{1}{\sqrt{m}} \sum_{r \in \bar{E}_{i,R}} \left\|\frac{\eta}{n}\left[\mathbf{Z}_{\mathbf{S}}(t)\right]_{[(r-1)d+1:rd]} \mathbf{P}^{(r_0)}\left(\widehat{\mathbf{y}}(t) - \mathbf{y}\right)\right\|_2 \overset{②}{\leq} \frac{c_{\mathbf{u}}}{\sqrt{m}} \sum_{r \in \bar{E}_{i,R}} \frac{\eta}{\sqrt{m}} = \eta c_{\mathbf{u}} \cdot \frac{\left|\bar{E}_{i,R}\right|}{m}.
\end{aligned} \tag{70}$$

Here ①, ② follow from (88) and (89) in the proof of Lemma 20.

Let m be sufficiently large such that $R \leq R_0$ for the absolute positive constant $R_0 < \kappa$ specified in Theorem 13. Since $\mathbf{W}(0) \in \mathcal{W}_0$, we have

$$\sup_{\mathbf{x} \in \mathcal{X}} \left|\widehat{v}_R(\mathbf{W}(0), \mathbf{x})\right| \leq \frac{2R}{\sqrt{2\pi\kappa}} + C_2(m/2, d, 1/n), \tag{71}$$

where $\widehat{v}_R(\mathbf{W}(0), \mathbf{x}) = \frac{1}{m}\sum_{r=1}^m \mathbb{1}_{\left\{\left|\vec{\mathbf{w}}_r(0)^\top \mathbf{x}\right| \leq R\right\}}$, so that $\widehat{v}_R(\mathbf{W}(0), \vec{\mathbf{x}}_i) = \left|\bar{E}_{i,R}\right|/m$. It follows from (70) and (71) that $\left|\mathbf{E}_i^{(1)}\right| \leq \eta c_{\mathbf{u}}\left(\frac{2R}{\sqrt{2\pi\kappa}} + C_2(m/2, d, 1/n)\right)$, so that $\left\|\mathbf{E}^{(1)}\right\|_2$ can be bounded by

$$\left\|\mathbf{E}^{(1)}\right\|_2 \leq \eta c_{\mathbf{u}} \sqrt{n}\left(\frac{2R}{\sqrt{2\pi\kappa}} + C_2(m/2, d, 1/n)\right). \tag{72}$$

$\mathbf{D}_i^{(1)}$ on the RHS of (69) is expressed by

$$\begin{aligned}
\mathbf{D}_i^{(1)} &= \frac{1}{\sqrt{m}} \sum_{r \in E_{i,R}} a_r \left(\sigma\left(\vec{\mathbf{w}}_{\mathbf{S},r}^\top(t+1)\vec{\mathbf{x}}_i\right) - \sigma\left(\vec{\mathbf{w}}_{\mathbf{S},r}^\top(t)\vec{\mathbf{x}}_i\right)\right) \\
&= \frac{1}{\sqrt{m}} \sum_{r \in E_{i,R}} a_r \mathbb{1}_{\left\{\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top \vec{\mathbf{x}}_i \geq 0\right\}} \left(\vec{\mathbf{w}}_{\mathbf{S},r}(t+1) - \vec{\mathbf{w}}_{\mathbf{S},r}(t)\right)^\top \vec{\mathbf{x}}_i \\
&= \frac{1}{\sqrt{m}} \sum_{r=1}^m a_r \mathbb{1}_{\left\{\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top \vec{\mathbf{x}}_i \geq 0\right\}} \left(-\frac{\eta}{n}\left[\mathbf{Z}_{\mathbf{S}}(t)\right]_{[(r-1)d+1:rd]} \mathbf{P}^{(r_0)}\left(\widehat{\mathbf{y}}(t) - \mathbf{y}\right)\right)^\top \vec{\mathbf{x}}_i \\
&\quad + \frac{1}{\sqrt{m}} \sum_{r \in \bar{E}_{i,R}} a_r \mathbb{1}_{\left\{\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top \vec{\mathbf{x}}_i \geq 0\right\}} \left(\frac{\eta}{n}\left[\mathbf{Z}_{\mathbf{S}}(t)\right]_{[(r-1)d+1:rd]} \mathbf{P}^{(r_0)}\left(\widehat{\mathbf{y}}(t) - \mathbf{y}\right)\right)^\top \vec{\mathbf{x}}_i \\
&= \underbrace{-\frac{\eta}{n}\left[\mathbf{H}^{(1)}(t)\right]_i \mathbf{P}^{(r_0)}\left(\widehat{\mathbf{y}}(t) - \mathbf{y}\right)}_{:=\mathbf{D}_i^{(2)}}
\end{aligned}$$

$$+\underbrace{\frac{1}{\sqrt{m}}\sum_{r\in\bar{E}_{i,R}} a_r \mathbb{I}_{\left\{\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top \vec{\mathbf{x}}_i\geq 0\right\}}\left(\frac{\eta}{n}\left[\mathbf{Z}_{\mathbf{S}}(t)\right]_{[(r-1)d+1:rd]}\mathbf{P}^{(r_0)}\left(\widehat{\mathbf{y}}(t)-\mathbf{y}\right)\right)^\top \vec{\mathbf{x}}_i}_{:=\mathbf{E}_i^{(2)}}$$

$$=\mathbf{D}_i^{(2)}+\mathbf{E}_i^{(2)}, \tag{73}$$

where $\mathbf{H}^{(1)}(t)\in\mathbb{R}^{n\times n}$ is a matrix specified by

$$\mathbf{H}_{pq}^{(1)}(t)=\frac{\vec{\mathbf{x}}_p^\top\vec{\mathbf{x}}_q}{m}\sum_{r=1}^{m}\mathbb{I}_{\left\{\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top\vec{\mathbf{x}}_p\geq 0\right\}}\mathbb{I}_{\left\{\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top\vec{\mathbf{x}}_q\geq 0\right\}},\quad \forall\, p\in[n], q\in[n].$$

Let $\mathbf{D}^{(2)},\mathbf{E}^{(2)}\in\mathbb{R}^n$ be a vector with their i-the element being $\mathbf{D}_i^{(2)}$ and $\mathbf{E}_i^{(2)}$ defined on the RHS of (73). $\mathbf{E}^{(2)}$ can be expressed by $\mathbf{E}^{(2)}=\frac{\eta}{n}\tilde{\mathbf{E}}^{(2)}\mathbf{P}^{(r_0)}\left(\widehat{\mathbf{y}}(t)-\mathbf{y}\right)$ with $\tilde{\mathbf{E}}^{(2)}\in\mathbb{R}^{n\times n}$ and

$$\tilde{\mathbf{E}}_{pq}^{(2)}=\frac{1}{m}\sum_{r\in\bar{E}_{i,R}}\mathbb{I}_{\left\{\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top\vec{\mathbf{x}}_p\geq 0\right\}}\mathbb{I}_{\left\{\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top\vec{\mathbf{x}}_q\geq 0\right\}}\vec{\mathbf{x}}_q^\top\vec{\mathbf{x}}_p\leq\frac{1}{m}\sum_{r\in\bar{E}_{i,R}}1=\cdot\frac{\left|\bar{E}_{i,R}\right|}{m}$$

for all $p\in[n],q\in[n]$. The spectral norm of $\tilde{\mathbf{E}}^{(2)}$ is bounded by

$$\left\|\tilde{\mathbf{E}}^{(2)}\right\|_2\leq\left\|\tilde{\mathbf{E}}^{(2)}\right\|_{\mathrm{F}}\leq n\frac{\left|\bar{E}_{i,R}\right|}{m}\overset{①}{\leq} n\left(\frac{2R}{\sqrt{2\pi\kappa}}+C_2(m/2,d,1/n)\right), \tag{74}$$

where ① follows from (71). It follows from (74) that $\left\|\mathbf{E}^{(2)}\right\|_2$ can be bounded by

$$\left\|\mathbf{E}^{(2)}\right\|_2\leq\frac{\eta}{n}\left\|\tilde{\mathbf{E}}^{(2)}\right\|_2\left\|\mathbf{P}^{(r_0)}\right\|_2\|\mathbf{y}(t)-\mathbf{y}\|_2\leq\eta c_{\mathbf{u}}\sqrt{n}\left(\frac{2R}{\sqrt{2\pi\kappa}}+C_2(m/2,d,1/n)\right). \tag{75}$$

$\mathbf{D}_i^{(2)}$ on the RHS of (73) is expressed by

$$\begin{aligned}\mathbf{D}^{(2)}&=-\frac{\eta}{n}\mathbf{H}^{(1)}(t)\mathbf{P}^{(r_0)}\left(\widehat{\mathbf{y}}(t)-\mathbf{y}\right)\\&=\underbrace{-\frac{\eta}{n}\mathbf{K}^{(1)}\mathbf{P}^{(r_0)}\left(\widehat{\mathbf{y}}(t)-\mathbf{y}\right)}_{:=\mathbf{D}^{(3)}}+\underbrace{\frac{\eta}{n}\left(\mathbf{K}^{(1)}-\mathbf{H}^{(1)}(0)\right)\mathbf{P}^{(r_0)}\left(\widehat{\mathbf{y}}(t)-\mathbf{y}\right)}_{:=\mathbf{E}^{(3)}}\\&\quad+\underbrace{\frac{\eta}{n}\left(\mathbf{H}^{(1)}(0)-\mathbf{H}^{(1)}(t)\right)\mathbf{P}^{(r_0)}\left(\widehat{\mathbf{y}}(t)-\mathbf{y}\right)}_{:=\mathbf{E}^{(4)}}\\&=\mathbf{D}^{(3)}+\mathbf{E}^{(3)}+\mathbf{E}^{(4)}.\end{aligned} \tag{76}$$

On the RHS of (76), $\mathbf{D}^{(3)},\mathbf{E}^{(3)},\mathbf{E}^{(4)}\in\mathbb{R}^n$ are vectors which are analyzed as follows. We have

$$\left\|\mathbf{K}^{(1)}-\mathbf{H}^{(1)}(0)\right\|_2\leq\left\|\mathbf{K}^{(1)}-\mathbf{H}^{(1)}(0)\right\|_F\leq nC_1(m/2,d,1/n), \tag{77}$$

where the last inequality is due to $\mathbf{W}(0)\in\mathcal{W}_0$.

In order to bound $\mathbf{E}^{(4)}$, we first estimate the upper bound for $\left|\mathbf{H}_{ij}^{(1)}(t)-\mathbf{H}_{ij}^{(1)}(0)\right|$ for all $i,j\in[n]$. We note that

$$\mathbb{1}_{\left\{\mathbb{1}_{\left\{\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top\vec{\mathbf{x}}_i\geq 0\right\}}\neq\mathbb{1}_{\left\{\mathbf{w}_r(0)^\top\vec{\mathbf{x}}_i\geq 0\right\}}\right\}}\leq\mathbb{1}_{\left\{\left|\mathbf{w}_r(0)^\top\vec{\mathbf{x}}_i\right|\leq R\right\}}+\mathbb{1}_{\left\{\left\|\mathbf{w}_{\mathbf{S},r}(t)-\vec{\mathbf{w}}_r(0)\right\|_2>R\right\}}. \tag{78}$$

It follows from (78) that

$$\begin{aligned}
&\left|\mathbf{H}_{ij}^{(1)}(t)-\mathbf{H}_{ij}^{(1)}(0)\right|\\
&=\left|\frac{\vec{\mathbf{x}}_i^\top\vec{\mathbf{x}}_j}{m}\sum_{r=1}^m\left(\mathbb{1}_{\left\{\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top\vec{\mathbf{x}}_i\geq 0\right\}}\mathbb{1}_{\left\{\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top\vec{\mathbf{x}}_j\geq 0\right\}}-\mathbb{1}_{\left\{\mathbf{w}_r(0)^\top\vec{\mathbf{x}}_i\geq 0\right\}}\mathbb{1}_{\left\{\mathbf{w}_r(0)^\top\vec{\mathbf{x}}_j\geq 0\right\}}\right)\right|\\
&\leq\frac{1}{m}\sum_{r=1}^m\left(\mathbb{1}_{\left\{\mathbb{1}_{\left\{\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top\vec{\mathbf{x}}_i\geq 0\right\}}\neq\mathbb{1}_{\left\{\vec{\mathbf{w}}_r(0)^\top\vec{\mathbf{x}}_i\geq 0\right\}}\right\}}+\mathbb{1}_{\left\{\mathbb{1}_{\left\{\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top\vec{\mathbf{x}}_j\geq 0\right\}}\neq\mathbb{1}_{\left\{\vec{\mathbf{w}}_r(0)^\top\vec{\mathbf{x}}_j\geq 0\right\}}\right\}}\right)\\
&\leq\frac{1}{m}\sum_{r=1}^m\left(\mathbb{1}_{\left\{\left|\vec{\mathbf{w}}_r(0)^\top\vec{\mathbf{x}}_i\right|\leq R\right\}}+\mathbb{1}_{\left\{\left|\vec{\mathbf{w}}_r(0)^\top\vec{\mathbf{x}}_j\right|\leq R\right\}}+2\mathbb{1}_{\left\{\left\|\mathbf{w}_{\mathbf{S},r}(t)-\vec{\mathbf{w}}_r(0)\right\|_2>R\right\}}\right)\\
&\leq 2v_R(\mathbf{W}(0),\vec{\mathbf{x}}_i)\overset{①}{\leq}\left(\frac{4R}{\sqrt{2\pi}\kappa}+2C_2(m/2,d,1/n)\right),
\end{aligned}\tag{79}$$

where ① follows from (71).

It follows from (77) and (79) that $\left\|\mathbf{E}^{(3)}\right\|_2$, $\left\|\mathbf{E}^{(4)}\right\|_2$ are bounded by

$$\left\|\mathbf{E}^{(3)}\right\|_2\leq\frac{\eta}{n}\left\|\mathbf{K}^{(1)}-\mathbf{H}^{(1)}(0)\right\|_2\left\|\mathbf{P}^{(r_0)}\right\|_2\|\widehat{\mathbf{y}}(t)-\mathbf{y}\|_2\leq\eta c_{\mathbf{u}}\sqrt{n}C_1(m/2,d,1/n), \tag{80}$$

$$\begin{aligned}
\left\|\mathbf{E}^{(4)}\right\|_2&\leq\frac{\eta}{n}\left\|\mathbf{H}^{(1)}(0)-\mathbf{H}^{(1)}(t)\right\|_2\left\|\mathbf{P}^{(r_0)}\right\|_2\|\widehat{\mathbf{y}}(t)-\mathbf{y}\|_2\\
&\leq\eta c_{\mathbf{u}}\sqrt{n}\left(\frac{4R}{\sqrt{2\pi}\kappa}+2C_2(m/2,d,1/n)\right).
\end{aligned}\tag{81}$$

It follows from (73) and (76) that

$$\mathbf{D}_i^{(1)}=\mathbf{D}_i^{(3)}+\mathbf{E}_i^{(2)}+\mathbf{E}_i^{(3)}+\mathbf{E}_i^{(4)}. \tag{82}$$

We also have

$$-\frac{\eta}{n}\mathbf{H}^{(0)}\mathbf{P}^{(r_0)}\left(\widehat{\mathbf{y}}(t)-\mathbf{y}\right)=\underbrace{-\frac{\eta}{n}\left(\mathbf{H}^{(0)}-\mathbf{K}^{(0)}\right)\mathbf{P}^{(r_0)}\left(\widehat{\mathbf{y}}(t)-\mathbf{y}\right)}_{:=\mathbf{E}^{(5)}}-\frac{\eta}{n}\mathbf{K}^{(0)}\mathbf{P}^{(r_0)}\left(\widehat{\mathbf{y}}(t)-\mathbf{y}\right). \tag{83}$$

Similar to (80), $\mathbf{E}^{(5)}$ is bounded by

$$\left\|\mathbf{E}^{(5)}\right\|_2\leq\frac{\eta}{n}\left\|\mathbf{H}^{(0)}-\mathbf{K}^{(0)}\right\|_2\left\|\mathbf{P}^{(r_0)}\right\|_2\|\widehat{\mathbf{y}}(t)-\mathbf{y}\|_2\leq\eta c_{\mathbf{u}}\sqrt{n}C_1(m/2,d,1/n). \tag{84}$$

It then follows from (69) and (85) that

$$\widehat{\mathbf{y}}_i(t+1)-\widehat{\mathbf{y}}_i(t)=\mathbf{D}_i^{(1)}+\mathbf{E}_i^{(1)}-\frac{\eta}{n}\left[\mathbf{H}^{(0)}\right]_i\left(\widehat{\mathbf{y}}(t)-\mathbf{y}\right)$$

$$= \mathbf{D}_i^{(3)} - \frac{\eta}{n}\left[\mathbf{K}^{(0)}\right]_i \mathbf{P}^{(r_0)}\left(\widehat{\mathbf{y}}(t) - \mathbf{y}\right) + \underbrace{\mathbf{E}_i^{(1)} + \mathbf{E}_i^{(2)} + \mathbf{E}_i^{(3)} + \mathbf{E}_i^{(4)} + \mathbf{E}_i^{(5)}}_{:=\mathbf{E}_i}$$

$$= -\frac{\eta}{n}\mathbf{K}\left(\widehat{\mathbf{y}}(t) - \mathbf{y}\right) + \mathbf{E}_i, \tag{85}$$

where $\mathbf{E} \in \mathbb{R}^n$ with its i-th element being $\mathbf{E}_i$, and $\mathbf{E} = \mathbf{E}^{(1)} + \mathbf{E}^{(2)} + \mathbf{E}^{(3)} + \mathbf{E}^{(4)} + \mathbf{E}^{(5)}$. It then follows from (72), (75), (80), (81), and (84) that

$$\|\mathbf{E}\|_2 \leq \eta c_{\mathbf{u}}\sqrt{n}\left(4\left(\frac{2R}{\sqrt{2\pi}\kappa} + C_2(m/2, d, 1/n)\right) + 2C_1(m/2, d, 1/n)\right). \tag{86}$$

Finally, (85) can be rewritten as

$$\widehat{\mathbf{y}}(t+1) - \mathbf{y} = \left(\mathbf{I} - \frac{\eta}{n}\mathbf{K}\right)\mathbf{P}^{(r_0)}\left(\widehat{\mathbf{y}}(t) - \mathbf{y}\right) + \mathbf{E}(t+1),$$

which proves (66) with the upper bound for $\|\mathbf{E}\|_2$ in (86). ■

Lemma 20 *Suppose that $t \in [0, T-1]$ for $T \geq 1$, and $\|\widehat{\mathbf{y}}(t') - \mathbf{y}\|_2 \leq \sqrt{n}c_{\mathbf{u}}$ holds for all $0 \leq t' \leq t$. Then*

$$\left\|\vec{\mathbf{w}}_{\mathbf{S},r}(t') - \vec{\mathbf{w}}_r(0)\right\|_2 \leq R, \quad \forall\, 0 \leq t' \leq t+1. \tag{87}$$

Proof Let $[\mathbf{Z}_{\mathbf{S}}(t)]_{[(r-1)d+1:rd]}$ denote the submatrix of $\mathbf{Z}_{\mathbf{S}}(t)$ formed by the the rows of $\mathbf{Z}_{\mathbf{Q}}(t)$ with row indices in $[(r-1)d+1 : rd]$. By the GD update rule we have for $t \in [0, T-1]$ that

$$\vec{\mathbf{w}}_{\mathbf{S},r}(t+1) - \vec{\mathbf{w}}_{\mathbf{S},r}(t) = -\frac{\eta}{n}\left[\mathbf{Z}_{\mathbf{S}}(t)\right]_{[(r-1)d+1:rd]}\mathbf{P}^{(r_0)}\left(\widehat{\mathbf{y}}(t) - \mathbf{y}\right), \tag{88}$$

We have $\left\|\left[\mathbf{Z}_{\mathbf{S}}(t)\right]_{[(r-1)d+1:rd]}\right\|_2 \leq \sqrt{n/m}$. It then follows from (88) that

$$\left\|\vec{\mathbf{w}}_{\mathbf{S},r}(t+1) - \vec{\mathbf{w}}_{\mathbf{S},r}(t)\right\|_2 \leq \frac{\eta}{n}\left\|\left[\mathbf{Z}_{\mathbf{S}}(t)\right]_{[(r-1)d+1:rd]}\right\|_2\left\|\mathbf{P}^{(r_0)}\right\|_2\|\widehat{\mathbf{y}}(t) - \mathbf{y}\|_2 \leq \frac{\eta c_{\mathbf{u}}}{\sqrt{m}}. \tag{89}$$

Note that (87) trivially holds for $t' = 0$. For $t' \in [1, t+1]$, it follows from (89) that

$$\left\|\vec{\mathbf{w}}_{\mathbf{S},r}(t') - \vec{\mathbf{w}}_r(0)\right\|_2 \leq \sum_{t''=0}^{t'-1}\left\|\vec{\mathbf{w}}_{\mathbf{S},r}(t''+1) - \vec{\mathbf{w}}_{\mathbf{S},r}(t'')\right\|_2 \leq \frac{\eta}{\sqrt{m}}\sum_{t''=0}^{t'-1} c_{\mathbf{u}} \leq \frac{\eta c_{\mathbf{u}} T}{\sqrt{m}} = R, \tag{90}$$

which completes the proof. ■

Lemma 21 *Suppose $n \geq \Theta(\log(2/\delta) \cdot d^{2k_0})$ and $\delta \in (0, 1/2)$. Let $h_t(\cdot) = \sum_{t'=0}^{t-1} h(\cdot, t')$ for $t \in [T]$, $T \leq \widehat{T}$ where*

$$h(\cdot, t') = v(\cdot, t') + \widehat{e}(\cdot, t'),$$

$$v(\cdot,t') - \frac{\eta}{n}\sum_{j=1}^{n} K(\cdot,\vec{\mathbf{x}}_j)\left[\mathbf{P}^{(r_0)}\mathbf{v}(t')\right]_j,$$
$$\widehat{e}(\cdot,t') = \frac{\eta}{n}\sum_{j=1}^{n} K(\cdot,\vec{\mathbf{x}}_j)\left[\mathbf{P}^{(r_0)}\mathbf{e}(t')\right]_j,$$

where $\mathbf{v}(t') \in \mathcal{V}_{t'}$, $\mathbf{e}(t') \in \mathcal{E}_{t',\tau}$ *for all* $0 \le t' \le t-1$. *Suppose that* $\tau \le 1/(\eta T)$, *then with probability at least* $1-\delta-\exp\left(-\Theta(r_0)\right)$ *over the random training features* $\mathbf{S}$ *and the random noise* $\mathbf{w}$,

$$\|h_t\|_{\mathcal{H}_K} \le B_h = \gamma_0 + \Theta(1), \tag{91}$$

where $r_0 = m_{k_0}$.

Proof We have $\mathbf{v}(t) = -\left(\mathbf{I}-\eta\mathbf{K}_n\mathbf{P}^{(r_0)}\right)^t f^*(\mathbf{S})$, $\mathbf{e}(t) = \vec{\mathbf{e}}_1(t)+\vec{\mathbf{e}}_2(t)$ with $\vec{\mathbf{e}}_1(t) = -\left(\mathbf{I}-\eta\mathbf{K}_n\mathbf{P}^{(r_0)}\right)^t \mathbf{w}$, $\left\|\vec{\mathbf{e}}_2(t)\right\|_2 \le \sqrt{n}\tau$. We define

$$\widehat{e}_1(\cdot,t') \coloneqq -\frac{\eta}{n}\sum_{j=1}^{n} K(\vec{\mathbf{x}}_j,\mathbf{x})\left[\mathbf{P}^{(r_0)}\vec{\mathbf{e}}_1(t')\right]_j, \quad \widehat{e}_2(\cdot,t') \coloneqq -\frac{\eta}{n}\sum_{j=1}^{n} K(\vec{\mathbf{x}}_j,\mathbf{x})\left[\mathbf{P}^{(r_0)}\vec{\mathbf{e}}_2(t')\right]_j, \tag{92}$$

Let $\mathbf{\Sigma}$ be the diagonal matrix containing eigenvalues of $\mathbf{K}_n$, we then have

$$\begin{aligned}\sum_{t'=0}^{t-1} v(\mathbf{x},t') &= \frac{\eta}{n}\sum_{j=1}^{n}\sum_{t'=0}^{t-1}\left[\mathbf{P}^{(r_0)}\left(\mathbf{I}-\eta\mathbf{K}_n\mathbf{P}^{(r_0)}\right)^{t'} f^*(\mathbf{S})\right]_j K(\vec{\mathbf{x}}_j,\mathbf{x}) \\ &= \frac{\eta}{n}\sum_{j=1}^{n}\sum_{t'=0}^{t-1}\left[\mathbf{P}^{(r_0)}\mathbf{U}\left(\mathbf{I}-\eta\mathbf{\Sigma}^{(r_0)}\right)^{t'}\mathbf{U}^\top f^*(\mathbf{S})\right]_j K(\vec{\mathbf{x}}_j,\mathbf{x}).\end{aligned} \tag{93}$$

We then have

$$\begin{aligned}&\left\|\sum_{t'=0}^{t-1} v(\cdot,t')\right\|_{\mathcal{H}_K}^2 \\ &= \frac{\eta^2}{n^2} f^*(\mathbf{S})^\top\mathbf{U}\sum_{t'=0}^{t-1}\left(\mathbf{I}-\eta\mathbf{\Sigma}^{(r_0)}\right)^{t'}\mathbf{U}^\top\mathbf{P}^{(r_0)}\mathbf{K}\mathbf{P}^{(r_0)}\mathbf{U}\sum_{t'=0}^{t-1}\left(\mathbf{I}-\eta\mathbf{\Sigma}^{(r_0)}\right)^{t'}\mathbf{U}^\top f^*(\mathbf{S}) \\ &= \frac{1}{n}\left\|\eta\left(\mathbf{K}_n\right)^{1/2}\mathbf{P}^{(r_0)}\mathbf{U}\sum_{t'=0}^{t-1}\left(\mathbf{I}-\eta\mathbf{\Sigma}^{(r_0)}\right)^{t'}\mathbf{U}^\top f^*(\mathbf{S})\right\|_2^2 \\ &\le \frac{1}{n}\sum_{i=1}^{r_0}\frac{\left(1-\left(1-\eta\widehat{\lambda}_i\right)^t\right)^2}{\widehat{\lambda}_i}\left[\mathbf{U}^\top f^*(\mathbf{S})\right]_i^2 \le \frac{1}{n}\sum_{i=1}^{n}\frac{\left(1-\left(1-\eta\widehat{\lambda}_i\right)^t\right)^2}{\widehat{\lambda}_i}\left[\mathbf{U}^\top f^*(\mathbf{S})\right]_i^2 \le \gamma_0^2,\end{aligned} \tag{94}$$

where the last inequality follows from Lemma 22 since $f^* \in \mathcal{F}^* \subseteq \mathcal{H}_{K^{(r_0)}}(\gamma_0) \subseteq \mathcal{H}_K(\gamma_0)$. We define $E_1 := \left\|\sum_{t'=0}^{t-1} \widehat{e}_1(\cdot, t')\right\|_{\mathcal{H}_K}^2$ and $E_2 := \left\|\sum_{t'=0}^{t-1} \widehat{e}_2(\cdot, t')\right\|_{\mathcal{H}_K}$. It follows from (58) in the proof of Theorem 4 that with probability at least $1 - \exp(-\Theta(r_0))$, $\left\|\mathbf{U}^{(r_0)\top}\mathbf{w}\right\|_2^2 \lesssim r_0 = \Theta(d^{k_0})$.

With $n \geq \Theta(\log(2/\delta) \cdot d^{2k_0})$ and $r \in [r_0]$, it follows from Lemma 25 that with probability $1-\delta$ over $\mathbf{S}$, we have

$$\widehat{\lambda}_r \geq \widehat{\lambda}_{r_0} \geq \lambda_{r-1} - 2\sqrt{\frac{2\log\frac{2}{\delta}}{n}} \geq \mu_{k_0} - 2\sqrt{\frac{2\log\frac{2}{\delta}}{n}} \geq \Theta(d^{-k_0}). \tag{95}$$

It then follows from (95) that

$$E_1 \leq \frac{1}{n}\sum_{i=1}^{r_0} \frac{\left(1 - \left(1 - \eta\widehat{\lambda}_i\right)^t\right)^2}{\widehat{\lambda}_i} \left[\mathbf{U}^\top \mathbf{w}\right]_i^2 \leq \frac{\Theta(d^{k_0})}{n} \cdot \Theta(d^{k_0}) \leq \Theta(1). \tag{96}$$

We now find the upper bound for E_2. We have

$$\left\|\widehat{e}_2(\cdot, t')\right\|_{\mathcal{H}_K}^2 \leq \frac{\eta^2}{n^2} \vec{\mathbf{e}}_2^{\,\top}(t') \mathbf{K} \vec{\mathbf{e}}_2(t') \leq \eta^2 \widehat{\lambda}_1 \tau^2,$$

so that

$$E_2 \leq \sum_{t'=0}^{t-1} \left\|\widehat{e}_2(\cdot, t')\right\|_{\mathcal{H}_K} \leq T\eta\sqrt{\widehat{\lambda}_1}\tau \leq 1, \tag{97}$$

if $\tau \leq 1/(\eta T)$ since $\widehat{\lambda}_1 \in (0, 1)$.

Finally, it follows from (93), (96), and (97) that

$$\|h_t\|_{\mathcal{H}_K} \leq \left\|\sum_{t'=0}^{t-1} \widehat{v}(\cdot, t')\right\|_{\mathcal{H}_K} + \left\|\sum_{t'=0}^{t-1} \widehat{e}_1(\cdot, t')\right\|_{\mathcal{H}_K} + \left\|\sum_{t'=0}^{t-1} \widehat{e}_2(\cdot, t')\right\|_{\mathcal{H}_K} \leq \gamma_0 + \Theta(1).$$

■

Lemma 22 (In the proof of [27, Lemma 8]) *For any $f \in \mathcal{H}_K(\gamma_0)$, we have*

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\left[\mathbf{U}^\top f(\mathbf{S}')\right]_i^2}{\widehat{\lambda}_i} \leq \gamma_0^2. \tag{98}$$

Lemma 23 *For any positive real number $a \in (0, 1)$ and natural number t, we have*

$$(1-a)^t \leq e^{-ta} \leq \frac{1}{eta}. \tag{99}$$

Proof The result follows from the facts that $\log(1-a) \le a$ for $a \in (0,1)$ and $\sup_{u\in\mathbb{R}} ue^{-u} \le 1/e$. ■

Background about the Integral Operator on $\mathcal{H}_{\mathbf{S}}$. Suppose K is a PSD kernel defined over $\mathcal{X}\times\mathcal{X}$ and let the empirical gram matrix computed by K on the training features $\mathbf{S}$ be $\mathbf{K}_n$ with the eigenvalues $\widehat{\lambda}_1 \ge \ldots \ge \widehat{\lambda}_n \ge 0$. We need the following background in the RKHS spanned by $\left\{K(\cdot, \vec{\mathbf{x}}_i)\right\}_{i=1}^n$ for the proof of Lemma 24. Herein we introduce the operator $T_n \colon \mathcal{H}_{\mathbf{S}} \to \mathcal{H}_{\mathbf{S}}$ which is defined by $T_n g \coloneqq \frac{1}{n}\sum_{i=1}^n K(\cdot, \vec{\mathbf{x}}_i) g(\vec{\mathbf{x}}_i)$ for every $g \in \mathcal{H}_{\mathbf{S}}$. It can be verified that the eigenvalues of T_n coincide with the eigenvalues of $\mathbf{K}_n$, that is, the eigenvalues of T_n are $\left\{\widehat{\lambda}_i\right\}_{i=1}^n$. By the spectral theorem, all the normalized eigenfunctions of T_n, denoted by $\left\{\Phi^{(k)}\right\}_{k=0}^{n-1}$ with $\Phi^{(k)} = 1/\sqrt{n\widehat{\lambda}_{k+1}} \cdot \sum_{j=1}^n K(\cdot, \vec{\mathbf{x}}_j)\left[\mathbf{U}^{[k+1]}\right]_j$ for $k \in [0\colon n-1]$, is an orthonormal basis of $\mathcal{H}_{\mathbf{S}}$. The eigenvalue of T_n corresponding to the eigenfunction $\Phi^{(k)}$ is $\widehat{\lambda}_{k+1}$ for $0 \le k \le n-1$. Since $\mathcal{H}_{\mathbf{S}} \subseteq \mathcal{H}_K$, we can complete $\left\{\Phi^{(k)}\right\}_{k=0}^{n-1}$ so that $\left\{\Phi^{(k)}\right\}_{k\ge 0}$ is an orthonormal basis of the RKHS $\mathcal{H}_K$.

Lemma 24 *Suppose $\delta \in (0, 1/2)$ and $n \ge \Theta(\log(2/\delta)\cdot d^{2k_0})$. Let $\mathbb{P}_{\mathbf{U}^{(-r_0)}} = \mathbb{P}_{\mathrm{Span}(\mathbf{U}^{(r_0)})^\perp}$. Then with probability at least $1-2\delta$ over the random training features $\mathbf{S}$,*

$$\left\|\mathbb{P}_{\mathbf{U}^{(-r_0)}}(f^*(\mathbf{S}))\right\|_2^2 \le n\gamma_0^2 \log\frac{2}{\delta}\cdot\Theta\left(\frac{d^{k_0}}{n}\right). \tag{100}$$

Proof We have $\mathbb{P}_{\mathcal{H}_{\mathbf{S}}}(f^*) = \sum_{k=0}^{n-1}\left\langle f^*, \Phi^{(k)}\right\rangle \Phi^{(k)}$, $\mathbb{P}_{\mathcal{H}_{\mathbf{S},r_0}}(f^*) = \sum_{k=0}^{r_0-1}\left\langle f^*, \Phi^{(k)}\right\rangle \Phi^{(k)}$, and define

$$\bar{f}^{*,r_0} \coloneqq \mathbb{P}_{\mathcal{H}_{\mathbf{S}}}(f^*) - \mathbb{P}_{\mathcal{H}_{\mathbf{S},r_0}}(f^*) = \sum_{q=r_0}^{n}\left\langle f^*, \Phi^{(q)}\right\rangle \Phi^{(q)}.$$

Let $\mathbf{U}^{(-r_0)} \in \mathbb{R}^{n\times(n-r_0)}$ be the submatrix formed by all the columns of $\mathbf{U}$ except for the top r_0 columns in $\mathbf{U}^{(r_0)}$. It follows by the introduction to the space $\mathcal{H}_{\mathbf{S}}$ before Lemma 24 that $\left\{\Phi^{(k)}\right\}_{k=0}^{n-1}$ is an orthonormal basis of $\mathcal{H}_{\mathbf{S}}$, and $\Phi^{(k)}$ is the eigenfunction of the operator T_n with the corresponding eigenvalue $\widehat{\lambda}_{k+1}$. Therefore, $\mathbf{U}^{(-r_0)}{\mathbf{U}^{(-r_0)}}^\top \Phi^{(k)}(\mathbf{S}) = 0$ for all $k \in [r_0-1]$. As a result, with probability at least $1-\delta$,

$$\frac{1}{n}\left\|\mathbf{U}^{(-r_0)}{\mathbf{U}^{(-r_0)}}^\top f^*(\mathbf{S})\right\|_2^2 = \frac{1}{n}\left\|\mathbf{U}^{(-r_0)}{\mathbf{U}^{(-r_0)}}^\top \left(\mathbb{P}_{\mathcal{H}_{\mathbf{S}}}(f^*)\right)(\mathbf{S})\right\|_2^2 = \frac{1}{n}\sum_{i=1}^n\left(\bar{f}^{*,r_0}(\vec{\mathbf{x}}_i)\right)^2. \tag{101}$$

We have

$$\left\langle T_n\bar{f}^{*,r_0}, \bar{f}^{*,r_0}\right\rangle = \left\langle \frac{1}{n}\sum_{i=1}^n K(\cdot,\vec{\mathbf{x}}_i)\bar{f}^{*,r_0}(\vec{\mathbf{x}}_i), \bar{f}^{*,r_0}\right\rangle_{\mathcal{H}_K} = \frac{1}{n}\sum_{i=1}^n\left(\bar{f}^{*,r_0}(\vec{\mathbf{x}}_i)\right)^2. \tag{102}$$

On the other hand, with probability $1-\delta$,

$$\begin{aligned}\left\langle T_n \bar{f}^{*,r_0}, \bar{f}^{*,r_0}\right\rangle &= \left\langle T_n \sum_{q=r_0}^{n} \left\langle \bar{f}^{*,r_0}, \Phi^{(q)}\right\rangle \Phi^{(q)}, \sum_{q=r_0}^{n} \left\langle \bar{f}^{*,r_0}, \Phi^{(q)}\right\rangle \Phi^{(q)} \right\rangle_{\mathcal{H}_K} \\ &= \sum_{q=r_0}^{n} \widehat{\lambda}_{q+1} \left\langle \bar{f}^{*,r_0}, \Phi^{(q)}\right\rangle^2 \le \widehat{\lambda}_{r_0+1} \sum_{q=r_0}^{n} \left\langle f^*, \Phi^{(q)}\right\rangle^2 \overset{①}{\le} \widehat{\lambda}_{r_0+1}\zeta_{n,\gamma_0,r_0,\delta}, \quad (103)\end{aligned}$$

where ① is due to Theorem 14. It follows from (101)-(103) that

$$\frac{1}{n}\left\|\mathbf{U}^{(-r_0)}\mathbf{U}^{(-r_0)^\top} f^*(\mathbf{S})\right\|_2^2 \le \widehat{\lambda}_{r_0+1}\zeta_{n,\gamma_0,r_0,\delta}. \quad (104)$$

We now find the upper bound for $\widehat{\lambda}_{r_0+1}$. It follows from Lemma 25 that $\left|\lambda_j - \widehat{\lambda}_j\right| \le 2\sqrt{\frac{2\log\frac{2}{\delta}}{n}}$ for all $j \in [n]$ with probability at least $1-\delta$. Furthermore, it follows from Theorem 31 that $\lambda_{r_0} = \mu_{k_0+1} = \Theta(d^{-k_0-1})$ with $r_0 = m_{k_0}$. As a result, we have

$$\widehat{\lambda}_{r_0+1} \le \lambda_{r_0} + 2\sqrt{\frac{2\log\frac{2}{\delta}}{n}} \le \Theta(d^{-k_0}), \quad (105)$$

where the last inequality holds with probability $1-\delta$ over $\mathbf{S}$ due to Lemma 25 and $n \ge \Theta(\log(2/\delta)\cdot d^{2k_0})$. It then follows from (104) and (105) that

$$\begin{aligned}\frac{1}{n}\left\|\mathbf{U}^{(-r_0)}\mathbf{U}^{(-r_0)^\top} f^*(\mathbf{S})\right\|_2^2 &\le \Theta(d^{-k_0})\cdot\zeta_{n,\gamma_0,r_0,\delta} = \Theta(d^{-k_0})\cdot\frac{32\gamma_0^2\log\frac{2}{\delta}}{\left(\mu_{k_0}-\mu_{k_0+1}\right)^2 n} \\ &= \Theta(d^{-k_0})\cdot\frac{32\gamma_0^2\log\frac{2}{\delta}}{\left(\Theta(d^{-k_0})-\Theta(d^{-k_0-1})\right)^2 n} \\ &= \gamma_0^2\log\frac{2}{\delta}\cdot\Theta\left(\frac{d^{k_0}}{n}\right),\end{aligned}$$

which proves (100). ■

Lemma 25 ([28, Proposition 10])
Let $\delta \in (0,1)$, then with probability $1-\delta$ over the training features $\mathbf{S}$, for all $j \in [n]$,

$$\left|\lambda_{j-1} - \widehat{\lambda}_j\right| \le 2\sqrt{\frac{2\log\frac{2}{\delta}}{n}}. \quad (106)$$

Remark 26 *We remark that the sequence $\{\lambda_j\}_{j\ge 0}$ starts with index 0, so that λ_{j-1} is in fact the j-th element in the extended enumeration of the distinct eigenvalues of T_K. The extended enumeration [28] of the distinct eigenvalues of T_K is a sequence where each nonzero eigenvalue of T_K appears as many times as its multiplicity and the other values (if any) are zero.*

Lemma 27 *Suppose $n \geq r_0$. Then with probability at least $1-\delta$ over the random training features $\mathbf{S}$, for every $r > 0$, we have*

$$\mathfrak{R}\left(\left\{f - f^* \colon f \in \mathcal{F}(B_h, w, \mathbf{S}, r_0), \mathbb{E}_P\left[(f-f^*)^2\right] \leq r\right\}\right) \leq \sqrt{\log\frac{2}{\delta}} \cdot \Theta\left(\frac{d^{k_0}}{n}\right) + \sqrt{\frac{rr_0}{n}} + 2w. \tag{107}$$

Proof Let $\mathcal{H}_{K,r_0} = \overline{\operatorname{Span}\{v_q\}_{q=0}^{r_0-1}}$ be the subspace in $\mathcal{H}_K$ spanned by $\{v_q\}_{q=0}^{r_0-1}$, and we define $\widehat{\mathcal{F}}_r \coloneqq \left\{f \in \mathcal{F}(B_h, w, \mathbf{S}, r_0), \mathbb{E}_P\left[(f-f^*)^2\right] \leq r\right\}$. For every $f \in \widehat{\mathcal{F}}_r$, we have $f = h + e$ such that $\|e\|_\infty \leq w$ and $h \in \mathcal{H}_K(B_h) \cap \mathcal{H}_{\mathbf{S},r_0}$, and $\mathbb{E}_P\left[(h-f^*)^2\right] \leq 2(r+w^2)$. Furthermore, we have $\mathbb{P}_{\mathcal{H}_{r_0}}(h) = \sum_{j=0}^{r_0-1}\alpha_j v_j$ with $\alpha_j = \langle h, v_j\rangle_{\mathcal{H}_K}$ for all $j \geq 0$. We define $\bar{h} = h - \mathbb{P}_{\mathcal{H}_{r_0}}(h)$, then $\bar{h} \in \mathcal{H}_K(B_h)$. We have

$$\begin{aligned}\mathbb{E}_P\left[\bar{h}^2\right] &= \mathbb{E}_P\left[\left(\sum_{j\geq r_0}\alpha_j v_j\right)^2\right] = \sum_{j\geq r_0}\alpha_j^2\lambda_j \leq \lambda_{r_0}\cdot\sum_{j\geq r_0}\alpha_j^2 \overset{①}{\leq} \lambda_{r_0}\zeta_{n,B_h,r_0,\delta} \\ &\overset{②}{\leq} \log\frac{2}{\delta}\cdot\Theta\left(\frac{d^{k_0}}{n}\right) \coloneqq r_{n,k_0,\delta},\end{aligned} \tag{108}$$

where ① holds with probability at least $1-\delta$ over $\mathbf{S}$ by (34) of Theorem 14, and ② follows from the similar argument in the last part of the proof of Lemma 24 with $\lambda_{r_0} = \mu_{k_0+1} = \Theta(d^{-k_0-1})$. It then follows from (108) and the Cauchy-Schwarz inequality that for every $f \in \widehat{\mathcal{F}}_r$,

$$\mathbb{E}_P\left[(\mathbb{P}_{\mathcal{H}_{r_0}}(h) - f^*)^2\right] \leq 2\mathbb{E}_P\left[(h-f^*)^2\right] + 2\mathbb{E}_P\left[\bar{h}^2\right] \leq 4(r+w^2) + 2r_{n,k_0,\delta}. \tag{109}$$

We then have

$$\begin{aligned}&\mathfrak{R}\left(\left\{\mathbb{P}_{\mathcal{H}_{r_0}}(h) - f^* \colon f \in \widehat{\mathcal{F}}_r\right\}\right) \\ &\overset{③}{\leq} \mathfrak{R}\left(\left\{\mathbb{P}_{\mathcal{H}_{r_0}}(h) - f^* \colon \mathbb{E}_P\left[(\mathbb{P}_{\mathcal{H}_{r_0}}(h) - f^*)^2\right] \leq 4(r+w^2) + 2r(n,k_0,\delta)\right\}\right) \\ &\overset{④}{\leq} 2\mathfrak{R}\left(\left\{f \in \mathcal{H}_{K^{(r_0)}}(B_h) \colon \mathbb{E}_P\left[f^2\right] \leq r + w^2 + \frac{r_{n,k_0,\delta}}{2}\right\}\right) \overset{⑤}{\leq} \sqrt{r + w^2 + \frac{r_{n,k_0,\delta}}{2}}\cdot\sqrt{\frac{r_0}{n}}.\end{aligned} \tag{110}$$

Here ③ follows from (109). Since $\mathbb{P}_{\mathcal{H}_{r_0}}(f), f^* \in \mathcal{H}_{r_0} \cap \mathcal{H}_K(B_h) \subseteq \mathcal{H}_{K^{(r_0)}}(B_h)$, we have $(\mathbb{P}_{\mathcal{H}_{r_0}}(f) - f^*)/2 \in \mathcal{H}_{K^{(r_0)}}(B_h)$ due to the fact that $\mathcal{H}_{K^{(r_0)}}(B_h)$ is symmetric and convex, and it follows that ④ holds. ⑤ follows from Lemma 28 with $Q = r_0$ in (115) of Lemma 28.

We then derive the upper bound for $\mathfrak{R}\left(\left\{\bar{h} \colon f \in \widehat{\mathcal{F}}_r\right\}\right)$. First, it follows from Theorem 14 and the argument similar to (108) that

$$\left\|\bar{h}\right\|_{\mathcal{H}_K}^2 = \sum_{j\geq r_0}\alpha_j^2 \leq \zeta_{n,B_h,r_0,\delta} \leq B_h^2\log\frac{2}{\delta}\cdot\Theta\left(\frac{d^{2k_0}}{n}\right) \coloneqq B_{\bar{h}}^2. \tag{111}$$

We then have

$$\mathfrak{R}\left(\left\{\bar{h} \colon f \in \widehat{\mathcal{F}}_r\right\}\right) = \mathbb{E}_{\left\{\vec{\mathbf{x}}_i\right\}_{i=1}^n, \{\sigma_i\}_{i=1}^n}\left[\sup_{\bar{h}\in\mathcal{H}_K(B_{\bar{h}})}\frac{1}{n}\sum_{i=1}^n\sigma_i\bar{h}(\vec{\mathbf{x}}_i)\right]$$

$$\leq \frac{B_{\bar{h}}}{n}\mathbb{E}_{\left\{\vec{\mathbf{x}}_i\right\}_{i=1}^n,\{\sigma_i\}_{i=1}^n}\left[\sup_{f\in\mathcal{F}}\left\|\sum_{i=1}^n \sigma_i K(\cdot,\vec{\mathbf{x}}_i)\right\|_{\mathcal{H}_K}\right] \leq \frac{B_{\bar{h}}}{\sqrt{n}} \leq \sqrt{\log\frac{2}{\delta}}\cdot\Theta\left(\frac{d^{k_0}}{n}\right). \tag{112}$$

Finally, it follows from (110) and (112) that

$$\begin{aligned}\mathfrak{R}\left(\left\{f-f^*\colon f\in\widehat{\mathcal{F}}_r\right\}\right) &\leq \mathfrak{R}\left(\left\{\mathbb{P}_{\mathcal{H}_{r_0}}(h)-f^*\colon f\in\widehat{\mathcal{F}}_r\right\}\right)+\mathfrak{R}\left(\left\{\bar{h}\colon f\in\widehat{\mathcal{F}}_r\right\}\right)+w\\ &\leq \sqrt{\log\frac{2}{\delta}}\cdot\Theta\left(\frac{d^{k_0}}{n}\right)+w\sqrt{\frac{r_0}{n}}+\sqrt{\frac{rr_0}{n}}+w \leq \sqrt{\log\frac{2}{\delta}}\cdot\Theta\left(\frac{d^{k_0}}{n}\right)+\sqrt{\frac{rr_0}{n}}+2w,\end{aligned} \tag{113}$$

which proves (107).

■

Lemma 28 ([37, Lemma C.9],[38, Lemma VI.4]) *For every $B, w > 0$, the function class $\mathcal{F}(B,w)$ is defined as $\mathcal{F}(B,w) \coloneqq \{f\colon f = h + e, h \in \mathcal{H}_K(B), \|e\|_\infty \leq w\}$. Then for every $r > 0$,*

$$\mathfrak{R}\left(\left\{f\in\mathcal{F}(B,w)\colon \mathbb{E}_P\left[f^2\right]\leq r\right\}\right) \leq \varphi_{B,w}(r), \tag{114}$$

where

$$\varphi_{B,w}(r) \coloneqq \min_{Q\colon Q\geq 0}\left((\sqrt{r}+w)\sqrt{\frac{Q}{n}}+B\left(\frac{\sum_{q=Q+1}^{\infty}\lambda_q}{n}\right)^{1/2}\right)+w. \tag{115}$$

Lemma 29 ([38, Lemma B.9]) *Suppose $\psi\colon [0,\infty)\to[0,\infty)$ is a sub-root function with the unique fixed point r^*. Then the following properties hold.*

(1) Let $a \geq 0$, then $\psi(r)+a$ as a function of r is also a sub-root function with fixed point r_a^*, and $r^* \leq r_a^* \leq r^* + 2a$.

(2) Let $b \geq 1$, $c \geq 0$ then $\psi(br+c)$ as a function of r is also a sub-root function with fixed point r_b^*, and $r_b^* \leq br^* + 2c/b$.

(3) Let $b \geq 1$, then $\psi_b(r) = b\psi(r)$ is also a sub-root function with fixed point r_b^*, and $r_b^* \leq b^2 r^*$.

D.5. Proofs of Theorem 14

Proof [Proof of Theorem 14] With probability at least $1-\delta$, we have

$$\begin{aligned}\sum_{q=r_0}^{\infty}\left\langle f^*,\Phi^{(q)}\right\rangle_{\mathcal{H}_K}^2 &= \sum_{q=r_0}^{\infty}\left\langle\sum_{j=0}^{r_0-1}\beta_j v_j,\Phi^{(q)}\right\rangle_{\mathcal{H}_K}^2 \leq \sum_{q=r_0}^{\infty}\sum_{j=0}^{r_0-1}\beta_j^2\cdot\sum_{j=0}^{r_0-1}\left\langle v_j,\Phi^{(q)}\right\rangle^2\\ &\leq \gamma_0^2\sum_{q=r_0}^{\infty}\sum_{j=0}^{r_0-1}\left\langle v_j,\Phi^{(q)}\right\rangle^2 \leq \frac{32\gamma_0^2\log\frac{2}{\delta}}{\left(\mu_{k_0}-\mu_{k_0+1}\right)^2 n}.\end{aligned}$$

Here the last inequality follows by Lemma 30 with $\tau_0^2 = 1$ and $m_{k_0} = r_0$, which proves (33). Since $f \in \mathcal{F}(B_h, w, \mathbf{S}, r_0)$, we have $f = \sum_{j=0}^{r_0-1} \alpha_j \Phi^{(j)}$ with $\alpha_j = \left\langle f, \Phi^{(j)} \right\rangle_{\mathcal{H}_K}$ for $j \in [0, r_0 - 1]$. Following a similar argument, we have

$$\sum_{q=r_0}^{\infty} \langle f, v_q \rangle_{\mathcal{H}_K}^2 = \sum_{q=r_0}^{\infty} \left\langle \sum_{j=0}^{r_0-1} \alpha_j \Phi^{(j)}, v_q \right\rangle_{\mathcal{H}_K}^2 \le \sum_{q=r_0}^{\infty} \sum_{j=0}^{r_0-1} \alpha_j^2 \cdot \sum_{j=0}^{r_0-1} \left\langle \Phi^{(j)}, v_q \right\rangle^2$$
$$\le B_h^2 \sum_{q=r_0}^{\infty} \sum_{j=0}^{r_0-1} \left\langle \Phi^{(j)}, v_q \right\rangle^2 \le \frac{32 B_h^2 \log \frac{2}{\delta}}{(\mu_{k_0} - \mu_{k_0+1})^2 n},$$

which proves (34). ■

Lemma 30 *Let* $\sup_{\mathbf{x} \in \mathcal{X}} K(\mathbf{x}, \mathbf{x}) = \tau_0^2$. *For any* $\delta \in (0, 1)$, *with probability at least* $1 - \delta$ *over the random training features* $\mathbf{S}$,

$$\sum_{i=0}^{m_{k_0}-1} \sum_{j \ge m_{k_0}} \left\langle \Phi^{(i)}, v_j \right\rangle_{\mathcal{H}}^2 + \sum_{i \ge m_{k_0}} \sum_{j=0}^{m_{k_0}-1} \left\langle \Phi^{(i)}, v_j \right\rangle_{\mathcal{H}}^2 \le \frac{32 \tau_0^4 \log \frac{2}{\delta}}{(\mu_{k_0} - \mu_{k_0+1})^2 n}. \tag{116}$$

Proof Define operator $T_n \colon \mathcal{H}_K \to \mathcal{H}_K$ by $T_n g = \frac{1}{n} \sum_{i=1}^n K(\cdot, \vec{\mathbf{x}}_i) g(\vec{\mathbf{x}}_i)$ as introduced before Lemma 24, and let $\left\{\Phi^{(k)}\right\}_{k \ge 0}$ be an orthonormal basis of the RKHS $\mathcal{H}_K$.

Let P_N^T be an orthogonal projection operator which projects any input onto the subspace spanned by eigenfunctions corresponding to the top N eigenvalues of the operator T, and T is defined on the RKHS $\mathcal{H}$.

We now work on the following two orthogonal projection operators, $P_{m_{k_0}}^{T_K}$ and $P_{m_{k_0}}^{T_n}$. Each of the two operators projects its input onto the space spanned by all the eigenfunctions of the corresponding operator, that is.

$$P_{m_{k_0}}^{T_K} h = \sum_{j=0}^{m_{k_0}-1} \langle h, v_j \rangle_{\mathcal{H}} v_j, \quad P_{m_{k_0}}^{T_n} h = \sum_{j=0}^{m_{k_0}-1} \left\langle h, \Phi^{(j)} \right\rangle_{\mathcal{H}} \Phi^{(j)}. \tag{117}$$

The Hilbert-Schmidt norm of $P_{m_{k_0}}^{T_K} - P_{m_{k_0}}^{T_n}$ is

$$\left\| P_{m_{k_0}}^{T_K} - P_{m_{k_0}}^{T_n} \right\|_{\text{HS}}^2 = \sum_{i \ge 0, j \ge 0} \left\langle \left(P_{m_{k_0}}^{T_K} - P_{m_{k_0}}^{T_n} \right) \Phi^{(i)}, v_j \right\rangle_{\mathcal{H}}^2, \tag{118}$$

which is due to the fact that both $\left\{\Phi^{(j)}\right\}_{j \ge 0}$ and $\{v_j\}_{j \ge 0}$ are orthonormal bases of $\mathcal{H}$. It can be verified that

$$\left\langle \left(P_{m_{k_0}}^{T_K} - P_{m_{k_0}}^{T_n} \right) \Phi^{(i)}, v_j \right\rangle_{\mathcal{H}} = \begin{cases} 0 & \text{if } i < m_{k_0}, j < m_{k_0}, \\ -\left\langle \Phi^{(i)}, v_j \right\rangle_{\mathcal{H}} & \text{if } i < m_{k_0}, j \ge m_{k_0}, \\ \left\langle \Phi^{(i)}, v_j \right\rangle_{\mathcal{H}} & \text{if } i \ge m_{k_0}, j < m_{k_0}, \\ 0 & \text{if } i \ge m_{k_0}, j \ge m_{k_0}, \end{cases} \tag{119}$$

and similar results are obtained in the proof of [28, Theorem 12].

Because T_K and T_n are Hilbert-Schmidt operators, by [28, Theorem 7], for all $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\|T_K - T_n\|_{\mathrm{HS}} \le \frac{2\sqrt{2}\tau_0^2\sqrt{\log\frac{2}{\delta}}}{\sqrt{n}}. \tag{120}$$

When $n \ge \frac{128\tau_0^4 \log\frac{2}{\delta}}{\left(\mu_{k_0}-\mu_{k_0+1}\right)^2}$, $\|T_K - T_n\|_{\mathrm{HS}} \le \frac{\mu_{k_0}-\mu_{k_0+1}}{4}$. It follows from [28, Proposition 6] and noting that the operator norm in [28, Proposition 6] can be replaced by the Hilbert-Schmidt norm,

$$\left\|P_{m_{k_0}}^{T_K} - P_{m_{k_0}}^{T_n}\right\|_{\mathrm{HS}}^2 \le \frac{4}{\left(\mu_{k_0}-\mu_{k_0+1}\right)^2}\|T_K - T_n\|_{\mathrm{HS}}^2 \le \frac{32\tau_0^4\log\frac{2}{\delta}}{\left(\mu_{k_0}-\mu_{k_0+1}\right)^2 n}. \tag{121}$$

Then (116) follows from (118), (119), and (121).

■

D.6. Results about Eigenvalues of the Integral Operators

The following theorem is a refined version of the Mercer's theorem on the PSD kernel K defined in (2), with the exact estimation about the decaying rate of the distinct eigenvalues $\{\mu_\ell\}_{\ell\ge 0}$.

Theorem 31 (Eigenvalue of The Integral Operator Associated with the NTK (2)) *Let the distinct eigenvalues of the integral operator T_K associated with the PSD kenrel K defined in (2) be $\{\mu_\ell : \ell \ge 0\}$ with $\mu_0 > \mu_1 > \ldots$, where μ_ℓ is the eigenvalue corresponding to $\mathcal{H}_\ell$. Suppose that $\bar{k}_0 = \Theta(1)$ and $d \ge \Theta(1)$. Then $\mu_k = \Theta(d^{-k})$ for $0 \le k \le \bar{k}_0$. Moreover, for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X} = \mathbb{S}^{d-1}$,*

$$K(\mathbf{x},\mathbf{x}') = \sum_{\ell\ge 0}\mu_\ell \sum_{j=1}^{N(d,\ell)} Y_{\ell,j}(\mathbf{x})Y_{\ell,j}(\mathbf{x}') = \sum_{\ell\ge 0}\mu_\ell N(d,\ell)P_\ell(\langle \mathbf{x},\mathbf{x}'\rangle), \tag{122}$$

where μ_ℓ is the eigenvalue of the integral operator T_K associated with K corresponding to $\mathcal{H}_\ell$, and $\{Y_{\ell,j}\}_{j=1}^{N(d,\ell)}$ are the eigenfunctions corresponding to the eigenvalue μ_ℓ. That is, $T_K Y_{\ell,j} = \mu_\ell Y_{\ell,j}$ for all $\ell \ge 0$ and $j \in [N(d,\ell)]$. The series on the RHS of (122) converges absolutely and uniformly on $\mathcal{X}\times\mathcal{X}$.

Proof (122) follows from the background about Harmonic Analysis on spheres in Section C and the Mercer's theorem. Since K is a continuous PSD kernel defined on the compact set $\mathcal{X}\times\mathcal{X}$, it follows from the Mercer's theorem again that the series on the RHS of (122) converges absolutely and uniformly on $\mathcal{X}\times\mathcal{X}$ to K.

We now set to compute the eigenvalues $\left\{\lambda_k : 0 \le k \le \bar{k}_0\right\}$. Let the distinct eigenvalues of the PSD kenrel $K^{(0)}$, which is defined in (2) and repeated below

$$K^{(0)}(\mathbf{x},\mathbf{x}') = \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}\left[\mathbb{1}_{\{\mathbf{x}^\top\mathbf{w}\ge 0\}}\mathbb{1}_{\{\mathbf{x}'^\top\mathbf{w}\ge 0\}}\right] = \frac{\pi - \arccos(\mathbf{x}^\top\mathbf{x}')}{2\pi}, \quad \forall \mathbf{x},\mathbf{x}' \in \mathcal{X} = \mathcal{X},$$

be $\{\lambda_{0,k} : k \ge 0\}$, where $\lambda_{0,k}$ is the eigenvalue corresponding to $\mathcal{H}_k$, the space of degree-ℓ homogeneous harmonic polynomials on $\mathcal{X} = \mathbb{S}^{d-1}$.

Define

$$s_k \coloneqq \frac{\omega_{d-2}}{\omega_{d-1}} \int_{-1}^{1} \mathbb{I}_{\{t\geq 0\}} P_k(t)(1-t^2)^{(d-3)/2}\mathrm{d}t,$$

It then follows by the computation in [3, Section D.2] that $s_0 = \Theta(1)$. Also, for all $t \in \mathbb{N}$, $s_{2t} = 0$, and

$$\begin{aligned} s_{2t-1} &= \frac{\omega_{d-2}}{\omega_{d-1}} \left(\frac{1}{2}\right)^{2t-1} (-1)^{t-1} \frac{\Gamma((d-1)/2)\Gamma(2t-1)}{\Gamma(t)\Gamma(t+(d-1)/2)} \\ &\overset{①}{\asymp} (-1)^{t-1}\sqrt{d}\frac{(d-1)^{\frac{d}{2}-1}(2t-1)^{2t-1.5}}{(2t)^{t-0.5}(2t+d-1)^{t+\frac{d}{2}-1}} \overset{②}{\asymp} \frac{1}{d^{t-0.5}}, \end{aligned}$$

where we used the approximation to the Gamma function [17] $\Gamma(x) \asymp x^{x-0.5}\exp(-x)\sqrt{2\pi}$ and the fact that $\frac{\omega_{d-2}}{\omega_{d-1}} \asymp \sqrt{d}$ in ①. ② is due to $t = \Theta(1)$.

It follows from [6] that $\lambda_{0,k} = s_k^2$ for all $k \geq 0$. When $k = 2t-1$ for $t \in \mathbb{N}$, we have $\lambda_{0,k} = s_k^2 = \Theta(d^{-(2t-1)}) = \Theta(d^{-k})$. Moreover, $\lambda_{0,k} = 0$ for all $k = 2t$ with $t \in \mathbb{N}$, and $\lambda_{0,0} = s_0^2 = \Theta(1)$. As a result, we have $\lambda_{0,0} = \Theta(1)$, and

$$\lambda_{0,k} = \begin{cases} 0 & k = 2t, t \in \mathbb{N}, k \leq \bar{k}_0, \\ \Theta(d^{-k}) & k = 2t-1, t \in \mathbb{N}, k \leq \bar{k}_0. \end{cases} \tag{123}$$

Let the distinct eigenvalues of the PSD kenrel $K^{(1)}$ which is also defined in (2) be $\{\lambda_{1,k} \colon k \geq 0\}$, where $\lambda_{0,k}$ is the eigenvalue corresponding to $\mathcal{H}_k$. Define $\kappa(t) = t\kappa^{(0)}(t)$ with $\kappa^{(0)}(t) \coloneqq \frac{\pi - \arccos t}{2\pi}$ for $t \in [-1,1]$. Then for $k \geq 1$ we have

$$\begin{aligned} \lambda_{1,k} &= \frac{\omega_{d-2}}{\omega_{d-1}} \int_{-1}^{1} \kappa(t)P_k(t)(1-t^2)^{(d-3)/2}\mathrm{d}t = \frac{\omega_{d-1}}{\omega_{d-2}} \int_{-1}^{1} \kappa^{(0)}(t)t(1-t^2)^{(d-3)/2}P_k(t)\mathrm{d}t \\ &= \frac{\omega_{d-2}}{\omega_{d-1}} \int_{-1}^{1} \kappa^{(0)}(t) \left(\frac{k}{2k+d-2}P_{k-1}(t) + \frac{k+d-2}{2k+d-2}P_{k+1}(t)\right)(1-t^2)^{(d-3)/2}\mathrm{d}t \\ &= \frac{k}{2k+d-2}\lambda_{0,k-1} + \frac{k+d-2}{2k+d-2}\lambda_{0,k+1}. \end{aligned} \tag{124}$$

Moreover,

$$\lambda_{1,0} = \frac{\omega_{d-2}}{\omega_{d-1}} \int_{-1}^{1} \kappa(t)P_0(t)(1-t^2)^{(d-3)/2}\mathrm{d}t = \frac{\omega_{d-2}}{\omega_{d-1}} \int_{-1}^{1} \kappa^{(0)}(t)P_1(t)(1-t^2)^{(d-3)/2}\mathrm{d}t = \lambda_{0,1}. \tag{125}$$

It follows from (123)-(125) that $\lambda_{1,0} = \Theta(1/d)$, $\lambda_{1,1} = \Theta(1/d)$. Moreover,

$$\lambda_{1,k} = \begin{cases} \Theta(d^{-k}) & k = 2t, t \in \mathbb{N}, k \leq \bar{k}_0, \\ 0 & k = 2t-1, t \in \mathbb{N}, t \geq 2, k \leq \bar{k}_0. \end{cases} \tag{126}$$

It then follows from (123) and (126) that $\mu_k = \lambda_{0,k} + \lambda_{1,k} = \Theta(d^{-k})$ for $0 \leq k \leq \bar{k}_0$ ■

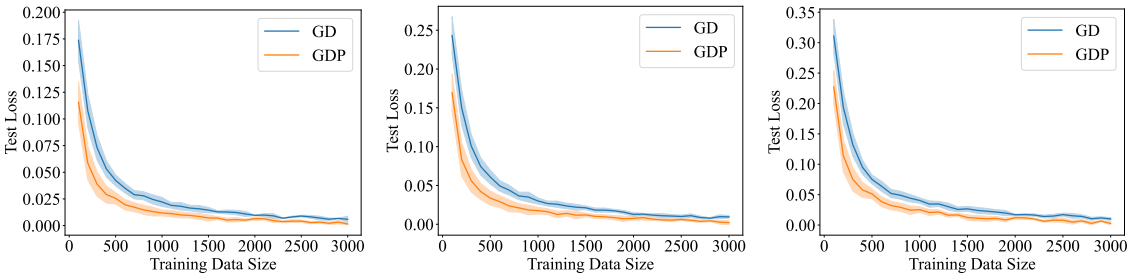

Figure 1: Test loss for target functions with degree $k_0 = 1, 2, 3$ for varying n in $[100, 3000]$ with a step size of 100. The shaded area in each plot indicates the standard deviation across 10 random initializations of the neural network.

Algorithm 2 Training the Two-Layer NN by GD

1: $\mathbf{W}(T) \leftarrow$ Training-by-GD$(T, \mathbf{W}(0))$
2: **input:** $T, \mathbf{W}(0)$
3: **for** $t = 1, \ldots, T$ **do**
4: Perform the t-th step of GD by

$$\text{vec}\left(\mathbf{W}(t+1)\right) - \text{vec}\left(\mathbf{W}(t)\right) = -\frac{\eta}{n}\mathbf{Z}_{\mathbf{S}}(t)(\widehat{\mathbf{y}}(t) - \mathbf{y})$$

5: **end for**
6: **return** $\mathbf{W}(T)$

Appendix E. Simulation Study

We present simulation results in this section. We randomly sample n points $\left\{\vec{\mathbf{x}}_i\right\}_{i=1}^n$ as a i.i.d. sample of random variables distributed uniformly on the unit sphere $\mathbb{S}^9$ in $\mathbb{R}^{10}$, and the variance of the noise is set to $\sigma_0^2 = 1$. n ranges within $[100, 3000]$ with a step size of 100. We set the target function as the degree-k_0 spherical polynomial as $f^*(\mathbf{x}) = (\mathbf{s}^\top \mathbf{x})^{k_0}$ where $\mathbf{x} \in \mathbb{S}^9$ and $\mathbf{s} \sim \mathrm{Unif}(\mathcal{X})$ is randomly sampled. We also uniformly and independently sample 1000 points on $\mathbb{S}^9$ as the test data. We train the two-layer NN (1) using GDP by Algoirthm 1 with $m = 10000$ on an NVIDIA A100 GPU card with a learning rate $\eta = 1$, and illustrate the test loss in Figure 1. It can be observed that GDP always demonstrates better generalization than the vanilla gradient descent (GD) described in Algoirthm 2 through lower test losses across different training data size. The experiments are performed for the target function f^* with $k_0 \in \{1, 2, 3\}$. Figure 1 illustrates the test loss for each n in $[100, 3000]$ with a step size of 100.