Submitted 7/23; Revised 1/24; Published 3/24

On the Eigenvalue Decay Rates of a Class of Neural-Network Related Kernel Functions Defined on General Domains

Yicheng Li liyc22@mails.tsinghua.edu.cn
Center for Statistical Science, Department of Industrial Engineering
Tsinghua University
Beijing, 100084, China

Zixiong Yu yuzx19@mails.tsinghua.edu.cn
Yau Mathematical Sciences Center, Department of Mathematical Sciences
Tsinghua University
Beijing, 100084, China

Guhan Chen chen-gh23@mails.tsinghua.edu.cn
Center for Statistical Science, Department of Industrial Engineering
Tsinghua University
Beijing, 100084, China

Qian Lin[*] qianlin@tsinghua.edu.cn
Center for Statistical Science, Department of Industrial Engineering
Tsinghua University
Beijing, 100084, China

Editor: Ohad Shamir

Abstract

In this paper, we provide a strategy to determine the eigenvalue decay rate (EDR) of a large class of kernel functions defined on a general domain rather than $\mathbb{S}^d$. This class of kernel functions include but are not limited to the neural tangent kernel associated with neural networks with different depths and various activation functions. After proving that the dynamics of training the wide neural networks uniformly approximated that of the neural tangent kernel regression on general domains, we can further illustrate the minimax optimality of the wide neural network provided that the underground truth function $f \in [\mathcal{H}_{\mathrm{NTK}}]^s$, an interpolation space associated with the RKHS $\mathcal{H}_{\mathrm{NTK}}$ of NTK. We also showed that the overfitted neural network can not generalize well. We believe our approach for determining the EDR of kernels might be also of independent interests.

Keywords: Neural tangent kernel, eigenvalue decay rate, early stopping, non-parametric regression, reproducing kernel Hilbert space

1. Introduction

Deep neural networks have achieved incredible success in a variety of areas, from image classification (He et al., 2016; Krizhevsky et al., 2017) to natural language processing (Devlin et al., 2019), generative models (Karras et al., 2019), and beyond. The number of parameters appearing in modern deep neural networks is often ten or hundreds of times

[*] Corresponding author.

larger than the sample size of the data. It is widely observed that large neural networks possess smaller generalization errors than traditional methods. This "benign overfitting phenomenon" brings challenges to the usual bias-variance trade-off doctrine in statistical learning theory. Understanding the mysterious generalization power of deep neural networks might be one of the most interesting statistical problems.

Although the training dynamics of neural networks is highly non-linear and non-convex, the celebrated neural tangent kernel (NTK) theory (Jacot et al., 2018) provides us a way to study the generalization ability of over-parametrized neural networks. It is shown that when the width of neural networks is sufficiently large (i.e., in the over-parameterized or lazy trained regime), the training dynamics of the neural network can be well approximated by a simpler kernel regression method with respect to the corresponding NTK. Consequently, it offers us a way to investigate the generalization ability of the over-parametrized neural network by means of the well established theory of generalization in kernel regression (Caponnetto and De Vito, 2007; Andreas Christmann, 2008; Lin et al., 2018).

However, to obtain the generalization results in kernel regression, the eigenvalue decay rate (EDR) of the kernel (see (3) and below) is an essential quantity that must be determined a priori. Considering the NTKs associated with two-layer and multilayer fully-connected ReLU neural networks, Bietti and Mairal (2019) and the subsequent work Bietti and Bach (2020) showed that the EDR of the NTKs is $i^{-(d+1)/d}$ when the inputs are uniformly distributed on $\mathbb{S}^d$. Consequently, Hu et al. (2021) and Suh et al. (2022) claimed that the neural network can achieve the minimax rate $n^{-(d+1)/(2d+1)}$ of the excess risk. However, their assumption on the input distribution is too restrictive, and can hardly be satisfied in practice, so it is of interest to determine the EDR of the NTKs for general input domains and distributions. As far as we know, few works have studied the EDR of the NTKs beyond the case of uniform distribution on $\mathbb{S}^d$. More recently, focusing on one dimensional data over an interval, Lai et al. (2023) showed that the EDR of the NTK associated with two-layer neural networks is i^{-2} and thus the neural network can achieve the minimax rate $n^{-2/3}$ of the excess risk. However, their approach of determining the EDR, which relies heavily on the closed form expression of the NTK, can not be generalized to d-dimensional inputs or the NTK associated with multilayer neural networks.

In this work, we study the EDR of the NTKs associated with multilayer fully-connected ReLU neural networks on a general domain in $\mathbb{R}^d$ with respect to a general input distribution μ satisfying mild assumptions. For this purpose, we develop a novel approach for determining the EDR of kernels by transformation and restriction. As a key contribution, we prove that the EDR of a dot-product kernel on the sphere remains the same if one restricts it to a subset of the sphere, which is a non-trivial generalization of the result in Widom (1963). Consequently, we can show that the EDR of the NTKs is $i^{-(d+1)/d}$ for general input domains and distributions. Moreover, after proving the uniform approximation of the over-parameterized neural network by the NTK regression, we show the statistical optimality of the over-parameterized neural network trained via gradient descent with proper early stopping. In comparison, we also show that the overfitted neural network can not generalize well.

1.1 Related works

The EDR of NTKs The spectral properties of NTK have been of particular interests to the community of theorists since Jacot et al. (2018) introduced the neural tangent kernel. For example, noticing that the NTKs associated with fully-connected ReLU networks are inner product kernels on the sphere, several works utilized the theory of the spherical harmonics (Dai and Xu, 2013; Azevedo and Menegatto, 2014) to study the eigen-decomposition of the NTK (Bietti and Mairal, 2019; Ronen et al., 2019; Geifman et al., 2020; Chen and Xu, 2020; Bietti and Bach, 2020). In particular, Bietti and Mairal (2019) and Bietti and Bach (2020) showed that the EDR of the NTKs associated with the two-layer and multilayer neural network is $i^{-(d+1)/d}$ if the inputs are uniformly distributed on $\mathbb{S}^d$. However, their analysis depends on the spherical harmonics theory on the sphere to derive the explicit expression of the eigenvalues, which cannot be extended to general input domains and distributions. Recently, considering two-layer ReLU neural networks on an interval, Lai et al. (2023) showed that the EDR of the corresponding NTK is i^{-2}. However, their technique relies heavily on the explicit expression of the NTK on $\mathbb{R}$ and can hardly be extended to NTKs defined on $\mathbb{R}^d$ or NTKs associated with multilayer wide networks.

The generalization performance of over-parameterized neural networks Though now it is a common strategy to study the generalization ability of over-parameterized neural networks through that of the NTK regression, few works state it explicitly or rigorously. For example, Du et al. (2018); Li and Liang (2018); Arora et al. (2019a) showed that the training trajectory of two-layer neural networks converges pointwisely to that of the NTK regressor; Du et al. (2019); Allen-Zhu et al. (2019a); Lee et al. (2019) further extended the results to the multilayer networks and ResNet. However, if one wants to approximate the generalization error of over-parameterized neural network by that of the NTK regressor, the approximation of the neural network by the kernel regressor has to be uniform. Unfortunately, the existing two works (Hu et al., 2021; Suh et al., 2022) studying the generalization error of over-parameterized neural networks overlooked the aforementioned subtle difference between the pointwise convergence and uniform convergence, so there might be some gaps in their claims. To the best of our knowledge, Lai et al. (2023) might be one of the first works who showed the two-layer wide ReLU neural networks converge uniformly to the corresponding NTK regressor.

The high-dimensional setting It should be also noted that several other works tried to consider the generalization error of NTK regression in the high-dimensional setting, where the dimension of the input diverges as the number of samples tends to infinity. These works include the eigenvalues of NTK, the "benign overfitting phenomenon", the "double descent phenomenon", and the generalization error. For example, Frei et al. (2022), Nakkiran et al. (2019) and Liang and Rakhlin (2020) have shown the benign overfitting and double descent phenomena, while Fan and Wang (2020) and Nguyen et al. (2021) have investigated the eigenvalue properties of NTK in the high-dimensional setting. Furthermore, recent works by Montanari and Zhong (2022) have examined the generalization performance of neural networks in the high-dimensional setting. However, it has been suggested by Rakhlin and Zhai (2018); Beaglehole et al. (2022) that there may be differences between the traditional fixed-dimensional setting and the high-dimensional setting. In this work, we focus solely on the fixed-dimensional setting.

1.2 Our contributions

The main contribution of this paper is that we determine the EDR of the NTKs associated with multilayer fully-connected ReLU neural networks on $\mathbb{R}^d$ with respect to a general input distribution μ satisfying mild assumptions. We develop a novel approach for determining the EDR of kernels by means of algebraic transformation and restriction to subsets: if the kernel can be transformed to a dot-product kernel on the sphere, its EDR on a general domain coincides with the EDR of the resulting dot-product kernel with respect to the uniform distribution over the entire sphere, while the latter can be determined more easily by the theory of spherical harmonics. In particular, we show that the EDR of the considered NTKs is $i^{-(d+1)/d}$, which coincides with that of the NTKs on the sphere. Besides, we also prove that the NTKs are strictly positive definite. As a key technical contribution, we prove that the EDR of a dot-product kernel on the sphere remains the same if one restricts it to a subset of the sphere, provided that the EDR of the kernel satisfies a very mild assumption. This result is a non-trivial generalization of the result on shift-invariant kernels in Widom (1963) and its proof involves fine-grained harmonic analysis on the sphere. We believe that our approach is also of independent interest in the research of kernel methods.

Another contribution of this paper is that we rigorously prove that the over-parameterized multilayer neural network trained by gradient descent can be approximated uniformly by the corresponding NTK regressor. Combined with the aforementioned EDR result, this uniform approximation allows us to characterize the generalization performance of the neural network through the well-established kernel regression theory. The theoretical results show that proper early stopping is essential for the generalization performance of the neural networks, which urges us to scrutinize the widely reported "benign overfitting phenomenon" in deep neural network literature.

1.3 Notations

For two sequences $a_n, b_n,\ n \geq 1$ of non-negative numbers, we write $a_n = O(b_n)$ (or $a_n = \Omega(b_n)$) if there exists absolute constant $C > 0$ such that $a_n \leq C b_n$ (or $a_n \geq C b_n$). We also denote $a_n \asymp b_n$ (or $a_n = \Theta(b_n)$) if $a_n = O(b_n)$ and $a_n = \Omega(b_n)$. For a function $f : \mathcal{X} \to \mathbb{R}$, we denote by $\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|$ the sup-norm of f. We denote by $L^p(\mathcal{X}, \mathrm{d}\mu)$ the Lebesgue L^p-space over $\mathcal{X}$ with respect to μ.

2. Analysis of Eigenvalue Decay Rate

The neural tangent kernel (NTK) theory (Jacot et al., 2018) has been widely used to explain the generalization ability of neural networks, which establishes a connection between neural networks and kernel methods (Caponnetto and De Vito, 2007; Bauer et al., 2007). In the framework of kernel methods, the spectral properties, in particular the eigenvalue decay rate, of the kernel function are crucial in the analysis of the generalization ability. Although there are several previous works (Bietti and Mairal, 2019; Chen and Xu, 2020; Geifman et al., 2020; Bietti and Bach, 2020) investigating the spectral properties of NTKs on the sphere, their results are limited to the case where the input distribution is uniform on the sphere. Therefore, we would like to determine the spectral properties of NTKs on a general domain with a general input distribution. In this section, we provide some general results on the asymptotic behavior of the eigenvalues of certain type of kernels. As a consequence, we are able to determine the eigenvalue decay rate of NTKs on a general domain.

2.1 The integral operator and the eigenvalues

Let $\mathcal{X}$ be a Hausdorff space and μ be a Borel measure on $\mathcal{X}$. In the following, we always consider a continuous positive definite kernel $k(x,x') : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that

$$\int_{\mathcal{X}} k(x,x) \mathrm{d}\mu(x) < \infty. \tag{1}$$

We denote by $L^2 = L^2(\mathcal{X}, \mathrm{d}\mu)$ and $\mathcal{H}$ the reproducing kernel Hilbert space (RKHS) associated with k. Introduce the integral operator $T = T_{k;\mathcal{X},\mu} : L^2 \to L^2$ by

$$(Tf)(x) = \int_{\mathcal{X}} k(x,x') f(x') \mathrm{d}\mu(x'). \tag{2}$$

It is well-known (Andreas Christmann, 2008; Steinwart and Scovel, 2012) that T is self-adjoint, positive and trace-class (hence compact). Consequently, we can derive the spectral decomposition of T and also the Mercer's decomposition of k as

$$T = \sum_{i \in N} \lambda_i \left\langle \cdot, e_i \right\rangle_{L^2} e_i, \qquad k(x,x') = \sum_{i \in N} \lambda_i e_i(x) e_i(x'), \tag{3}$$

where $N \subseteq \mathbb{N}$ is an index set ($N = \mathbb{N}$ if the space is infinite dimensional), $(\lambda_i)_{i \in N}$ is the set of positive eigenvalues (counting multiplicities) of T in descending order and $(e_i)_{i \in N}$ are the corresponding eigenfunction, which are an orthonormal set in $L^2(\mathcal{X}, \mathrm{d}\mu)$. To emphasize the dependence of the eigenvalues on the kernel and the measure, we also denote by $\lambda_i(k; \mathcal{X}, \mathrm{d}\mu) = \lambda_i$. We refer to the asymptotic rate of λ_i as i tends to infinity as the eigenvalue decay rate (EDR) of k with respect to $\mathcal{X}$ and μ.

In the kernel regression literature, the EDR of the kernel is closely related to the capacity condition of the corresponding reproducing kernel Hilbert space (RKHS) and affects the rate of convergence of the kernel regression estimator (see, e.g., Caponnetto and De Vito (2007); Lin et al. (2018)). Particularly, a power-law decay that $\lambda_i \asymp i^{-\beta}$ is often assumed in the literature and the corresponding minimax optimal rate depends on the exponent β. Therefore, it would be helpful to determine such decay rate for a kernel of interest.

2.2 Preliminary results on the eigenvalues

In this subsection, we present some preliminary results on the eigenvalues of T, which allow us to manipulate the kernel with algebraic transformations to simplify the analysis. Let us first define the scaled kernel $(\rho \odot k)(x,x') = \rho(x) k(x,x') \rho(x')$ for some function $\rho : \mathcal{X} \to \mathbb{R}$. It is easy to see the following:

Proposition 1 *Let $\rho : \mathcal{X} \to \mathbb{R}$ be a measurable function such that $\rho \odot k$ satisfies* (1)*. Then,*

$$\lambda_i(\rho \odot k; \mathcal{X}, \mathrm{d}\mu) = \lambda_i(k; \mathcal{X}, \rho^2 \mathrm{d}\mu). \tag{4}$$

Furthermore, if ρ is bounded, we can further estimate the eigenvalues using the minimax principle on the eigenvalues of self-adjoint compact positive operators.

Lemma 2 *Let a measurable function $\rho : \mathcal{X} \to \mathbb{R}$ satisfy $0 \leq c \leq \rho^2(x) \leq C$. Then,*

$$c\lambda_i(k;\mathcal{X},\mathrm{d}\mu) \leq \lambda_i(\rho \odot k;\mathcal{X},\mathrm{d}\mu) \leq C\lambda_i(k;\mathcal{X},\mathrm{d}\mu), \quad \forall i = 1,2,\dots.$$

Consequently, if ν is another measure on $\mathcal{X}$ such that $0 \leq c \leq \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \leq C$, then

$$c\lambda_i(k;\mathcal{X},\mathrm{d}\mu) \leq \lambda_i(k;\mathcal{X},\mathrm{d}\nu) \leq C\lambda_i(k;\mathcal{X},\mathrm{d}\mu), \quad \forall i = 1,2,\dots. \tag{5}$$

Now, we consider the transformation of the kernel. Let $\mathcal{X}_1, \mathcal{X}_2$ be two sets, $\varphi : \mathcal{X}_1 \to \mathcal{X}_2$ be a bijection and k_2 be a kernel over $\mathcal{X}_2$. We define the pull-back kernel $\varphi^* k_2$ over $\mathcal{X}_1$ by

$$(\varphi^* k_2)(x_1, x_1') = k_2(\varphi(x_1), \varphi(x_1')).$$

Moreover, suppose $\mathcal{X}_1$ is a measurable space with measure μ_1, we define the push-forward measure $\mu_2 = \varphi_* \mu_1$ on $\mathcal{X}_2$ by $\mu_2(A) = \mu_1(\varphi^{-1}(A))$. Then, it is easy to see that:

Proposition 3 *Let $\mathcal{X}_1, \mathcal{X}_2$ be two measurable spaces, $\varphi : \mathcal{X}_1 \to \mathcal{X}_2$ be a measurable injection, μ_1 be a measure on $\mathcal{X}_1$ and $\mu_2 = \varphi_* \mu_1$. Suppose k_2 is a kernel over $\mathcal{X}_2$ and $k_1 = \varphi^* k_2$ satisfies* (1). *Then,*

$$\lambda_i(k_1;\mathcal{X}_1,\mathrm{d}\mu_1) = \lambda_i(k_2;\mathcal{X}_2,\mathrm{d}\mu_2). \tag{6}$$

Finally, this lemma deals with the case of the sum of two kernels of different EDRs, which is a direct consequence of Lemma 50.

Lemma 4 *Let k_1, k_2 be two positive definite kernels on $\mathcal{X}$. Suppose $\lambda_i(k_1;\mathcal{X},\mathrm{d}\mu) \asymp \lambda_{2i}(k_1;\mathcal{X},\mathrm{d}\mu)$ and $\lambda_i(k_2;\mathcal{X},\mathrm{d}\mu) = O\left(\lambda_i(k_1;\mathcal{X},\mathrm{d}\mu)\right)$ as $i \to \infty$. Then,*

$$\lambda_i(k_1 + k_2;\mathcal{X},\mathrm{d}\mu) \asymp \lambda_i(k_1;\mathcal{X},\mathrm{d}\mu).$$

2.3 Eigenvalues of kernels restricted on a subdomain

Suppose we are interested in $\lambda_i(k_1;\mathcal{X}_1,\mathrm{d}\mu_1)$. If $k_1 = \varphi^* k_2$ for some transformation φ and the EDR of k_2 with respect to some measure σ on $\mathcal{X}_2$ is known or can be easily obtained, Then, it is tempting to combine Proposition 3 and Lemma 2 to obtain the EDR of k_1 with respect to μ_1. However, in many cases $\varphi(\mathcal{X}_1)$ is a proper subset of $\mathcal{X}_2$ and $\mu_2 = \varphi_* \mu_1$ is only supported on $\varphi(\mathcal{X}_1)$, so the Radon derivative $\frac{\mathrm{d}\mu_2}{\mathrm{d}\sigma}$ is not bounded from below (that is, $c = 0$) and the lower bound in (5) vanishes, which is exactly the case of the NTK that we are interested in. Fortunately, we can still provide such a lower bound if the kernel satisfies an appropriate invariance property. Considering translation invariant kernels (that is, $k(x,x') = g(x - x')$), the following result based on Widom (1963) is very inspiring.

Proposition 5 (Widom (1963)) *Let $\mathbb{T}^d = [-\pi,\pi)^d$ be the d-dimensional torus and*

$$k(x,x') = \sum_{\boldsymbol{n} \in \mathbb{Z}^d} c_{\boldsymbol{n}} e^{i\boldsymbol{n}\cdot x} e^{-i\boldsymbol{n}\cdot x'}$$

be a translation invariant kernel on $\mathbb{T}^d$. Suppose further that $c_{\boldsymbol{n}}$ satisfies (i) $c_{\boldsymbol{n}} \geq 0$; (ii) with all n_i fixed but n_{i_0}, $c_{\boldsymbol{n}}$, as a function of n_{i_0}, is nondecreasing between $-\infty$ and some

$\bar{n} = \bar{n}(i_0)$ and nonincreasing between $\bar{n}$ and ∞; (iii) if $|\boldsymbol{n}|, |\boldsymbol{m}| \to \infty$ and $|\boldsymbol{n}| = O(|\boldsymbol{m}|)$, then $c_{\boldsymbol{m}} = O(c_{\boldsymbol{n}})$; (iv) if $|\boldsymbol{n}|, |\boldsymbol{m}| \to \infty$ and $|\boldsymbol{n}| = o(|\boldsymbol{m}|)$, then $c_{\boldsymbol{m}} = o(c_{\boldsymbol{n}})$. Then, for a bounded non-zero Riemann-integrable function ρ, we have

$$\lambda_i(k; \mathbb{T}^d, \rho^2 \mathrm{d}x) \asymp \lambda_i(k; \mathbb{T}^d, \mathrm{d}x).$$

However, the above result is not applicable to our case since the NTKs we are interested in is not translation invariant on the torus, but rotation invariant on the sphere. Nevertheless, inspired by this result, we establish a similar result for dot-product kernels on the sphere as one of our main contribution. Let $\mathbb{S}^d \subset \mathbb{R}^{d+1}$ be the d-dimensional unit sphere and σ be the Lebesgue measure on $\mathbb{S}^d$. We recall that a dot-product kernel $k(x, x')$ is a kernel that depends only on the dot product $u = \langle x, x' \rangle$ of the inputs. Thanks to the theory of spherical harmonics (Dai and Xu, 2013), the eigenfunctions of the integral operator T and also the Mercer's decomposition of k can be explicitly given by

$$k(x, x') = \sum_{n=0}^{\infty} \mu_n \sum_{l=1}^{a_n} Y_{n,l}(x) Y_{n,l}(x'), \tag{7}$$

where $\{Y_{n,l}, n \geq 0,\ l = 1, \dots, a_n\}$ is an orthonormal basis formed by spherical harmonics, $a_n = \binom{n+d}{n} - \binom{n-2+d}{n-2}$ is the dimension of the space of order-n spherical harmonics, and μ_n an eigenvalue of T with multiplicity a_n. To state our result, let us first introduce the following condition on the asymptotic decay rate of the eigenvalues.

Condition 6 *Let $(\mu_n)_{n \geq 0}$ be a decreasing sequence of positive numbers.*

(a) Define $N(\varepsilon) = \max\{n : \mu_n > \varepsilon\}$. For any fixed constant $c > 0$, $N(c\varepsilon) = \Theta(N(\varepsilon))$ as $\varepsilon \to 0$; suppose $\varepsilon, \delta \to 0$ with $\varepsilon = o(\delta)$, then $N(\delta) = o(N(\varepsilon))$.

(b) $\triangle^{d+1} \mu_n \geq 0$ for all n, where $\triangle$ is the forward difference operator in Definition 55.

(c) There is some constant $q \in \mathbb{N}_+$ and $D > 0$ such that for any $n \geq 0$,

$$\sum_{l=0}^{d} \binom{\tilde{n} + l}{l} \triangle^l \mu_{\tilde{n}} \leq D \mu_n, \quad \text{where} \quad \tilde{n} = qn. \tag{8}$$

Remark 7 *Condition 6 is a mild condition on the decay rate and Theorem 8 only requires that Condition 6 holds in the asymptotic sense, so this requirement is quite general. For instance, this requirement is satisfied if*

- *$\mu_n \asymp n^{-\beta}$ for some $\beta > d$.*
- *$\mu_n \asymp \exp\left(-c_1 n^{\beta}\right)$ for some $c_1, \beta > 0$.*
- *$\mu_n \asymp n^{-\beta} (\ln n)^p$ for $c_0 > 0$, $\beta > d$ and $p \in \mathbb{R}$, or $\beta = d$ and $p > 1$.*

Furthermore, our decay rate condition aligns with existing theory, as similar conditions (ii)-(iv) are also needed in Widom (1963).

Theorem 8 *Let $k(x,x')$ be a dot-product kernel on $\mathbb{S}^d$ whose corresponding eigenvalues in the decomposition* (7) *are $(\mu_n)_{n\geq 0}$. Assume that there is a sequence $(\tilde{\mu}_n)_{n\geq 0}$ satisfying Condition* 6 *such that $\mu_n \asymp \tilde{\mu}_n$. Then, for a bounded non-zero Riemann-integrable function ρ on $\mathbb{S}^d$, we have*

$$\lambda_i(k;\mathbb{S}^d,\rho^2\mathrm{d}\sigma) \asymp \lambda_i(k;\mathbb{S}^d,\mathrm{d}\sigma). \tag{9}$$

As our main technical contribution, this theorem is a non-trivial generalization of the result in Widom (1963), adapting it from the torus to the sphere. Following the basic idea of Widom (1963), we establish the theorem by proving first the main lemma (Lemma 22), but now the approach of Widom (1963) is not applicable since the eigen-system differs greatly. To prove the main lemma, we utilize refined harmonic analysis on the sphere, incorporating the technique of Cesaro summation and the left extrapolation of eigenvalues, which necessitates the subtle requirement of Condition 6. Detailed proof can be found in Section 4.

Theorem 8 shows that the EDR of a dot-product kernel with respect to a general measure is the same as that of the kernel with respect to the uniform measure. Combined with the results in Section 2.2, it provides a new approach to determine the EDR of a kernel on a general domain. One could first transform the kernel to a dot-product kernel on the sphere with respect to some measure; then use Theorem 8 to show that the decay rate of the resulting dot-product kernel remains the same if we consider the uniform measure on the sphere instead; and finally determine the decay rate of the dot-product kernel on the entire sphere by some analytic tools. This approach enables us to determine the EDR of the NTKs corresponding to multilayer neural networks on a general domain.

2.4 EDR of NTK on a general domain

A bunch of previous literature (Bietti and Mairal, 2019; Chen and Xu, 2020; Geifman et al., 2020; Bietti and Bach, 2020) have analyzed the RKHS as well as the spectral properties of the NTKs on the sphere by means of the theory of spherical harmonics. However, these results require the inputs to be uniformly distributed on the sphere and hence do not apply to general domains with general input distribution. Therefore, it is of our interest to investigate the eigenvalue properties of the NTKs on a general domain with a general input distribution since it is more realistic. To the best of our knowledge, only Lai et al. (2023) considered a non-spherical case of an interval on $\mathbb{R}$ and the EDR of the NTK corresponding to a two-layer neural network, but their techniques are very restrictive and can not be extended to higher dimensions or multilayer neural networks. Thanks to the results established in previous subsections, we can determine the EDR of the NTKs on a general domain using the established results on their spectral properties on the whole sphere.

Let us focus on the following explicit formula of the NTK, which corresponds to a multilayer neural network defined later in Section 3.1. Introduce the arc-cosine kernels (Cho and Saul, 2009) by

$$\kappa_0(u) = \frac{1}{\pi}\left(\pi - \arccos u\right), \quad \kappa_1(u) = \frac{1}{\pi}\left[\sqrt{1-u^2} + u(\pi - \arccos u)\right]. \tag{10}$$

Then, we define the kernel K^{NT} on $\mathbb{R}^d$ by

$$K^{\mathrm{NT}}(x,x') = \|\tilde{x}\|\|\tilde{x}'\| \sum_{r=0}^{L} \kappa_1^{(r)}(\bar{u}) \prod_{s=r}^{L-1} \kappa_0(\kappa_1^{(s)}(\bar{u})) + 1, \tag{11}$$

where $L \geq 2$ is the number of hidden layers, $\tilde{x} = (x,1)/\|(x,1)\|$, $\bar{u} = \langle \tilde{x}, \tilde{x}' \rangle$ and $\kappa_1^{(r)}$ represents r-times composition of κ_1, see, e.g., Jacot et al. (2018); Bietti and Bach (2020). First, we show that K^{NT} is strictly positive definite, the proof of which is deferred to Section B.1.

Proposition 9 *K^{NT} is strictly positive definite on $\mathbb{R}^d$, that is, for distinct points $x_1, \dots, x_n \in \mathbb{R}^d$, the kernel matrix's smallest eigenvalue $\lambda_{\min}\big(K^{\mathrm{NT}}(x_i,x_j)\big)_{n\times n} > 0$.*

Theorem 10 *Let μ be a probability measure on $\mathbb{R}^d$ with Riemann-integrable density $p(x)$ such that $p(x) \leq C(1+\|x\|^2)^{-(d+3)/2}$ for some constant C. Then, the EDR of K^{NT} on $\mathbb{R}^d$ with respect to μ is*

$$\lambda_i(K^{\mathrm{NT}}; \mathbb{R}^d, \mathrm{d}\mu) \asymp i^{-\frac{d+1}{d}}. \tag{12}$$

Remark 11 *The condition on the density $p(x)$ is satisfied by many common distributions, such as sub-Gaussian distributions or distributions with bounded support. Moreover, the result on the EDR can also be established for the NTKs corresponding to other activations (including homogeneous activations such as $\mathrm{ReLU}^\alpha(x) = \max(x,0)^\alpha$ and leaky ReLU) and other network architectures (such as residual neural networks), as long as the corresponding kernel can be transformed to a dot-product kernel on the sphere.*

Proof [of Theorem 10] Let us denote $\mathbb{S}^d_+ = \big\{ y = (y_1, \dots, y_{d+1}) \in \mathbb{S}^d : y_{d+1} > 0 \big\}$ and introduce the homeomorphism $\Phi : \mathbb{R}^d \to \mathbb{S}^d_+$ by $x \mapsto \tilde{x}/\|\tilde{x}\|$, where $\tilde{x} = (x,1) \in \mathbb{R}^{d+1}$. It is easy to show that the Jacobian and the Gram matrix are given by

$$J\Phi = \frac{1}{\|\tilde{x}\|}\begin{pmatrix} I_d \\ 0 \end{pmatrix} - \frac{\tilde{x}x^T}{\|\tilde{x}\|^3}, \quad G = (J\Phi)^T J\Phi = \frac{1}{\|\tilde{x}\|^2} I_d - \frac{xx^T}{\|\tilde{x}\|^4}, \quad \det G = \|\tilde{x}\|^{-2(d+1)}.$$

Defining the homogeneous NTK K_0^{NT} on $\mathbb{S}^d$ by

$$K_0^{\mathrm{NT}}(y,y') \coloneqq \sum_{r=0}^{L} \kappa_1^{(r)}(u) \prod_{s=r}^{L-1} \kappa_0(\kappa_1^{(s)}(u)), \quad u = \langle y, y' \rangle, \tag{13}$$

it is easy to verify that

$$K_1(x,x') \coloneqq \Phi^* K_0^{\mathrm{NT}} = \sum_{r=0}^{l} \kappa_1^{(r)}(\bar{u}) \prod_{s=r}^{l-1} \kappa_0(\kappa_1^{(s)}(\bar{u})), \quad K^{\mathrm{NT}} = \|x\| \odot K_1 + 1.$$

Therefore, Proposition 1 and then Proposition 3 yields

$$\lambda_i(\|x\| \odot K_1; \mathcal{X}, \mathrm{d}\mu) = \lambda_i(K_1; \mathcal{X}, \|\tilde{x}\|^2 \mathrm{d}\mu) = \lambda_i\left(K_0^{\mathrm{NT}}; \mathbb{S}^d, \Phi_*(\|\tilde{x}\|^2 \mathrm{d}\mu)\right).$$

Moreover, denoting $\tilde{\sigma} = \Phi_*(\|\tilde{x}\|^2 \mathrm{d}\mu)$ and $p(x) = \frac{\mathrm{d}\mu}{\mathrm{d}x}$, we have $\mathrm{d}\tilde{\sigma} = p(x)\|\tilde{x}\|^2 \Phi_*(\mathrm{d}x)$. On the other hand, the canonical uniform measure σ on $\mathbb{S}^d_+$ is given by $\mathrm{d}\sigma = |\det G|^{\frac{1}{2}} \Phi_*(\mathrm{d}x)$, so we have

$$q(y) \coloneqq \frac{\mathrm{d}\tilde{\sigma}}{\mathrm{d}\sigma} = |\det G|^{-\frac{1}{2}} \|\tilde{x}\|^2 p(x) = \|\tilde{x}\|^{d+3} p(x), \quad y \in \mathbb{S}^d_+.$$

Therefore, the condition on $p(x)$ implies that $q(y)$ is Riemann-integrable and upper bounded. Now, the EDR of the dot-product kernel K_0^{NT} on $\mathbb{S}^d$ with respect to $\mathrm{d}\sigma$ is already established in Bietti and Bach (2020) that $\lambda_i(K_0^{\mathrm{NT}}; \mathbb{S}^d, \mathrm{d}\sigma) \asymp i^{-\frac{d+1}{d}}$, so Theorem 8 shows that $\lambda_i\left(K_0^{\mathrm{NT}}; \mathbb{S}^d, \mathrm{d}\tilde{\sigma}\right) \asymp i^{-\frac{d+1}{d}}$. Finally, the proof is completed by applying Lemma 4 to show that the extra constant does not affect the EDR. ■

3. Application: Optimal Rates of Over-parameterized Neural Networks

In this section, using the spectral properties of the NTK obtained in the previous section, we derive the optimal rates of over-parameterized neural networks by combining the NTK theory and the kernel regression theory. Let d be fixed, $\mathcal{X} \subseteq \mathbb{R}^d$ and μ be a sub-Gaussian[1] probability distribution supported on $\mathcal{X}$ with upper bounded Riemann-integrable density. Suppose we are given i.i.d. samples $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n) \in \mathcal{X} \times \mathbb{R}$ generated from the model $y = f^*(x) + \varepsilon$, where $x \sim \mu$, $f^* : \mathcal{X} \to \mathbb{R}$ is an unknown regression function and the independent noise ε is sub-Gaussian.

In terms of notations, we denote $\boldsymbol{X} = (x_1, \dots, x_n)$ and $\boldsymbol{y} = (y_1, \dots, y_n)^T$. For a kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, we write $k(x, \boldsymbol{X}) = (k(x, x_1), \dots, k(x, x_n))$ and $k(\boldsymbol{X}, \boldsymbol{X}) = \big(k(x_i, x_j)\big)_{n \times n}$.

3.1 Setting of the neural network

We are interested in the following fully connected ReLU neural network $f(x; \boldsymbol{\theta})$ with L-hidden layers of widths $m_1, m_2, \dots, m_L$, where $L \geq 2$ is fixed. The network includes bias terms on the first and the last layers. To ensure that the final predictor corresponds to the kernel regressor, we consider a special mirrored architecture. In detail, the network model is given by the following:

$$\begin{aligned}
\boldsymbol{\alpha}^{(1,p)}(x) &= \sqrt{\tfrac{2}{m_1}} \sigma\Big(\boldsymbol{A}^{(p)} x + \boldsymbol{b}^{(0,p)}\Big) \in \mathbb{R}^{m_1},\ p \in \{1, 2\}, \\
\boldsymbol{\alpha}^{(l,p)}(x) &= \sqrt{\tfrac{2}{m_l}} \sigma\Big(\boldsymbol{W}^{(l-1,p)} \boldsymbol{\alpha}^{(l-1,p)}(x)\Big) \in \mathbb{R}^{m_l},\ l \in \{2, 3, \dots, L\},\ p \in \{1, 2\}, \\
g^{(p)}(x; \boldsymbol{\theta}) &= \boldsymbol{W}^{(L,p)} \boldsymbol{\alpha}^{(L,p)}(x) + b^{(L,p)} \in \mathbb{R},\ p \in \{1, 2\}, \\
f(x; \boldsymbol{\theta}) &= \frac{\sqrt{2}}{2} \Big[g^{(1)}(x; \boldsymbol{\theta}) - g^{(2)}(x; \boldsymbol{\theta})\Big] \in \mathbb{R}.
\end{aligned}$$

Here, $\boldsymbol{\alpha}^{(l,p)}$ represents the hidden layers; $l \in \{1, 2, \dots, L\}$, $p \in \{1, 2\}$ stand for the index of layers and parity respectively; $\sigma(x) \coloneqq \max(x, 0)$ is the ReLU activation (applied elementwise); parameters $\boldsymbol{A}^{(p)} \in \mathbb{R}^{m_1 \times d}$, $\boldsymbol{W}^{(l,p)} \in \mathbb{R}^{m_{l+1} \times m_l}$, $\boldsymbol{b}^{(0,p)} \in \mathbb{R}^{m_1}$, $b^{(L,p)} \in \mathbb{R}$, where we set

[1] That is, $\mu\big(\{x \in \mathbb{R}^d : \|x\| \geq t\}\big) \leq 2\exp\big(-t^2/C^2\big)$, $\forall t \geq 0$ for some constant $C > 0$.

$m_{L+1} = 1$; and we use $\boldsymbol{\theta}$ to represent the collection of all parameters flatten as a column vector. Letting $m = \min(m_1, m_2, \dots, m_L)$, we assume that $\max(m_1, m_2, \dots, m_L) \leq C_{\text{width}} m$ for some constant C_{width}.

Initialization Considering the mirrored architecture, we initialize the parameters in one parity to be i.i.d. normal and set the parameters in the other parity be the same as the corresponding ones. More precisely,

$$\begin{aligned}
&\boldsymbol{A}^{(1)}_{i,j}, \boldsymbol{W}^{(l,1)}_{i,j}, \boldsymbol{b}^{(0,1)}_{i}, b^{(L,1)} \overset{\text{i.i.d.}}{\sim} N(0,1), \quad \text{for } l = 0, 1, \dots, L, \\
&\boldsymbol{W}^{(l,2)} = \boldsymbol{W}^{(l,1)}, \quad \boldsymbol{A}^{(2)} = \boldsymbol{A}^{(1)}, \quad \boldsymbol{b}^{(0,2)} = \boldsymbol{b}^{(0,1)}, \quad b^{(L,2)} = b^{(L,1)}.
\end{aligned}$$

Such kind of "mirror initialization" ensures that the model output is always zero at initialization, which is also considered in Lai et al. (2023).

Training Neural networks are often trained by the gradient descent (or its variants) with respect to the empirical loss $\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2n}\sum_{i=1}^{n}(f(x_i;\boldsymbol{\theta}) - y_i)^2$. For simplicity, we consider the continuous version of gradient descent, namely the gradient flow for the training process. Denote by $\boldsymbol{\theta}_t$ the parameter at the time $t \geq 0$, the gradient flow is given by

$$\dot{\boldsymbol{\theta}}_t = -\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}_t) = -\frac{1}{n} \nabla_{\boldsymbol{\theta}} f(\boldsymbol{X};\boldsymbol{\theta}_t)(f(\boldsymbol{X};\boldsymbol{\theta}_t) - \boldsymbol{y}) \tag{14}$$

where $f(\boldsymbol{X};\boldsymbol{\theta}_t) = (f(x_1;\boldsymbol{\theta}_t), \dots, f(x_n;\boldsymbol{\theta}_t))^T$ and $\nabla_{\boldsymbol{\theta}} f(\boldsymbol{X};\boldsymbol{\theta}_t)$ is an $M \times n$ matrix where M is the number of parameters. Finally, let us denote by $\hat{f}^{\text{NN}}_t(x) \coloneqq f(x;\boldsymbol{\theta}_t)$ the resulting neural network predictor.

3.2 Uniform convergence to kernel regression

Although the gradient flow (14) is a highly non-linear and hard to analyze, the celebrated neural tangent kernel (NTK) theory (Jacot et al., 2018) provides a way to approximate the gradient flow by a kernel regressor when the width of the network tends to infinity, which is also referred to as the *lazy training regime*. Introducing a random kernel function $K_t(x,x') = \langle \nabla_{\boldsymbol{\theta}} f(x;\boldsymbol{\theta}_t), \nabla_{\boldsymbol{\theta}} f(x';\boldsymbol{\theta}_t) \rangle$, it is shown that $K_t(x,x')$ concentrates in probability to a deterministic kernel K^{NT} called the neural tangent kernel (NTK). Consequently, the predictor $\hat{f}^{\text{NN}}_t(x)$ is well approximated by the kernel regressor $\hat{f}^{\text{NTK}}_t(x)$ given by the following gradient flow:

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{f}^{\text{NTK}}_t(x) = -K^{\text{NT}}(x,\boldsymbol{X})(\hat{f}^{\text{NTK}}_t(\boldsymbol{X}) - \boldsymbol{y}), \tag{15}$$

where $\hat{f}^{\text{NTK}}_t(\boldsymbol{X}) = (\hat{f}^{\text{NTK}}_t(x_1), \dots, \hat{f}^{\text{NTK}}_t(x_n))^T$. Thanks to the mirrored architecture, we have $\hat{f}^{\text{NN}}_t(x) \equiv 0$ at initialization and thus we also have $\hat{f}^{\text{NTK}}_t(x) \equiv 0$. The recursive formula of the NTK also enables us to give explicitly (Jacot et al., 2018; Bietti and Bach, 2020) the formula of K^{NT} in (11).

Although previous works (Lee et al., 2019; Arora et al., 2019b; Allen-Zhu et al., 2019b) showed that the neural network regressor $\hat{f}^{\text{NN}}_t(x)$ can be approximated by $\hat{f}^{\text{NTK}}_t(x)$, most of these results are established pointwisely, namely, for fixed x, $\sup_{t \geq 0} \left| \hat{f}^{\text{NTK}}_t(x) - \hat{f}^{\text{NN}}_t(x) \right|$ is small with high probability. However, to analyze the generalization performance of $\hat{f}^{\text{NN}}_t(x)$, the convergence is further needed to be uniform over $x \in \mathcal{X}$. Consider the simple case of two-layer neural network, Lai et al. (2023) rigorously showed such uniform convergence. With more complicated analysis, we prove the uniform convergence of $\hat{f}_t^{\mathrm{NN}}(x)$ to $\hat{f}_t^{\mathrm{NTK}}(x)$ for multilayer neural networks. To state our result, let us denote by $\lambda_0 = \lambda_{\min}\left(K^{\mathrm{NT}}(\boldsymbol{X}, \boldsymbol{X})\right)$ the minimal eigenvalue of the kernel matrix, which, by Proposition 9, can be assumed to be positive in the following.

Lemma 12 *Denote $M_{\boldsymbol{X}} = \sum_{i=1}^n \|x_i\|_2$ and $B_r = \left\{x \in \mathbb{R}^d : \|x\| \leq r\right\}$ for $r \geq 1$. There exists a polynomial* $\operatorname{poly}(\cdot)$ *such that for any $\delta \in (0,1)$ and $k > 0$, when $m \geq \operatorname{poly}(n, M_{\boldsymbol{X}}, \lambda_0^{-1}, \|\boldsymbol{y}\|, \ln(1/\delta), k)$ and $m \geq r^k$, with probability at least $1 - \delta$ with respect to random initialization, we have*

$$\sup_{t \geq 0} \sup_{x \in B_r} \left| \hat{f}_t^{\mathrm{NTK}}(x) - \hat{f}_t^{\mathrm{NN}}(x) \right| \leq O(r^2 m^{-\frac{1}{12}} \sqrt{\ln m}).$$

Lemma 12 shows that as m tends to infinity, $\hat{f}_t^{\mathrm{NN}}(x)$ can be approximated uniformly by $\hat{f}_t^{\mathrm{NTK}}(x)$ on a bounded set, which is also allowed to grow with m. Consequently, we can study the generalization performance of the neural network in the lazy training regime by that of the corresponding kernel regressor.

To establish Lemma 12, it is essential to demonstrate the uniform convergence of the kernel $K_t(x, x')$ towards $K^{\mathrm{NT}}(x, x')$. This is achieved by first establishing the Hölder continuity of $K_t(x, x')$ and $K^{\mathrm{NT}}(x, x')$, and then applying an ϵ-net argument in conjunction with the pointwise convergence. Since the detailed proof is laborious, it is deferred to Section A.

3.3 The optimal rates of the over-parameterized neural network

With the uniform convergence of the neural network to the kernel regressor established and the eigenvalue decay rate of the NTK determined, we can now derive the optimal rates of the over-parameterized neural network. Let us denote by $\mathcal{H} = \mathcal{H}_{\mathrm{NTK}}$ the RKHS associated with the NTK (11) on $\mathcal{X}$. We introduce the integral operator T in (2) and recall its spectral decomposition in (3). The kernel regression literature often introduce the interpolation spaces of the RKHS to characterize the regularity of the regression function (Steinwart and Scovel, 2012; Fischer and Steinwart, 2020). For $s \geq 0$, we define the interpolation space $[\mathcal{H}]^s$ by

$$[\mathcal{H}]^s = \left\{ \sum_{i=1}^\infty a_i \lambda_i^{s/2} e_i \;\middle|\; \sum_{i=1}^\infty a_i^2 < \infty \right\} \subseteq L^2, \tag{16}$$

which is equipped with the norm $\left\| \sum_{i=1}^\infty a_i \lambda_i^{s/2} e_i \right\|_{[\mathcal{H}]^s} := \left(\sum_{i=1}^\infty a_i^2\right)^{1/2}$. It can be seen that $[\mathcal{H}]^s$ is a separable Hilbert space with $\left(\lambda_i^{s/2} e_i\right)_{i \geq 1}$ as its orthonormal basis. We also have $[\mathcal{H}]^0 = L^2$ and $[\mathcal{H}]^1 = \mathcal{H}$. Moreover, when $s \in (0,1)$, the space $[\mathcal{H}]^s$ also coincides with the space $(L^2, \mathcal{H})_{s,2}$ defined by real interpolation (Steinwart and Scovel, 2012). We also denote by $B_R([\mathcal{H}]^s) = \left\{ f \in [\mathcal{H}]^s \mid \|f\|_{[\mathcal{H}]^s}^2 \leq R \right\}$. Then, we derive the following optimal rates of the neural network from the optimality result in the kernel regression (Lin et al., 2018).

Proposition 13 *Suppose $f^* \in B_R([\mathcal{H}]^s) \cap L^\infty$ for constants $s > \frac{1}{d+1}$ and $R > 0$. Let us choose $t_{\mathrm{op}} = t_{\mathrm{op}}(n) \asymp n^{(d+1)/[s(d+1)+d]}$. Then, there exists a polynomial $\mathrm{poly}(\cdot)$ such that for any $\delta \in (0,1)$, when n is sufficiently large and the width $m \geq \mathrm{poly}(n, \ln(1/\delta), \lambda_0^{-1})$, with probability at least $1-\delta$ with respect to random samples and random initialization,*

$$\left\| \hat{f}^{\mathrm{NN}}_{t_{\mathrm{op}}} - f^* \right\|_{L^2}^2 \leq C \left(\ln \frac{12}{\delta} \right)^2 n^{-\frac{s(d+1)}{s(d+1)+d}}, \tag{17}$$

where the constant $C > 0$ is independent of δ, n. Moreover, the convergence rate in (17) achieves the optimal rate in $B_R([\mathcal{H}]^s)$.

The results in the kernel regression literature also allows us to provide the following sup-norm learning rate.

Proposition 14 *Under the settings of Proposition 13, suppose further that $s \geq 1$ and $\mathcal{X}$ is bounded. Then, when n is sufficiently large, with probability at least $1-\delta$,*

$$\left\| \hat{f}^{\mathrm{NN}}_{t_{\mathrm{op}}} - f^* \right\|_{\infty}^2 \leq C \left(\ln \frac{12}{\delta} \right)^2 n^{-\frac{(s-1)(d+1)}{s(d+1)+d}},$$

where the constant $C > 0$ is independent of δ, n.

Remark 15 *Proposition 13 shows the minimax optimality of wide neural networks, where optimal rate is also adaptive to the relative smoothness of the regression function to the NTK. Our result extends the result in Lai et al. (2023) to the scenario of $d > 1$ and $L > 1$, and also distinguishes with Hu et al. (2021); Suh et al. (2022) in the following aspects: (1) The critical uniform convergence (Lemma 12) is not well-supported in these two works, as pointed out in Lai et al. (2023); (2) They have to assume the data distribution is uniform on the sphere, while we allow $\mathcal{X}$ to be a general domain; (3) They introduce an explicit ℓ_2 regularization in the gradient descent and approximate the training dynamics by kernel ridge regression (KRR), while we consider directly the kernel gradient flow and early stopping serves as an implicit regularization, which is more natural. Moreover, our gradient method can adapt to higher order smoothness of the regression function and do not saturate as KRR (Li et al., 2023b) or consequently their ℓ_2-regularized neural networks.*

Moreover, using the idea in Caponnetto and Yao (2010), we can also show that the cross validation can be used to choose the optimal stopping time. Let us further assume that $\operatorname{Supp} \mu$ is bounded and $y \in [-M, M]$ almost surely for some M and introduce the truncation $L_M(a) = \min\{|a|, M\} \operatorname{sgn}(a)$. Suppose now we have $\tilde{n}$ extra independent samples $(\tilde{x}_1, \tilde{y}_1), \dots, (\tilde{x}_{\tilde{n}}, \tilde{y}_{\tilde{n}})$, where $\tilde{n} \geq c_{\mathrm{v}} n$ for some constant $c_{\mathrm{v}} > 0$. Let T_n be a set of stopping time candidates, we can choose the empirical stopping time by cross validation

$$\hat{t}_{\mathrm{cv}} = \operatorname*{arg\,min}_{t \in T_n} \sum_{i=1}^{\tilde{n}} \left[L_M \left(\hat{f}^{\mathrm{NN}}_t(\tilde{x}_i) \right) - \tilde{y}_i \right]^2. \tag{18}$$

Proposition 16 *Under the settings of Proposition 13 and the further assumptions given above, let $T_n = \left\{1, Q, \dots, Q^{\lfloor \ln_Q n \rfloor}\right\}$ for arbitrary fixed $Q > 1$ and $\hat{t}_{\mathrm{cv}}$ be chosen from (18).*

Define $\hat{f}^{\mathrm{NN}}_{\mathrm{cv}}(x) = L_M\left(\hat{f}^{\mathrm{NN}}_{\hat{t}_{\mathrm{cv}}}(x)\right)$*. Then, there exists a polynomial* $\mathrm{poly}(\cdot)$ *such that when* n *is sufficiently large and* $m \geq \mathrm{poly}(n, \ln(1/\delta), \lambda_0^{-1})$*, one has*

$$\left\| \hat{f}^{\mathrm{NN}}_{\mathrm{cv}} - f^* \right\|^2_{L^2} \leq C \left(\ln \frac{12}{\delta} \right)^2 n^{-\frac{s(d+1)}{s(d+1)+d}}$$

with probability at least $1 - \delta$ *with respect to random samples and initialization, where the constant* $C > 0$ *is independent of* δ, n*.*

Early stopping, as an implicit regularization, is necessary for the generalization of neural networks. The following proposition, which is a consequence of the result in Li et al. (2023a), shows overfitted multilayer neural networks generalize poorly.

Proposition 17 *Suppose further that the samples are distributed uniformly on* $\mathbb{S}^d$ *and the noise is non-zero. Then, for any* $\varepsilon > 0$ *and* $\delta \in (0, 1)$*, there is some* $c > 0$ *such that when* n *and* m *is sufficiently large, one has that*

$$\mathbb{E}\left[\liminf_{t \to \infty} \left\| \hat{f}^{\mathrm{NN}}_t - f^* \right\|^2_{L^2} \;\middle|\; \boldsymbol{X} \right] \geq c n^{-\varepsilon}$$

holds with probability at least $1 - \delta$*.*

Remark 18 *Proposition* 17 *seems to contradict with the "benign overfitting" phenomenon (e.g., Bartlett et al. (2020); Frei et al. (2022)). However, we point out that in these works the dimension* d *of the input diverges with the sample size* n*, while in our case* d *is fixed, so the setting is different. In fact, in the fixed-*d *scenario, several works have argued that overfitting is harmful (Rakhlin and Zhai, 2018; Beaglehole et al., 2022; Li et al., 2023a) and our result is consistent with theirs.*

Remark 19 *The requirement of uniformly distributed samples on the sphere is due to the technical condition of the embedding index in Li et al. (2023a), which is critical for more refined analysis in the kernel regression (Fischer and Steinwart, 2020). With this condition, the requirement of* s *in Proposition* 13 *can further be relaxed to* $s > 0$*. We hypothesize that this embedding index condition is also satisfied for the NTK on a general domain, but we would like to leave it to future work since more techniques on function theory are needed.*

4. Proof of the Result on the Eigenvalues

In this section we provide the proof of our key result, Theorem 8. The proof idea follows the same line as Widom (1963): we first establish the key lemma (Lemma 22) and then use the decomposition of domains on the sphere to show Theorem 8. The key technical contribution here lies in the proof of Lemma 22 where we apply a refined analysis on bounding the spherical harmonics using the Cesaro summation.

For a compact self-adjoint operator T, we denote by $N^{\pm}(\varepsilon, T)$ the count of eigenvalues of T that is strictly greater (smaller) than ε $(-\varepsilon)$. We denote by P_Ω the operator of the multiplication of the characteristic function $\mathbf{1}_\Omega$. For convenience, we will use C to represent some positive constant that may vary in each appearance in the proof.

4.1 Spherical harmonics

Let us first introduce spherical harmonics and some properties that will be used. We refer to Dai and Xu (2013) for more details. Let σ be the Lebesgue measure on $\mathbb{S}^d$ and $L^2(\mathbb{S}^d)$ be the (real) Hilbert space equipped with the inner product

$$\langle f, g\rangle_{L^2(\mathbb{S}^d)} = \frac{1}{\omega_d}\int_{\mathbb{S}^d} fg \, \mathrm{d}\sigma.$$

By the theory of spherical harmonics, the eigen-system of the Laplace-Beltrami operator $\Delta_{\mathbb{S}^d}$, the spherical Laplacian, gives an orthogonal direct sum decomposition $L^2(\mathbb{S}^d) = \bigoplus_{n=0}^{\infty} \mathcal{H}_n^d(\mathbb{S}^d)$, where $\mathcal{H}_n^d(\mathbb{S}^d)$ is the restriction of n-degree homogeneous harmonic polynomials with $d+1$ variables on $\mathbb{S}^d$ and each element in $\mathcal{H}_n^d(\mathbb{S}^d)$ is an eigen-function of $\Delta_{\mathbb{S}^d}$ with eigenvalue $-n(n+d-1)$. This gives an orthonormal basis

$$\{Y_{n,l}, \ l = 1, \dots, a_n, \ n = 1, 2, \dots\}$$

of $L^2(\mathbb{S}^d)$, where $a_n = \binom{n+d}{n} - \binom{n-2+d}{n-2} \asymp n^{d-1}$ is the dimension of $\mathcal{H}_n^d(\mathbb{S}^d)$ and $Y_{n,l} \in \mathcal{H}_n^d(\mathbb{S}^d)$. We also notice that

$$\sum_{n\leq N} a_n = C_{N+d}^{N} + C_{N-1+d}^{N-1} \asymp N^d. \tag{19}$$

Moreover, the summation

$$Z_n(x, y) = \sum_{l=1}^{a_n} Y_{n,l}(x) Y_{n,l}(y) \tag{20}$$

is invariant of selection of orthonormal basis $Y_{n,l}$ and Z_n's are called zonal polynomials. When $d \geq 2$, we have

$$Z_n(x, y) = \frac{n+\lambda}{\lambda} C_n^{\lambda}(u), \quad u = \langle x, y\rangle, \ \lambda = \frac{d-1}{2}, \tag{21}$$

where C_n^{λ} is the Gegenbauer polynomial.

The key property of spherical harmonics is the following Funk-Hecke formula (Dai and Xu, 2013, Theorem 1.2.9).

Proposition 20 (Funk-Hecke formula) *Let $d \geq 3$ and f be an integrable function such that $\int_{-1}^{1} |f(t)|(1-t^2)^{d/2-1} \mathrm{d}t$ is finite. Then for every $Y_n \in \mathcal{H}_n^d(\mathbb{S}^d)$,*

$$\frac{1}{\omega_d}\int_{\mathbb{S}^d} f(\langle x, y\rangle) Y_n(y) \mathrm{d}\sigma(y) = \mu_n(f) Y_n(x), \quad \forall x \in \mathbb{S}^d, \tag{22}$$

where $\mu_n(f)$ is a constant defined by $\mu_n(f) = \omega_d \int_{-1}^{1} f(t) \frac{C_n^{\lambda}(t)}{C_n^{\lambda}(1)} (1-t^2)^{\frac{d-2}{2}} \mathrm{d}t$.

We also need the following theorem relating to the Cesaro sum of zonal polynomials. Readers may refer to Section D.3 for a definition of the Cesaro sum.

Proposition 21 *Let*

$$K_n = \frac{1}{A_n^d} \sum_{k=0}^{n} A_{n-k}^d \frac{k+\lambda}{\lambda} C_k^\lambda(u), \tag{23}$$

be the d-Cesaro sum of $\frac{k+\lambda}{\lambda} C_k^\lambda$*. Then,*

$$0 \le K_n(u) \le C n^{-1} (1 - u + n^{-2})^{-(\lambda+1)}, \quad \forall n \ge 1 \tag{24}$$

for some positive constant C*.*

Proof Please refer to Dai and Xu (2013, Theorem 2.4.3 and Lemma 2.4.6). ■

4.2 Dot-product kernel on the sphere

Comparing (22) with (2), the Funk-Hecke formula shows that Y_n is an eigenfunction of any dot-product kernel $k(x,y) = f(\langle x,y \rangle)$ on the sphere. Therefore, a dot-product kernel $k(x,y)$ always admits the following Mercer and spectral decompositions

$$k(x,y) = \sum_{n=0}^{\infty} \mu_n \sum_{l=1}^{a_n} Y_{n,l}(x) Y_{n,l}(y), \quad T = \sum_{n=0}^{\infty} \mu_n \sum_{l=1}^{a_n} Y_{n,l} \otimes Y_{n,l}. \tag{25}$$

Here we notice that μ_n is an eigenvalue having multiplicity a_n and it should not be confused with λ_i where multiplicity are counted. In the view of (25), we may connect a dot-product kernel as well as the corresponding integral operator with the sequence $(\mu_n)_{n \ge 0}$.

Moreover, since each μ_n is of multiplicity a_n, (19) gives

$$N^+(\varepsilon, T) = \sum_{n \le N(\varepsilon)} a_n \asymp N(\varepsilon)^d, \tag{26}$$

where $N(\varepsilon) = \max\{n : \mu_n > \varepsilon\}$ as defined in Condition 6 (a). This gives a simple relation between the asymptotic rates λ_i and μ_n.

4.3 The main lemma

The following main lemma is essential in the proof of our final result, which is a spherical version of the main lemma in Widom (1963). Since the eigen-system is now given by the spherical harmonics, the approach in Widom (1963) can not be applied. The proof is now based on refined harmonic analysis on the sphere with the technique of Cesaro summation and the left extrapolation of eigenvalues.

Lemma 22 *Let* T *be given by* (25) *with the descending eigenvalues* $\boldsymbol{\mu} = (\mu_n)_{n \ge 0}$*. Suppose further that* $(\mu_n)_{n \ge 0}$ *satisfies Condition* 6*. Let* Ω_1, Ω_2 *be two disjoint domains with piecewise smooth boundary. Then, we have*

$$N^{\pm}(\varepsilon, P_{\Omega_1} T P_{\Omega_2} + P_{\Omega_2} T P_{\Omega_1}) = o(N^+(\varepsilon, T)), \quad \textit{as} \quad \varepsilon \to 0. \tag{27}$$

Proof

Let $\delta > \varepsilon > 0$ and δ will be determined later. Take $M_\delta = \min\{n : \mu_n \leq \delta\} \leq M_\varepsilon = \min\{n : \mu_n \leq \varepsilon\}$. Using Lemma 61 for $p = d+1$ with Condition 6, we can first construct a sequence $\boldsymbol{\mu}^{(1)}$ as the left extrapolation of $\boldsymbol{\mu}$ at qM_ε, then construct a sequence $\boldsymbol{\mu}^{(2)}$ as the left extrapolation of the residual sequence $\boldsymbol{\mu} - \boldsymbol{\mu}^{(1)}$ at qM_δ, and denote $\boldsymbol{\mu}^{(3)} = \boldsymbol{\mu} - \boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}$, where q is the integer specified in Condition 6 Then, the three sequences satisfy

$$\begin{aligned}
&\mu_n = \mu_n^{(1)} + \mu_n^{(2)} + \mu_n^{(3)}, \quad \triangle^{d+1}\mu_n^{(i)} \geq 0,\ i = 1,2,3;\\
&\mu_0^{(1)} = \mathcal{L}_{qM_\varepsilon}^{d+1}\boldsymbol{\mu} \leq D\mu_{M_\varepsilon} \leq D\varepsilon;\\
&\mu_n^{(2)} = 0,\ \forall n \geq qM_\varepsilon, \quad \mu_0^{(2)} = \mathcal{L}_{qM_\delta}^{d+1}(\boldsymbol{\mu} - \boldsymbol{\mu}^{(1)}) \leq \mathcal{L}_{qM_\delta}^{d+1}\boldsymbol{\mu} \leq D\delta;\\
&\mu_n^{(3)} = 0,\ \forall n \geq qM_\delta,
\end{aligned} \tag{28}$$

where the control $\mathcal{L}_{qM_\delta}^{d+1}\boldsymbol{\mu} \leq C\mu_M$ comes from (8). Now, we define T_i to be the integral operator associated with $\boldsymbol{\mu}^{(i)}$, that is, $T_i = \sum_{n=0}^\infty \mu_n^{(i)} \sum_{l=1}^{a_n} Y_{n,l} \otimes Y_{n,l}$. Let $N_i^+(\varepsilon)$ be the count of eigenvalues of $P_{\Omega_1}T_iP_{\Omega_2} + P_{\Omega_2}T_iP_{\Omega_1}$ greater than ε. By Lemma 51 we have

$$N^+((2D+1)\varepsilon, P_{\Omega_1}TP_{\Omega_2} + P_{\Omega_2}TP_{\Omega_1}) \leq N_1^+(2D\varepsilon) + N_2^+(\varepsilon) + N_3^+(0). \tag{29}$$

For $N_1^+(2D\varepsilon)$, we notice that $\|T_1\| \leq D\varepsilon$ and hence

$$\|P_{\Omega_1}T_1P_{\Omega_2} + P_{\Omega_2}T_1P_{\Omega_1}\| \leq 2D\varepsilon,$$

which implies that

$$N_1^+(2D\varepsilon) = 0. \tag{30}$$

For $N_3^+(0)$, since $\mu_n^{(3)} \neq 0$ only when $n < qM_\delta$, so

$$N_3^+(0) \leq 2\mathrm{Rank}(T_3) \leq 2\sum_{n<qM_\delta} a_n \leq 2Cq^d \sum_{n<M_\delta} a_n = 2CN^+(\delta, T), \tag{31}$$

where we use (19) in the third inequality and $\sum_{n<M_\delta} a_n = N^+(\delta, T)$ (notice that μ_n is an eigenvalue of multiplicity a_n).

It remains to bound $N_2^+(\varepsilon)$. First, by definition of the HS-norm and Simon (2015, Theorem 3.8.5), we have

$$\varepsilon^2 N_2^+(\varepsilon) \leq \|P_{\Omega_1}T_2P_{\Omega_2} + P_{\Omega_2}T_2P_{\Omega_1}\|_{\mathrm{HS}}^2 = 2\int_{\Omega_1}\int_{\Omega_2}\left|\sum_n \mu_n^{(2)} Z_n(x,y)\right|^2 \mathrm{d}y\mathrm{d}x =: 2I. \tag{32}$$

Fixing an interior point e of Ω_2, we introduce an isometric transform $R_{e,x}$ such that $R_{e,x}e = x$. It can be taken to be the rotation over the plane spanned by e, x if they are not parallel, to be the identity map if $x = e$ and to be reflection if $x = -e$. Then, since $R_{e,x}$ is isometric and $Z_n(x,y)$ depends only on $\langle x, y\rangle$, we have

$$I = \int_{\Omega_1}\int_{\Omega_2}\left|\sum_n \mu_n^{(2)} Z_n(x,y)\right|^2 \mathrm{d}y\mathrm{d}x = \int_{\Omega_1}\int_{\Omega_2}\left|\sum_n \mu_n^{(2)} Z_n(R_{e,x}^{-1}x, R_{e,x}^{-1}y)\right|^2 \mathrm{d}y\mathrm{d}x$$

$$
\begin{aligned}
&= \int_{\Omega_1}\int_{\Omega_2}\left|\sum_n \mu_n^{(2)} Z_n(e, R_{e,x}^{-1} y)\right|^2 \mathrm{d}y\mathrm{d}x = \int_{\Omega_1}\int_{R_{e,x}^{-1}\Omega_2}\left|\sum_n \mu_n^{(2)} Z_n(e,y)\right|^2 \mathrm{d}y\mathrm{d}x \\
&= \iint_{\mathbb{S}^d\times\mathbb{S}^d} \mathbf{1}\left\{x\in\Omega_1,\ y\in R_{e,x}^{-1}\Omega_2\right\}\left|\sum_n \mu_n^{(2)} Z_n(e,y)\right|^2 \mathrm{d}x\mathrm{d}y \\
&= \int_{\mathbb{S}^d}\left(\int_{\mathbb{S}^d}\mathbf{1}\left\{x\in\Omega_1,\ R_{e,x}y\in\Omega_2\right\}\mathrm{d}x\right)\left|\sum_n \mu_n^{(2)} Z_n(e,y)\right|^2 \mathrm{d}y \\
&= \int_{\mathbb{S}^d}\left|\{x\in\Omega_1 : R_{e,x}y\in\Omega_2\}\right|\left|\sum_n \mu_n^{(2)} Z_n(e,y)\right|^2 \mathrm{d}y \\
&\le \int_{\mathbb{S}^d}\left|\{x\in\Omega_1 : R_{e,x}y\in\Omega_2\}\right|\left|\sum_n \mu_n^{(2)} Z_n(e,y)\right|^2 \mathrm{d}y \\
&\le C\int_{\mathbb{S}^d}\arccos\langle y,e\rangle\left|\sum_n \mu_n^{(2)} Z_n(e,y)\right|^2 \mathrm{d}y,
\end{aligned}
$$

where the last inequality comes from Proposition 52. Let $\eta > 0$ (which will be determined later), we decompose the last integral into two parts:

$$
I = I_1 + I_2 = \int_{\langle y,e\rangle > 1-\eta} + \int_{\langle y,e\rangle < 1-\eta}\arccos\langle y,e\rangle\left|\sum_n \mu_n^{(2)} Z_n(e,y)\right|^2 \mathrm{d}y.
$$

For I_1, using the estimation $\arccos u \le C\sqrt{1-u}$, we obtain

$$
I_1 \le \int_{\langle y,e\rangle > 1-\eta} C\eta^{\frac{1}{2}}\left|\sum_n \mu_n^{(2)} Z_n(e,y)\right|^2 \mathrm{d}y \le C\eta^{\frac{1}{2}}\int_{\mathbb{S}^d}\left|\sum_n \mu_n Z_n(e,y)\right|^2 \mathrm{d}y = C\eta^{\frac{1}{2}}\sum_n a_n(\mu_n^{(2)})^2.
$$

Using (28), we get

$$
\begin{aligned}
I_1 &\le C\eta^{\frac{1}{2}}\sum_n a_n(\mu_n^{(2)})^2 \le C\eta^{\frac{1}{2}}\sum_{n<qM_\varepsilon} a_n(\mu_0^{(2)})^2 \le C\eta^{\frac{1}{2}} q^d\sum_{n<M_\varepsilon} a_n(\mu_0^{(2)})^2 \\
&\le C\eta^{\frac{1}{2}} N^+(\varepsilon,T)\delta^2, \qquad (33)
\end{aligned}
$$

where we use (19) again in the third inequality.

For I_2, recalling (21) and denoting $u = \langle y,e\rangle$, we have

$$
I_2 = \int_{-1}^{1-\eta}\left|\sum_n \mu_n^{(2)}\frac{n+\lambda}{\lambda}C_n^\lambda(u)\right|^2\left(1-u^2\right)^{\frac{d-2}{2}}\arccos u\mathrm{d}u,
$$

where $\lambda = \frac{d-1}{2}$. Using summation by parts (Proposition 58), we obtain

$$
\sum_n \mu_n^{(2)}\frac{k+\lambda}{\lambda}C_n^\lambda(u) = \sum_n \triangle^{d+1}\mu_n^{(2)} A_n^d K_n(u),
$$

where K_n is the d-Cesaro sum of $C_k^\lambda(u)$ as in (23). Moreover, (24) in Proposition 21 yields

$$\begin{aligned}\left|\sum_n \mu_n \frac{k+\lambda}{\lambda} C_n^\lambda(u)\right| &= \sum_n \triangle^{d+1}\mu_n^{(2)} A_n^d K_n(u) \leq C(1-u)^{-(\lambda+1)} \sum_n A_n^d \triangle^{d+1} \mu_n^{(2)} \\ &= C(1-u)^{-(\lambda+1)}\mu_{0,2} \leq C(1-u)^{-(\lambda+1)}\delta,\end{aligned}$$

where the last but second equality comes from Proposition 57. Plugging this estimation back into I_2, we obtain

$$I_2 \leq C\delta^2 \int_{-1}^{1-\eta} (1-u)^{-2(\lambda+1)} \left(1-u^2\right)^{\frac{d-2}{2}} (1-u)^{1/2} \mathrm{d}u \leq C\delta^2 \eta^{-\frac{d-1}{2}}. \tag{34}$$

Now we obtain the estimations (33) and (34). Taking $\eta = N^+(\delta,T)^{-\frac{2}{d}}$, we have

$$I \leq I_1 + I_2 \leq C\delta^2 N^+(\varepsilon,T)^{\frac{d-1}{d}},$$

so (32) yields

$$N_2^+(\varepsilon) \leq C\left(\frac{\delta}{\varepsilon}\right)^2 N^+(\varepsilon,T)^{\frac{d-1}{d}}. \tag{35}$$

Finally, plugging (30), (35) and (31) into (29), we have

$$N^+((2D+1)\varepsilon, P_{\Omega_1}TP_{\Omega_2} + P_{\Omega_2}TP_{\Omega_1}) \leq 2N^+(\delta,T) + C\left(\frac{\delta}{\varepsilon}\right)^2 N^+(\varepsilon,T)^{\frac{d-1}{d}}.$$

Now, (26) allows us to derive a similar condition on $N^+(\varepsilon,T)$ as (a) in Condition 6. Therefore, taking $\delta = \varepsilon N^+(\varepsilon,T)^{\frac{1}{4d}}$ so that $\varepsilon = o(\delta)$, we obtain $N^+(\delta,T) = o\left(N^+(\varepsilon,T)\right)$, so

$$N^+((2D+1)\varepsilon, P_{\Omega_1}TP_{\Omega_2} + P_{\Omega_2}TP_{\Omega_1}) \leq o\left(N^+(\varepsilon,T)\right) + CN^+(\varepsilon,T)^{\frac{2d-1}{2d}} = o\left(N^+(\varepsilon,T)\right).$$

Since D is a fixed constant and $N^+(\varepsilon/(2D+1),T) = \Theta(N^+(\varepsilon,T))$, replacing ε by $\varepsilon/(2D+1)$ yields the desired result.

The proof of the case $N^-(\varepsilon, P_{\Omega_1}TP_{\Omega_2} + P_{\Omega_2}TP_{\Omega_1})$ is similar. ■

4.4 The main result

The following is a direct corollary of Lemma 2. We present it here since it will be frequently used later.

Corollary 23 *Suppose $\Omega_1 \subseteq \Omega_2$. Then, $N^+(\varepsilon, P_{\Omega_1}TP_{\Omega_1}) \leq N^+(\varepsilon, P_{\Omega_2}TP_{\Omega_2})$.*

We first prove the following lemma about dividing a domain into isometric subdomains, which will be used recursively in the proof later.

Lemma 24 *Let T be the same in Lemma 22. Let $S \subseteq \mathbb{S}^d$ and suppose $N^+(\varepsilon, P_S T P_S) \asymp N^+(\varepsilon, T)$ as $\varepsilon \to 0$. Suppose further that $\Omega \subseteq \mathbb{S}^d$ is a subdomain with piecewise smooth boundary and there exists isometric copies $\Omega_1, \dots, \Omega_m$ of Ω such that their disjoint union (with a difference of a null-set) is S. Then, there is some constant $c > 0$ such that for small ε,*

$$cN^+(\varepsilon, T) \le N^+(\varepsilon, P_\Omega T P_\Omega) \le N^+(\varepsilon, T). \tag{36}$$

Proof The upper bound follows from Corollary 23. Now we consider the lower bound. Since $\Omega_1, \dots, \Omega_m$ form a disjoint cover of S, we have

$$P_S T P_S = (\sum_{i=1}^m P_{\Omega_i}) T (\sum_{j=1}^m P_{\Omega_j}) = \sum_i P_{\Omega_i} T P_{\Omega_i} + \sum_{i<j} \left(P_{\Omega_i} T P_{\Omega_j} + P_{\Omega_j} T P_{\Omega_i} \right).$$

Using Lemma 51, we get

$$N^+(2\varepsilon, T) \le N^+(\varepsilon, \sum_i P_{\Omega_i} T P_{\Omega_i}) + \sum_{i<j} N^+ \left(\frac{1}{C_m^2} \varepsilon, P_{\Omega_i} T P_{\Omega_j} + P_{\Omega_j} T P_{\Omega_i} \right),$$

and thus

$$N^+(\varepsilon, \sum_i P_{\Omega_i} T P_{\Omega_i}) \ge N^+(2\varepsilon, P_S T P_S) - \sum_{i<j} N^+ \left(\frac{1}{C_m^2} \varepsilon, P_{\Omega_i} T P_{\Omega_j} + P_{\Omega_j} T P_{\Omega_i} \right).$$

Noticing the fact that Ω_i are disjoint and isometric with Ω, for the left hand side we obtain

$$N^+(\varepsilon, \sum_i P_{\Omega_i} T P_{\Omega_i}) = \sum_i N^+(\varepsilon, P_{\Omega_i} T P_{\Omega_i}) = mN^+(\varepsilon, P_\Omega T P_\Omega).$$

On the other hand, by Lemma 22,

$$N^+ \left(\frac{1}{C_m^2} \varepsilon, P_{\Omega_i} T P_{\Omega_j} + P_{\Omega_j} T P_{\Omega_i} \right) = o \left(N^+ \left(\frac{1}{C_m^2} \varepsilon, T \right) \right) = o \left(N^+(\varepsilon, T) \right),$$

where we notice that $N^+(c\varepsilon, T) \asymp N^+(\varepsilon, T)$ for fixed $c > 0$ by (a) in Condition 6. Plugging in the two estimation and using $N^+(\varepsilon, T) \asymp N^+(\varepsilon, P_S T P_S)$, we obtain

$$\begin{aligned} N^+(\varepsilon, P_\Omega T P_\Omega) &\ge \frac{1}{m} N^+(2\varepsilon, P_S T P_S) - o \left(N^+(\varepsilon, T) \right) \ge cN^+(2\varepsilon, T) - o \left(N^+(\varepsilon, T) \right) \\ &\ge cN^+(\varepsilon, T) - o \left(N^+(\varepsilon, T) \right), \end{aligned}$$

which proves the desired lower bound. ■

After all these preparation, we can prove Theorem 8:

Proof of Theorem 8 Let T be the integral operator associated with k. We start with the case of $\rho = \mathbf{1}_S$ is the indicator of an open set S and $\mu_n = \tilde{\mu}_n$. Then from Proposition 1 it suffices to consider $P_S T P_S$. Since the asymptotic behavior of $N^+(\varepsilon, A)$ determines uniquely $\lambda_i(A)$, it suffices to prove that $N^+(\varepsilon, P_S T P_S) \asymp N^+(\varepsilon, T)$.

We take the sequence $U_0, V_0, U_1, V_1, \cdots \subseteq \mathbb{S}^d$ of subdomains given in Proposition 53 and prove that $N^+(\varepsilon, P_{U_ii} T P_{U_i}) \asymp N^+(\varepsilon, T)$ by induction. The initial case follows from $U_0 = \mathbb{S}^d$. Suppose $N^+(\varepsilon, U_i) \asymp N^+(\varepsilon, T)$, by Lemma 24 and the fact that there are isometric copies of V_i whose disjoint union is U_i, we obtain $N^+(\varepsilon, P_{V_i} T P_{V_i}) \asymp N^+(\varepsilon, T)$. Moreover, since $V_i \subseteq U_{i+1}$, by Corollary 23 again, we have

$$N^+(\varepsilon, P_{V_i} T P_{V_i}) \leq N^+(\varepsilon, P_{U_{i+1}} T P_{U_{i+1}}) \leq N^+(\varepsilon, T)$$

and thus $N^+(\varepsilon, P_{U_{i+1}} T P_{U_{i+1}}) \asymp N^+(\varepsilon, T)$.

Now we have shown that $N^+(\varepsilon, P_{U_i} T P_{U_i}) \asymp N^+(\varepsilon, T)$. Since S is an open set and diam $U_i \to 0$, we can find some $U_i \subseteq S$, and hence $N^+(\varepsilon, P_S T P_S) \asymp N^+(\varepsilon, T)$ by Corollary 23.

For the general case of μ_n, let T_- and T_+ be the integral operators defined similarly to (25) by the sequences $c_1 \tilde{\mu}_n$ and $c_2 \tilde{\mu}_n$ respectively. Then, $T_- \preceq T \preceq T_+$ and thus $P_\Omega T_- P_\Omega \preceq P_\Omega T P_\Omega \preceq P_\Omega T_+ P_\Omega$, implying that

$$\lambda_i \left(P_S T_- P_S\right) \leq \lambda_i \left(P_S T P_S\right) \leq \lambda_i \left(P_S T_+ P_S\right)$$

and the results are obtained immediately from the previous case.

Finally, suppose ρ is a bounded Riemann-integrable function that is non-zero. The upper bound is proven by Lemma 2 with boundedness of ρ. For the lower bound, we assert that there is an open set Ω such that $\rho(x)^2 \geq c > 0$ on Ω. Then, by Lemma 2 again, we conclude that

$$\lambda_i(k; \mathbb{S}^d, \rho^2 \mathrm{d}\sigma) \geq \lambda_i(k; \Omega, \rho^2 \mathrm{d}\sigma) \geq c \lambda_i(k; \Omega, \mathrm{d}\sigma) \asymp \lambda_i(k; \mathbb{S}^d, \mathrm{d}\sigma)$$

Now we prove the assertion. Since ρ is Riemann-integrable, the set of discontinuity is a null-set. If $\rho(x) = 0$ for all the continuity point, then $\rho(x) = 0$, a.e., which contradicts to the assumption that ρ is non-zero. So there is a continuity point x_0 such that $\rho(x_0) > 0$, and Ω can be taken as a small neighbour of x_0.

4.5 Discussion on Condition 6

In this subsection, we discuss some sufficient conditions that Condition 6 holds. The first proposition shows the basic relation between the difference and the derivative if μ_n is given by a function f, which is a direct consequences of (58)

Proposition 25 *Suppose $\mu_n = f(n)$ for some function $f(x)$ defined on $\mathbb{R}_{\geq 0}$. Then,*

(1) If $(-1)^p f^{(p)}(x) \geq 0,\ \forall x \geq N_0$, then $\triangle^p \mu_n \geq 0,\ \forall n \geq N_0$.

(2) If $(-1)^{p+1} f^{(p+1)}(x) \geq 0,\ \forall x \geq N_0$, then $\triangle^p \mu_n \leq f^{(p)}(n),\ \forall n \geq N_0$.

The next lemma shows that bounding the highest order term in (8) is sufficient.

Lemma 26 *If $\binom{n+d}{d}\triangle^d \mu_n \leq B_n$ holds for a decreasing sequence B_n for all $n \geq 0$. Then, $\binom{n+l}{l}\triangle^l \mu_n \leq \frac{d}{l} B_n$ holds for all $1 \leq l \leq d$. Consequently, if $B_{qn} \leq D' \mu_n$ for some $q \in \mathbb{N}_+$ and $D' > 0$, then* (8) *holds.*

Proof We prove the result by induction. Suppose the statement holds for $l+1$, then

$$\begin{aligned}\binom{n+l}{l}\triangle^l \mu_n &= \binom{n+l}{l}\sum_{k\geq n}\triangle^{l+1}\mu_k \leq \binom{n+l}{l}\sum_{k\geq n}\frac{d}{l+1}B_k\binom{k+l+1}{l+1}^{-1} \\ &\leq \binom{n+l}{l}\frac{d}{l+1}B_n\sum_{k\geq n}\binom{k+l+1}{l+1}^{-1} \\ &= \binom{n+l}{l}\frac{d}{l+1}B_n\frac{n!(l+1)!}{l(n+l)!} = \frac{d}{l}B_n.\end{aligned}$$

■

Combining the previous two results yields the following corollary.

Corollary 27 *Suppose $\mu_n = f(n)$ for some function $f(x)$ defined on $\mathbb{R}_{\geq 0}$. Then a sufficient condition that (b,c) in Condition* 6 *holds for $n \geq N_0$ is that*

$$(-1)^{d+1} f^{(d+1)}(x) \geq 0 \quad \textit{and} \quad (-1)^d x^d f^{(d)}(qx) \leq D' f(x), \quad \forall x \geq N_0 \tag{37}$$

for some $q \in \mathbb{N}_+$ and $D' > 0$.

Proposition 28 *For each of the following formulations of μ_n, there is a sequence $(\mu_n)_{n\geq 0}$ that Condition* 6 *is satisfied and the formulation holds when n is sufficiently large.*

- *$\mu_n = c_0 n^{-\beta}$ for $c_0 > 0$ and $\beta > d$;*
- *$\mu_n = c_0 \exp\left(-c_1 n^\beta\right)$ for $c_0, c_1, \beta > 0$;*
- *$\mu_n = c_0 n^{-\beta}(\ln n)^p$ for $c_0 > 0$, $\beta > d$ and $p \in \mathbb{R}$, or $\beta = d$ and $p > 1$.*

Proof The condition (a) is obviously satisfied by these asymptotic rates. We verify (b,c) by Corollary 27 when $n \geq N_0$ for some large N_0 and take a left extrapolation of μ_n as in Lemma 61 so the conditions hold for all $n \geq 0$.

- For $\mu_n = f(n) = c_0 n^{-\beta}$, we have $(-1)^p f^{(p)}(x) = c_0(\beta)_p x^{-(\beta+p)}$, where $(\beta)_p = \beta(\beta+1)\cdots(\beta+p-1)$, so (37) holds for $q = 1$ and $D' = (\beta)_p$.
- For $\mu_n = f(n) = c_0 \exp\left(-c_1 n^\beta\right)$, it is easy to show that

  $$(-1)^p f^{(p)}(x) \asymp c_0 (c_1\beta)^p x^{p(\beta-1)} \exp\left(-c_1 x^\beta\right) \quad \text{as} \quad x \to \infty,$$

  so (37) holds if we take $q = 2$ since the exponential term is dominating.

- For $\mu_n = f(n) = c_0 n^{-\beta}(\ln n)^p$, we have

  $$(-1)^p f^{(p)}(x) \asymp c_0 (\beta)_p x^{-(\beta+p)}(\ln x)^p \quad \text{as} \quad x \to \infty,$$

  so (37) still holds for $q = 1$.

■

5. Conclusion

In this paper, we develop a novel approach for determining the eigenvalue decay rate (EDR) of certain kernels using transformation and restriction. Using this approach, we determine the EDR of the NTKs associated with multilayer fully-connected ReLU neural networks on a general domain. Combining this result with the uniform approximation of the neural network by the NTK regression, we determine the generalization performance of the over-parameterized neural network through the kernel regression theory. The theoretical results show that proper early stopping is essential for the generalization performance of the neural networks, which urges us to scrutinize the widely reported "benign overfitting phenomenon" in deep neural network literature.

For future directions, it is natural to extend our results to the NTKs associated with other neural network architectures, such as convolutional neural networks and residual neural networks. Also, it would be of great interest to see if these results can be extended to the large dimensional data where $d \propto n^s$ for some $s > 0$ instead of the fixed d here.

Acknowledgments

The authors thank the anonymous reviewers for their valuable suggestions. Lin's research is supported in part by the National Natural Science Foundation of China (Grant 92370122, Grant 11971257)

Appendix A. Uniform Convergence of the Neural Network to Kernel Regression

In this section we will prove Lemma 12. Applying Proposition 3.2 and Proposition 3.3 in Lai et al. (2023), it suffices to show Proposition 45, that is, the kernel K_t converges uniformly to K^{NT}. This rest of this section is organized as follows: We first introduce some more preliminaries; in Section A.1, we discuss some properties of the network at initialization; in Section A.2, we analyze the effect of small perturbation during training process of the network; we then prove the lazy regime approximation of the neural network in Section A.3; we also show the Hölder continuity of K^{NT} in Section A.4; finally, we prove the kernel uniform convergence in Section A.5.

Further notations Let us denote $\tilde{B}_R = \left\{x \in \mathbb{R}^d : \tilde{x} \leq R\right\}$ for $R \geq 1$. For a vector $\boldsymbol{v} = (v_1, v_2, \cdots, v_m) \in \mathbb{R}^m$, we use $\|\boldsymbol{v}\|_2$ (or simply $\|\boldsymbol{v}\|$) to represent the Euclidean norm. Additionally, if we have a univariate function $f : \mathbb{R} \to \mathbb{R}$, we define $f(\boldsymbol{v}) = (f(v_1), f(v_2), \cdots, f(v_m)) \in \mathbb{R}^m$. We denote by $\|\boldsymbol{M}\|_2$ and $\|\boldsymbol{M}\|_{\mathrm{F}}$ the spectral and Frobenius norm of a matrix $\boldsymbol{M}$ respectively. Also, we use $\|\cdot\|_0$ to represent the number of non-zero elements of a vector or matrix. For matrices $\boldsymbol{A} \in \mathbb{R}^{n_1 \times n_2}$ and $\boldsymbol{B} \in \mathbb{R}^{n_2 \times n_1}$, we define $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \mathrm{Tr}\big(\boldsymbol{A}\boldsymbol{B}^T\big)$. We remind that $\langle \boldsymbol{M}, \boldsymbol{M} \rangle = \|\boldsymbol{M}\|_{\mathrm{F}}^2$ in this way.

Network Architecture Let us recall the neural network in the main text. Since it can be shown easily that the bias term $\boldsymbol{b}^{(0,p)}$ in the first layer can be absorbed into $\boldsymbol{A}^{(p)}$ if we append an 1 at the last coordinate of x, we denote $\boldsymbol{W}^{(0,p)} = (\boldsymbol{A}^{(p)}\ \boldsymbol{b}^{(0,p)})$, $\tilde{x} = (x^T, 1)^T \in \mathbb{R}^d \times \{1\} \subset \mathbb{R}^{d+1}$ and consider the following equivalent neural network:

$$\begin{aligned}
\boldsymbol{\alpha}^{(0,p)}(x) &= \widetilde{\boldsymbol{\alpha}}^{(0,p)}(x) = \tilde{x} \in \mathbb{R}^{d+1}, \\
\widetilde{\boldsymbol{\alpha}}^{(l,p)}(x) &= \sqrt{\frac{2}{m_l}} \boldsymbol{W}^{(l-1,p)} \boldsymbol{\alpha}^{(l-1,p)}(x) \in \mathbb{R}^{m_l}, \\
\boldsymbol{\alpha}^{(l,p)}(x) &= \sigma\Big(\widetilde{\boldsymbol{\alpha}}^{(l,p)}(x)\Big) \in \mathbb{R}^{m_l}, \\
g^{(p)}(x; \boldsymbol{\theta}) &= \boldsymbol{W}^{(L,p)} \boldsymbol{\alpha}^{(L,p)}(x) + b^{(L,p)} \in \mathbb{R}, \\
f(x; \boldsymbol{\theta}) &= \frac{\sqrt{2}}{2} \Big[g^{(1)}(x; \boldsymbol{\theta}) - g^{(2)}(x; \boldsymbol{\theta})\Big] \in \mathbb{R}.
\end{aligned} \tag{38}$$

for $p = 1, 2$ and $l = 1, \cdots, L$. Recall that the integers $m_1, m_2, \cdots, m_L$ are the width of L-hidden layers and $m_{L+1} = 1$ is the width of output layer. Additionally, we have set $m = \min(m_1, m_2, \cdots, m_L)$ and made the assumption that $\max(m_1, m_2, \ldots, m_L) \leq C_{\mathrm{m}} m$ for some absolute constant C_{m}. By setting $m_0 = d + 1$ for convenience, we have $\boldsymbol{W}^{(l,p)} \in \mathbb{R}^{m_{l+1} \times m_l}$ for $l \in \{0, 1, 2, \cdots, L\}$.

For $p = 1, 2$, we define $g_t^{(p)}(x) = g^{(p)}(x; \boldsymbol{\theta}_t)$ and $f_t(x) = f(x; \boldsymbol{\theta}_t)$. Similarly, we also add a subscript t for all the related quantities (including those defined afterwards) to indicate their values at time t during the training process.

The neural tangent kernel Let us consider the following neural network kernel

$$K_t(x, x') = \left\langle \nabla_{\boldsymbol{\theta}} f(x; \boldsymbol{\theta}_t), \nabla_{\boldsymbol{\theta}} f(x'; \boldsymbol{\theta}_t) \right\rangle.$$

Then, the gradient flow can be written as (Jacot et al., 2018)

$$\dot{f}(x;\boldsymbol{\theta}_t) = -\frac{1}{n}K_t(x,\boldsymbol{X})\left(f(\boldsymbol{X};\boldsymbol{\theta}_t)-\boldsymbol{y}\right).$$

As shown in Jacot et al. (2018), as the width m goes to infinity, the kernel K_t converges to the deterministic neural tangent kernel K^{NT}. With the mirrored architecture given by (38), we can express the kernel $K_t(x,x')$ as follows:

$$\begin{aligned}K_t(x,x') &= \langle \nabla_{\boldsymbol{\theta}} f(x;\boldsymbol{\theta}_t), \nabla_{\boldsymbol{\theta}} f(x;\boldsymbol{\theta}_t)\rangle = \sum_{p=1}^{2}\langle \nabla_{\boldsymbol{\theta}^{(p)}} f(x;\boldsymbol{\theta}_t), \nabla_{\boldsymbol{\theta}^{(p)}} f(x;\boldsymbol{\theta}_t)\rangle \\ &= \frac{1}{2}\sum_{p=1}^{2}\left\langle \nabla_{\boldsymbol{\theta}^{(p)}} g_t^{(p)}(x), \nabla_{\boldsymbol{\theta}^{(p)}} g_t^{(p)}(x')\right\rangle = \frac{1}{2}\left(K_t^{(1)}(x,x')+K_t^{(2)}(x,x')\right),\end{aligned}$$

where $K_t^{(p)}(x,x') = \left\langle \nabla_{\boldsymbol{\theta}^{(p)}} g_t^{(p)}(x), \nabla_{\boldsymbol{\theta}^{(p)}} g_t^{(p)}(x')\right\rangle$ is the neural network kernel of $g_t^{(p)}$ for $p=1,2$, which is a vanilla neural network. Consequently, due to the mirror initialization, we have $K_0^{(1)}(x,x') = K_0^{(2)}(x,x') = K_0(x,x')$.

An expanded matrix form Sometimes it is convenient to write the neural network (38) in an expanded matrix form as introduced in Allen-Zhu et al. (2019b). Let us define the activation matrix

$$\boldsymbol{D}_x^{(l,p)} = \begin{cases}\boldsymbol{I}_{d+1}, & l=0;\\ \operatorname{diag}\left(\dot{\sigma}\left(\widetilde{\boldsymbol{\alpha}}^{(l,p)}(x)\right)\right), & l\geq 1,\end{cases} \quad \in \mathbb{R}^{m_l\times m_l}$$

for $p=1,2$, where $\boldsymbol{I}_{d+1}$ is the identity matrix. Then, we have

$$\widetilde{\boldsymbol{\alpha}}^{(l,p)}(x) = \sqrt{\frac{2}{m_l}}\boldsymbol{W}^{(l-1,p)}\boldsymbol{D}_x^{(l-1,p)}\widetilde{\boldsymbol{\alpha}}^{(l-1,p)}(x), \quad \boldsymbol{\alpha}^{(l,p)}(x) = \sqrt{\frac{2}{m_l}}\boldsymbol{D}_x^{(l,p)}\boldsymbol{W}^{(l-1,p)}\boldsymbol{\alpha}^{(l-1,p)}(x).$$

To further write it as product of matrices, we first introduce the following notation to avoid confusion since matrix product is not commutative. For matrices $\boldsymbol{A}_0,\boldsymbol{A}_1,\cdots,\boldsymbol{A}_L$, we define the left multiplication product

$$\prod_{i=a}^{b}\boldsymbol{A}_i = \begin{cases}1, & 0\leq b<a\leq L;\\ \boldsymbol{A}_b\boldsymbol{A}_{b-1}\cdots\boldsymbol{A}_{a+1}\boldsymbol{A}_a, & 0\leq a\leq b\leq L.\end{cases}$$

Since real number multiplication is commutative, the notation introduced above is compatible with the traditional usage when $\boldsymbol{A}_0,\boldsymbol{A}_1,\cdots,\boldsymbol{A}_L$ degenerate into real numbers. In this way, we have

$$\widetilde{\boldsymbol{\alpha}}^{(l,p)}(x) = \left(\prod_{r=1}^{l}\sqrt{\frac{2}{m_r}}\boldsymbol{W}^{(r-1,p)}\boldsymbol{D}_x^{(r-1,p)}\right)\tilde{x}, \quad \boldsymbol{\alpha}^{(l,p)}(x) = \left(\prod_{r=1}^{l}\sqrt{\frac{2}{m_r}}\boldsymbol{D}_x^{(r,p)}\boldsymbol{W}^{(r-1,p)}\right)\tilde{x}. \tag{39}$$

Using the above expressions, we can obtain

$$g^{(p)}(x) = \boldsymbol{W}^{(L,p)}\boldsymbol{\alpha}^{(L,p)}(x) + b^{(L,p)} = \boldsymbol{W}^{(L,p)}\left(\prod_{r=l+1}^{L}\sqrt{\frac{2}{m_r}}\boldsymbol{D}_x^{(r,p)}\boldsymbol{W}^{(r-1,p)}\right)\boldsymbol{\alpha}^{(l,p)}(x) + b^{(L,p)}$$
$$= \left(\prod_{r=l+1}^{L}\sqrt{\frac{2}{m_r}}\boldsymbol{W}^{(r,p)}\boldsymbol{D}_x^{(r,p)}\right)\boldsymbol{W}^{(l,p)}\boldsymbol{\alpha}^{(l,p)}(x) + b^{(L,p)}.$$

Finally, we use the above results to calculate the gradient $\nabla_{\boldsymbol{W}^{(l,p)}} g^{(p)}(x)$. To simplify notation, we define:

$$\widetilde{\boldsymbol{\alpha}}_x^{(l,p)} = \widetilde{\boldsymbol{\alpha}}^{(l,p)}(x), \quad \boldsymbol{\alpha}_x^{(l,p)} = \boldsymbol{\alpha}^{(l,p)}(x), \quad \boldsymbol{\gamma}_x^{(l,p)} = \left(\prod_{r=l+1}^{L}\sqrt{\frac{2}{m_r}}\boldsymbol{W}^{(r,p)}\boldsymbol{D}_x^{(r,p)}\right)^T \in \mathbb{R}^{m_{l+1}}. \tag{40}$$

Then we can obtain

$$g^{(p)}(x) = \boldsymbol{\gamma}_x^{(l,p),T}\boldsymbol{W}^{(l,p)}\boldsymbol{\alpha}_x^{(l,p)} + b^{(L,p)},$$

which can lead to

$$\nabla_{\boldsymbol{W}^{(l,p)}} g^{(p)}(x) = \boldsymbol{\gamma}_x^{(l,p)}\boldsymbol{\alpha}_x^{(l,p),T}, \quad l = 0, 1, \cdots, L, \ p = 1, 2. \tag{41}$$

Also, it is worth noting that for two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, we have

$$\left\|\boldsymbol{a}\boldsymbol{b}^T\right\|_{\mathrm{F}}^2 = \mathrm{Tr}\big(\boldsymbol{a}\boldsymbol{b}^T\boldsymbol{b}\boldsymbol{a}^T\big) = \mathrm{Tr}\big(\boldsymbol{a}^T\boldsymbol{a}\boldsymbol{b}^T\boldsymbol{b}\big) = \|\boldsymbol{a}\|_2^2\|\boldsymbol{b}\|_2^2.$$

Consequently, we can get

$$\left\|\nabla_{\boldsymbol{W}^{(l,p)}} g^{(p)}(x)\right\|_{\mathrm{F}} = \left\|\boldsymbol{\gamma}_x^{(l,p)}\right\|_2\left\|\boldsymbol{\alpha}_x^{(l,p)}\right\|_2. \tag{42}$$

A.1 Initialization

Since our neural network is mirrored, we can focus only on one part $g^{(p)}(x)$ of the network at initialization. For notational simplicity, we omit the superscript p for $\boldsymbol{W}_t^{(l,p)}$ and other notations in the following if there is no ambiguity. And unless otherwise stated, it is understood that the conclusions hold for both $p = 1$ and $p = 2$.

Since $K_0^{(p)}$ corresponds to the tangent kernel of a vanilla fully connected neural network, Arora et al. (2019b, Theorem 3.1) shows the following convergence result.

Lemma 29 (Convergence to the NTK at initialization) *There exist some positive absolute constants $C_1 > 0$ and $C_2 \geq 1$ such that if $\varepsilon \in (0, 1)$, $\delta \in (0, 1)$ and $m \geq C_1\varepsilon^{-4}\ln(C_2/\delta)$, then for any fixed $\boldsymbol{z}, \boldsymbol{z}'$ such that $\|\boldsymbol{z}\| \leq 1$ and $\|\boldsymbol{z}'\| \leq 1$, with probability at least $1 - \delta$ with respect to the initialization, we have*

$$\left|K_0^{(p)}\big(\boldsymbol{z}, \boldsymbol{z}'\big) - K^{\mathrm{NT}}\big(\boldsymbol{z}, \boldsymbol{z}'\big)\right| \leq \varepsilon.$$

Letting $\varepsilon = m^{-1/5}$ in the previous lemma, we can get the following corollary:

Corollary 30 *There exist some positive absolute constants $C_1 > 0$ and $C_2 \geq 1$ such that if $\delta \in (0,1)$ and $m \geq C_1 \left(\ln(C_2/\delta)\right)^5$, then for any fixed $\boldsymbol{z}, \boldsymbol{z}' \in \tilde{B}_R$, with probability at least $1-\delta$ with respect to the initialization, we have*

$$\left| K_0^{(p)}(\boldsymbol{z}, \boldsymbol{z}') - K^{\mathrm{NT}}(\boldsymbol{z}, \boldsymbol{z}') \right| = O\Big(R^2 m^{-1/5}\Big).$$

Now we further provide some bounds about the magnitudes of weight matrices and layer outputs. The following is a standard estimation of Gaussian random matrix, which is a direct consequence of Vershynin (2010, Corollary 5.35).

Lemma 31 *At initialization, there exists a positive absolute constant C, such that when $m \geq C$, with probability at least $1 - \exp(-\Omega(m))$ with respect to the initialization, we have*

$$\left\| \boldsymbol{W}_0^{(l)} \right\|_2 = O\big(\sqrt{m}\big), \quad l \in \{0, 1, \ldots, L\}.$$

Noticing that $\left\| \boldsymbol{D}_x^{(l)} \right\|_2 \leq 1$ and combining Lemma 31 with (39), (40) and (42), we have:

Lemma 32 *There exists a positive absolute constant C, such that when $m \geq C$, with probability at least $1-\exp(-\Omega(m))$ with respect to the initialization, for any $l \in \{0, 1, \cdots, L\}$ and $x \in \tilde{B}_R$, we have*

$$\left\| \tilde{\boldsymbol{\alpha}}_{x,0}^{(l)} \right\|_2 = O(R), \quad \left\| \boldsymbol{\alpha}_{x,0}^{(l)} \right\|_2 = O(R), \quad \left\| \boldsymbol{\gamma}_{x,0}^{(l)} \right\|_2 = O(1) \quad \text{and} \quad \left\| \nabla_{\boldsymbol{W}^{(l)}} g_0(x) \right\|_{\mathrm{F}} = O(R).$$

Lemma 31 and Lemma 32 provide some upper bounds, and the subsequent lemma provides a lower bound. It is important to note that the previous lemma holds uniformly for $x \in \tilde{B}_R$, while the following lemma only holds pointwisely.

Lemma 33 (Lemma 7.1 in Allen-Zhu et al. (2019b)) *There exists a positive absolute constant C such that when $m \geq C$, for any fixed $\boldsymbol{z} \in \tilde{B}_R$, with probability at least $1 - \exp(-\Omega(m))$ with respect to the initialization, we have $\left\| \boldsymbol{\alpha}_{\boldsymbol{z},0}^{(l)} \right\|_2 = \Theta(R)$ for $l \in \{0, 1, \cdots, L\}$.*

A.2 The training process

In this subsection we will show that as long as the parameters and input do not change much, some the relative quantities can also be bounded. We still focus on one parity in this subsection and suppress the superscript p for convenience.

The most crucial result we will obtain in this subsection is the following proposition:

Proposition 34 *Fix $\boldsymbol{z}, \boldsymbol{z}' \in \tilde{B}_R$ and $T \subseteq [0, \infty)$. Suppose that $\left\| \boldsymbol{W}_t^{(l)} - \boldsymbol{W}_0^{(l)} \right\|_{\mathrm{F}} = O(m^{1/4})$ holds for all $t \in T$ and $l \in \{0, 1, \cdots, L\}$. Then there exists a positive absolute constant C such that when $m \geq C$, with probability at least $1 - \exp\big(-\Omega(m^{5/6})\big)$, for any $x, x' \in \mathcal{X}$ such that $\|x - z\|_2, \|x' - z'\|_2 \leq O(1/m)$, we have*

$$\sup_{t \in T} \left| K_t^{(p)}(x, x') - K_0^{(p)}(z, z') \right| = O\Big(R^2 m^{-1/12} \sqrt{\ln m}\Big).$$

The proof of this proposition will be presented at the end of this subsection. Combining this proposition with Corollary 30, we can derive the following corollary:

Corollary 35 *Fix $\boldsymbol{z}, \boldsymbol{z}' \in \tilde{B}_R$ and let $\delta \in (0,1)$, $T \subseteq [0,\infty)$. Suppose that $\left\|\boldsymbol{W}_t^{(l)} - \boldsymbol{W}_0^{(l)}\right\|_{\mathrm{F}} = O(m^{1/4})$ holds for all $t \in T$ and $l \in \{0,1,\cdots,L\}$. Then there exist some positive absolute constants $C_1 > 0$ and $C_2 \geq 1$ such that with probability at least $1-\delta$, for any $x, x' \in \mathcal{X}$ such that $\|x - z\|_2, \|x' - z'\|_2 \leq O(1/m)$, we have*

$$\sup_{t\in T}\left|K_t^{(p)}(x,x') - K^{\mathrm{NT}}(z,z')\right| = O\Big(R^2 m^{-1/12}\sqrt{\ln m}\Big), \text{ when } m \geq C_1\left(\ln(C_2/\delta)\right)^5.$$

To prove Proposition 34, we need to introduce some necessary lemmas. In Lemma 32, we have provided upper bounds for the norms of $\tilde{\boldsymbol{\alpha}}_{x,0}^{(l)}$, $\boldsymbol{\alpha}_{x,0}^{(l)}$, $\boldsymbol{\gamma}_{x,0}^{(l)}$ and $\nabla_{\boldsymbol{W}^{(l)}} g_0(x)$ at initialization. Next, we aim to prove that under perturbations in the parameters and the input, the corresponding changes in these quantities will also be small.

In fact, similar lemmas can be found in Allen-Zhu et al. (2019b), although they have different conditions compared to the lemmas in this paper. For example, in Allen-Zhu et al. (2019b), the input points are constrained to lie on a sphere, the input and output layers are not involved in training, and each hidden layer has the same width.

However, the most crucial point is that Allen-Zhu et al. (2019b) did not consider the impact of small perturbations in the input, which is vital for proving uniform convergence. In fact, the slight perturbation between x and z can be regarded as taking a slight perturbation on $\boldsymbol{W}^{(0)}$, with other $\boldsymbol{W}^{(l)}$ fixed. Additionally, since this paper fixes the number of layers L, there is no need to consider the impact of L on the bounds. This simplifies the proof of the corresponding conclusions.

Inspired by Allen-Zhu et al. (2019b, Lemma 8.2), we can prove the following lemma:

Lemma 36 *Let $\Delta = O(1/\sqrt{m})$, $\tau \in \left[\Delta\sqrt{m}, O\big(\sqrt{m}/(\ln m)^3\big)\right]$, $T \subseteq [0,\infty)$ and fix $\boldsymbol{z} \in \tilde{B}_R$. Suppose that $\left\|\boldsymbol{W}_t^{(l)} - \boldsymbol{W}_0^{(l)}\right\|_{\mathrm{F}} \leq \tau$ holds for all $t \in T$ and $l \in \{0,1,\cdots,L\}$. Then there exists a positive absolute constant C such that with probability at least $1-\exp\big(-\Omega(m^{2/3}\tau^{2/3})\big)$, for all $t \in T$, $l \in \{0,1,\cdots,L\}$ and $x \in \tilde{B}_R$ such that $\|x - \boldsymbol{z}\|_2 \leq \Delta$, we have*

(a) $\left\|\tilde{\boldsymbol{\alpha}}_{x,t}^{(l)} - \tilde{\boldsymbol{\alpha}}_{z,0}^{(l)}\right\|_2 = O(R\tau/\sqrt{m})$ and thus $\left\|\tilde{\boldsymbol{\alpha}}_{x,t}^{(l)}\right\|_2 = O(R)$;

(b) $\left\|\boldsymbol{D}_{x,t}^{(l)} - \boldsymbol{D}_{\boldsymbol{z},0}^{(l)}\right\|_0 = O(m^{2/3}\tau^{2/3})$ and $\left\|\left(\boldsymbol{D}_{x,t}^{(l)} - \boldsymbol{D}_{\boldsymbol{z},0}^{(l)}\right)\tilde{\boldsymbol{\alpha}}_{x,t}^{(l)}\right\|_2 = O(R\tau/\sqrt{m})$;

(c) $\left\|\boldsymbol{\alpha}_{x,t}^{(l)} - \boldsymbol{\alpha}_{z,0}^{(l)}\right\|_2 = O(R\tau/\sqrt{m})$ and thus $\left\|\boldsymbol{\alpha}_{x,t}^{(l)}\right\|_2 = O(R)$,

when m is greater than the positive constant C.

Proof We use mathematical induction to prove this lemma. When $l = 0$, it can be easily verified that $\boldsymbol{D}_{\boldsymbol{x},t}^{(0)} - \boldsymbol{D}_{\boldsymbol{z},0}^{(0)} = \boldsymbol{O}$ and $\left\|\tilde{\boldsymbol{\alpha}}_{x,t}^{(l)} - \tilde{\boldsymbol{\alpha}}_{z,0}^{(l)}\right\|_2 = \left\|\boldsymbol{\alpha}_{x,t}^{(0)} - \boldsymbol{\alpha}_{\boldsymbol{z},0}^{(0)}\right\|_2 = \|\tilde{x} - \tilde{z}\|_2 = \|x - \boldsymbol{z}\|_2 \leq \Delta \leq \tau/\sqrt{m}$, where $\boldsymbol{O}$ represents the zero matrix, $\tilde{x}$ and $\tilde{\boldsymbol{z}}$ are extended vectors with an additional coordinate of 1. Thus, all the statements hold for $l = 0$. Now we assume that this lemma holds for $l = k \in \{0,1,\cdots,L-1\}$.

(a) First of all, we can decompose $\widetilde{\boldsymbol{\alpha}}_{x,t}^{(k+1)} - \widetilde{\boldsymbol{\alpha}}_{z,0}^{(k+1)}$ as following:

$$\begin{aligned}\widetilde{\boldsymbol{\alpha}}_{x,t}^{(k+1)} - \widetilde{\boldsymbol{\alpha}}_{z,0}^{(k+1)} &= \sqrt{\frac{2}{m_{k+1}}}\boldsymbol{W}_t^{(k)}\boldsymbol{\alpha}_{x,t}^{(k)} - \sqrt{\frac{2}{m_{k+1}}}\boldsymbol{W}_0^{(k)}\boldsymbol{\alpha}_{z,0}^{(k)} \\ &= \sqrt{\frac{2}{m_{k+1}}}\left(\boldsymbol{W}_t^{(k)} - \boldsymbol{W}_0^{(k)}\right)\boldsymbol{\alpha}_{x,t}^{(k)} + \sqrt{\frac{2}{m_{k+1}}}\boldsymbol{W}_0^{(k)}\left(\boldsymbol{\alpha}_{x,t}^{(k)} - \boldsymbol{\alpha}_{z,0}^{(k)}\right).\end{aligned}$$

From the above equation, we can deduce that (a) holds for $l = k+1$ by the induction hypothesis and Lemma 32.

(b) Let us consider the following choices in Lemma 37:

$$\boldsymbol{g} = \frac{\widetilde{\boldsymbol{\alpha}}_{z,0}^{(k+1)}}{\sqrt{2/m_{k+1}}\left\|\boldsymbol{\alpha}_{z,0}^{(k)}\right\|_2} = \frac{\boldsymbol{W}_0^{(k)}\boldsymbol{\alpha}_{z,0}^{(k)}}{\left\|\boldsymbol{\alpha}_{z,0}^{(k)}\right\|_2}, \qquad \boldsymbol{g}' = \frac{\widetilde{\boldsymbol{\alpha}}_{x,t}^{(k+1)} - \widetilde{\boldsymbol{\alpha}}_{z,0}^{(k+1)}}{\sqrt{2/m_{k+1}}\left\|\boldsymbol{\alpha}_{z,0}^{(k)}\right\|_2}.$$

It follows that $\boldsymbol{g} \sim N(0, \boldsymbol{I})$ if we fix $\boldsymbol{\alpha}_{z,0}^{(k)}$ and only consider the randomness of $\boldsymbol{W}_0^{(k)}$. Also, we have $\|\boldsymbol{g}'\|_2 \leq O(\tau/\sqrt{m}) \cdot O(\sqrt{m}) \leq O(\tau)$ holds for all $x \in \tilde{B}_R$ such that $\|x - \boldsymbol{z}\|_2 \leq \Delta$ since we have previously shown that $\left\|\boldsymbol{\alpha}_{z,0}^{(k)}\right\|_2 = \Theta(R)$ in Lemma 33. Therefore, we can choose $\delta = \Theta(\tau)$ such that $\|\boldsymbol{g}'\|_2 \leq \delta$. Then, we can obtain

$$\boldsymbol{g} + \boldsymbol{g}' = \frac{\widetilde{\boldsymbol{\alpha}}_{x,t}^{(k+1)}}{\sqrt{\frac{2}{m_{k+1}}}\left\|\boldsymbol{\alpha}_{z,0}^{(k)}\right\|_2}, \quad \boldsymbol{D}' = \boldsymbol{D}_{x,t}^{(k+1)} - \boldsymbol{D}_{\boldsymbol{z},0}^{(k+1)} \quad \text{and} \quad \boldsymbol{u} = \frac{\left(\boldsymbol{D}_{x,t}^{(k+1)} - \boldsymbol{D}_{\boldsymbol{z},0}^{(k+1)}\right)\widetilde{\boldsymbol{\alpha}}_{x,t}^{(k+1)}}{\sqrt{\frac{2}{m_{k+1}}}\left\|\boldsymbol{\alpha}_{z,0}^{(k)}\right\|_2},$$

where $\boldsymbol{D}'$ and $\boldsymbol{u}$ are defined as in Lemma 37. By applying Lemma 37, we can get $\|\boldsymbol{D}'\|_0 \leq O(m^{2/3}\tau^{2/3})$ and $\|\boldsymbol{u}\|_2 \leq O(\delta)$, which establish the conclusion of part (b).

(c) Further, the third statement can be directly obtained from the following inequality:

$$\left\|\boldsymbol{\alpha}_{x,t}^{(k+1)} - \boldsymbol{\alpha}_{\boldsymbol{z},0}^{(k+1)}\right\|_2 \leq \left\|\boldsymbol{D}_{\boldsymbol{z},0}^{(k+1)}\left(\widetilde{\boldsymbol{\alpha}}_{x,t}^{(k+1)} - \widetilde{\boldsymbol{\alpha}}_{\boldsymbol{z},0}^{(k+1)}\right)\right\|_2 + \left\|\left(\boldsymbol{D}_{x,t}^{(k+1)} - \boldsymbol{D}_{\boldsymbol{z},0}^{(k+1)}\right)\widetilde{\boldsymbol{\alpha}}_{x,t}^{(k+1)}\right\|_2.$$

Thus, the proof of this lemma is complete. ■

In the proof of Lemma 36, we use the following result:

Lemma 37 (Allen-Zhu et al. (2019b) Claim 8.3) *Suppose each entry of $\boldsymbol{g} \in \mathbb{R}^m$ follows $g_i \overset{\text{i.i.d.}}{\sim} N(0,1)$. For any $\delta > 0$, with probability at least $1 - \exp\left(-\Omega(m^{2/3}\delta^{2/3})\right)$, the following proposition holds:*

Select $\boldsymbol{g}' \in \mathbb{R}^m$ such that $\|\boldsymbol{g}'\|_2 \leq \delta$. Let $\boldsymbol{D}' = \operatorname{diag}(D'_{kk})$ be a diagonal matrix, where the k-th diagonal element D'_{kk} follows

$$D'_{kk} = \boldsymbol{1}\left\{\left(g + g'\right)_k \geq 0\right\} - \boldsymbol{1}\left\{g_k \geq 0\right\}, \quad k \in [m].$$

If we define $\boldsymbol{u} = \boldsymbol{D}'(\boldsymbol{g} + \boldsymbol{g}')$, then it satisfies the following inequalities:

$$\|\boldsymbol{u}\|_0 \leq \left\|\boldsymbol{D}'\right\|_0 \leq O\left(m^{2/3}\delta^{2/3}\right) \quad \textit{and} \quad \|\boldsymbol{u}\|_2 \leq O(\delta).$$

Inspired by Allen-Zhu et al. (2019b, Lemma 8.7), we then have the following lemma:

Lemma 38 *Let $\Delta = O(1/\sqrt{m})$, $\tau \in \big[\Delta\sqrt{m}, O\big(\sqrt{m}/(\ln m)^3\big)\big]$, $T \subseteq [0,\infty)$ and fix $\boldsymbol{z} \in \tilde{B}_R$. Suppose that $\left\|\boldsymbol{W}_t^{(l)} - \boldsymbol{W}_0^{(l)}\right\|_{\mathrm{F}} \le \tau$ holds for all $t \in T$ and $l \in \{0,1,\cdots,L\}$. Then there exists a positive absolute constant C such that with probability at least $1-\exp\big(-\Omega(m^{2/3}\tau^{2/3})\big)$, for all $t \in T$, $l \in \{0,1,\cdots,L\}$ and $x \in \tilde{B}_R$ such that $\|x-\boldsymbol{z}\|_2 \le \Delta$, we have*

$$\left\|\boldsymbol{\gamma}_{x,t}^{(l)} - \boldsymbol{\gamma}_{z,0}^{(l)}\right\|_2 = O\Big(m^{-1/6}\tau^{1/3}\sqrt{\ln m}\Big) \qquad \textit{and thus} \qquad \left\|\boldsymbol{\gamma}_{x,t}^{(l)}\right\|_2 = O(1)$$

when m is greater than the positive constant C.

By using Lemma 36 and Lemma 38, we can prove the following lemma:

Lemma 39 *Let $\Delta = O(1/\sqrt{m})$, $\tau \in \big[\Delta\sqrt{m}, O\big(\sqrt{m}/(\ln m)^3\big)\big]$, $T \subseteq [0,\infty)$ and fix $\boldsymbol{z} \in \tilde{B}_R$. Suppose that $\left\|\boldsymbol{W}_t^{(l)} - \boldsymbol{W}_0^{(l)}\right\|_F \le \tau$ holds for all $t \in T$ and $l \in \{0,1,\cdots,L\}$.*

Then there exists a positive absolute constant C such that with probability at least $1 - \exp\big(-\Omega(m^{2/3}\tau^{2/3})\big)$, for all $t \in T$, $l \in \{0,1,\cdots,L\}$ and $x \in \tilde{B}_R$ such that $\|x-\boldsymbol{z}\|_2 \le \Delta$, we have

$$\|\nabla_{\boldsymbol{W}^{(l)}} g_t(x) - \nabla_{\boldsymbol{W}^{(l)}} g_0(z)\|_{\mathrm{F}} = O\Big(Rm^{-1/6}\tau^{1/3}\sqrt{\ln m}\Big) \ \textit{and thus}\ \|\nabla_{\boldsymbol{W}^{(l)}} g_t(x)\|_{\mathrm{F}} = O(R).$$

when m is greater than the positive constant C.

Proof Recalling (41), we have $\nabla_{\boldsymbol{W}^{(l)}} g_t(x) = \boldsymbol{\gamma}_{x,t}^{(l)}\ \boldsymbol{\alpha}_{x,t}^{(l),T}$. Then, we have

$$\begin{aligned}\|\nabla_{\boldsymbol{W}^{(l)}} g_t(x) - \nabla_{\boldsymbol{W}^{(l)}} g_0(z)\|_{\mathrm{F}} &= \left\|\boldsymbol{\gamma}_{x,t}^{(l)}\boldsymbol{\alpha}_{x,t}^{(l),T} - \boldsymbol{\gamma}_{z,0}^{(l)}\boldsymbol{\alpha}_{z,0}^{(l),T}\right\|_{\mathrm{F}} \\ &\le \left\|\boldsymbol{\gamma}_{x,t}^{(l)} - \boldsymbol{\gamma}_{z,0}^{(l)}\right\|_2\left\|\boldsymbol{\alpha}_{x,t}^{(l)}\right\|_2 + \left\|\boldsymbol{\gamma}_{z,0}^{(l)}\right\|_2\left\|\boldsymbol{\alpha}_{x,t}^{(l)} - \boldsymbol{\alpha}_{z,0}^{(l)}\right\|_2 \le O\Big(Rm^{-1/6}\tau^{1/3}\sqrt{\ln m}\Big)\end{aligned}$$

by Lemma 38, Lemma 36 and Lemma 32.

■

After preparing these tools, now we are ready to give the proof of Proposition 34.

Proof of Proposition 34 By applying Lemma 39 with $\Delta = O(1/m) \le O(1/\sqrt{m})$ and $\tau \asymp m^{1/4} \ge \Delta\sqrt{m}$, with probability at least $1-\exp\big(-\Omega(m^{5/6})\big)$, for all $x \in \tilde{B}_R$ such that $\|x-z\|_2 \le O(1/m)$, we can obtain the following result

$$\|\nabla_{\boldsymbol{W}^{(l)}} g_t(x) - \nabla_{\boldsymbol{W}^{(l)}} g_0(z)\|_{\mathrm{F}} = O\Big(Rm^{-1/12}\sqrt{\ln m}\Big) \text{ and } \|\nabla_{\boldsymbol{W}^{(l)}} g_t(x)\|_{\mathrm{F}} = O(R).$$

The same conclusion holds if we replace x and z with x' and z'. Thus, we have

$$\begin{aligned}\left|K_t^{(p)}(x,x') - K_0^{(p)}(z,z')\right| &\le \sum_{l=0}^{L} \left|\big\langle \nabla_{\boldsymbol{W}^{(l)}} g_t(x), \nabla_{\boldsymbol{W}^{(l)}} g_t(x')\big\rangle - \big\langle \nabla_{\boldsymbol{W}^{(l)}} g_0(z), \nabla_{\boldsymbol{W}^{(l)}} g_0(z')\big\rangle\right| \\ &\le \sum_{l=0}^{L}\Big[\|\nabla_{\boldsymbol{W}^{(l)}} g_t(x)\|_{\mathrm{F}}\big\|\nabla_{\boldsymbol{W}^{(l)}} g_t(x') - \nabla_{\boldsymbol{W}^{(l)}} g_0(z')\big\|_{\mathrm{F}} \\ &\qquad + \|\nabla_{\boldsymbol{W}^{(l)}} g_t(x) - \nabla_{\boldsymbol{W}^{(l)}} g_0(z)\|_{\mathrm{F}}\big\|\nabla_{\boldsymbol{W}^{(l)}} g_0(z')\big\|_{\mathrm{F}}\Big] = O\Big(Rm^{-1/12}\sqrt{\ln m}\Big)\end{aligned}$$

holds for all $t \in T$ with probability at least $1-\exp\big(-\Omega(m^{5/6})\big)$ when m is large enough.

A.3 Lazy regime

In this subsection, we will prove that during the process of gradient descent training, the parameters do not change much. Therefore, the conditions for the lemmas stated in the previous subsection are satisfied. Since training relies on the structure of the neural network, in this subsection, we will no longer omit the superscript p (although the corresponding conclusions also hold for non-mirror neural networks).

Let $\lambda_0 = \lambda_{\min}(K^{\mathrm{NT}}(\boldsymbol{X},\boldsymbol{X}))$ and $\boldsymbol{u}(t) = f_t(\boldsymbol{X}) - \boldsymbol{y}$. Denote $\tilde{M}_{\boldsymbol{X}} = \sum_{i=1}^n \|\tilde{x_i}\|_2$. Since we will show the positive definiteness of the NTK in Proposition 9, we can assume that $\lambda_0 > 0$ hereafter. Similar to Lemma F.8 and Lemma F.7 in Arora et al. (2019b), we have the following lemmas:

Lemma 40 *Let $\delta \in (0,1)$ and $t \in [0,\infty)$. Suppose that $\left\|\boldsymbol{W}_s^{(l,p)} - \boldsymbol{W}_0^{(l,p)}\right\|_{\mathrm{F}} = O(m^{1/4})$ holds for all $s \in [0,t]$, $l \in \{0,1,\cdots,L\}$ and $p \in \{1,2\}$. Then there exists a polynomial* $\operatorname{poly}(\cdot)$ *such that when $m \geq \operatorname{poly}\big(n, \lambda_0^{-1}, \ln(1/\delta)\big)$, with probability at least $1-\delta$, we have*

$$\|\boldsymbol{u}(s)\|^2 \leq \exp\left(-\frac{\lambda_0}{n}s\right)\|\boldsymbol{u}(0)\|^2 = \exp\left(-\frac{\lambda_0}{n}s\right)\|\boldsymbol{y}\|^2, \quad \textit{for all } s \in [0,t]. \tag{43}$$

Proof Denote $\tilde{\lambda}_0(s) = \lambda_{\min}\big(K_s(\boldsymbol{X},\boldsymbol{X})\big)$. By Weyl's inequality, we can get

$$\begin{aligned}\left|\tilde{\lambda}_0(s) - \lambda_0\right| \leq \left\|K_s(\boldsymbol{X},\boldsymbol{X}) - K^{\mathrm{NT}}(\boldsymbol{X},\boldsymbol{X})\right\|_2 &\leq \left\|K_s(\boldsymbol{X},\boldsymbol{X}) - K^{\mathrm{NT}}(\boldsymbol{X},\boldsymbol{X})\right\|_{\mathrm{F}} \\ &\leq \frac{1}{2}\sum_{p=1}^{2}\sum_{i,j=1}^{n}\left|K_s^{(p)}(x_i,x_j) - K^{\mathrm{NT}}(x_i,x_j)\right|.\end{aligned}$$

Applying Corollary 35 with $\delta' = \delta/(2n^2)$ to each difference, with probability at least $1 - 2n^2\delta' = 1-\delta$, we can obtain the following bound for all $s \in [0,t]$:

$$\left|\tilde{\lambda}_0(s) - \lambda_0\right| \leq n^2 O\Big(m^{-1/12}\sqrt{\ln m}\Big) \leq n^2 O\Big(m^{-1/15}\Big) \leq \frac{\lambda_0}{2},$$

when $m \geq C_1\left[\left(n^2\lambda_0^{-1}\right)^{15} + \left(\ln\left(C_2 n^2/\delta\right)\right)^5\right]$ for some positive absolute constants $C_1 > 0$ and $C_2 \geq 1$. This implies that $\tilde{\lambda}_0(s) \geq \lambda_0/2$ holds for all $s \in [0,t]$. Therefore, we have

$$\frac{\mathrm{d}}{\mathrm{d}s}\|\boldsymbol{u}(s)\|^2 = -\frac{2}{n}\boldsymbol{u}(s)^T K_s(\boldsymbol{X},\boldsymbol{X})\boldsymbol{u}(s) \leq -\frac{\lambda_0}{n}\|\boldsymbol{u}(s)\|^2,$$

which implies (43) by standard ODE theory. Finally, by choosing

$$\operatorname{poly}\big(n,\lambda_0^{-1},\ln(1/\delta)\big) = C_1\Big[\left(n^2\lambda_0^{-1}\right)^{15} + \left(2n + \ln(1/\delta) + \ln C_2\right)^5\Big]$$

we can complete the proof of this lemma. ■

Lemma 41 *Fix $l \in \{0, 1, \cdots, L\}$, $p \in \{1, 2\}$ and let $\delta \in (0, 1)$, $t \in [0, \infty)$. Suppose that for $s \in [0, t]$, we have*

$$\|f_s(\boldsymbol{X}) - \boldsymbol{y}\|_2 \le \exp\left(\frac{-\lambda_0}{4n} s\right) \|\boldsymbol{y}\|_2 \qquad \text{and}$$
$$\left\|\boldsymbol{W}_s^{(l',p')} - \boldsymbol{W}_0^{(l',p')}\right\|_{\mathrm{F}} \le \frac{\sqrt{m}}{(\ln m)^3}, \qquad \text{for all } (l', p') \neq (l, p).$$

Then there exists a polynomial $\operatorname{poly}(\cdot)$ *such that when* $m \ge \operatorname{poly}\left(n, \tilde{M}_{\boldsymbol{X}}, \|\boldsymbol{y}\|_2, \lambda_0^{-1}, \ln(1/\delta)\right)$, *with probability at least* $1 - \delta$, *we have* $\sup_{s \in [0,t]} \left\|\boldsymbol{W}_s^{(l,p)} - \boldsymbol{W}_0^{(l,p)}\right\|_{\mathrm{F}} = O\left(n\tilde{M}_{\boldsymbol{X}} \|\boldsymbol{y}\|_2 / \lambda_0\right)$.

Proof First of all, recalling (14) we have

$$\begin{aligned}\left\|\boldsymbol{W}_{t_0}^{(l,p)} - \boldsymbol{W}_0^{(l,p)}\right\|_{\mathrm{F}} = \left\|\int_0^{t_0} \mathrm{d}\boldsymbol{W}_s^{(l,p)}\right\|_{\mathrm{F}} &\le \int_0^{t_0} \left\|\frac{1}{n\sqrt{2}} \sum_{i=1}^n (f_s(x_i) - y_i) \nabla_{\boldsymbol{W}^{(l,p)}} g_s^{(p)}(x_i)\right\|_{\mathrm{F}} \mathrm{d}s \\ &\le \frac{1}{n\sqrt{2}} \sum_{i=1}^n \sup_{0 \le s \le t_0} \left\|\nabla_{\boldsymbol{W}^{(l,p)}} g_s^{(p)}(x_i)\right\|_{\mathrm{F}} \int_0^{t_0} \|f_s(\boldsymbol{X}) - \boldsymbol{y}\|_2 \,\mathrm{d}s \\ &\le O\left(\frac{\|\boldsymbol{y}\|}{\lambda_0}\right) \sum_{i=1}^n \sup_{0 \le s \le t_0} \left\|\nabla_{\boldsymbol{W}^{(l,p)}} g_s^{(p)}(x_i)\right\|_{\mathrm{F}}. \end{aligned} \tag{44}$$

for all $t_0 \in [0, t]$. Suppose that

$$s_0 = \min\left\{s \in [0, t] : \left\|\boldsymbol{W}_s^{(l,p)} - \boldsymbol{W}_0^{(l,p)}\right\|_{\mathrm{F}} \ge \sqrt{m}/(\ln m)^3\right\}$$

exists, then $\left\|\boldsymbol{W}_s^{(l',p')} - \boldsymbol{W}_0^{(l',p')}\right\|_{\mathrm{F}} \le \sqrt{m}/(\ln m)^3$ holds for all $s \in [0, s_0]$, $l' \in \{0, 1, \cdots, L\}$ and $p' \in \{1, 2\}$. Applying Lemma 39 with $\Delta = 0$ and $\tau = \sqrt{m}/(\ln m)^3$, we know that for any $i \in [n]$, with probability at least $1 - \exp\left(-\Omega\left(m/(\ln m)^2\right)\right) \ge 1 - \exp\left(-\Omega(m^{5/6})\right)$, we have

$$\sum_{i=1}^n \sup_{s \in [0, s_0]} \left\|\nabla_{\boldsymbol{W}^{(l,p)}} g_s^{(p)}(x_i)\right\|_{\mathrm{F}} = O(\tilde{M}_{\boldsymbol{X}}).$$

Plugging it back to (44), with probability at least $1 - n\exp\left(-\Omega(m^{5/6})\right)$, we obtain

$$\left\|\boldsymbol{W}_{s_0}^{(l,p)} - \boldsymbol{W}_0^{(l,p)}\right\|_{\mathrm{F}} = O(n\tilde{M}_{\boldsymbol{X}} \|\boldsymbol{y}\| / \lambda_0),$$

which contradicts to $\left\|\boldsymbol{W}_{s_0}^{(l,p)} - \boldsymbol{W}_0^{(l,p)}\right\|_{\mathrm{F}} \ge \sqrt{m}/(\ln m)^3$ when $m \ge C_1 \left(n\tilde{M}_{\boldsymbol{X}} \|\boldsymbol{y}\|_2 \lambda_0^{-1}\right)^5$ for some positive constant C_1.

Now we show that $\left\|\boldsymbol{W}_s^{(l',p')} - \boldsymbol{W}_0^{(l',p')}\right\|_{\mathrm{F}} \le \sqrt{m}/(\ln m)^3$ holds for all $s \in [0, t]$ and any (l', p'). The desired bound then follows from applying Lemma 39 again for the interval $[0, t]$.

Also, it is easy to check that there exists a positive absolute constant C such that when $m \ge C_2 \ln(n/\delta)^{6/5}$, we have $1 - n\exp\left(-\Omega(m^{5/6})\right) \ge 1 - \delta$. Finally, by choosing

$$\operatorname{poly}\left(n, \lambda_0^{-1}, \ln(1/\delta)\right) = C\left[\left(n\tilde{M}_{\boldsymbol{X}} \|\boldsymbol{y}\|_2 \lambda_0^{-1}\right)^5 + (n + \ln(1/\delta))^2 + 1\right]$$

for some positive absolute constant $C > 0$, we can complete the proof. ■

The following lemma is the key lemma that we aim to prove in this subsection. It serves as the prerequisite for establishing the conclusions of the lemmas in the preceding and subsequent subsections.

Lemma 42 (Lazy regime) *There exists a polynomial* $\operatorname{poly}(\cdot)$ *such that for any* $\delta \in (0,1)$*, with probability at least* $1-\delta$*, for all* $p \in \{1,2\}$ *and* $l \in \{0,1,\cdots,L\}$*, we have*

$$\sup_{t\geq 0}\left\|\boldsymbol{W}_t^{(l,p)} - \boldsymbol{W}_0^{(l,p)}\right\|_{\mathrm{F}} = O(m^{1/4}).$$

when $m \geq \operatorname{poly}\Big(n, \tilde{M}_{\boldsymbol{X}}, \|\boldsymbol{y}\|_2, \lambda_0^{-1}, \ln(1/\delta)\Big)$.

Proof Let us assume that

$$t_0 = \min\Big\{t \geq 0 : \exists l, p \text{ such that } \left\|\boldsymbol{W}_t^{(l,p)} - \boldsymbol{W}_0^{(l,p)}\right\|_{\mathrm{F}} \geq m^{1/4} \text{ or } \|\boldsymbol{u}(t)\| \geq \exp\Big(\tfrac{-\lambda_0}{4n}t\Big)\|\boldsymbol{y}\|\Big\}$$

exists. Then, for all $t \in [0, t_0]$, we have

$$\|\boldsymbol{u}(t)\| \leq \exp\left(\frac{-\lambda_0}{4n}t\right)\|\boldsymbol{y}\| \qquad \text{and} \qquad \left\|\boldsymbol{W}_t^{(l,p)} - \boldsymbol{W}_0^{(l,p)}\right\|_{\mathrm{F}} \leq m^{1/4} \text{ for all } l, p.$$

According to Lemma 41 and Lemma 40, we know that there exists a polynomial $\operatorname{poly}(\cdot)$ such that with probability at least $1-\delta$, we have

$$\|\boldsymbol{u}(t_0)\| \leq \exp\left(\frac{-\lambda_0}{2n}t_0\right)\|\boldsymbol{y}\| \qquad \text{and} \qquad \left\|\boldsymbol{W}_{t_0}^{(l,p)} - \boldsymbol{W}_0^{(l,p)}\right\|_{\mathrm{F}} = O\left(\frac{n\tilde{M}_{\boldsymbol{X}}\|\boldsymbol{y}\|}{\lambda_0}\right) \text{ for all } l, p$$

when $m \geq \operatorname{poly}\big(n, \|\boldsymbol{y}\|_2, \lambda_0^{-1}, \ln(1/\delta)\big)$, which contradicts to the definition of t_0 when $m \geq C\Big(n\tilde{M}_{\boldsymbol{X}}\|\boldsymbol{y}\|_2\lambda_0^{-1}\Big)^5$ for some positive absolute constant $C > 0$. ■

We also have a simple corollary:

Corollary 43 *There exists a positive absolute constant* M *and* C *such that when* $m \geq M$*, with probability at least* $1 - \exp\big(-\Omega(m^{5/6})\big)$*,*

$$|f_t^m(x)| \leq C\|\tilde{x}\|, \quad \forall x \in \mathbb{R}^d,\ \forall t \geq 0.$$

Proof Recall that

$$f_t^m(x) = \frac{\sqrt{2}}{2}[\boldsymbol{W}_t^{(L,1)}\boldsymbol{\alpha}_t^{(L,1)}(x) - \boldsymbol{W}_t^{(L,2)}\boldsymbol{\alpha}_t^{(L,2)}(x)].$$

Since with probability at least $1 - \exp\big(-\Omega(m^{5/6})\big)$ we have

$$\left\|\boldsymbol{W}_0^{(l,p)}\right\|_2 \leq O(\sqrt{m}), \quad \left\|\boldsymbol{D}_{x,t}^{(l,p)}\right\|_2 \leq 1, \quad \sup_{t\geq 0}\left\|\boldsymbol{W}_t^{(l,p)} - \boldsymbol{W}_0^{(l,p)}\right\|_2 \leq O(m^{\frac{1}{4}}),$$

the corollary is proved by Lemma 31 and Lemma 42 ■

A.4 Hölder continuity of K^{NT}

For convenience, let us first introduce the following definition of Hölder spaces (Adams and Fournier, 2003). For an open set $\Omega \subset \mathbb{R}^p$ and a real number $\alpha \in [0,1]$, let us define a semi-norm for $f : \Omega \to \mathbb{R}$ by

$$|f|_{0,\alpha} = \sup_{x,y\in\Omega,\ x\neq y} |f(x) - f(y)|/\|x-y\|^{\alpha}$$

and define the Hölder space by $C^{0,\alpha}(\Omega) = \Big\{f \in C(\Omega) : |f|_{0,\alpha} < \infty\Big\}$, which is equipped with norm $\|f\|_{C^{0,\alpha}(\Omega)} = \sup_{x\in\Omega}|f(x)| + |f|_{\alpha}$. Then it is easy to show that

(*a*) $C^{0,\alpha}(\Omega) \subseteq C^{0,\beta}(\Omega)$ if $\beta \leq \alpha$;

(*b*) if $f, g \in C^{0,\alpha}(\Omega)$, then $f+g,\ fg \in C^{\alpha}(\Omega)$;

(*c*) if $f \in C^{0,\alpha}(\Omega_1)$ and $g \in C^{0,\beta}(\Omega_2)$ with $\operatorname{Ran} g \subseteq \Omega_1$, then $f\circ g \in C^{0,\alpha\beta}(\Omega_2)$.

Consequently, using the formula (11), we can show the following proposition:

Proposition 44 *We have $K^{\mathrm{NT}} \in C^{0,s}(\tilde{B}_R \times \tilde{B}_R)$ with $s = 2^{-L}$. Particularly, there is some absolute constant $C > 0$ such that for any $x, x', z, z' \in \tilde{B}_R$,*

$$\left|K^{\mathrm{NT}}(x,x') - K^{\mathrm{NT}}(z,z')\right| \leq CR^2\left\|(x,x') - (z,z')\right\|^s. \tag{45}$$

Proof Let us recall (11). Since K^{NT} is symmetric, by triangle inequality it suffices to prove that $K^{\mathrm{NT}}(x_0,\cdot) \in C^{0,s}(\tilde{B}_R)$ with $\left|K^{\mathrm{NT}}(x_0,\cdot)\right|_{0,s}$ bounded by a constant independent of x_0. Now, the latter is proven by

(*a*) $x \mapsto \bar{u} = \langle \tilde{x}/\|\tilde{x}\|, \tilde{x}_0/\|\tilde{x}_0\|\rangle \in C^{0,1}(\tilde{B}_R)$, where the bound of the Hölder norm is independent of x_0;

(*b*) as functions of $\bar{u}$, both $\sqrt{1-\bar{u}^2}$ and $\arccos\bar{u}$ belong to $C^{0,1/2}([-1,1])$, and thus $\kappa_0, \kappa_1 \in C^{0,1/2}([-1,1])$;

(*c*) the expression of NTK together with the properties of Hölder functions.

■

A.5 The kernel uniform convergence

Proposition 45 (Kernel uniform convergence) *Denote $B_r = \{x \in \mathbb{R}^d : 1 \leq \|\tilde{x}\| \leq r\}$. There exists a polynomial* $\operatorname{poly}(\cdot)$ *such that for any $\delta \in (0,1)$, as long as $m \geq$* $\operatorname{poly}\Big(n, \tilde{M}_{\boldsymbol{X}}, \lambda_0^{-1}, \|\boldsymbol{y}\|, \ln(1/\delta), k\Big)$ *and $m \geq r^k$, with probability at least $1-\delta$ we have*

$$\sup_{t\geq 0}\sup_{x,x'\in B_r}\left|K_t(x,x') - K^{\mathrm{NT}}(x,x')\right| \leq O\Big(r^2 m^{-\frac{1}{12}}\sqrt{\ln m}\Big).$$

Proof First, by Lemma 42, we know that there exists a polynomial $\mathrm{poly}_1(\cdot)$ such that for any $\delta \in (0,1)$, when $m \geq \mathrm{poly}_1\Big(n, \tilde{M}_{\boldsymbol{X}}, \|\boldsymbol{y}\|, \lambda_0^{-1}, \ln(1/\delta)\Big)$, then with probability at least $1-\delta/2$, for all $p \in \{1,2\}$ and $l \in \{0,1,\cdots,L\}$, we have

$$\sup_{t\geq 0}\left\|\boldsymbol{W}_t^{(l,p)} - \boldsymbol{W}_0^{(l,p)}\right\|_{\mathrm{F}} = O(m^{1/4}).$$

Next, we condition on this event happens.

Since $B_r \subset \mathbb{R}^d$ is bounded, for any $\varepsilon > 0$ we have an ε-net $\mathcal{N}_\varepsilon$ (with respect to $\|\cdot\|_2$) of $\tilde{B}_R$ such that the cardinality $|\mathcal{N}_\varepsilon| = O(r^d\varepsilon^{-d})$ (Vershynin, 2018, Section 4.2). Specifically, we choose $\varepsilon = m^{-2^L}$ and thus $\ln|\mathcal{N}_\varepsilon| = O(\ln m)$ if $m \geq r^k$ and $m \geq \mathrm{poly}_3(k)$. Denote by

$$B_{z,z'}(\varepsilon) = \Big\{(x,x') : \|x - z\| \leq \varepsilon,\ \left\|x' - z'\right\| \leq \varepsilon,\ x, x' \in \tilde{B}_R\Big\}.$$

Then, fixing $z, z' \in \mathcal{N}_\varepsilon$, for any $(x,x') \in B_{z,z'}(\varepsilon)$, we have

$$\left|K_t(x,x') - K^{\mathrm{NT}}(x,x')\right| \leq \left|K_t(x,x') - K^{\mathrm{NT}}(z,z')\right| + \left|K^{\mathrm{NT}}(z,z') - K^{\mathrm{NT}}(x,x')\right|.$$

Then, noticing that $K_t = (K_t^{(1)} + K_t^{(2)})/2$, we control the two terms on the right hand side by Corollary 35 and Proposition 44 respectively, deriving that with probability at least $1 - \delta/\left(2|\mathcal{N}_\varepsilon|^2\right)$, for all $t \geq 0$, we have

$$\sup_{(x,x')\in B_{z,z'}(\varepsilon)} |K_t(x,x') - K^{\mathrm{NT}}(z,z')| = O\Big(r^2 m^{-1/12}\sqrt{\ln m}\Big),$$
$$\left|K^{\mathrm{NT}}(z,z') - K^{\mathrm{NT}}(x,x')\right| = O(r^2\varepsilon^{2^{-L}}) = O(r^2 m^{-1}),$$

if $m \geq C_1 \ln\!\big(C_2|\mathcal{N}_\varepsilon|^2/\delta\big)^5$ for some positive absolute constants $C_1 > 0$ and $C_2 \geq 1$.

And there also exists a polynomial $\mathrm{poly}_2(\cdot)$ such that when $m \geq \mathrm{poly}_2(\ln(1/\delta))$, we have $m \geq C_1 \ln\!\big(C_2|\mathcal{N}_\varepsilon|^2/\delta\big)^5$, since $\ln|\mathcal{N}_\varepsilon| = O(\ln m)$. Combining these two terms, we have

$$\sup_{t\geq 0}\ \sup_{(x,x')\in B_{z,z'}(\varepsilon)} \left|K_t(x,x') - K^{\mathrm{NT}}(x,x')\right| = O\Big(r^2 m^{-1/12}\sqrt{\ln m}\Big)$$

if $m \geq \mathrm{poly}_2(\ln(1/\delta))$.

Combining all of the above results and applying the union bound for all pair $\boldsymbol{z}, \boldsymbol{z}' \in \mathcal{N}_\varepsilon$, with probability at least $1-\delta$, we have

$$\sup_{t\geq 0}\ \sup_{x,x'\in B_r} \left|K_t(x,x') - K^{\mathrm{NT}}(x,x')\right| = O\Big(r^2 m^{-1/12}\sqrt{\ln m}\Big)$$

if $m \geq \mathrm{poly}_1\Big(n, \tilde{M}_{\boldsymbol{X}}, \|\boldsymbol{y}\|_2, \lambda_0^{-1}, \ln(1/\delta)\Big) + \mathrm{poly}_2(\ln(1/\delta)) + \mathrm{poly}_3(k)$. ■

	$d=3$			$d=4$			$d=5$		
Distribution	$L=2$	$L=3$	$L=4$	$L=2$	$L=3$	$L=4$	$L=2$	$L=3$	$L=4$
$U(-1,1)$	1.31	1.31	1.30	1.25	1.24	1.22	1.23	1.20	1.17
$U(0,1)$	1.33	1.33	1.32	1.26	1.26	1.25	1.14	1.13	1.12
Triangular	1.34	1.33	1.32	1.21	1.23	1.22	1.22	1.16	1.13
Clipped normal	1.28	1.30	1.28	1.26	1.24	1.21	1.11	1.09	1.06

Table 1: Eigenvalue decay rates of NTK, where each entry of x is drawn independently from multiple distributions. Triangular: $p(x) = 1 + x$, $x \in [-1, 0]$, $p(x) = 1 - x$, $x \in [0, 1]$; Clipped normal: standard normal clipped into $(-10, 10)$.

Appendix B. Auxiliary Results on the NTK

B.1 Positive definiteness

The following proposition is an elementary result on the power series expansion of the arc-cosine kernels.

Proposition 46 *We have the following power series expansion for κ_0 and κ_1 in* (10)*:*

$$\kappa_0(u) = \frac{1}{2} + \frac{1}{\pi}\sum_{r=0}^{\infty} \frac{\left(\frac{1}{2}\right)_r}{(2r+1)r!} u^{2r+1}, \quad \kappa_1(u) = \frac{1}{\pi} + \frac{1}{2}u + \frac{1}{\pi}\sum_{r=1}^{\infty} \frac{\left(\frac{1}{2}\right)_{r-1}}{2(2r-1)r!} u^{2r}, \tag{46}$$

where $(a)_p \coloneqq a(a+1)\dots(a+p-1)$ represents the rising factorial and both series converge absolutely for $u \in [-1, 1]$.

Lemma 47 (Lemma B.1 in Lai et al. (2023)) *Let $f : [-1, 1] \to \mathbb{R}$ be a continuous function with the expansion $f(u) = \sum_{n=0}^{\infty} a_n u^n, \quad u \in [-1, 1]$ and $k(x, y) = f(\langle x, y\rangle)$ be the associated inner-product kernel on $\mathbb{S}^d$. If $a_n \geq 0$ for all $n \geq 0$ and there are infinitely many $a_n > 0$, then k is positive definite on $\mathbb{S}^d_+$.*

Proof [of Proposition 9] Following the proof of Theorem 10, we introduce the transformation Φ and the homogeneous NTK K_0^{NT}. Plugging Proposition 46 into (13), by Lemma 47 we can show that K_0^{NT} is strictly positive definite on $\mathbb{S}^d_+ = \Phi(\mathbb{R}^d)$. Consequently, the positive definiteness of K^{NT} follows from the fact that Φ is bijective and $\|\tilde{x}\| \geq 1$. ■

B.2 Numerical experiments

We provide some numerical experiments on the eigenvalue decay of the neural tangent kernel. We approximate the eigenvalue λ_i by the eigenvalue $\lambda_i(K)$ of the regularized sample kernel matrix for n much larger than i. Then, we estimate eigenvalue decay by fitting a log least-square $\ln \lambda_i = r \ln i + b$. We skip the first several eigenvalues since they do not reflect the asymptotic decay. We report the results in Figure 1 on page 37 and Table 1 on page 36. The results match our theoretical prediction and justify our theory.

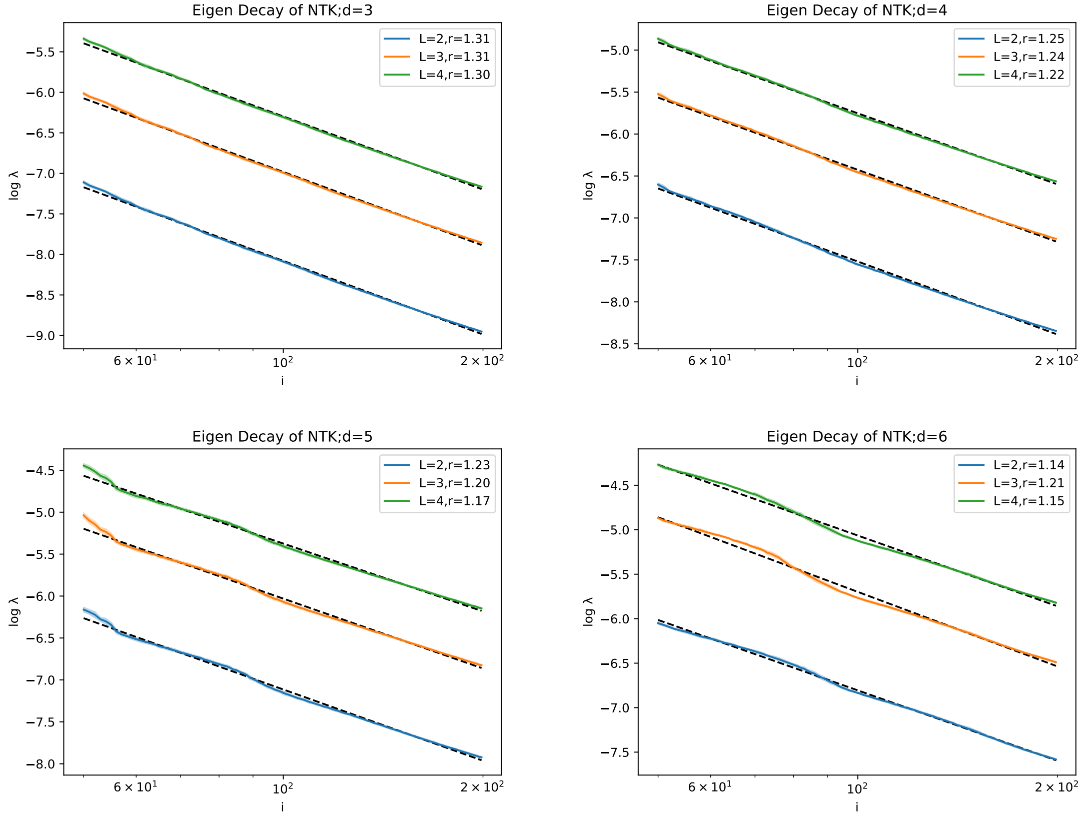

Figure 1: Eigenvalue decay of NTK under uniform distribution on $[-1,1]^d$, where i is selected in $[50,200]$ and $n=1000$. The dashed black line represents the log least-square fit and the decay rates r are reported.

Appendix C. Omitted proofs

Proof of Proposition 13 Theorem 10 shows the eigenvalue decay rate of K^{NT} is $(d+1)/d$. Therefore, the results in Lin et al. (2018) implies the lower rate and that the gradient flow of NTK satisfies

$$\left\| \hat{f}^{\mathrm{NTK}}_{t_{\mathrm{op}}} - f^* \right\|_{L^2} \leq C \left(\ln \frac{6}{\delta} \right) n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}} \tag{47}$$

with probability at least $1-\delta$, where $\beta = (d+1)/d$.

On the other hand, since μ is sub-Gaussian, $\sum_{i=1}^n \|x_i\|_2 \leq Cn^2$ for probability at least $1-\delta$ if $n \geq \mathrm{poly}(\ln(1/\delta))$. From $y_i = f^*(x_i) + \varepsilon_i$, $f^* \in L^\infty$ and ε_i is sub-Gaussian, we have $\|\boldsymbol{y}\| \leq 2Cn$ for probability at least $1-\delta$ as long as $n \geq \mathrm{poly}(\ln(1/\delta))$. Then, taking $k = 1/48$ and $r = m^k$ in Lemma 12, when $m \geq \mathrm{poly}(n, \lambda_0^{-1}, \ln(1/\delta))$, with probability $1-3\delta$ we have

$$\sup_{t\geq 0} \sup_{x\in B_r} \left| \hat{f}^{\mathrm{NTK}}_t(x) - \hat{f}^{\mathrm{NN}}_t(x) \right| \leq C m^{-\frac{1}{24}} \sqrt{\ln m} \leq C n^{-1}$$

as long as we take a larger power of n in the requirement of m. Consequently,

$$\left\|(\hat{f}^{\mathrm{NN}}_{t_{\mathrm{op}}} - f^*)\mathbf{1}_{B_r}\right\|_{L^2} \leq \left\|(\hat{f}^{\mathrm{NN}}_{t_{\mathrm{op}}} - \hat{f}^{\mathrm{NTK}}_{t_{\mathrm{op}}})\mathbf{1}_{B_r}\right\|_{L^2} + \left\|(\hat{f}^{\mathrm{NTK}}_{t_{\mathrm{op}}} - f^*)\mathbf{1}_{B_r}\right\|_{L^2} \leq \frac{1}{n} + C\left(\ln\frac{12}{\delta}\right) n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}}.$$

Now,

$$\left\|\hat{f}^{\mathrm{NN}}_{t_{\mathrm{op}}} - f^*\right\|_{L^2} \leq \left\|(\hat{f}^{\mathrm{NN}}_{t_{\mathrm{op}}} - f^*)\mathbf{1}_{B_r}\right\|_{L^2} + \left\|\hat{f}^{\mathrm{NN}}_{t_{\mathrm{op}}}\mathbf{1}_{B_r^{\complement}}\right\|_{L^2} + \left\|f^*\mathbf{1}_{B_r^{\complement}}\right\|_{L^2},$$

where the first term is already bounded. Noticing that μ is sub-Gaussian and $r = m^{1/48}$, by Corollary 43 we bound the second term by

$$\left\|\hat{f}^{\mathrm{NN}}_{t_{\mathrm{op}}}\mathbf{1}_{B_r^{\complement}}\right\|_{L^2} \leq \left\|Cm\|\tilde{x}\|\mathbf{1}_{B_r^{\complement}}\right\|_{L^2} \leq Cm^{-1} \leq Cn^{-1}$$

and the third term by

$$\left\|f^*\mathbf{1}_{B_r^{\complement}}\right\|_{L^2} \leq \|f^*\|_{L^\infty}\mu(B_r^{\complement})^{1/2} \leq Cn^{-1}.$$

Plugging these bounds into the above inequality, we finish the proof.

C.1 Choosing stopping time with cross validation

Before proving Proposition 16, we introduce a modified version of Caponnetto and Yao (2010, Theorem 3).

Proposition 48 *Let $\delta \in (0,1)$ and $\varepsilon > 0$. Suppose $\hat{f}_t$ is a family of estimators indexed by $t \in T_n$ such that with probability at least $1-\delta$, it holds that $\left\|\hat{f}_{t_n} - f^*\right\|_{L^2} \leq \varepsilon$ for some $t_n \in T_n$. Then, by choosing $\hat{t}_{\mathrm{cv}}$ by cross validation* (18)*, with probability at least $1-2\delta$, it holds that*

$$\left\|\hat{f}_{\hat{t}_{\mathrm{cv}}} - f^*\right\|_{L^2} \leq 2\varepsilon + \left(\frac{160M^2}{\tilde{n}}\ln\frac{2|T_n|}{\delta}\right)^{1/2}. \tag{48}$$

Proof of Proposition 16 The choice of T_n guarantees that there is $t_n \in T_n$ such that $t_{\mathrm{op}} \leq t_n \leq Qt_{\mathrm{op}}$ for $t_{\mathrm{op}} = n^{(d+1)/[s(d+1)+d]}$ and that $|T_n| \leq \log_Q n + 1 \leq C\ln n$. Then, by Proposition 13 we know that

$$\left\|\hat{f}_{t_n} - f^*\right\|_{L^2} \leq C\left(\ln\frac{12}{\delta}\right) n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}}.$$

Consequently, by Proposition 48, we conclude that

$$\left\|\hat{f}_{\hat{t}_{\mathrm{cv}}} - f^*\right\|_{L^2} \leq C\left(\ln\frac{12}{\delta}\right) n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}} + \left(\frac{160M^2}{c_{\mathrm{v}}n}\ln\frac{C\ln n}{\delta}\right)^{1/2} \leq C\left(\ln\frac{12}{\delta}\right) n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}}$$

as long as n is sufficiently large.

Appendix D. Auxiliary Results

D.1 Self-adjoint compact operator

For a self-adjoint compact positive operator A on a Hilbert space, we denote by $\lambda_n(A)$ the n-th largest eigenvalue of A. The following minimax principle is a classic result in functional analysis.

Lemma 49 (Minimax principle) *Let A be a self-adjoint compact positive operator. Then*

$$\lambda_n(A) = \sup_{\substack{V \subseteq H \\ \dim V = n}} \inf_{\substack{x \in V \\ \|x\|=1}} \langle Ax, x \rangle .$$

Lemma 50 (Weyl's inequality for operators) *Let A, B be self-adjoint compact positive operators. Then*

$$\lambda_{i+j-1}(A+B) \leq \lambda_i(A) + \lambda_j(B), \quad i,j \geq 1. \tag{49}$$

Lemma 51 *Let $A_1, \dots, A_k$ be self-adjoint and compact. Suppose $\varepsilon = \sum_{i=1}^k \varepsilon_i$. Denote by $N^{\pm}(\varepsilon, T)$ the count of eigenvalues of T that is strictly greater(smaller) than ε $(-\varepsilon)$. We have*

$$N^{\pm}(\varepsilon, \sum_{i=1}^k A_i) \leq \sum_{i=1}^k N^{\pm}(\varepsilon_i, A_i), \tag{50}$$

Proof Widom (1963, Lemma 5). ■

D.2 Subdomains on the sphere

Let $d(x,y) = \arccos \langle x, y \rangle$ be the geodesic distance on the sphere. The first proposition deals with the "overlapping area" after rotation of two subdomains.

Proposition 52 *Let $\Omega_1, \Omega_2 \subset \mathbb{S}^d$ be two disjoint domains with piecewise smooth boundary. Fix two points $e, y \in \mathbb{S}^d$. Suppose that for any $x \in \mathbb{S}^d$, $R_{e,x}$ is an isometric transformation such that $R_{e,x}e = x$. Then, there exists some M such that*

$$|\{x \in \Omega_1 : R_{e,x}y \in \Omega_2\}| \leq M d(y,e) = M \arccos \langle y, e \rangle , \tag{51}$$

Proof Let $r = d(y,e)$. Since $R_{e,x}$ is isometric, we have

$$r = d(y,e) = d(R_{e,x}y, R_{e,x}e) = d(R_{e,x}y, x).$$

Therefore, if $r < d(x, \partial\Omega_1)$, noticing that $x \in \Omega_1$, we have $R_{e,x}y \in \overline{B}_x(r) \subset \Omega_1$, and hence $R_{e,x}y \notin \Omega_2$. Therefore,

$$\{x \in \Omega_1 : R_{e,x}y \in \Omega_2\} \subseteq \{x \in \Omega_1 : d(x, \partial\Omega_1) \leq r\} .$$

The latter is a tube of radius r of $\partial\Omega_1$ as defined in Weyl (1939). Moreover, since $\partial\Omega_1$ is piecewise smooth, the results in Weyl (1939) show that there is some constant M such that

$$\left|\{x \in \Omega_1 : d(x, \partial\Omega_1) \leq r\}\right| \leq Mr,$$

giving the desired estimation. ■

This following proposition provides a decomposition of the sphere.

Proposition 53 *There exists a sequence of subdomains $U_0, V_0, U_1, V_1, \cdots \subseteq \mathbb{S}^d$ with piecewise smooth boundary such that*

(1) $U_0 = \mathbb{S}^d$;

(2) There are disjoint isometric copies $V_{i,1}, \ldots, V_{i,n_i}$ of V_i such that $U_i = \bigcup_{j=1}^{n_i} V_{i,n_i} \cup Z$, where Z is a null-set;

(3) $V_i \subseteq U_{i+1}$ after some isometric transformation;

(4) diam $V_i \to 0$.

Proof Let us denote by $S_{\boldsymbol{p},r} = \left\{x \in \mathbb{S}^d \mid \langle x, \boldsymbol{p}\rangle > \cos r\right\}$ the spherical cap centered at p with radius r, and $S_r = S_{e_{d+1},r}$, where e_{d+1} is the unit vector for the last coordinate.

First, let $V_0 = \left\{x = (x_1, \ldots, x_{d+1}) \in \mathbb{S}^d \;\middle|\; x_i > 0,\ i = 1, \ldots, d+1\right\}$. Then, by reflection, there are 2^{d+1} isometric copies of Ω such that their disjoint union is whole sphere minus equators, which is a null set.

To proceed, taking $\boldsymbol{p} = \frac{1}{\sqrt{d+1}}(1, \ldots, 1)$, for any points $x \in V_0$, we have

$$\langle x, \boldsymbol{p}\rangle = \frac{1}{\sqrt{d+1}}\left(x_1 + \cdots + x_{d+1}\right) \geq \frac{1}{\sqrt{d+1}}.$$

Therefore, $V_0 \subset S_{\boldsymbol{p},r_1}$ for $r_1 = \arccos\frac{1}{\sqrt{d+1}} < \frac{\pi}{2}$ and we may take $U_1 = S_{r_1}$.

Now suppose we have $U_i = S_{r_i}$ with $r_i < \frac{\pi}{2}$. Using polar coordinate, we have the parametrization

$$x_1 = \sin\theta_1 \cdots \sin\theta_d, \quad x_2 = \sin\theta_1 \cdots \cos\theta_d, \ldots, x_d = \sin\theta_1 \cos\theta_2, \quad x_{d+1} = \cos\theta_1, \tag{52}$$

where $\theta_d \in [0, 2\pi]$ and $\theta_j \in [0, \pi]$, $j = 1, \ldots, d-1$. Then, the spherical cap is given by $S_r = \{x \mid \theta_1 < r\}$. Let us consider the slice

$$V_i = \left\{x \in \mathbb{S}^d \;\middle|\; \theta_1 < r, \theta_j \in (0, \frac{\pi}{2}),\ j = 2, \ldots, d-1,\ \theta_d \in \left(-\frac{\pi}{4}, \frac{\pi}{4}\right)\right\} \subset S_r,$$

Then, by rotation over θ_d and reflection over θ_j, $j = 2, \ldots, d-1$ we can find 2^d isometric copies of V_i such that their disjoint union is only different with S_{r_i} by the union of the boundaries of V_i's, which a null-set.

Now, we find some $r_{i+1} < r_i$ such that $V_i \subset U_{i+1} = S_{\boldsymbol{p},r_{i+1}}$. Let us take the point $\boldsymbol{p} = (p_1, \ldots, p_{d+1})$ by

$$p_{d+1} = \cos\eta, \quad p_d = \cdots = p_2 = \frac{1}{\sqrt{d-1}}\sin\eta, \quad p_1 = 0,$$

where $\eta \in (0, r_i)$ will be determined later. Suppose now $x \in V_i$ is given by (52). We obtain that $\langle \boldsymbol{p}, x \rangle = \cos\eta\cos\theta_1 + \frac{\sin\eta}{\sqrt{d-1}}(x_d + \cdots + x_2)$. Noticing that $\theta_j \in [0, \frac{\pi}{2}]$ and $|\theta_d| \geq \frac{\pi}{4}$, we have $x_d + \cdots + x_2 \geq \sin\theta_1 \cos\theta_d \geq \frac{1}{\sqrt{2}}\sin\theta_1$. Therefore,

$$\begin{aligned}
\langle \boldsymbol{p}, x \rangle &\geq \cos\eta\cos\theta_1 + a\sin\eta\sin\theta_1, \quad a = \frac{1}{\sqrt{2(d-1)}}, \\
&\geq \min\left(\cos\eta,\ \cos\eta\cos r_i + a\sin\eta\sin r_i\right), \quad \text{since } \theta_1 \in (0, r_i), \\
&= \begin{cases} \cos\eta, & \tan\eta > \frac{1-\cos r_i}{a\sin r_i}, \\ \cos r_i\cos\eta + (a\sin r_i)\sin\eta, & \text{otherwise.} \end{cases}
\end{aligned}$$

We know that the second term is maximized by $\tan\eta_0 = a\tan r_i$ and

$$\cos r_i \cos\eta_0 + a\sin r_i \sin\eta_0 = \sqrt{\cos^2 r_i + a^2\sin^2 r_i}.$$

On one hand, if $\tan\eta_0 \leq \frac{1-\cos r_i}{a\sin r_i}$, we take $\eta = \eta_0$ and the minimum is taken by the second term, so $\langle \boldsymbol{p}, x \rangle \geq \sqrt{\cos^2 r_i + a^2\sin^2 r_i}$ and $V_i \subset S_{\boldsymbol{p}, r_{i+1}}$ for $r_{i+1} = \arccos\sqrt{\cos^2 r_i + a^2\sin^2 r_i}$. In this case, we have

$$\sin^2 r_{i+1} = 1 - (\cos^2 r_i + a^2\sin^2 r_i) = (1-a^2)\sin^2 r_i. \tag{53}$$

On the other hand, if $\tan\eta_0 = a\tan r_i > \frac{1-\cos r_i}{a\sin r_i}$, we take $\eta = \arctan\frac{1-\cos r_i}{a\sin r_i}$. Then, the minimum is taken by the first term and $\langle \boldsymbol{p}, x \rangle \geq \cos\eta$, implying $V \subset S_{\boldsymbol{p}, r_{i+1}}$ for $r_{i+1} = \eta$. In this case, we have

$$\tan r_{i+1} = \tan\eta = \frac{1-\cos r_i}{a\sin r_i} < a\tan r_i. \tag{54}$$

Considering both cases (53),(54) and noticing that $a = \frac{1}{\sqrt{2(d-1)}} \in (0,1)$, we conclude that $r_{i+1} < r_i$ and $r_i \to 0$. ∎

D.3 Cesaro sum

We will use Cesaro sum in our analysis of dot-product kernels. We also refer to Dai and Xu (2013, Section A.4).

Definition 54 *Let $p \geq 0$. The p-Cesaro sum s_n of a sequence a_k is defined by*

$$s_n^p = \frac{1}{A_n^p}\sum_{k=0}^n A_{n-k}^p a_k, \qquad A_k^p := \binom{k+p}{k}. \tag{55}$$

Definition 55 (Difference) *Let $\boldsymbol{a} = (a_k)_{k\geq 0}$ be a sequence. We define the difference operator on sequence by*

$$(\triangle^0 \boldsymbol{a})_k = a_k, \quad (\triangle \boldsymbol{a})_k = a_k - a_{k+1}, \quad \triangle^{p+1}\boldsymbol{a} = \triangle(\triangle^p \boldsymbol{a}). \tag{56}$$

We often write $(\triangle^p \boldsymbol{a})_k = \triangle^p a_k$. It is easy to see that $\triangle^p a_k = \sum_{r=0}^p \binom{k}{r}(-1)^r a_{k+r}$.

Definition 56 (Tail sum) *Let $\boldsymbol{a} = (a_k)_{k\geq 0}$ be a sequence. Assuming all the following summations are absolutely convergent, we define the tail sum operator on sequence by*

$$(S^0\boldsymbol{a})_k = a_k, \quad (S\boldsymbol{a})_k = \sum_{r\geq k} a_r, \quad S^{p+1}\boldsymbol{a} = S(S^p\boldsymbol{a}). \tag{57}$$

We often write $(S^p\boldsymbol{a})_k = S^p a_k$.

The following is an elementary proposition about the connection between tail sum and difference.

Proposition 57 *We have (a) $S^p a_n = \sum_{k=0}^{\infty} A_k^{p-1} a_{n+k}$; (b) $S\triangle a_n = \triangle S a_n = a_n$; (c) Consequently, $\sum_{k=0}^{\infty} A_k^p \triangle^{p+1} a_{n+k} = (S^{p+1}\triangle^{p+1} a)_n = a_n$.*

We have the following summation by parts formula, see also Dai and Xu (2013, (A.4.8)).

Proposition 58 (Summation by parts) *Let a_k, b_k be two sequence and $p \in \mathbb{N}$. Then,*

$$\sum_{k=0}^{\infty} a_k b_k = \sum_{k=0}^{\infty} \triangle^{p+1} b_k \sum_{j=0}^{k} A_{k-j}^p a_j = \sum_{k=0}^{\infty} \triangle^{p+1} b_k A_k^p s_k^p,$$

where s_k^p is the p-Cesaro mean of a_k.

For a function $f : [a, b] \to \mathbb{R}$, we can define similarly $\triangle f(x) = f(x) - f(x+1)$ and $\triangle^{p+1} f(x) = \triangle^p f(x) - \triangle^p f(x+1)$. The following elementary lemma provides a connection between the difference and the derivative of the function.

Lemma 59 *Let $p \in \mathbb{N}$. Suppose $f \in C^p([a, a+p])$, then*

$$\triangle^p f(a) = (-1)^p \int_{[0,1]^p} f^{(p)}(a + t_1 + \cdots + t_p) \mathrm{d}t_1 \cdots \mathrm{d}t_p. \tag{58}$$

Proposition 60 *Suppose that $\mu_k = c_0(k+1)^{-\beta},\ k \geq 0$. Letting $(\beta)_p := \beta(\beta+1)\cdots(\beta+p-1)$, then*

$$0 < \triangle^p \mu_k \leq c_0(\beta)_p (k+1)^{-(\beta+p)}.$$

Proof Apply the previous lemma with $f(x) = c(x+1)^{-\beta}$ and $f^{(p)}(x) = (-1)^p c_0(\beta)_p (x+1)^{-(\beta+p)}$. ■

Lemma 61 *Let $\boldsymbol{\mu} = (\mu_k)_{k\geq 0}$ be a sequence such that $\triangle^p \mu_k \geq 0,\ \forall k \geq 0$ for some $p \geq 0$. Given $N \geq 0$, we can construct a left extrapolation sequence $(\tilde{\mu}_k)_{k\geq 0}$ such that*

(1) $\tilde{\mu}_k = \mu_k$ for $k \geq N$ and $\tilde{\mu}_k \leq \mu_k$ for $k < N$;

(2) $\triangle^p \tilde{\mu}_k \geq 0$, $\forall k \geq 0$;

(3) Let $\bar{\mu}_k = \mu_k - \tilde{\mu}_k$ be the residual sequence, then $\triangle^p \bar{\mu}_k \geq 0$, $\forall k \geq 0$;

(4) The leading term satisfies

$$\tilde{\mu}_0 = \mathcal{L}_N^p \boldsymbol{\mu} := \sum_{l=0}^{p-1} A_N^l \triangle^l \mu_N.$$

We remark that the $\mathcal{L}_N^p$ is in the same form as the LHS of (8) *in Condition* 6*.*

Proof We define $\tilde{\mu}_k$ recursively by its p-differences. Let $\triangle^p \tilde{\mu}_k = 0$ for $k < N$ and $\triangle^p \tilde{\mu}_k = \triangle^p \mu_k$ for $k \geq N$. Then, summing up the terms iteratively yields (1), (2) and also the recursive formula:

$$\triangle^{p-s} \tilde{\mu}_{N-r} = \sum_{l=0}^{s-1} A_r^l \triangle^{p-s+l} \mu_N, \quad s = 1, \dots, p, \quad r = 0, \dots, N-1,$$

which gives (4). The last statement (3) follows from the fact that $\triangle^p \bar{\mu}_k = \triangle^p \mu_k - \triangle^p \tilde{\mu}_k \geq 0$. ■

References

Robert A Adams and John JF Fournier. *Sobolev Spaces*. Elsevier, 2003.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. On the convergence rate of training recurrent neural networks. *Advances in neural information processing systems*, 32, 2019a.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization, June 2019b. URL `http://arxiv.org/abs/1811.03962`.

Ingo Steinwart (auth.) Andreas Christmann. *Support Vector Machines*. Information Science and Statistics. Springer-Verlag New York, New York, NY, 1 edition, 2008. ISBN 0-387-77242-1 0-387-77241-3 978-0-387-77241-7 978-0-387-77242-4. doi: 10.1007/978-0-387-77242-4.

Sanjeev Arora, Simon 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*, pages 322–332. PMLR, 2019a.

Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019b. URL `https://proceedings.neurips.cc/paper/2019/hash/dbc4d84bfcfe2284ba11beffb853a8c4-Abstract.html`.

D. Azevedo and V.A. Menegatto. Sharp estimates for eigenvalues of integral operators generated by dot product kernels on the sphere. *Journal of Approximation Theory*, 177: 57–68, January 2014. ISSN 00219045. doi: 10.1016/j.jat.2013.10.002.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

F. Bauer, S. Pereverzyev, and L. Rosasco. On regularization algorithms in learning theory. *Journal of complexity*, 23(1):52–72, 2007. doi: 10.1016/j.jco.2006.07.001.

Daniel Beaglehole, Mikhail Belkin, and Parthe Pandit. Kernel ridgeless regression is inconsistent in low dimensions, June 2022.

Alberto Bietti and Francis Bach. Deep equals shallow for ReLU networks in kernel regimes. *arXiv preprint arXiv:2009.14397*, 2020.

Alberto Bietti and Julien Mairal. On the inductive bias of neural tangent kernels. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

A. Caponnetto and Y. Yao. Cross-validation based adaptation for regularization operators in learning theory. *Analysis and Applications*, 08:161–183, 2010. doi: 10.1142/S0219530510001564.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007. doi: 10.1007/s10208-006-0196-8.

Lin Chen and Sheng Xu. Deep neural tangent kernel and laplace kernel have the same RKHS. *arXiv preprint arXiv:2009.10683*, 2020.

Youngmin Cho and Lawrence Saul. Kernel methods for deep learning. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009. URL `https://proceedings.neurips.cc/paper/2009/file/5751ec3e9a4feab575962e78e006250d-Paper.pdf`.

Feng Dai and Yuan Xu. *Approximation Theory and Harmonic Analysis on Spheres and Balls*. Springer Monographs in Mathematics. Springer New York, New York, NY, 2013. ISBN 978-1-4614-6659-8 978-1-4614-6660-4. doi: 10.1007/978-1-4614-6660-4.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding, May 2019.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, September 2018. URL `https://openreview.net/forum?id=S1eK3i09YQ`.

Simon S. Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1675–1685. PMLR, May 2019. URL `https://proceedings.mlr.press/v97/du19c.html`.

Zhou Fan and Zhichao Wang. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. *Advances in neural information processing systems*, 33: 7710–7721, 2020.

Simon-Raphael Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research*, 21:205:1–205:38, 2020. URL `https://www.semanticscholar.org/paper/248fb62f75dac19f02f683cecc2bf4929f3fcf6d`.

Spencer Frei, Niladri S. Chatterji, and Peter Bartlett. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. In *Proceedings of Thirty Fifth Conference on Learning Theory*, pages 2668–2703. PMLR, June 2022. URL `https://proceedings.mlr.press/v178/frei22a.html`.

Amnon Geifman, Abhay Yadav, Yoni Kasten, Meirav Galun, David Jacobs, and Basri Ronen. On the similarity between the Laplace and neural tangent kernels. In *Advances in Neural Information Processing Systems*, volume 33, pages 1451–1461, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Tianyang Hu, Wenjia Wang, Cong Lin, and Guang Cheng. Regularization matters: A nonparametric perspective on overparametrized neural network. In *International Conference on Artificial Intelligence and Statistics*, pages 829–837. PMLR, 2021.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf`.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Jianfa Lai, Manyun Xu, Rui Chen, and Qian Lin. Generalization ability of wide neural networks on R, February 2023.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/hash/0d1a9651497a38d8b1c3871c84528bd4-Abstract.html`.

Yicheng Li, Haobo Zhang, and Qian Lin. Kernel interpolation generalizes poorly. *Biometrika*, page asad048, August 2023a. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/asad048.

Yicheng Li, Haobo Zhang, and Qian Lin. On the saturation effect of kernel ridge regression. In *International Conference on Learning Representations*, February 2023b. URL `https://openreview.net/forum?id=tFvr-kYWs_Y`.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. *Advances in neural information processing systems*, 31, 2018.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel ”ridgeless” regression can generalize. *The Annals of Statistics*, 48(3), June 2020. ISSN 0090-5364. doi: 10.1214/19-AOS1849.

Junhong Lin, Alessandro Rudi, L. Rosasco, and V. Cevher. Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces. *Applied and Computational Harmonic Analysis*, 48:868–890, 2018. doi: 10.1016/j.acha.2018.09.009.

Andrea Montanari and Yiqiao Zhong. The interpolation phase transition in neural networks: Memorization and generalization under lazy training. *The Annals of Statistics*, 50(5): 2816–2847, 2022.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, September 2019. URL `https://openreview.net/forum?id=B1g5sA4twr`.

Quynh Nguyen, Marco Mondelli, and Guido F. Montufar. Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep ReLU networks. In *International Conference on Machine Learning*, pages 8119–8129. PMLR, 2021.

Alexander Rakhlin and Xiyu Zhai. Consistency of interpolation with Laplace kernels is a high-dimensional phenomenon, December 2018. URL `http://arxiv.org/abs/1812.11167`.

Basri Ronen, David Jacobs, Yoni Kasten, and Shira Kritchman. The convergence rate of neural networks for learned functions of different frequencies. *Advances in Neural Information Processing Systems*, 32, 2019.

Barry Simon. *Operator Theory*. American Mathematical Society, Providence, Rhode Island, November 2015. ISBN 978-1-4704-1103-9 978-1-4704-2763-4. doi: 10.1090/simon/004.

Ingo Steinwart and C. Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35(3):363–417, 2012. doi: 10.1007/S00365-012-9153-3.

Namjoon Suh, Hyunouk Ko, and Xiaoming Huo. A non-parametric regression viewpoint: Generalization of overparametrized deep ReLU network under noisy observations. In *International Conference on Learning Representations*, May 2022. URL `https://openreview.net/forum?id=bZJbzaj_IlP`.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge university press, 2018. ISBN 1-108-24454-8.

Hermann Weyl. On the volume of tubes. *American Journal of Mathematics*, 61(2):461–472, 1939.

Harold Widom. Asymptotic behavior of the eigenvalues of certain integral equations. *Transactions of the American Mathematical Society*, 109(2):278–295, 1963. ISSN 0002-9947. doi: 10.2307/1993907.