Journal of Machine Learning Research 25 (2024) 1-49 Submitted 5/22; Published 4/24

Spectral Analysis of the Neural Tangent Kernel for Deep Residual Networks

Yuval Belfer* YUVAL.BELFER@WEIZMANN.AC.IL
Department of Math and CS
Weizmann Institute of Science
Rehovot, Israel

Amnon Geifman* AMNON.GEIFMAN@WEIZMANN.AC.IL
Department of Math and CS
Weizmann Institute of Science
Rehovot, Israel

Meirav Galun MEIRAV.GALUN@WEIZMANN.AC.IL
Department of Math and CS
Weizmann Institute of Science
Rehovot, Israel

Ronen Basri RONEN.BASRI@WEIZMANN.AC.IL
Department of Math and CS
Weizmann Institute of Science
Rehovot, Israel

Editor: Moritz Hardt

Abstract

Deep residual network architectures have been shown to achieve superior accuracy over classical feed-forward networks, yet their success is still not fully understood. Focusing on massively over-parameterized, fully connected residual networks with ReLU activation through their respective neural tangent kernels (ResNTK), we provide here a spectral analysis of these kernels. Specifically, we show that, much like NTK for fully connected networks (FC-NTK), for input distributed uniformly on the hypersphere $\mathbb{S}^{d-1}$, the eigenvalues of ResNTK corresponding to their spherical harmonics eigenfunctions decay polynomially with frequency k as k^{-d}. These in turn imply that the set of functions in their Reproducing Kernel Hilbert Space are identical to those of both FC-NTK as well as the standard Laplace kernel. Our spectral analysis allows us to highlight several additional properties of ResNTK, which depend on the choice of a hyper-parameter that balances between the skip and residual connections. Specifically, (1) with no bias, deep ResNTK is significantly biased toward even frequency functions; (2) unlike FC-NTK for deep networks, which is spiky and therefore yields poor generalization, ResNTK is stable and yields small generalization errors. We finally demonstrate these with experiments showing further that these phenomena arise in real networks.

Keywords: Neural Tangent Kernel, Residual Neural Network, Spectral Bias, Deep learning, Kernel Methods

. *Equal contributors

1. Introduction

Deep residual networks (ResNets), first introduced in He et al. (2016a), are to date amongst the most effective network architectures for image understanding as well as other tasks Howard et al. (2019); Radosavovic et al. (2020); Tan et al. (2019); Greenfeld et al. (2019); Siravenha et al. (2019). Residual networks use a sequence of block operations of the form

$$\mathbf{x}_\ell = \mathbf{x}_{\ell-1} + \alpha G(\mathbf{x}_{\ell-1}, \theta_{\ell-1}) \tag{1}$$

in which the input to each block, denoted $x_{\ell-1}$, is added to its output $G(x_{\ell-1}, \theta_{\ell-1})$ (called the *residual*), and their sum is passed to the next block. ($\theta_{\ell-1}$ denote the block parameters.) The scalar hyper-parameter α balances between the residual and skip connections. While He et al. (2016a)'s implementation uses $\alpha = 1$, Huang et al. (2020); Du et al. (2019); Hayou et al. (2021) suggested to set this constant according to $\alpha = L^{-\gamma}$ with $0.5 \leq \gamma \leq 1$ and L denotes the total number of hidden layers in the network. Our analysis in this paper examines the range $0 \leq \gamma \leq 1$.

The introduction of skip connections in ResNets allowed researchers to train networks with hundreds, and even thousands of layers and to achieve unprecedentedly accurate classification results on the competitive ImageNet dataset He et al. (2016a,b). The reasons for the advantage of residual over classical feed-forward architectures are not yet fully understood. Several papers argue that skip connections alleviate the problem of *vanishing gradients*, which is prevalent in classical deep architectures Balduzzi et al. (2017); Veit et al. (2016). Subsequent work showed that ResNets can avoid spurious local minima Liu et al. (2019), while Li et al. (2018) showed, by empirically visualizing the loss landscape, that skip connections make the loss smoother.

In this work, we examine residual networks from the perspective of the neural tangent kernels. As with many existing network models, residual network applications are typically over-parameterized. He et al. (2016a)'s implementation, for example, trains a network with roughly 60M trainable parameters on the 1.2M images of ImageNet. Recent work Jacot et al. (2018) suggested that massively overparameterized neural networks behave similarly to kernel regressors with a family of kernels called *Neural Tangent Kernels* (NTKs). Huang et al. (2020); Tirer et al. (2021) proved that fully connected residual networks of infinite width converge to such kernel, which we here call *ResNTK*, and provided a closed form derivation.

Kernel regression is characterized by the set of functions in the corresponding Reproducing Kernel Hilbert Space (RKHS) and by the norm induced in this space. These in turn are determined by the eigenfunctions and eigenvalues of the respective kernel under some measure, with the decay rate of the eigenvalues playing a particularly important role. (Note that the RKHS structure of a kernel is independent of the data distribution, see remark 4.3 in Kanagawa et al. (2018).) Previous work Hayou et al. (2021) showed that with data distributed uniformly on the hypersphere $\mathbb{S}^{d-1}$, the eigenfunctions of ResNTK are the spherical harmonics. Here we prove that the eigenvalues of ResNTK decay polynomially with frequency k at the rate of k^{-d}, thus characterizing the set of functions in the corresponding RKHS. We conclude that this set of functions is identical to the functions in the RKHS of NTK of classical, fully connected networks (denoted *FC-NTK*) Geifman et al. (2020); Bietti and Bach (2021); Chen and Xu (2020), and, as is implied by this previous work, also to those of the standard Laplace kernel, restricted to $\mathbb{S}^{d-1}$.

Our analysis reveals further connections and differences between ResNTK and FC-NTK. These appear to critically depend on the choice of hyperparameter α, which balances between the residual and skip connections (1). In particular, we prove for $\alpha = L^{-\gamma}$ with $0.5 < \gamma \leq 1$ that when L (the network depth) tends to infinity ResNTK converges *uniformly* to a two-layer FC-NTK in the interval. (A weaker result showing point-wise convergence with $\gamma = 1$ was provided in Huang et al. (2020).) This implies on one hand that in this parameter regime, in contrast to FC-NTK, deep Res-NTK is not spiky, and consequently, as we establish by proving generalization bounds, deep ResNTK can generalize to new data, while deep FC-NTK cannot. On the other hand, we prove that in the same parameter regime deep, bias-free ResNTK is prone to parity imbalance. That is, with deep ResNTK, eigenfunctions of odd frequencies $k \geq 3$ have significantly lower eigenvalues than those of even frequencies, indicating that odd frequencies are difficult to learn. Lastly, with $0 \leq \gamma < 0.5$, there is no significant parity imbalance, but, much like FC-NTK, ResNTK becomes spiky with deep architectures.

We finally provide experiments with real datasets (CIFAR-10 and SVHN) showing the behavior of ResNTK with different settings of α as well as experiments that show the same behavior in real neural networks.

In summary our paper makes the following four contributions.

1. We prove the decay rate of the eigenvalues of ResNTK, thus providing a full characterization of its RKHS.

2. We prove the *uniform* convergence of deep ResNTK to FC-NTK of two layers in the entire interval $\frac{1}{L} \leq \alpha < \frac{1}{\sqrt{L}}$.

3. We use this uniform convergence to show that without bias, ResNTK for deep networks suffers from a parity imbalance. In this, it is significantly inferior to the bias-free FC-NTK.

4. We leverage our spectral characterization to contrast the generalization properties of ResNTK to those of FC-NTK.

2. Previous work

Existing neural network models are typically applied with many more learnable parameters than training data items, yet somewhat counter-intuitively they successfully generalize to unseen data. Attempting to explain this phenomenon, Jacot et al. (2018) showed that infinite width networks whose parameters are initialized sufficiently close to zero behave like kernel regression with novel kernels called Neural Tangent Kernels. Specifically, for an input $\mathbf{x} \in \mathbb{R}^d$ and learnable parameters $\theta \in \mathbb{R}^m$, denote the network by $f(\mathbf{x}, \theta)$, then the corresponding NTK is given by $\mathbb{E}_{\theta \sim \mathcal{P}} \left\langle \frac{\partial f(\mathbf{x}_i, \theta)}{\partial \theta}, \frac{\partial f(\mathbf{x}_j, \theta)}{\partial \theta} \right\rangle$, where $\mathbf{x}_i$ and $\mathbf{x}_j$ is a training pair, and the expectation is over the distribution $\mathcal{P}$ with which θ is initialized (typically the standard normal distribution).

Subsequent work showed that very wide networks of finite width converge to a global minimum Du et al. (2019); Allen-Zhu et al. (2019); Chizat et al. (2019) and further characterized the speed of convergence as a function of the data distribution and the frequency

of the target function Arora et al. (2019a); Basri et al. (2019, 2020). In particular, for data distributed uniformly in the hypersphere $\mathbb{S}^{d-1}$, it was shown that the eigenfunctions of FC-NTK with any depth are the spherical harmonics and the eigenvalues decay at the rate of k^{-d}, where k denotes frequency Bietti and Bach (2021); Chen and Xu (2020) (Basri et al. (2019); Bietti and Mairal (2019); Cao et al. (2019); Su and Yang (2019) show analogous results in the case of FC-NTK for two-layer networks, while Fan and Wang (2020) studied the restricted case in which the input samples are approximately orthogonal). This completely characterizes the set of functions in the RKHS of FC-NTK. Recent work showed that the set of functions in the RKHS of FC-NTK is identical to the respective set of functions of the classical Laplace kernel Geifman et al. (2020); Bietti and Bach (2021); Chen and Xu (2020). Our paper extends these results to NTK of residual networks of any depth.

Understanding the spectrum of a kernel is useful for a number of objectives. It indicates whether a kernel exhibits a frequency bias Cao et al. (2019); Rahaman et al. (2019); Xu et al. (2019), it provides an estimate of the number of gradient descent iterations needed to learn certain target functions Basri et al. (2019, 2020) (accordingly, the number of iterations to learn an eigenfunction of a certain frequency is inversely proportional to the corresponding eigenvalue). Finally, the characterization of the spectrum can be used to estimate the generalization error obtained by using the kernel as a minimum interpolant regressor (ridge-less kernel regression). For example, Liang et al. (2020, 2019); Pagliana et al. (2020) analyzed the bias-variance interplay of minimum norm interpolation with a growing number of samples when the dimension is either fixed or growing at the same rate.

Several recent studies examined the behavior of over-parameterized residual networks. Du et al. (2019); Zhang et al. (2019b) showed that very wide ResNets of finite size converge to their global minima. Huang et al. (2020); Tirer et al. (2021) derived a formula for ResNTK. The formula uses a parameter α that balances between the skip and residual connections. Huang et al. (2020); Du et al. (2019); Hayou et al. (2021) argued that α should be properly scaled with depth. Hayou et al. (2021) showed that the spherical harmonics form the eigenfunctions of ResNTK on the hypersphere $\mathbb{S}^{d-1}$ and proved its universality (for networks with bias) in the infinite depth limit. Tirer et al. (2021)'s analysis further suggested that ResNTK tends to generate smoother functions than FC-NTK. Huang et al. (2020) showed that FC-NTK becomes spiky for deep networks, indicating that learning with these kernels becomes degenerate, while ResNTK (with $\alpha = 1/L$) remains stable with depth.

Our analysis, in contrast, (1) calculates the decay rate of the eigenvalues of ResNTK, thus characterizing its RKHS structure and establishing that the functions in the RKHS of both ResNTK and FC-NTK have the same smoothness properties. (2) It shows a parity imbalance in deep, bias-free ResNTK. (3) It shows that depending on hyperparameters, deep ResNTK may or may not become spiky. Moreover, we use the spectral properties of these kernels to prove generalization bounds for both ResNTK and FC-NTK.

It has been a subject of debate whether results about NTK will help us to understand the behavior of real neural networks, that are not so massively overparameterized. In particular, it was argued that with real NNs, weights do not stay near their initialization, and therefore linear models ("lazy training") do not properly capture their dynamics Chizat et al. (2019); Tachella et al. (2020). Moreover, several recent papers have crafted learning problems that can be solved with real neural networks, but that linear models, including NTK, are unable to solve Daniely and Malach (2020); Ghorbani et al. (2020); Yehudai and Shamir

(2019). Nevertheless, previous studies have shown experimentally that simple FC networks are subject to a similar inductive bias and training dynamics as predicted by analysis of FC-NTK Basri et al. (2019, 2020). Moreover, Arora et al. (2020); Lee et al. (2020) have shown that kernel regression using FC-NTK performs similarly, and a bit better than real FC networks on a large number of machine learning datasets. We further refer the reader to recent findings on this subject in Malach et al. (2021). Our experiments indicate that properties of FC-NTK and ResNTK indeed show up in real networks.

3. Preliminaries

We consider positive definite kernels $\boldsymbol{k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ over inputs $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$. $\boldsymbol{k}$ is called zonal if when $\mathbf{x}, \mathbf{z}$ are restricted to the hypersphere $\mathbb{S}^{d-1}$, $\boldsymbol{k}$ can be expressed as a function of $\mathbf{x}^T\mathbf{z}$. In such case we overload our definition of $\boldsymbol{k}$, defining also $\boldsymbol{k} : [-1, 1] \to \mathbb{R}$ by letting $u = \mathbf{x}^T\mathbf{z}$ and writing $\boldsymbol{k}(\mathbf{x}, \mathbf{z}) = \boldsymbol{k}(u)$. To avoid unnecessary scaling and to allow comparison of different kernels and kernels constructed for networks with different depths, a good practice is to normalize the kernel by a constant multiplicative factor such that $\boldsymbol{k}(1) = 1$. Such normalization does not alter the RKHS of the kernel. The eigenfunctions and eigenvalues derived in this paper are with respect to the uniform measure on the hypersphere $\mathbb{S}^{d-1}$. Note, however, that the resulting RKHS definition is independent of data distribution. The kernels we use in this paper are ResNTK and FC-NTK, denoted respectively by $\boldsymbol{r}$ and $\boldsymbol{k}$, as well as the Laplace kernel (denoted $\boldsymbol{k}_{Lap}$), with superscripts denoting the number of hidden layers, e.g. $\boldsymbol{k}^{(L)}$, i.e., $L = 1$ corresponds to a network with one hidden layer (i.e., a two-layer network). All proofs are deferred to the supplementary material.

3.1 NTK for FC Networks

A fully-connected neural network (also called multi-layer perceptron, MLP) with L hidden layers and m units in each hidden layer is expressed as

$$f(\theta, \mathbf{x}) = \mathbf{v}^T\mathbf{x}_L, \quad \mathbf{x}_\ell = \sqrt{\frac{c_\sigma}{m}}\sigma\left(W^{(l)}\mathbf{x}_{\ell-1}\right), \quad \ell \in [L]$$

and $\mathbf{x}_0 = \mathbf{x}$. The network parameters θ include $W^{(1)}, W^{(2)}, ..., W^{(L)}$, where $W^{(1)} \in \mathbb{R}^{d\times m}$, $W^{(\ell)} \in \mathbb{R}^{m\times m}$ $(2 \le \ell \le L)$, and $\mathbf{v} \in \mathbb{R}^m$. We denote by σ the ReLU activation function and by $c_\sigma = 1/\left(\mathbb{E}_{z\sim\mathcal{N}(0,1)}[\sigma(z)^2]\right) = 2$. The network parameters are initialized randomly with $\mathcal{N}(0, I)$.

Jacot et al. (2018) showed that when the width $m \to \infty$ the network behaves like kernel regression with the neural tangent kernel. Bietti and Mairal (2019) showed that this kernel, denoted for $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$ by $\boldsymbol{k}^{(L)}(\mathbf{x}, \mathbf{z})$, is homogeneous of degree 1 and zonal, so that $\boldsymbol{k}^{(L)}(\mathbf{x}, \mathbf{z}) = \|\mathbf{x}\|\|\mathbf{z}\|\boldsymbol{k}^{(L)}(u)$, where $u = \frac{\mathbf{x}^T\mathbf{z}}{\|\mathbf{x}\|\|\mathbf{z}\|} \in [-1, 1]$. The (normalized) kernel is defined by

$$\boldsymbol{k}^{(L)}(u) = \frac{1}{L+1}\tilde{\boldsymbol{k}}^{(L)}(u)$$

with the recursive formula

$$\begin{aligned} \tilde{\boldsymbol{k}}^{(\ell)}(u) &= \tilde{\boldsymbol{k}}^{(\ell-1)}(u)\kappa_0(\Sigma^{(\ell-1)}(u)) + \Sigma^{(\ell)}(u) \\ \Sigma^{(\ell)}(u) &= \kappa_1(\Sigma^{(\ell-1)}(u)), \quad \ell \in [L]. \end{aligned} \tag{2}$$

It can be ready shown that with this definition, $\boldsymbol{k}^{(L)}(1) = 1$ for all values of L. The functions κ_1, κ_0 are the arc-cosine kernels Cho and Saul (2009), defined as

$$\kappa_0(u) = \frac{1}{\pi}(\pi - acos(u)) \tag{3}$$

$$\kappa_1(u) = \frac{1}{\pi}\left(u \cdot (\pi - acos(u)) + \sqrt{1-u^2}\right), \tag{4}$$

and $\tilde{\boldsymbol{k}}^{(0)}(u) = \Sigma^{(0)}(u) = u$.

3.2 NTK for residual networks

For the definition of a fully connected residual network we follow the formulation of Huang et al. (2020); Tirer et al. (2021).

$$g(\mathbf{x}, \theta) = \mathbf{v}^T \mathbf{x}_L$$
$$\mathbf{x}_\ell = \mathbf{x}_{\ell-1} + \alpha\sqrt{\frac{1}{m}} V_\ell \, \sigma\left(\sqrt{\frac{2}{m}} W_\ell \mathbf{x}_{\ell-1} + \tau \mathbf{b}_\ell\right), \ \ell \in [L],$$

and $\mathbf{x}_0 = \sqrt{\frac{1}{m}} A\mathbf{x}$. The parameters include $A \in \mathbb{R}^{m\times d}$, $V_\ell, W_\ell \in \mathbb{R}^{m\times m}$, $\mathbf{v} \in \mathbb{R}^m$, and $\sigma(\cdot)$ is the ReLU function. α is a constant hyperparameter. Huang et al. (2020); Du et al. (2019); Hayou et al. (2021) suggested to set this constant according to $\alpha = L^{-\gamma}$ with $0.5 \leq \gamma \leq 1$. In contrast, He et al. (2016a)'s implementation uses $\alpha = 1$ (and an additional ReLU function applied to $V_\ell \, \sigma(.)$). Recent work argued that setting α to decay with depth is enforced in practice through suitable small initialization of the residual parameters or by applying normalization blocks Zhang et al. (2019a). Our analysis below examines the range $0 \leq \gamma \leq 1$.

Adopting Huang et al. (2020)'s derivation, we assume that both A and $\mathbf{v}$ are fixed at their initial values and that V_ℓ, W_ℓ, and $\mathbf{b}$ are learned, with all parameters initialized with the standard normal distribution except for the bias terms $\mathbf{b}_\ell$, which are initialized at 0. Let $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$. The respective NTK, denoted $\boldsymbol{r}^{(L)}(\mathbf{x}, \mathbf{z})$, is given by

$$\begin{aligned}\boldsymbol{r}^{(L)}(\mathbf{x}, \mathbf{z}) = C_\tau \sum_{\ell=1}^{L} B_{\ell+1}(\mathbf{x}, \mathbf{z}) \big[& v_{\ell-1}(\mathbf{x}, \mathbf{z}) \kappa_1(u_{\ell-1}(\mathbf{x}, \mathbf{z})) \\ & + (K_{\ell-1}(\mathbf{x}, \mathbf{z}) + \tau^2) \kappa_0(u_{\ell-1}(\mathbf{x}, \mathbf{z}))\big],\end{aligned} \tag{5}$$

where for $\ell \in [L]$ we let

$$\begin{aligned} v_\ell(\mathbf{x}, \mathbf{z}) &= \sqrt{K_\ell(\mathbf{x}, \mathbf{x}) K_\ell(\mathbf{z}, \mathbf{z})}, \quad u_\ell(\mathbf{x}, \mathbf{z}) = \frac{K_\ell(\mathbf{x}, \mathbf{z})}{v_\ell(\mathbf{x}, \mathbf{z})} \\ K_\ell(\mathbf{x}, \mathbf{z}) &= K_{\ell-1}(\mathbf{x}, \mathbf{z}) + \alpha^2 v_{\ell-1}(\mathbf{x}, \mathbf{z}) \kappa_1(u_{\ell-1}) \\ B_\ell(\mathbf{x}, \mathbf{z}) &= B_{\ell+1}(\mathbf{x}, \mathbf{z})[1 + \alpha^2 \kappa_0(u_{\ell-1})] \\ K_0(\mathbf{x}, \mathbf{z}) &= \mathbf{x}^T \mathbf{z}, \quad B_{L+1}(\mathbf{x}, \mathbf{z}) = 1 \\ C_\tau &= \left(2L(1+\alpha^2)^{L-1} + \tau^2 \frac{(1+\alpha^2)^L - 1}{\alpha^2}\right)^{-1}, \end{aligned}$$

and κ_0 and κ_1 are defined in (3)-(4). The expression of C_τ is set so as to obtain $\boldsymbol{r}(1) = 1$. This is essential, as it was argued in Huang et al. (2020) to prevent the kernel from diverging or vanishing as the depth L tends to infinity. We note that with this model, with $L = 1$ ResNTK is equal to FC-NTK, i.e., $\boldsymbol{r}^{(1)} = \boldsymbol{k}^{(1)}$.

4. Analysis of ResNTK

In this section, we analyze properties of the ResNTK kernel. We prove three main results:

1. We prove the decay rate of the eigenvalues of ResNTK and compare its RKHS structure to those of FC-NTK and the Laplace kernel.

2. We show a significant parity imbalance in the bias-free, deep ResNTK.

3. We prove generalization bounds for ResNTK and compare them to those of FC-NTK.

4.1 RKHS Structure of ResNTK

We next characterize the RKHS structure of ResNTK. Hayou et al. (2021) has shown that ResNTK is zonal, and therefore its eigenfunctions under the uniform measure in the hypersphere $\mathbb{S}^{d-1}$ consist of the spherical harmonics. Our main result in this section establishes that the eigenvalues of ResNTK decay with frequency k at the rate of k^{-d}.

Theorem 1 *The eigenvalues λ_k of ResNTK, $\boldsymbol{r}(\mathbf{x}, \mathbf{z})$, for $\mathbf{x}, \mathbf{z} \in \mathbb{S}^{d-1}$ corresponding to a spherical harmonic eigenfunction $Y_{kl}(\mathbf{x})$ with frequency $k \geq 0$ and phase $1 \leq l \leq N(d, k)$ decay when $k \to \infty$ under the uniform measure as*

$$\lambda_k = \begin{cases} \bar{C}(c_1 + c_{-1})k^{-d} \ if \ k \ is \ even \\ \bar{C}(c_1 - c_{-1})k^{-d} \ if \ k \ is \ odd, \end{cases}$$

where $\bar{C}, c_1, c_{-1}$ are constants that depend on α, L, τ.

The proof of this theorem, given in the supplementary material, relies on an analysis of the infinitesimal tendency of ResNTK near ± 1. Specifically, we prove that for input in $\mathbb{S}^{d-1}$ with $\alpha > 0$, for $L \geq 1$,

1. Near $+1$, $\boldsymbol{r}^{(L)}(1 - t) = 1 + c_1 t^{1/2} + o(t^{1/2})$, with

$$c_1 = -C_\tau \frac{\sqrt{2}}{\pi} L(1 + \alpha^2)^{L-2} \left((1 + \alpha^2 L) + \tau^2 (1 + \alpha^2) \right).$$

2. Near -1, $\boldsymbol{r}^{(L)}(-1 + t) = p_{-1}(t) + c_{-1} t^{1/2} + o(t^{1/2})$, with $|c_{-1}| < \max\{\tau^2, 1\} C_\tau \frac{\sqrt{2}}{\pi} (1 + \alpha^2)^{L-1}$, where p_{-1} is a finite degree polynomial.

Consequently,

$$|c_{-1}| < \max\{\tau^2, 1\} C_\tau \frac{\sqrt{2}}{\pi}(1+\alpha^2)^{L-1} \leq$$
$$C_\tau \frac{\sqrt{2}}{\pi}(1+\alpha^2)^{L-1} \max\left\{L\tau^2, \frac{1+\alpha^2 L}{1+\alpha^2}\right\} \leq |c_1|, .$$

Together, these results satisfy the conditions of a theorem proved in Bietti and Bach (2021) from which it follows that the eigenvalues of ResNTK decay at the rate of k^{-d} with coefficients as specified in Theorem 1.

Theorems 1 provides a full characterization of the set of functions in the reproducing kernel Hilbert space of ResNTK, denoted $\mathcal{H}_{\boldsymbol{r}}$, defined in $\mathbb{S}^{d-1}$ as

$$\mathcal{H}_{\boldsymbol{r}} = \left\{ f(\mathbf{x}) = \sum_{\substack{k\geq 0 \\ \lambda_k \neq 0}} \sum_{j=1}^{N(d,k)} a_{kj} Y_{kj}(\mathbf{x}) \;\; \text{s.t.} \; \|f\|_{\mathcal{H}_{\boldsymbol{r}}} < \infty \right\},$$

where λ_k are the eigenvalues of $\boldsymbol{r}$, $N(d,k)$ is the number of Harmonics of frequency k in $\mathbb{R}^d$ and j is the phase. The coefficient a_{kj} is the projection of f onto the Spherical Harmonic $Y_{kj}(\mathbf{x})$ and

$$\|f\|_{\mathcal{H}_{\boldsymbol{r}}} = \sum_{\substack{k\geq 0 \\ \lambda_k \neq 0}} \sum_{j=1}^{N(d,k)} \frac{a_{kj}^2}{\lambda_k}. \tag{6}$$

Theorem 1 holds for data sampled uniformly from $\mathbb{S}^{d-1}$. It can however be easily extended to any radial distribution in $\mathbb{R}^d$, by applying Theorem 5 in Geifman et al. (2020). Specifically, (5) establishes that ResNTK is the sum of homogeneous kernels of order zero and order one and thus satisfies the conditions of that theorem.

As the eigenvalues of ResNTK decay at the same rate as those of both FC-NTK and the standard Laplace kernel Bietti and Bach (2021); Chen and Xu (2020); Geifman et al. (2020), their RKHSs are closely related. This is summarized in the following corollary.

Corollary 2 *Let $\boldsymbol{k}$ and $\boldsymbol{r}$ respectively denote the FC-NTK and ResNTK kernels. Denote by $\mathcal{H}_{\boldsymbol{k}}$ and $\mathcal{H}_{\boldsymbol{r}}$ the set of functions in the RKHS of the FC-NTK $\boldsymbol{k}$ and ResNTK $\boldsymbol{r}$ in $\mathbb{S}^{d-1}$ Then,*

$$\mathcal{H}_{\boldsymbol{k}} = \mathcal{H}_{\boldsymbol{r}} = \mathcal{H}_{\boldsymbol{k}_{Lap}},$$

where for $\mathbf{x}, \mathbf{z} \in \mathbb{S}^{d-1}$, $\boldsymbol{k}_{Lap}$ denotes the standard Laplace kernel defined by

$$\boldsymbol{k}_{Lap}(\mathbf{x}, \mathbf{z}) = e^{-c\|\mathbf{x}-\mathbf{z}\|} = e^{-c\sqrt{2(1-\mathbf{x}^T\mathbf{z})}}. \tag{7}$$

A consequence of Corollary 2 is that the three kernels, ResNTK, FC-NTK, and the Laplace kernel generate functions of the same smoothness properties, i.e., all three RKHSs include functions that have weak derivatives up to order $d/2$ Narcowich et al. (2007). However,

the structure of the RKHSs is not identical; while the set of functions in the three RKHSs are identical, each kernel is associated with a different RKHS norm. In particular, while the eigenvalues decay at the same rate, they are not identical across kernels, or even across different depths for the same kernel, producing different RKHS norms (6). This, in turn, implies that when applied to the same regression problem, the kernels may produce somewhat different outcomes. For example, as we prove below, with deep architectures the bias-free ResNTK will be biased to interpolate functions with even frequencies, while with bias it will be agnostic to parity. Also, Tirer et al. (2021) showed that under a suitable measure, with low values of α ResNTK tends to produce smoother interpolations. A close examination of their experiments however reveals that also with small values of α their interpolations are only piecewise smooth, consistent with the structure of the respective RKHS derived here.

4.2 Parity Imbalance

Theorem 1 indicates that the rate of decay for all frequencies is k^{-d}. However, for the bias-free ResNTK according to Theorem 1, the leading coefficient for the even and odd frequencies, respectively $\bar{C}(c_1 \pm c_{-1})$, differ. In fact, if the hyperparameter α, which relates between the residual and the skip connections, decays sufficiently fast with network depth, then the eigenvalues for the odd frequencies become extremely small compared to those for the even frequencies. This in fact happens when α is chosen according to Huang et al. (2020); Du et al. (2019); Hayou et al. (2021), i.e., when $\alpha = L^{-\gamma}$ with $0.5 < \gamma \leq 1$, see Figure 1(top left).

To prove the parity imbalance, we will need the following theorem, which establishes that with $0.5 < \gamma \leq 1$, the bias-free ResNTK converges *uniformly* to the two-layer, bias-free FC-NTK for which, as is shown in Basri et al. (2019), the eigenvalues corresponding to the odd frequency $k \geq 3$ are zero. Our theorem below extends over a weaker theorem by Huang et al. (2020), which proved point-wise convergence for $\gamma = 1$ in the open interval $\mathbf{x}^T\mathbf{z} \in (-1, 1)$ only.

Theorem 3 *ResNTK $\boldsymbol{r}^{(L)}$ for residual networks with $L \in \mathbb{N}$ layers and the hyperparameter $\alpha = L^{-\gamma}$, $0.5 < \gamma \leq 1$, approaches the 2-layer FC-NTK uniformly in the interval $\mathbf{x}^T\mathbf{z} \in [-1, 1]$, where $\mathbf{x}, \mathbf{z} \in \mathbb{S}^{d-1}$; that is, let $\epsilon > 0$, $\forall L > c(\epsilon, \gamma)$*

$$|\boldsymbol{r}^{(L)}(\mathbf{x}, \mathbf{z}) - \boldsymbol{k}^{(1)}(\mathbf{x}, \mathbf{z})| \leq \epsilon.$$

For the proof, given in the supplementary material, we first prove uniform convergence in the interval $[-1 + \delta, 1 - \delta]$ for all $\delta > 0$. We then extend this to the full range $[-1, 1]$ by utilizing the Taylor expansion of both kernels at the endpoints, relying on our asymptotic analysis for ResNTK.

Based on this theorem we obtain

Corollary 4 *With $\alpha = L^{-\gamma}$, $0.5 < \gamma \leq 1$, and $L \to \infty$ the eigenvalues of the bias-free ResNTK (i.e setting $\tau = 0$) of odd frequencies $k \geq 3$ vanish.*

The proof follows directly from Theorem 3, since for FC-NTK the eigenvalues corresponding to odd frequency functions are zero Basri et al. (2019). This is also evident from the expansion near ± 1. We show in the supplementary material that when $\alpha^2 L \ll 1$,

$c_1 \xrightarrow{\alpha^2 L \to 0} -\frac{1}{\sqrt{2}\pi} = c_{-1}$, indicating (using Bietti and Bach (2021)) that the eigenvalues corresponding to odd frequencies vanish. It is important to note that according to Theorem 1 for every finite L the eigenvalues for the odd frequencies are non-zero and decay asymptotically as k^{-d}. The eigenvalues vanish at the limit when $L \to \infty$.

With finite L and a training set of n samples, the uniform convergence and the Wielandt-Hoffman inequality Golub and Van Loan (1996) implies that the eigenvalues associated with the odd frequencies are at most $O(nL^{1-2\gamma})$. Therefore, ResNTK differs from FC-NTK, for which in all depths except $L = 1$ the eigenvalues of odd and even frequencies have similar values.

Figure 1(top left) shows the eigenvalues of ResNTK for various depth values as a function of frequency. It can be seen that as depth increases the eigenvalues of odd frequencies considerably decrease, compared to those of the even frequencies. We note finally that this parity difference disappears if we choose $\gamma = 0.5$, i.e., $\alpha = 1/\sqrt{L}$, or if we include bias ($\tau > 0$), as can be seen in Figure 1 (top right, bottom left). The same phenomenon is observed with real networks. We used a bias-free residual network with 50 layers and $\alpha = 1/L$ to regress sinusoidal functions for data in $\mathbb{S}^1$. Consistent with our findings, fitting both the odd and even frequency functions took $O(k^2)$ epochs, i.e., the number of epochs is indeed inversely proportional to the corresponding eigenvalue, as indicated in Basri et al. (2019), but fitting the odd frequencies required significantly more time than fitting the even frequency ones. Convergence times are shown in Figure 1 (bottom right).

4.3 Generalization

Our analysis also allows us to determine how sharp ResNTK is near 1. In particular, the expansion of the Laplace kernel (7) near 1, derived in Bietti and Bach (2021), is given by

$$\boldsymbol{k}_{Lap}(1-t) = 1 - c\sqrt{2t} + O(t).$$

Therefore, the coefficient of $t^{1/2}$ indicates how steep a kernel is near 1. Specifically, for the bias-free ResNTK, with $c = \frac{1+\alpha^2 L}{2\pi(1+\alpha^2)}$, where α is the balancing parameter defined in Section 3.2, we have that $\boldsymbol{r}^{(L)}(1-t) - \boldsymbol{k}_{Lap}(1-t) = o(t^{1/2})$ (see supplementary material). Therefore, with $0.5 \leq \gamma \leq 1$ the steepness of ResNTK is bounded, i.e.,

$$c^{\mathrm{RES}}(L) = \frac{(1+\alpha^2 L)}{2\pi(1+\alpha^2)} \xrightarrow{L\to\infty} \begin{cases} \frac{1}{\pi}, & \gamma = 0.5 \\ \frac{1}{2\pi}, & 0.5 < \gamma \leq 1. \end{cases}$$

In contrast, with $0 \leq \gamma < 0.5$, $c^{\mathrm{RES}}(L)$ either grows linearly (if $\gamma = 0$) or sublinearly with L. This is similar to FC-NTK, for which with $c = \frac{L}{2\pi}$, $\boldsymbol{k}^{(L)}(1-t) - \boldsymbol{k}_{Lap}(1-t) = o(t^{1/2})$, implying that deep FC-NTK becomes steeper near 1. This is consistent with Huang et al. (2020); Xiao et al. (2020) who proved that, except near $u = \mathbf{x}^T\mathbf{z} = 1$, as the depth L tends to infinity FC-NTK approaches the constant 0.25. Therefore, with deep architectures, FC-NTK forms a spike. Figure 2 shows the shape of both FC-NTK and ResNTK for three choices of network depths.

Indeed, the instability of FC-NTK and of ResNTK with $0 \leq \gamma < 0.5$ badly affects their generalization, while the stability of ResNTK with $0.5 < \gamma \leq 1$ allows it to learn target

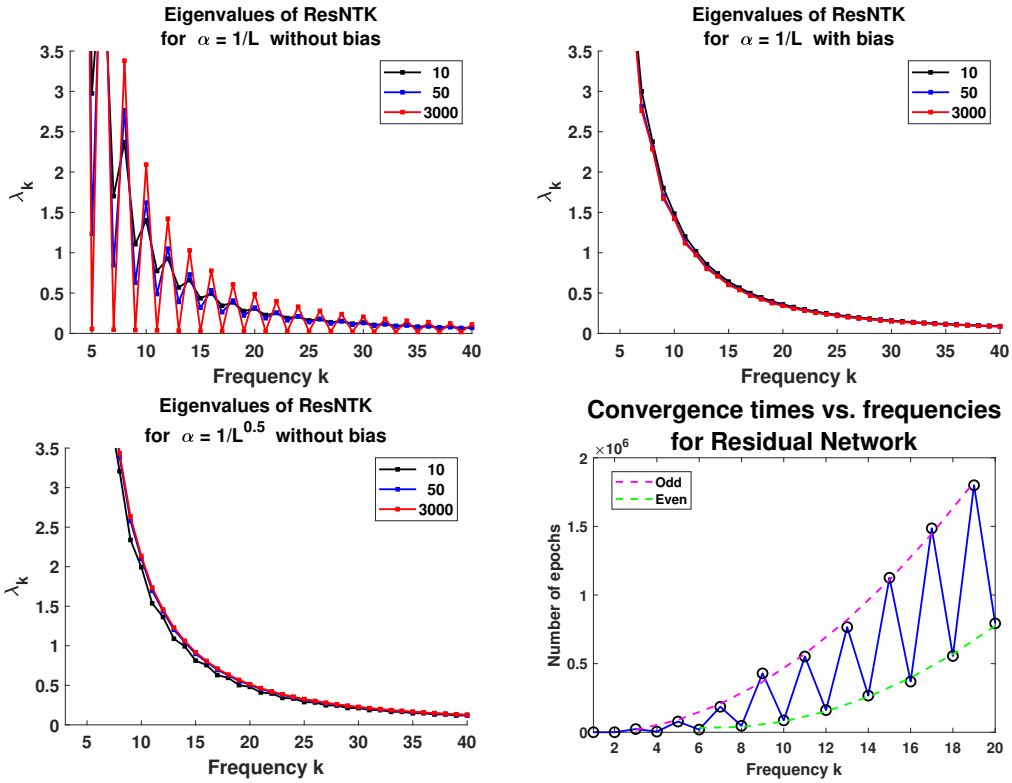

Figure 1: Top left: the eigenvalues of ResNTK without bias ($\gamma = 1$, i.e., $\alpha = 1/L$) as a function of frequency for different depths L. The eigenvalues for the odd frequencies as L grows are significantly lower than the eigenvalues for the nearby even frequencies, approaching 0 at $L \to \infty$. This parity difference disappears with bias (top right) or if the hyperparameter γ is set to 0.5 or lower ($\gamma = 0.5$, i.e., $\alpha = 1/\sqrt{L}$, no bias, bottom left), in which case the eigenvalues decrease monotonically with frequency. Bottom right: a real residual network with $L = 50$ and $\alpha = 1/50$ was trained to regress the function $\sin(kx)$ for input $(\cos x, \sin x) \in \mathbb{S}^1$. The figure shows the number of epochs to convergence for each frequency k (execution was stopped when a fit with 5% error was achieved). Consistent with our findings for ResNTK, convergence times grow quadratically with k, but convergence for the odd frequency target functions is significantly slower than for the even frequencies.

functions with a small generalization error. This is established in the following theorem. Denote the space of band limited functions

$$\left\{ y(\mathbf{x}) | y(\mathbf{x}) = \sum_{k=0}^{r} \sum_{j=1}^{N(d,k)} \alpha_{kj} Y_{kj}(\mathbf{x}),\ \mathbf{x} \in \mathbb{S}^{d-1} \right\}$$

by $\mathcal{H}_r(\mathbb{S}^{d-1})$. The theorem states that with the truncated ℓ_2 loss $\mathcal{L}$ and $\mathbf{x}$ drawn from the uniform distribution, deep ResNTK with $0.5 < \gamma \leq 1$ can learn functions $y(\mathbf{x}) \in \mathcal{H}_r(\mathbb{S}^{d-1})$ with generalization error that approaches zero as the number of training samples increases, whereas, in contrast, FC-NTK of sufficient depth achieves poor generalization for such functions. Below we denote by f_{ResNTK} and $f_{\text{FC-NTK}}$ the predictions of the ResNTK and FC-NTK kernel regressors.

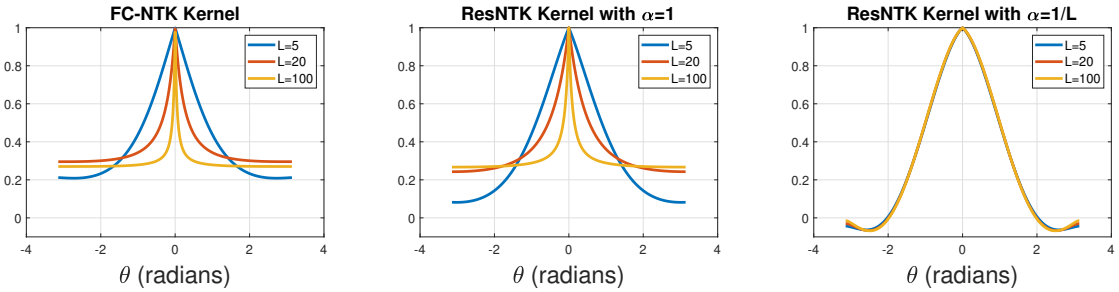

Figure 2: FC-NTK (left) and ResNTK (center $\alpha = 1$, right $\alpha = 1/L$) for networks of different depths, $L = 5, 20, 100$. For FC-NTK and ResNTK with $\alpha = 1$, the kernel becomes spiky with depth. With $\alpha = 1/L$ ResNTK remains stable for all depths.

Theorem 5 *Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be n i.i.d samples such that $\{\mathbf{x}_i\}_{i=1}^n$ are drawn from the uniform distribution on $\mathbb{S}^{d-1}$ and assuming that $y \in \mathcal{H}_r(\mathbb{S}^{d-1})$. Then,*

1. *There exists L_0 such that $\forall L > L_0$ it holds that with probability at least $1 - \delta$, the expected risk of the ResNTK with depth L and $0.5 < \gamma \leq 1$ is upper bounded by*

$$\mathbb{E}(\mathcal{L}(f_{\text{ResNTK}}(\mathbf{x}), y)) \leq O\left(\frac{r^{\frac{3d-2}{2}}}{\sqrt{n}} + \sqrt{\frac{log(1/\delta)}{n}}\right).$$

2. *$\forall \epsilon > 0$ there exists $L_0 = L(\epsilon, n)$, such that $\forall L > L_0$, the NTK predictor for $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ and depth L satisfies almost surely*

$$\mathbb{E}(\mathcal{L}(f_{\text{FC-NTK}}(\mathbf{x}), y)) = \int_{\mathbb{S}^{d-1}} (f_{\text{FC-NTK}}(\mathbf{x}) - y(\mathbf{x}))^2 d\mathbf{x} \geq 1 - O(\epsilon),$$

where the expectation is taken over the data distribution stated above.

The main contribution in Theorem 5 is in proving that with a proper set of parameters deep ResNTK generalizes well, extending existing generalization results Arora et al. (2019a); Nitanda and Suzuki (2020); Huang et al. (2020); Xiao et al. (2020) to ResNTK. We note that Arora et al. (2019b); Nitanda and Suzuki (2020) show better generalization bounds for FC-NTK, but their results are applicable to two layers Neural Networks only. Xiao et al. (2020); Huang et al. (2020) show that with large depth, FC-NTK becomes degenerate. The second part of Thm. 5 leverages their results to establish a trivial MSE for such FC-NTKs.

Our experiments in Section 5 below indeed indicate that while for FC-NTK and ResNTK with $0 \leq \gamma < 0.5$ learning accuracy degrades with depth, for ResNTK with $0.5 \leq \gamma \leq 1$ learning accuracy is stable across depth. This is also observed in the corresponding fully connected and residual networks.

5. Experiments

We performed a number of experiments to show the effect of depth on ResNTK and to compare it to FC-NTK. We further tested these effects also on actual networks. In the

Table 1: Classification accuracies (percent) obtained by applying FC-NTK and ResNTK on the CIFAR-10 dataset with $\alpha \in \{\frac{1}{L^{5/4}}, 1/L, 1/\sqrt{L}, \frac{1}{\sqrt[4]{L}}, 1\}$.

		ResNTK				
L	FC-NTK	$\alpha = \frac{1}{L^{5/4}}$	$\alpha = \frac{1}{L}$	$\alpha = \frac{1}{\sqrt{L}}$	$\alpha = \frac{1}{\sqrt[4]{L}}$	$\alpha = 1$
5	58.29	57.89	58.23	58.32	58.37	58.31
25	54.33	57.44	57.72	58.33	57.92	55.91
50	51.42	57.38	57.58	58.34	57.63	54.18
100	48.27	57.34	57.53	58.34	57.01	51.39

Table 2: Classification accuracies (percent) obtained by applying on the SVHN dataset FC-NTK and ResNTK with $\alpha \in \{1/L, 1/\sqrt{L}, 1\}$.

		ResNTK				
L	FC-NTK	$\alpha = \frac{1}{L^{5/4}}$	$\alpha = \frac{1}{L}$	$\alpha = \frac{1}{\sqrt{L}}$	$\alpha = \frac{1}{\sqrt[4]{L}}$	$\alpha = 1$
5	74.44	79.4	73.62	78.36	77.9	77.72
25	48.75	79.8	74.73	78.17	77.3	76.03
50	33.69	79.8	74.89	78.14	77	74.16
100	21.12	79.7	74.91	78.13	76.4	71.12

kernel experiments, we minimized the ridge regression formula

$$\min_{f \in \mathcal{H}} \sum_{i=1}^{n} (f(\mathbf{x}_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}},$$

with a constant ridge parameter $\lambda \geq 0$. We used in these experiments the bias-free FC-NTK and ResNTK. In the network experiment, we minimized the cross-entropy loss with no regularization.

CIFAR-10 We applied both ResNTK and FC-NTK to the CIFAR-10 dataset. Note that the kernels we applied correspond to classical and residual fully connected architectures and are not convolutional. We normalized the pixels in each image to zero mean and unit variance and used kernel regression with $\lambda = 0$. Table 1 (left) shows classification accuracies with FC-NTK and ResNTK with $\alpha \in \{\frac{1}{L^{5/4}}, \frac{1}{L}, \frac{1}{\sqrt{L}}, \frac{1}{\sqrt[4]{L}}, 1\}$. It can be seen that test accuracies for FC-NTK degrade from 58.28% with 5 layers to 48.27% with 100 layers. Likewise, test accuracies for ResNTK with $\alpha = 1$ degrade from 58.31% with 5 layers to 51.39% with 100 layers. In contrast, consistent with our theory, ResNTK with $\alpha \in \{1/L, 1/\sqrt{L}\}$ maintains an accuracy of 57.5%-58.3% across depth.

SVHN We repeated the same experiments on the SVHN dataset, see Table 2 (right). Here too we normalized the pixels in each image to zero mean and unit variance but used regression with $\lambda = 1e^{-5}$. The differences between FC-NTK and ResNTK are even more extreme in this experiment. FC-NTK degrades from an accuracy of 74.44% with 5 layers to 21.12% with 100 layers. ResNTK with $\alpha = 1$ degrades more modestly, from 77.72% with 5

Table 3: Network application. Classification accuracies (percent) on the CIFAR-10 dataset obtained by applying a residual network with $\alpha \in \{\frac{1}{L^{5/4}}, \frac{1}{L}, \frac{1}{\sqrt{L}}, \frac{1}{\sqrt[4]{L}}\}$ and with a standard fully-connected network (FC-Net).

		ResNet			
L	FC-Net	$\alpha = \frac{1}{L^{5/4}}$	$\alpha = \frac{1}{L}$	$\alpha = \frac{1}{\sqrt{L}}$	$\alpha = \frac{1}{\sqrt[4]{L}}$
5	52.32	53.17	53.15	53.58	52.17
25	41.08	53.07	53.49	52.09	47.32
50	36.76	52.63	53.49	52.89	45.34
100	36.73	52.0	53.49	52.05	36.48

layers to 71.12%, while with $\alpha = 1/L$ and $\alpha = 1/\sqrt{L}$ it maintains respectively a 74-75% and 78% accuracy for all tested depths as our theory predicts.

CIFAR-10: Real networks To further examine the relevance of these results to real networks, we applied a fully-connected network (FC-Net) and a residual, fully-connected network (ResNet), both with hidden layers of width $m = 2000$. We optimized the networks with SGD with no momentum and used a constant learning rate of $5e^{-5}$. The results shown in Table 3 are similar to those obtained with the respective kernels. Specifically, both FC-Net and ResNet with $\alpha = L^{-1/4}$ degrade from 52% with 5 layers to 36% with deeper layers. Consistent with our kernel analysis, ResNet with $\gamma \in \{1, 0.5\}$ maintains an accuracy of 52-53% for all depths.

6. Conclusion

We have provided derivations to determine the RKHS structure of NTK for residual networks and analyzed the shape of the kernel and its generalization properties with different hyperparameters. Our analysis indicates that similar to NTK for classical, fully connected networks, the eigenvalues of ResNTK corresponding to its spherical harmonic eigenfunctions decay polynomially with frequency k at the rate of k^{-d}. These in turn imply that the set of functions in its RKHS are identical to those of both FC-NTK and the Laplace kernel restricted to the hypersphere $\mathbb{S}^{d-1}$. Our results imply that all three kernels produce functions of similar smoothness properties. We have shown further that, depending on the choice of α, which balances between the residual and skip connections, deep bias-free ResNTK is significantly biased toward the even frequencies. Finally, we saw that ResNTK can be controlled to become spiky with depth, as is the case with FC-NTK, or maintain a stable shape, enabling superior generalization with deep networks.

Our results suggest that NTK provides only a partial explanation to the success of residual networks. Indeed it appears that classification with FC-NTK degrades with depth, while classification with ResNTK can be made stable with a proper choice of a balancing hyper-parameter. However, our experiments suggest that with an optimal choice of depth, classification results with FC-NTK and ResNTK are similar, most likely due to their similar RKHS structures. This is somewhat in contrast to actual implementations, in which residual networks seem to significantly outperform classical feed-forward networks. This difference

may be attributed to optimization issues, or to the possible invalidity of the assumptions of NTK to real networks of finite width. It is also possible that differences between residual and classical kernels are more significant in convolutional architectures.

Acknowledgments

This research was supported in part by the Israel Science Foundation, grant No. 1639/19, by the Israeli Council for Higher Education (CHE) via the Weizmann Data Science Research Center, by the MBZUAI-WIS Joint Program for Artificial Intelligence Research and by research grants from the Estates of Bernice Bernath and Marni Josephs Grossman; Joel B. Levey; Tully and Michele Plesser and the Anita James Rosen and Harry Schutzman Foundations.

Appendix A. Decay rate of eigenvalues of ResNTK without bias

In this section we prove Theorem 1 for a bias-free ResNTK (i.e. $\tau = 0$). We extend this to $\tau \geq 0$ in Section B.

Theorem A.1 *The eigenvalues λ_k of ResNTK, $\boldsymbol{r}(\mathbf{x}, \mathbf{z})$, for $\mathbf{x}, \mathbf{z} \in \mathbb{S}^{d-1}$ corresponding to a spherical harmonic eigenfunction $Y_{kl}(\mathbf{x})$ with frequency $k \geq 0$ and phase $1 \leq l \leq N(d,k)$ decay when $k \to \infty$ under the uniform measure as*

$$\lambda_k = \begin{cases} \bar{C}(c_1 + c_{-1})k^{-d} \ if \ k \ is \ even \\ \bar{C}(c_1 - c_{-1})k^{-d} \ if \ k \ is \ odd, \end{cases}$$

where $\bar{C}, c_1, c_{-1}$ are constants that depend on α, L, τ.

Before proving the theorem, we will lay out some notations and prove related lemmas. We assume that $\mathbf{x}, \mathbf{z} \in \mathbb{S}^{d-1}$ and let $u = \mathbf{x}^T\mathbf{z}$. We use the following lemma to simplify ResNTK for such input.

Lemma A.2 *Huang et al. (2020) For inputs in $\mathbb{S}^{d-1}$ and with no bias ($\tau = 0$), $K_\ell(\mathbf{x}, \mathbf{x}) = (1+\alpha^2)^\ell$.*

Using the lemma above, bias-free ResNTK can be expressed as follows

$$\boldsymbol{r}^{(L)}(u) = C\sum_{\ell=1}^{L} B_{\ell+1}(u)\left[(1+\alpha^2)^{\ell-1}\kappa_1\left(\frac{K_{\ell-1}(u)}{(1+\alpha^2)^{\ell-1}}\right) + K_{\ell-1}(u)\kappa_0\left(\frac{K_{\ell-1}(u)}{(1+\alpha^2)^{\ell-1}}\right)\right], \quad (8)$$

where $K_0(u) = u$, $B_{L+1}(u) = 1$, $C = \frac{1}{2L(1+\alpha^2)^{L-1}}$ and

$$K_\ell(u) = K_{\ell-1}(u) + \alpha^2(1-\alpha^2)^{l-1}\kappa_1\left(\frac{K_{\ell-1}(u)}{(1+\alpha^2)^{\ell-1}}\right), \quad \ell = 1, ..., L-1 \quad (9)$$

$$B_\ell(u) = B_{\ell+1}(u)\left[1 + \alpha^2\kappa_0\left(\frac{K_{\ell-1}(u)}{(1+\alpha^2)^{\ell-1}}\right)\right], \quad \ell = L, ..., 2, \quad (10)$$

and κ_0 and κ_1 are defined as

$$\kappa_0(u) = \frac{1}{\pi}(\pi - acos(u)) \quad (11)$$

$$\kappa_1(u) = \frac{1}{\pi}\left(u \cdot (\pi - acos(u)) + \sqrt{1-u^2}\right). \quad (12)$$

We further define the following variables, to be used in Section A.2. Near -1 (small $t > 0$):

$$\nu_\ell = \frac{K_{\ell-1}(-1+t)}{(1+\alpha^2)^{\ell-1}} \quad (13)$$

$$\beta_\ell = \kappa_1(\nu_\ell) \quad (14)$$

$$\eta_\ell = \kappa_0(\nu_\ell) \quad (15)$$

for $\ell = 1, 2, ...$, and $\beta_0 = \eta_L = 0$. Note that $\beta_\ell, \eta_\ell \in [0, 1]$ due to the image of the arc-cosine kernels.

The proof of Theorem 1 proceeds by calculating the asymptotes of ResNTK near ± 1 and applying a result from Bietti and Bach (2021), which for certain zonal kernels relates the decay rate of the eigenvalues of a kernel to its infinitesimal tendency near ± 1. Below we review the theorem and provide additional lemmas, which together allow us to prove Theorem 1.

Theorem A.3 (Bietti and Bach (2021)) *Let $\kappa : [-1, 1] \to \mathbb{R}$ be a C^∞ function on $(-1, 1)$ that has the following asymptotic expansions around ± 1*

$$\kappa(1-t) = p_1(t) + c_1 t^\nu + o(t^\nu) \tag{16}$$
$$\kappa(-1+t) = p_{-1}(t) + c_{-1} t^\nu + o(t^\nu) \tag{17}$$

for $t \geq 0$, where p_1, p_{-1} are polynomials and $\nu > 0$ is not an integer. Let λ_k denote an eigenvalue of κ corresponding to a spherical harmonic eigenfunction of frequency k. Then, there is an absolute constant $C(d, \nu)$ depending on d and ν such that

- *For k even, if $c_1 \neq -c_{-1}$:*
  $\lambda_k \sim (c_1 + c_{-1}) C(d, \nu) k^{-d-2\nu+1}$.
- *For k odd, if $c_1 \neq c_{-1}$:*
  $\lambda_k \sim (c_1 - c_{-1}) C(d, \nu) k^{-d-2\nu+1}$.

In the case $|c_1| = |c_{-1}|$, we have $\lambda_k = o(k^{-d-2\nu+1})$ for one of the two parities (or both if $c_1 = c_{-1} = 0$). If κ is infinitely differentiable on $[-1, 1]$ so that no such ν exists, then λ_k decays faster than any polynomial.

In the following sections, we provide the full Taylor expansions around ± 1 for ResNTK. We further calculate the fractional power ν and show it equals 0.5. In addition, we prove that $c_1 \neq c_{-1}$, except with a special choice of the parameter α and large L. These allow us to deduce that the eigenvalues decay as $O(k^{-d})$.

A.1 Expansion near 1

The expansion is given by the following lemma (Theorem 1, part 1 with $\tau = 0$ in the paper).

Lemma A.4 *For inputs in $\mathbb{S}^{d-1}$ and near +1, if $\alpha > 0$ and $L \geq 1$*

$$\boldsymbol{r}^{(L)}(1-t) = 1 + c_1 t^{1/2} + o(t^{1/2})$$

where

$$c_1 = -\frac{1 + \alpha^2 L}{\sqrt{2}\pi(1+\alpha^2)}.$$

We prove this using the following lemmas.

Lemma A.5 *Bietti and Bach (2021) The arc-cosine kernels near 1 satisfy*

$$\kappa_0(1-t) = 1 - \frac{\sqrt{2}}{\pi} t^{1/2} + \mathcal{O}(t^{3/2}) \tag{18}$$
$$\kappa_1(1-t) = 1 - t + \frac{2\sqrt{2}}{3\pi} t^{3/2} + \mathcal{O}(t^{5/2}). \tag{19}$$

Lemma A.6 *For small $t > 0$, $K_\ell(1-t) = (1+\alpha^2)^\ell(1-t) + o(t)$, where K_ℓ is defined in* (9).

Proof We prove this by induction. For $\ell = 0$, $K_0(1-t) = 1-t$, trivially satisfying the lemma. Suppose the lemma holds for $K_{\ell-1}(1-t)$, using (9)

$$\begin{aligned} K_\ell(1-t) &= K_{\ell-1}(1-t) + \alpha^2(1+\alpha^2)^{\ell-1}\kappa_1\left(\frac{K_{\ell-1}(1-t)}{(1+\alpha^2)^{\ell-1}}\right) \\ &= (1+\alpha^2)^{\ell-1}(1-t) + o(t) + \alpha^2(1+\alpha^2)^{\ell-1}\kappa_1\left(\frac{(1+\alpha^2)^{\ell-1}(1-t)+o(t)}{(1+\alpha^2)^{\ell-1}}\right) \\ &= (1+\alpha^2)^{\ell-1}(1-t) + o(t) + \alpha^2(1+\alpha^2)^{\ell-1}\kappa_1(1-t+o(t)) \\ &= (1+\alpha^2)^{\ell-1}(1-t) + \alpha^2(1+\alpha^2)^{\ell-1}(1-t) + o(t) = (1+\alpha^2)^\ell(1-t) + o(t), \end{aligned}$$

where the leftmost equality in the last line is due to (19). ■

Lemma A.7 *With small $t > 0$,*

$$\begin{aligned} \kappa_0\left(\frac{K_{\ell-1}(1-t)}{(1+\alpha^2)^{\ell-1}}\right) &= 1 - \frac{\sqrt{2}}{\pi}t^{1/2} + o(t) \\ \kappa_1\left(\frac{K_{\ell-1}(1-t)}{(1+\alpha^2)^{\ell-1}}\right) &= 1 - t + o(t). \end{aligned}$$

Proof Using Lemma A.6, for small $t > 0$,

$$\frac{K_{\ell-1}(1-t)}{(1+\alpha^2)^{\ell-1}} = \frac{(1+\alpha^2)^{\ell-1}(1-t)+o(t)}{(1+\alpha^2)^{\ell-1}} = 1 - t + o(t).$$

Next, using (18)

$$\kappa_0\left(\frac{K_{\ell-1}(1-t)}{(1+\alpha^2)^{\ell-1}}\right) = \kappa_0(1-t+o(t)) = 1 - \frac{\sqrt{2}}{\pi}t^{1/2} + o(t),$$

and using (19)

$$\kappa_1\left(\frac{K_{\ell-1}(1-t)}{(1+\alpha^2)^{\ell-1}}\right) = \kappa_1(1-t+o(t)) = 1 - t + o(t).$$

■

Lemma A.8 *With small $t > 0$,*

$$B_{\ell+1}(1-t) = (1+\alpha^2)^{L-\ell} - \frac{\sqrt{2}\,\alpha^2}{\pi}(1+\alpha^2)^{L-\ell-1}(L-\ell)\,t^{1/2} + \mathcal{O}(t),$$

where B_ℓ is defined in (10).

Proof With small $t > 0$, we use Lemma A.7 to simplify (10) as follows:

$$B_\ell(1-t) = B_{\ell+1}(1-t)\left[1+\alpha^2\left(1-\frac{\sqrt{2}}{\pi}t^{1/2}+o(t)\right)\right].$$

Since $B_{L+1} = 1$, resolving the recursion yields

$$B_{\ell+1}(1-t) = \left(1+\alpha^2-\frac{\sqrt{2}\,\alpha^2}{\pi}t^{1/2}+\mathcal{O}(t^{3/2})\right)^{L-\ell}.$$

This can be simplified as follows

$$B_{\ell+1}(1-t) = \sum_{i=0}^{L-\ell}\binom{L-\ell}{i}\left(1+\alpha^2+\mathcal{O}(t^{3/2})\right)^{L-\ell-i}\left(-\frac{\sqrt{2}\,\alpha^2}{\pi}t^{1/2}+\mathcal{O}(t^{3/2})\right)^{i}.$$

Grouping together all $\mathcal{O}(t)$ terms, we finally obtain

$$B_{\ell+1}(1-t) = (1+\alpha^2)^{L-\ell} - \frac{\sqrt{2}\,\alpha^2}{\pi}(1+\alpha^2)^{L-\ell-1}(L-\ell)\,t^{1/2}+\mathcal{O}(t).$$

■

We next prove Lemma A.4.

Proof (Lemma A.4) Rewrite (8) as $\boldsymbol{r}^{(L)}(1-t) = C\sum_{\ell=1}^{L} X_\ell Y_\ell$, where:

$$\begin{aligned} C &= \frac{1}{2L(1+\alpha^2)^{L-1}} \\ X_\ell &= (1+\alpha^2)^{\ell-1}\kappa_1\left(\frac{K_{\ell-1}(1-t)}{(1+\alpha^2)^{\ell-1}}\right) + K_{\ell-1}(1-t)\kappa_0\left(\frac{K_{\ell-1}(1-t)}{(1+\alpha^2)^{\ell-1}}\right) \\ Y_\ell &= B_{\ell+1}(1-t). \end{aligned}$$

Using Lemmas A.6 and A.7, for small $t > 0$,

$$\begin{aligned} X_\ell &= (1+\alpha^2)^{\ell-1}(1-t+o(t)) + ((1+\alpha^2)^{\ell-1}(1-t)+o(t))\left(1-\frac{\sqrt{2}}{\pi}t^{1/2}+o(t)\right) \\ &= (1+\alpha^2)^{\ell-1}(1-t) + (1+\alpha^2)^{\ell-1}(1-t)\left(1-\frac{\sqrt{2}}{\pi}t^{1/2}\right)+\mathcal{O}(t) \\ &= (1+\alpha^2)^{\ell-1}(1-t)\left(2-\frac{\sqrt{2}}{\pi}t^{1/2}\right)+\mathcal{O}(t) = (1+\alpha^2)^{\ell-1}\left(2-\frac{\sqrt{2}}{\pi}t^{1/2}\right)+o(t^{1/2}). \end{aligned}$$

Using Lemma A.8 each term in the sum can be written as

$$\begin{aligned} X_\ell Y_\ell &= \left[(1+\alpha^2)^{\ell-1}\left(2-\frac{\sqrt{2}}{\pi}t^{1/2}\right)\right]\left[(1+\alpha^2)^{L-\ell}-\frac{\alpha^2\sqrt{2}}{\pi}(1+\alpha^2)^{L-\ell-1}(L-\ell)\,t^{1/2}\right]+\mathcal{O}(t) \\ &= \left[2(1+\alpha^2)^{L-1}-\frac{\sqrt{2}}{\pi}\left(2\alpha^2(1+\alpha^2)^{L-2}(L-\ell)+(1+\alpha^2)^{L-1}\right)t^{1/2}\right]+\mathcal{O}(t) \\ &= (1+\alpha^2)^{L-1}\left[2-\frac{\sqrt{2}}{\pi}\left(\frac{2\alpha^2(L-\ell)}{1+\alpha^2}+1\right)t^{1/2}\right]+\mathcal{O}(t). \end{aligned}$$

Recall that $C = \frac{1}{2L(1+\alpha^2)^{L-1}}$

$$CX_\ell Y_\ell = \frac{1}{2L}\left[2 - \frac{\sqrt{2}}{\pi}\left(\frac{2\alpha^2(L-\ell)}{1+\alpha^2}+1\right)t^{1/2}\right] + \mathcal{O}(t).$$

Summing over the layers

$$\begin{aligned}\boldsymbol{r}^{(L)}(1-t) &= C\sum_{\ell=1}^{L} X_\ell Y_\ell = 1 - \frac{1}{\sqrt{2}\pi L}\left[\frac{\alpha^2 L(L-1)}{1+\alpha^2}+L\right]t^{1/2} + \mathcal{O}(t) \\ &= 1 - \frac{1+\alpha^2 L}{\sqrt{2}\pi(1+\alpha^2)}t^{1/2} + o(t^{1/2}). \end{aligned} \tag{20}$$

■

A.2 Expansion near -1

Here we investigate the expansion of ResNTK near -1. We consider two cases. First, with $\alpha > 0$ such that $\alpha^2 L$ does not vanish as L grows, and secondly, with $\alpha > 0$ and $\alpha^2 L \ll 1$.

A.2.1 $\alpha > 0$ such that $\alpha^2 L \not\ll 1$

The expansion is given by the following lemma (Theorem 1, part 2 with $\tau = 0$ in the paper).

Lemma A.9 *For inputs in $\mathbb{S}^{d-1}$ and near -1, if $\alpha > 0$ and $L \geq 2$ then*

$$\boldsymbol{r}^{(L)}(-1+t) = p_{-1}(t) + c_{-1}t^{1/2} + o(t^{1/2}),$$

with

$$|c_{-1}| \leq \frac{1}{\sqrt{2}\pi(1+\alpha^2)L}.$$

We prove this using the following lemmas.

Lemma A.10 *Bietti and Bach (2021) The arc-cosine kernels near -1 satisfy*

$$\kappa_0(-1+t) = \frac{\sqrt{2}}{\pi}t^{1/2} + \mathcal{O}(t^{3/2}) \tag{21}$$

$$\kappa_1(-1+t) = \frac{2\sqrt{2}}{3\pi}t^{3/2} + \mathcal{O}(t^{5/2}). \tag{22}$$

Lemma A.11 *With small $t > 0$,*

$$K_\ell(-1+t) = -1 + t + \alpha^2\sum_{j=0}^{\ell}(1+\alpha^2)^{j-1}\beta_j + \mathcal{O}(t^{3/2}),$$

where β_ℓ as defined in (14).

Proof With $\ell = 0$, $K_0(-1+t) = -1+t$, trivially satisfying the lemma. Suppose the lemma holds for $K_{\ell-1}(-1+t)$. Then, using (9) and (14)

$$\begin{aligned} K_\ell(-1+t) &= K_{\ell-1}(-1+t) + \alpha^2(1+\alpha^2)^{\ell-1}\kappa_1\left(\frac{K_{\ell-1}(-1+t)}{(1+\alpha^2)^{\ell-1}}\right) \\ &= K_{\ell-1}(-1+t) + \alpha^2(1+\alpha^2)^{\ell-1}\beta_\ell. \end{aligned}$$

By the induction assumption

$$\begin{aligned} K_\ell(-1+t) &= -1+t+\alpha^2\sum_{j=0}^{\ell-1}(1+\alpha^2)^{j-1}\beta_j + \alpha^2(1+\alpha^2)^{\ell-1}\beta_\ell + \mathcal{O}(t^{3/2}) \\ &= -1+t+\alpha^2\sum_{j=0}^{\ell}(1+\alpha^2)^{j-1}\beta_j + \mathcal{O}(t^{3/2}). \end{aligned}$$

■

The next Lemma ensures that β_ℓ is well defined (since κ_1 takes input in $[-1,1]$).

Lemma A.12 *Let ν_ℓ as defined in* (13)*. Then, $\forall \ell \geq 1, \quad |\nu_\ell| \leq 1$.*

Proof Using (13) and Lemma A.11 we have

$$\nu_\ell = \frac{-1+t+\alpha^2\sum_{j=0}^{\ell-1}(1+\alpha^2)^{j-1}\beta_j}{(1+\alpha^2)^{\ell-1}}. \tag{23}$$

Since $\beta_0 = 0$, with $\ell = 1$ $|\nu_1| = |-1+t| \leq 1$. With $\ell > 1$ using triangle inequality,

$$|\nu_\ell| \leq \left|\frac{-1+t+\alpha^2\sum_{j=0}^{\ell-2}(1+\alpha^2)^{j-1}\beta_j}{(1+\alpha^2)^{\ell-1}}\right| + \left|\frac{\alpha^2(1+\alpha^2)^{\ell-2}\beta_{\ell-1}}{(1+\alpha^2)^{\ell-1}}\right|.$$

Noting that the first term is $\left|\frac{\nu_{\ell-1}}{1+\alpha^2}\right|$, and assuming by induction that the lemma is satisfied for $\nu_{\ell-1}$, then

$$|\nu_\ell| \leq \frac{1}{1+\alpha^2} + \frac{\alpha^2\beta_{\ell-1}}{1+\alpha^2} \leq \frac{1}{1+\alpha^2} + \frac{\alpha^2}{1+\alpha^2} = 1,$$

where the rightmost inequality is because by definition $\beta_\ell \in [0,1]$. ■

Lemma A.13 *Let $\delta_\ell = \frac{-1+\alpha^2\sum_{j=0}^{\ell-1}(1+\alpha^2)^{j-1}\beta_j}{(1+\alpha^2)^{\ell-1}}$. Then, $\forall \ell \geq 2, \quad |\delta_\ell| < 1$.*

Proof For $\ell = 2$ we have $|\delta_2| = \left|\frac{-1+\alpha^2\beta_1}{1+\alpha^2}\right| \leq \max\{\frac{1}{1+\alpha^2}, \frac{\alpha^2-1}{1+\alpha^2}\} < 1$. Assume the lemma holds for $\ell - 1$. We prove for ℓ:

$$|\delta_\ell| = \left|\frac{-1+\alpha^2\sum_{j=0}^{\ell-1}(1+\alpha^2)^{j-1}\beta_j}{(1+\alpha^2)^{\ell-1}}\right| = \left|\frac{-1+\alpha^2\sum_{j=0}^{\ell-2}(1+\alpha^2)^{j-1}\beta_j + \alpha^2(1+\alpha^2)^{\ell-1}\beta_\ell}{(1+\alpha^2)^{\ell-2}(1+\alpha^2)}\right| =$$

$$\left|\frac{\delta_{\ell-1}}{(1+\alpha^2)} + \frac{\alpha^2(1+\alpha^2)^{\ell-2}\beta_\ell}{(1+\alpha^2)^{\ell-2}(1+\alpha^2)}\right| = \left|\frac{\delta_{\ell-1}}{(1+\alpha^2)} + \frac{\alpha^2\beta_\ell}{(1+\alpha^2)}\right| \overset{1}{\leq} \left|\frac{\delta_{\ell-1}}{(1+\alpha^2)}\right| + \left|\frac{\alpha^2\beta_\ell}{(1+\alpha^2)}\right| \overset{2}{<}$$

$$\frac{1}{(1+\alpha^2)} + \frac{\alpha^2}{(1+\alpha^2)} = 1,$$

where $\leq^1$ uses the triangle inequality, and $<^2$ is due to the induction hypothesis and the fact that $\forall \ell, \beta_\ell \in [0,1]$. ■

Lemma A.14 *With small $t > 0$, $\forall \ell \in [L-1]$*

$$\beta_\ell = \kappa_1 \left(\frac{-1 + \alpha^2 \sum_{j=0}^{\ell-1} (1+\alpha^2)^{j-1} \beta_j}{(1+\alpha^2)^{\ell-1}} \right) + \mathcal{O}(t).$$

Proof First, note that for $\ell = 1$ we get this directly from Lemma 22. For $\ell \geq 2$, using Lemma A.11 and the definition in (14):

$$\begin{aligned} \beta_\ell &= \kappa_1 \left(\frac{-1 + t + \alpha^2 \sum_{j=0}^{\ell-1} (1+\alpha^2)^{j-1} \beta_j}{(1+\alpha^2)^{\ell-1}} \right) \\ &= \kappa_1 \left(\frac{-1 + \alpha^2 \sum_{j=0}^{\ell-1} (1+\alpha^2)^{j-1} \beta_j}{(1+\alpha^2)^{\ell-1}} + \mathcal{O}(t) \right) = \kappa_1 \left(\delta_\ell + \mathcal{O}(t) \right), \end{aligned}$$

where δ_ℓ is defined in Lemma A.13. Note that from this lemma, $-1 < \delta_\ell < 1$. In this domain, κ_1 is infinitely differentiable, hence we get:

$$\beta_\ell = \kappa_1 \left(\delta_\ell \right) + \mathcal{O}(t) = \kappa_1 \left(\frac{-1 + \alpha^2 \sum_{j=0}^{\ell-1} (1+\alpha^2)^{j-1} \beta_j}{(1+\alpha^2)^{\ell-1}} \right) + \mathcal{O}(t).$$

■

Lemma A.15 *With small $t > 0$, $\forall \ell \in [L-1]$*

$$\beta_\ell = \tilde{c}_\ell + \mathcal{O}(t),$$

where $\tilde{c}_\ell \in [0,1]$ does not depend on t.

Proof The proof is by induction. For $\ell = 1$ we have from Lemma A.14

$$\beta_1 = \kappa_1 \left(\frac{-1}{(1+\alpha^2)} \right) + \mathcal{O}(t) = \tilde{c_1} + \mathcal{O}(t).$$

Suppose the lemma holds for $\beta_{\ell-1}$ and show for β_ℓ

$$\begin{aligned} \beta_\ell &= \kappa_1 \left(\frac{-1 + \alpha^2 \sum_{j=0}^{\ell-1} (1+\alpha^2)^{j-1} \beta_j}{(1+\alpha^2)^{\ell-1}} + \mathcal{O}(t) \right) \\ &= \kappa_1 \left(\frac{-1 + \alpha^2 \sum_{j=0}^{\ell-1} (1+\alpha^2)^{j-1} (\tilde{c_j} + \mathcal{O}(t))}{(1+\alpha^2)^{\ell-1}} + \mathcal{O}(t) \right) \\ &= \kappa_1 \left(\frac{-1 + \alpha^2 \sum_{j=0}^{\ell-1} (1+\alpha^2)^{j-1} \tilde{c_j}}{(1+\alpha^2)^{\ell-1}} + \mathcal{O}(t) \right) \\ &= \kappa_1 \left(\frac{-1 + \alpha^2 \sum_{j=0}^{\ell-1} (1+\alpha^2)^{j-1} \tilde{c_j}}{(1+\alpha^2)^{\ell-1}} \right) + \mathcal{O}(t) = \tilde{c}_\ell + \mathcal{O}(t), \end{aligned}$$

where the leftmost equality in the second line is from Lemma A.14. The definition of $\tilde{c}_\ell$ directly implies that $\tilde{c}_\ell \in [0,1]$. ■

Lemma A.16 *With small $t > 0$, and for $\ell = 1$,*

$$\eta_1 = \frac{\sqrt{2}}{\pi} t^{1/2} + \mathcal{O}(t^{3/2}).$$

For $\ell \geq 2$,

$$\eta_\ell = \kappa_0 \left(\frac{-1 + \alpha^2 \sum_{j=0}^{\ell-1} (1+\alpha^2)^{j-1} \beta_j}{(1+\alpha^2)^{\ell-1}} \right) + \mathcal{O}(t).$$

where η_ℓ is defined in (15)*.*

Proof First, note that for $\ell = 1$ we get this directly from Lemma 21. For $\ell \geq 2$, using Lemma A.11 and the definition (15):

$$\begin{aligned} \eta_\ell &= \kappa_0 \left(\frac{-1 + t + \alpha^2 \sum_{j=0}^{\ell-1} (1+\alpha^2)^{j-1} \beta_j}{(1+\alpha^2)^{\ell-1}} \right) \\ &= \kappa_0 \left(\frac{-1 + \alpha^2 \sum_{j=0}^{\ell-1} (1+\alpha^2)^{j-1} \beta_j}{(1+\alpha^2)^{\ell-1}} + \mathcal{O}(t) \right) = \kappa_0 \left(\delta_\ell + \mathcal{O}(t) \right). \end{aligned}$$

where δ_ℓ is defined in Lemma A.13. Note that from this lemma, $-1 < \delta_\ell < 1$. In this domain, κ_0 is infinitely differentiable, hence we get:

$$\eta_\ell = \kappa_0 \left(\delta_\ell \right) + \mathcal{O}(t) = \kappa_0 \left(\frac{-1 + \alpha^2 \sum_{j=0}^{\ell-1} (1+\alpha^2)^{j-1} \beta_j}{(1+\alpha^2)^{\ell-1}} \right) + \mathcal{O}(t).$$

■

Lemma A.17 *With small $t > 0$, $\forall \ell \geq 2$*

$$\eta_\ell = \tilde{d}_\ell + \mathcal{O}(t),$$

where $\tilde{d}_\ell \in [0,1]$ does not depend on t.

Proof The proof is by induction. For $\ell = 2$ we have from Lemma A.16

$$\eta_2 = \kappa_0 \left(\frac{-1}{(1+\alpha^2)} \right) + \mathcal{O}(t) = \tilde{d}_2 + \mathcal{O}(t).$$

Suppose the lemma holds for $\eta_{\ell-1}$ and show for η_ℓ

$$\begin{aligned}
\eta_\ell &= \kappa_0\left(\frac{-1+\alpha^2\sum_{j=0}^{\ell-1}(1+\alpha^2)^{j-1}\beta_j}{(1+\alpha^2)^{\ell-1}}+\mathcal{O}(t)\right)\\
&= \kappa_0\left(\frac{-1+\alpha^2\sum_{j=0}^{\ell-1}(1+\alpha^2)^{j-1}(\tilde{c_j}+\mathcal{O}(t))}{(1+\alpha^2)^{\ell-1}}+\mathcal{O}(t)\right)\\
&= \kappa_0\left(\frac{-1+\alpha^2\sum_{j=0}^{\ell-1}(1+\alpha^2)^{j-1}\tilde{c_j}}{(1+\alpha^2)^{\ell-1}}+\mathcal{O}(t)\right)\\
&= \kappa_0\left(\frac{-1+\alpha^2\sum_{j=0}^{\ell-1}(1+\alpha^2)^{j-1}\tilde{c_j}}{(1+\alpha^2)^{\ell-1}}\right)+\mathcal{O}(t)=\tilde{d}_\ell+\mathcal{O}(t),
\end{aligned}$$

where the leftmost equality in the second line is from Lemma A.16. The definition of $\tilde{d}_\ell$ directly implies that $\tilde{d}_\ell \in [0,1]$. ■

Lemma A.18 *With small $t>0$,*

$$B_{\ell+1}(-1+t)=\prod_{i=\ell+1}^{L}(1+\alpha^2\eta_i),$$

where η_ℓ is defined in (15)*.*

Proof Since $B_{L+1}=1$ and using (10)

$$B_{\ell+1}(-1+t)=\prod_{i=\ell+1}^{L}\left[1+\alpha^2\kappa_0\left(\frac{K_{i-1}(-1+t)}{(1+\alpha^2)^{i-1}}\right)\right]=\prod_{i=\ell+1}^{L}\left[1+\alpha^2\eta_\ell\right]$$

■

We next prove Lemma A.9.

Proof (Lemma A.9) Rewrite (8) as $\boldsymbol{r}^{(L)}(-1+t)=C\sum_{\ell=1}^{L}X_\ell Y_\ell$, where

$$\begin{aligned}
C &= \frac{1}{2L(1+\alpha^2)^{L-1}}\\
X_\ell &= (1+\alpha^2)^{\ell-1}\kappa_1\left(\frac{K_{\ell-1}(-1+t)}{(1+\alpha^2)^{\ell-1}}\right)+K_{\ell-1}(-1+t)\kappa_0\left(\frac{K_{\ell-1}(-1+t)}{(1+\alpha^2)^{\ell-1}}\right)\\
&= (1+\alpha^2)^{\ell-1}\beta_\ell+K_{\ell-1}(-1+t)\eta_\ell\\
Y_\ell &= B_{\ell+1}(-1+t).
\end{aligned}$$

By plugging Lemma A.11 into the definition of X_ℓ we have

$$X_\ell = (1+\alpha^2)^{\ell-1}\beta_\ell+\left(-1+\alpha^2\sum_{j=0}^{\ell-1}(1+\alpha^2)^{j-1}\beta_j\right)\eta_\ell+\mathcal{O}(t).$$

Using Lemma A.18 the sum can be written as

$$\sum_{\ell=1}^{L} X_\ell Y_\ell = \sum_{\ell=1}^{L} \left((1+\alpha^2)^{\ell-1}\beta_\ell + \left(-1 + \alpha^2 \sum_{j=0}^{\ell-1} (1+\alpha^2)^{j-1}\beta_j \right) \eta_\ell \right) \prod_{i=\ell+1}^{L} (1+\alpha^2\eta_i) + \mathcal{O}(t).$$

From Lemma A.16, there is a difference between $\ell = 1$ and $\ell \geq 2$. For $\ell = 1$:

$$\begin{aligned} X_1 Y_1 &= \left((1+\alpha^2)^0 \beta_1 + \left(-1 + \alpha^2 \sum_{j=0}^{0} (1+\alpha^2)^{j-1}\beta_j \right) \eta_1 \right) \prod_{i=1+1}^{L} (1+\alpha^2\eta_i) + \mathcal{O}(t) \\ &= -\eta_1 \prod_{i=2}^{L} (1+\alpha^2\eta_i) + \mathcal{O}(t) = -\left(\prod_{i=2}^{L} (1+\alpha^2\eta_i) \right) \frac{\sqrt{2}}{\pi} t^{1/2} + \mathcal{O}(t). \end{aligned}$$

Using Lemma A.17 this simplifies to

$$X_1 Y_1 = -\left(\prod_{i=2}^{L} (1+\alpha^2(\tilde{d}_i + \mathcal{O}(t))) \right) \frac{\sqrt{2}}{\pi} t^{1/2} + \mathcal{O}(t) = -\left(\prod_{i=2}^{L} (1+\alpha^2\tilde{d}_i) \right) \frac{\sqrt{2}}{\pi} t^{1/2} + \mathcal{O}(t).$$

For $\ell \geq 2$, using Lemmas A.15 and A.17

$$\begin{aligned} X_\ell Y_\ell &= \left((1+\alpha^2)^{\ell-1}\beta_\ell + \left(-1 + \alpha^2 \sum_{j=0}^{\ell-1} (1+\alpha^2)^{j-1}\beta_j \right) \eta_\ell \right) \prod_{i=\ell+1}^{L} (1+\alpha^2\eta_i) + \mathcal{O}(t) \\ &= \left((1+\alpha^2)^{\ell-1}(\tilde{c}_\ell + \mathcal{O}(t)) + \left(-1 + \alpha^2 \sum_{j=0}^{\ell-1} (1+\alpha^2)^{j-1}(\tilde{c}_j + \mathcal{O}(t)) \right) (\tilde{d}_\ell + \mathcal{O}(t)) \right) \\ &\quad \prod_{i=\ell+1}^{L} (1+\alpha^2(\tilde{d}_i + \mathcal{O}(t))) + \mathcal{O}(t) \\ &= \left((1+\alpha^2)^{\ell-1}\tilde{c}_\ell + \left(-1 + \alpha^2 \sum_{j=0}^{\ell-1} (1+\alpha^2)^{j-1}\tilde{c}_j \right) \tilde{d}_\ell \right) \prod_{i=\ell+1}^{L} (1+\alpha^2\tilde{d}_i) + \mathcal{O}(t)/ \end{aligned}$$

The sum can be rewritten as

$$\begin{aligned} \sum_{\ell=1}^{L} X_\ell Y_\ell &= \left(\sum_{\ell=2}^{L} \left((1+\alpha^2)^{\ell-1}\tilde{c}_\ell + \left(-1 + \alpha^2 \sum_{j=0}^{\ell-1} (1+\alpha^2)^{j-1}\tilde{c}_j \right) \tilde{d}_\ell \right) \prod_{i=\ell+1}^{L} (1+\alpha^2\tilde{d}_i) \right) \\ &\quad - \left(\prod_{i=2}^{L} (1+\alpha^2\tilde{d}_i) \right) \frac{\sqrt{2}}{\pi} t^{1/2} + \mathcal{O}(t). \end{aligned}$$

Multiplying this by the normalization factor C we have

$$\boldsymbol{r}^{(L)}(-1+t) = C \sum_{\ell=1}^{L} X_\ell Y_\ell = \frac{1}{2L(1+\alpha^2)^{L-1}} \sum_{\ell=1}^{L} X_\ell Y_\ell = p_{-1}(t) + c_{-1} t^{1/2} + o(t^{1/2}),$$

where

$$p_{-1}(t) = \frac{1}{2L(1+\alpha^2)^{L-1}} \left(\sum_{\ell=2}^{L} \left((1+\alpha^2)^{\ell-1} \tilde{c}_\ell + \left(-1 + \alpha^2 \sum_{j=0}^{\ell-1} (1+\alpha^2)^{j-1} \tilde{c}_j \right) \tilde{d}_\ell \right) \prod_{i=\ell+1}^{L} (1+\alpha^2 \tilde{d}_i) \right)$$

$$c_{-1} = -\frac{1}{2L(1+\alpha^2)^{L-1}} \left(\prod_{i=2}^{L} (1+\alpha^2 \tilde{d}_i) \right) \frac{\sqrt{2}}{\pi}.$$

From Lemma A.17,

$$\begin{aligned} |c_{-1}| &= \left| \frac{1}{2L(1+\alpha^2)^{L-1}} \left(\prod_{i=2}^{L} (1+\alpha^2 \tilde{d}_i) \right) \frac{\sqrt{2}}{\pi} \right| = \left| \frac{1}{\sqrt{2}\pi L(1+\alpha^2)^{L-1}} \left(\prod_{i=2}^{L} (1+\alpha^2 \tilde{d}_i) \right) \right| \\ &\leq \frac{\sqrt{2}(1+\alpha^2)^{L-2}}{2\pi L(1+\alpha^2)^{L-1}} = \frac{1}{\sqrt{2}\pi(1+\alpha^2)L}. \end{aligned}$$

■

A.2.2 Vanishing regime $\alpha^2 L \ll 1$

For the case where $\alpha^2 L \to 0$ with $L \to \infty$ (which implies $(1+\alpha^2)^j \approx 1, \forall j \in [L]$), the analysis takes the following form. The next Lemma is analogous to Lemma A.11.

Lemma A.19 *With small $t > 0$ and $\alpha^2 L \ll 1$,*

$$K_\ell(-1+t) = -1 + t + \mathcal{O}(t^{3/2}).$$

Proof With $\ell = 0$, $K_0(-1+t) = -1+t$, trivially satisfying the lemma. Suppose the lemma holds for $K_{\ell-1}(-1+t)$. Then, using (9) and (14)

$$\begin{aligned} K_\ell(-1+t) &= K_{\ell-1}(-1+t) + \alpha^2 (1+\alpha^2)^{\ell-1} \kappa_1 \left(\frac{K_{\ell-1}(-1+t)}{(1+\alpha^2)^{\ell-1}} \right) \\ &= K_{\ell-1}(-1+t) + \alpha^2 \kappa_1 \left(K_{\ell-1}(-1+t) \right). \end{aligned}$$

Where the last equality is from $\alpha^2 \ll 1$. By the induction assumption

$$K_\ell(-1+t) = (-1+t+\mathcal{O}(t^{3/2})) + \alpha^2 \kappa_1 \left(-1+t+\mathcal{O}(t^{3/2}) \right) = -1+t+\mathcal{O}(t^{3/2}),$$

where the last equality is directly from Lemma A.10. ■

The next Lemma is analogous to Lemma A.12.

Lemma A.20 *Let ν_ℓ as defined in* (13)*. Then, for $\alpha^2 L \ll 1$, $\forall \ell \geq 1$, $\nu_\ell = -1 + \mathcal{O}(t)$.*

Proof Using (23), with $\ell = 1$, $\nu_1 = -1 + t$. Assume the lemma is satisfied for $\nu_{\ell-1}$. Then, for $1 \leq j \leq \ell - 1$,

$$\beta_j = \kappa_1(\nu_j) = \kappa_1(-1 + \mathcal{O}(t)) = \mathcal{O}(t),$$

where the rightmost equality is due to (22). Therefore, using (23) and $(1+\alpha^2)^{\ell-1} \approx 1$ we obtain

$$\nu_\ell = \frac{-1+t+\alpha^2\sum_{j=0}^{\ell-1}(1+\alpha^2)^{j-1}\beta_j}{(1+\alpha^2)^{\ell-1}} = -1+t+\alpha^2\sum_{j=0}^{\ell-1}\mathcal{O}(t) = -1+\mathcal{O}(t).$$

■

Combining this lemma with lemma A.10 we get the following lemmas (analogous to A.14, A.16):

Lemma A.21 *With $\alpha^2 L \to 0$, $\forall \ell \in [L-1]$, $\beta_\ell = \kappa_1(\nu_\ell) = \kappa_1(-1+t) = \mathcal{O}(t)$.*

Lemma A.22 *With $\alpha^2 L \to 0$, $\forall \ell \in [L-1]$, $\eta_\ell = \kappa_0(\nu_\ell) = \kappa_0(-1+t) = \frac{\sqrt{2}}{\pi}t^{1/2} + \mathcal{O}(t)$.*

Lemma A.23 *With $\alpha^2 L \to 0$, $\forall \ell \in [L-1]$,*

$$B_{\ell+1}(-1+t) = 1 + (L-\ell)\frac{\sqrt{2}\alpha^2}{\pi}t^{1/2} + \mathcal{O}(t).$$

Proof Using lemma A.18, the expansion of B around -1 can be written in this regime as:

$$\begin{aligned} B_{\ell+1}(-1+t) &= \prod_{i=\ell+1}^{L}(1+\alpha^2\eta_i) = \prod_{i=\ell+1}^{L}\left(1+\frac{\sqrt{2}\alpha^2}{\pi}t^{1/2}\right) + \mathcal{O}(t) \\ &= \left(1+\frac{\sqrt{2}\alpha^2}{\pi}t^{1/2}\right)^{L-\ell} + \mathcal{O}(t) = 1 + (L-\ell)\frac{\sqrt{2}\alpha^2}{\pi}t^{1/2} + \mathcal{O}(t). \end{aligned}$$

■

We next prove the analogous to lemma A.9 for the "vanishing regime".

Lemma A.24 *For inputs in $\mathbb{S}^{d-1}$ and near -1, if $\alpha^2 L \ll 1$ then*

$$\boldsymbol{r}^{(L)}(-1+t) = c_{-1}t^{1/2} + o(t^{1/2})$$

with

$$c_{-1} = -\frac{1}{\sqrt{2}\pi}$$

Proof Rewrite (8) $\boldsymbol{r}^{(L)}(-1+t) = C\sum_{\ell=1}^{L} X_\ell Y_\ell$, where:

$$\begin{aligned} C &= \frac{1}{2L(1+\alpha^2)^{L-1}} \approx \frac{1}{2L} \\ X_\ell &= (1+\alpha^2)^{\ell-1}\kappa_1\left(\frac{K_{\ell-1}(-1+t)}{(1+\alpha^2)^{\ell-1}}\right) + K_{\ell-1}(-1+t)\kappa_0\left(\frac{K_{\ell-1}(-1+t)}{(1+\alpha^2)^{\ell-1}}\right) \\ &= (1+\alpha^2)^{\ell-1}\beta_\ell + K_{\ell-1}(-1+t)\eta_\ell \\ Y_\ell &= B_{\ell+1}(-1+t). \end{aligned}$$

Using $(1+\alpha^2) \approx 1$ and Lemmas A.19, A.21 and A.22

$$X_\ell = (1+\alpha^2)^{\ell-1}\beta_\ell + K_{\ell-1}(-1+t)\eta_\ell = -\frac{\sqrt{2}}{\pi}t^{1/2} + \mathcal{O}(t).$$

Using the above and Lemma A.23, we have

$$X_\ell Y_\ell = \left(-\frac{\sqrt{2}}{\pi}t^{1/2} + \mathcal{O}(t)\right)\left(1 + (L-\ell)\frac{\sqrt{2}\alpha^2}{\pi}t^{1/2} + \mathcal{O}(t)\right) = -\frac{\sqrt{2}}{\pi}t^{1/2} + \mathcal{O}(t).$$

Consequently,

$$\begin{aligned}\boldsymbol{r}^{(L)}(-1+t) &= C\sum_{\ell=1}^{L} X_\ell Y_\ell = C\sum_{\ell=1}^{L}\left(-\frac{\sqrt{2}}{\pi}t^{1/2} + \mathcal{O}(t)\right) \\ &= \frac{1}{2L}\left(-\frac{\sqrt{2}L}{\pi}t^{1/2}\right) + \mathcal{O}(t) = -\frac{1}{\sqrt{2}\pi}t^{1/2} + \mathcal{O}(t) = -\frac{1}{\sqrt{2}\pi}t^{1/2} + o(t^{1/2}).\end{aligned}$$

■

Note that with the conditions of $\alpha^2 L \to 0$ with $L \to \infty$, using Lemma A.4,

$$c_1 = -\frac{1+\alpha^2 L}{\sqrt{2}\pi(1+\alpha^2)} \xrightarrow{L\to\infty} -\frac{1}{\sqrt{2}\pi}.$$

This is indeed the case when $\alpha = L^{-\gamma}$ with $0.5 < \gamma \leq 1$. In this case we have from Lemma A.24 that $c_1 = c_{-1}$, implying that the odd frequencies decay faster than $\mathcal{O}(k^{-d})$. If however $\alpha = L^{-1/2}$ then for all L, $\alpha^2 L = 1$ and c_1 approaches $-\sqrt{2}/\pi$ and all the frequencies decay exactly at the rate of $\mathcal{O}(k^{-d})$. This also reflects the convergence of ResNTK as $L \to \infty$ to FC-NTK with $L = 1$. Note that this common value of c_1 and c_{-1} in the limit when $\alpha^2 L \to 0$ is identical to the value of the coefficients in the expansion of $\boldsymbol{k}^{(1)}$ near ± 1 for $L = 1$.

A.3 Proof of Theorem 1 in the bias-free case

We are now ready to prove Theorem 1.

Proof Lemmas A.4 and A.9 establish that for $L \geq 1$ ResNTK near ± 1 (outside the "vanishing regime") takes the forms of (16) and (17) with $\nu = 1/2$, satisfying the conditions of Theorem A.3. Moreover, clearly from these lemmas

$$|c_{-1}| \leq \frac{1}{\sqrt{2}\pi(1+\alpha^2)L} < \frac{1+\alpha^2 L}{\sqrt{2}\pi(1+\alpha^2)} = |c_1|.$$

Hence, $|c_1| \neq |c_{-1}|$. The eigenvalues of ResNTK, therefore, decay at the rate of k^{-d} both for the odd and even frequencies, proving Theorem 1. ■

Appendix B. Decay rate of eigenvalues of ResNTK with bias

Next we extend our analysis to the case that $\tau > 0$, thus proving Theorem 1 from the paper. Before proving the theorem, we will lay out some notations.

B.1 Normalization factor of ResNTK with bias

As mentioned before, it is a common practice to normalize the kernel such that $\boldsymbol{r}(1) = 1$. The kernel on the sphere is written as

$$\begin{aligned}\boldsymbol{r}^{(L)}(\mathbf{x},\mathbf{z}) = C_\tau \sum_{\ell=1}^{L} B_{\ell+1}(\mathbf{x},\mathbf{z}) \Bigg[&(1+\alpha^2)^{\ell-1}\kappa_1\left(\frac{K_{\ell-1}(u)}{(1+\alpha^2)^{\ell-1}}\right) \\ &+ (K_{\ell-1}(u)+\tau^2)\kappa_0\left(\frac{K_{\ell-1}(u)}{(1+\alpha^2)^{\ell-1}}\right)\Bigg]\end{aligned}$$

For $u = 1$:

$$\begin{aligned}K_\ell(1) &= (1+\alpha^2)^\ell \\ B_{L+1-\ell}(1) &= (1+\alpha^2)^\ell \iff B_{\ell+1}(1) = (1+\alpha^2)^{L-\ell} \\ \boldsymbol{r}^{(L)}(1) &= \sum_{\ell=1}^{L} B_{\ell+1}(1)\left[(1+\alpha^2)^{\ell-1}\kappa_1(1) + (K_{\ell-1}(1)+\tau^2)\kappa_0(1)\right] \\ &= \sum_{\ell=1}^{L}(1+\alpha^2)^{L-\ell}\left[(1+\alpha^2)^{\ell-1} + (1+\alpha^2)^{\ell-1} + \tau^2\right] \\ &= \sum_{\ell=1}^{L}[2(1+\alpha^2)^{L-1} + (1+\alpha^2)^{L-\ell}\tau^2] \\ &= 2L(1+\alpha^2)^{L-1} + \tau^2\frac{(1+\alpha^2)^L - 1}{\alpha^2}.\end{aligned}$$

Hence, to obtain $\boldsymbol{r}(1) = 1$, we multiply the kernel by $C_\tau = \left(2L(1+\alpha^2)^{L-1} + \tau^2\frac{(1+\alpha^2)^L-1}{\alpha^2}\right)^{-1}$.

B.2 Expansions around +1

Note that both $K_\ell(u)$ and $B_\ell(u)$ are not affected by bias. Hence, their asymptotic analysis does not change, compared to the bias-free case. The following lemma generalizes Lemma A.4 for the case with bias.

Lemma B.1 *For inputs in $\mathbb{S}^{d-1}$ and near +1, if $\alpha > 0$ and $L \geq 1$*

$$\boldsymbol{r}^{(L)}(1-t) = 1 + c_1 t^{1/2} + o(t^{1/2}),$$

where

$$c_1 = C_\tau\left(-\frac{\sqrt{2}}{\pi}L(1+\alpha^2)^{L-2}\left((1+\alpha^2 L) + \tau^2(1+\alpha^2)\right)\right).$$

Proof Similarly to the proof of lemma A.4, we rewrite $\boldsymbol{r}^{(L)}(1-t) = C_\tau\sum_{\ell=1}^{L} X_\ell Y_\ell$, where:

$$\begin{aligned}X_\ell &= (1+\alpha^2)^{\ell-1}\kappa_1\left(\frac{K_{\ell-1}(1-t)}{(1+\alpha^2)^{\ell-1}}\right) + K_{\ell-1}(1-t)\kappa_0\left(\frac{K_{\ell-1}(1-t)}{(1+\alpha^2)^{\ell-1}}\right) + \tau^2\kappa_0\left(\frac{K_{\ell-1}(1-t)}{(1+\alpha^2)^{\ell-1}}\right) \\ Y_\ell &= B_{\ell+1}(1-t).\end{aligned}$$

Using Lemmas A.6 and A.7, for small $t > 0$,

$$\begin{aligned}
X_\ell &= (1+\alpha^2)^{\ell-1}(1-t+o(t)) + ((1+\alpha^2)^{\ell-1}(1-t)+o(t))\left(1-\frac{\sqrt{2}}{\pi}t^{1/2}+o(t)\right) \\
&\quad + \tau^2\left(1-\frac{\sqrt{2}}{\pi}t^{1/2}+o(t)\right) \\
&= (1+\alpha^2)^{\ell-1}(1-t) + (1+\alpha^2)^{\ell-1}(1-t)\left(1-\frac{\sqrt{2}}{\pi}t^{1/2}\right) + \mathcal{O}(t) + \tau^2\left(1-\frac{\sqrt{2}}{\pi}t^{1/2}+o(t)\right) \\
&= (1+\alpha^2)^{\ell-1}(1-t)\left(2-\frac{\sqrt{2}}{\pi}t^{1/2}\right) + \mathcal{O}(t) + \tau^2\left(1-\frac{\sqrt{2}}{\pi}t^{1/2}+o(t)\right) \\
&= \left(2(1+\alpha^2)^{\ell-1}+\tau^2\right) - \frac{\sqrt{2}}{\pi}((1+\alpha^2)^{\ell-1}+\tau^2)t^{1/2} + o(t^{1/2}).
\end{aligned}$$

Using Lemma A.8 each term in the sum can be written as

$$\begin{aligned}
X_\ell Y_\ell &= \left[\left(2(1+\alpha^2)^{\ell-1}+\tau^2\right) - \frac{\sqrt{2}}{\pi}((1+\alpha^2)^{\ell-1}+\tau^2)t^{1/2}\right] \\
&\quad \left[(1+\alpha^2)^{L-\ell} - \frac{\alpha^2\sqrt{2}}{\pi}(1+\alpha^2)^{L-\ell-1}(L-\ell)\,t^{1/2}\right] + o(t^{1/2}) \\
&= \left(2(1+\alpha^2)^{\ell-1}+\tau^2\right)(1+\alpha^2)^{L-\ell} - \frac{\sqrt{2}}{\pi}[(2(1+\alpha^2)^{\ell-1}+\tau^2)\alpha^2(1+\alpha^2)^{L-\ell-1}(L-\ell) \\
&\quad +((1+\alpha^2)^{\ell-1}+\tau^2)(1+\alpha^2)^{L-\ell}]t^{1/2} + o(t^{1/2}) \\
&= \left(2(1+\alpha^2)^{L-1}+\tau^2(1+\alpha^2)^{L-\ell}\right) - \frac{\sqrt{2}}{\pi}\left(\tau^2\alpha^2(1+\alpha^2)^{L-\ell-1}(L-\ell)\right. \\
&\quad \left.+2\alpha^2(1+\alpha^2)^{L-2}(L-\ell) + (1+\alpha^2)^{L-1}+\tau^2(1+\alpha^2)^{L-\ell})t^{1/2}\right) + o(t^{1/2}) \\
&= \left(2(1+\alpha^2)^{L-1}+\tau^2(1+\alpha^2)^{L-\ell}\right) - \frac{\sqrt{2}}{\pi}\left(\tau^2\left[\alpha^2(1+\alpha^2)^{L-\ell-1}(L-\ell) + (1+\alpha^2)^{L-\ell}\right]\right. \\
&\quad \left.+2\alpha^2(1+\alpha^2)^{L-2}(L-\ell) + (1+\alpha^2)^{L-1}\right)t^{1/2} + o(t^{1/2}).
\end{aligned}$$

Summing over the layers, starting with the free term

$$C_\tau\sum_{\ell=1}^{L}\left(2(1+\alpha^2)^{L-1}+\tau^2(1+\alpha^2)^{L-\ell}\right) = C_\tau\left(2L(1+\alpha^2)^{L-1}+\tau^2\frac{(1+\alpha^2)^L-1}{\alpha^2}\right) = 1.$$

The term that includes $t^{1/2}$ is therefore

$$C_\tau\left(-\frac{\sqrt{2}}{\pi}[\alpha^2\tau^2(L-\ell)(1+\alpha^2)^{L-\ell-1} + (1+\alpha^2)^{L-\ell}((1+\alpha^2)^{\ell-1}+\tau^2) + 2\alpha^2(L-\ell)(1+\alpha^2)^{L-2}]\right)t^{1/2}$$

Summing over the layers we obtain

$$\begin{aligned} c_1 &= C_\tau \left(-\frac{\sqrt{2}}{\pi} \left((1+\alpha^2)^{L-1} \frac{L(1+\alpha^2 L)}{(1+\alpha^2)} + \tau^2 L (1+\alpha^2)^{L-1} \right) \right) \\ &= C_\tau \left(-\frac{\sqrt{2}}{\pi} L \left((1+\alpha^2)^{L-2} (1+\alpha^2 L) + \tau^2 (1+\alpha^2)^{L-1} \right) \right) \\ &= C_\tau \left(-\frac{\sqrt{2}}{\pi} L (1+\alpha^2)^{L-2} \left((1+\alpha^2 L) + \tau^2 (1+\alpha^2) \right) \right). \end{aligned}$$

■

B.3 Expansions around -1

Note that both $K_\ell(u)$ and $B_\ell(u)$ are not affected by bias. Hence, their asymptotic analysis does not change, compared to the no-bias case. As before, we need to distinguish between two regimes, (1) with $\alpha > 0$ such that $\alpha^2 L$ does not vanish as L grows, and (2) with $\alpha > 0$ and $\alpha^2 L \ll 1$.

B.3.1 $\alpha > 0$ such that $\alpha^2 L \not\ll 1$

The following lemma generalizes lemma A.9 for the case with bias.

Lemma B.2 *For inputs in $\mathbb{S}^{d-1}$ and near -1, if $\alpha > 0$ and $L \geq 2$ then*

$$\boldsymbol{r}^{(L)}(-1+t) = p_{-1}(t) + c_{-1} t^{1/2} + o(t^{1/2}),$$

with

$$|c_{-1}| < |c_1|.$$

Proof Similarly to the proof of lemma A.9, we rewrite $\boldsymbol{r}^{(L)}(-1+t) = C_\tau \sum_{\ell=1}^{L} X_\ell Y_\ell$, where

$$\begin{aligned} X_\ell &= (1+\alpha^2)^{\ell-1} \kappa_1 \left(\frac{K_{\ell-1}(-1+t)}{(1+\alpha^2)^{\ell-1}} \right) + K_{\ell-1}(-1+t) \kappa_0 \left(\frac{K_{\ell-1}(-1+t)}{(1+\alpha^2)^{\ell-1}} \right) + \tau^2 \kappa_0 \left(\frac{K_{\ell-1}(-1+t)}{(1+\alpha^2)^{\ell-1}} \right) \\ &= (1+\alpha^2)^{\ell-1} \beta_\ell + K_{\ell-1}(-1+t) \eta_\ell + \tau^2 \eta_\ell = (1+\alpha^2)^{\ell-1} \beta_\ell + \left(-1 + \tau^2 + \alpha^2 \sum_{j=0}^{\ell-1} (1+\alpha^2)^{j-1} \beta_j \right) \eta_\ell \\ Y_\ell &= B_{\ell+1}(-1+t). \end{aligned}$$

Using Lemma A.18 the sum can be written as

$$\sum_{\ell=1}^{L} X_\ell Y_\ell = \sum_{\ell=1}^{L} \left((1+\alpha^2)^{\ell-1} \beta_\ell + \left(-1 + \tau^2 + \alpha^2 \sum_{j=0}^{\ell-1} (1+\alpha^2)^{j-1} \beta_j \right) \eta_\ell \right) \prod_{i=\ell+1}^{L} (1+\alpha^2 \eta_i) + \mathcal{O}(t).$$

From Lemma A.16, there is a difference between $\ell = 1$ and $\ell \geq 2$. For $\ell = 1$:

$$X_1 Y_1 = \left((1+\alpha^2)^0 \beta_1 + \left(-1 + \tau^2 + \alpha^2 \sum_{j=0}^{0} (1+\alpha^2)^{j-1} \beta_j \right) \eta_1 \right) \prod_{i=1+1}^{L} (1+\alpha^2 \eta_i) + \mathcal{O}(t) =$$
$$(-1+\tau^2)\eta_1 \prod_{i=2}^{L} (1+\alpha^2\eta_i) + \mathcal{O}(t) = (-1+\tau^2) \left(\prod_{i=2}^{L} (1+\alpha^2\eta_i) \right) \frac{\sqrt{2}}{\pi} t^{1/2} + \mathcal{O}(t).$$

Using Lemma A.17 this simplifies to

$$\begin{aligned} X_1 Y_1 &= (-1+\tau^2) \left(\prod_{i=2}^{L} (1+\alpha^2(\tilde{d}_i + \mathcal{O}(t))) \right) \frac{\sqrt{2}}{\pi} t^{1/2} + \mathcal{O}(t) \\ &= (-1+\tau^2) \left(\prod_{i=2}^{L} (1+\alpha^2 \tilde{d}_i) \right) \frac{\sqrt{2}}{\pi} t^{1/2} + \mathcal{O}(t). \end{aligned}$$

For $\ell \geq 2$, using Lemmas A.15 and A.17

$$\begin{aligned} X_\ell Y_\ell &= \left((1+\alpha^2)^{\ell-1} \beta_\ell + \left(-1 + \tau^2 + \alpha^2 \sum_{j=0}^{\ell-1} (1+\alpha^2)^{j-1} \beta_j \right) \eta_\ell \right) \prod_{i=\ell+1}^{L} (1+\alpha^2 \eta_i) + \mathcal{O}(t) \\ &= \left((1+\alpha^2)^{\ell-1} (\tilde{c}_\ell + \mathcal{O}(t)) + \left(-1 + \tau^2 + \alpha^2 \sum_{j=0}^{\ell-1} (1+\alpha^2)^{j-1} (\tilde{c}_j + \mathcal{O}(t)) \right) (\tilde{d}_\ell + \mathcal{O}(t)) \right) \\ &\quad \prod_{i=\ell+1}^{L} (1+\alpha^2(\tilde{d}_i + \mathcal{O}(t))) + \mathcal{O}(t) \\ &= \left((1+\alpha^2)^{\ell-1} \tilde{c}_\ell + \left(-1 + \tau^2 + \alpha^2 \sum_{j=0}^{\ell-1} (1+\alpha^2)^{j-1} \tilde{c}_j \right) \tilde{d}_\ell \right) \prod_{i=\ell+1}^{L} (1+\alpha^2 \tilde{d}_i) + \mathcal{O}(t). \end{aligned}$$

The sum can be rewritten as

$$\sum_{\ell=1}^{L} X_\ell Y_\ell = \left(\sum_{\ell=2}^{L} \left((1+\alpha^2)^{\ell-1} \tilde{c}_\ell + \left(-1 + \tau^2 + \alpha^2 \sum_{j=0}^{\ell-1} (1+\alpha^2)^{j-1} \tilde{c}_j \right) \tilde{d}_\ell \right) \prod_{i=\ell+1}^{L} (1+\alpha^2 \tilde{d}_i) \right)$$
$$+ (-1+\tau^2) \left(\prod_{i=2}^{L} (1+\alpha^2 \tilde{d}_i) \right) \frac{\sqrt{2}}{\pi} t^{1/2} + \mathcal{O}(t).$$

The value of c_{-1}:

$$|c_{-1}| = \left| C_\tau (-1+\tau^2) \left(\prod_{i=2}^{L} (1+\alpha^2 \tilde{d}_i) \right) \frac{\sqrt{2}}{\pi} \right| \leq \left| C_\tau (-1+\tau^2) \left(\prod_{i=2}^{L} (1+\alpha^2) \right) \frac{\sqrt{2}}{\pi} \right| = C_\tau \left| (-1+\tau^2) \right| \frac{\sqrt{2}}{\pi} (1+\alpha^2)^{L-}$$

We want to show that $|c_{-1}| < |c_1|$ (note that $L \geq 1$). Recall that $c_1 = C_\tau (-\frac{\sqrt{2}}{\pi} L (1 + \alpha^2)^{L-2} \left((1+\alpha^2 L) + \tau^2 (1+\alpha^2) \right))$.

- $|\tau| > 1$:

$$C_\tau \left| (-1+\tau^2) \right| \frac{\sqrt{2}}{\pi} (1+\alpha^2)^{L-1} < C_\tau \tau^2 \frac{\sqrt{2}}{\pi} (1+\alpha^2)^{L-1} \leq C_\tau L \tau^2 \frac{\sqrt{2}}{\pi} (1+\alpha^2)^{L-1} \leq c_1.$$

- $|\tau| < 1$:

  $$C_\tau \left|(-1+\tau^2)\right| \frac{\sqrt{2}}{\pi}(1+\alpha^2)^{L-1} < C_\tau \frac{\sqrt{2}}{\pi}(1+\alpha^2)^{L-1} \leq C_\tau \frac{\sqrt{2}}{\pi}(1+\alpha^2)^{L-1}\frac{1+\alpha^2 L}{1+\alpha^2} \leq c_1.$$

  Where for $|\tau| = 1$ we get $c_{-1} = 0 \neq c_1$.

■

B.3.2 (Non) Vanishing regime $\alpha^2 L \ll 1$

Note that with bias, the odd eigenvalues do not vanish. For the case where $\alpha^2 L \to 0$ with $L \to \infty$ (which implies $(1+\alpha^2)^j \approx 1, \forall j \in [L]$), the analysis takes the following form. The next lemma generalizes lemma A.24.

Lemma B.3 *For inputs in $\mathbb{S}^{d-1}$ and near -1, if $\alpha^2 L \ll 1$ then*

$$\boldsymbol{r}^{(L)}(-1+t) = c_{-1} t^{1/2} + o(t^{1/2})$$

with

$$c_{-1} = C_\tau L \frac{1}{\sqrt{2}\pi}(-1+\tau^2)$$

Proof Similarly to the proof of lemma B.2, we rewrite $\boldsymbol{r}^{(L)}(-1+t) = C_\tau \sum_{\ell=1}^{L} X_\ell Y_\ell$, where:

$$\begin{aligned} X_\ell &= (1+\alpha^2)^{\ell-1}\kappa_1\left(\frac{K_{\ell-1}(-1+t)}{(1+\alpha^2)^{\ell-1}}\right) + K_{\ell-1}(-1+t)\kappa_0\left(\frac{K_{\ell-1}(-1+t)}{(1+\alpha^2)^{\ell-1}}\right) + \tau^2\kappa_0\left(\frac{K_{\ell-1}(-1+t)}{(1+\alpha^2)^{\ell-1}}\right) \\ &= (1+\alpha^2)^{\ell-1}\beta_\ell + K_{\ell-1}(-1+t)\eta_\ell + \tau^2\eta_\ell \\ Y_\ell &= B_{\ell+1}(-1+t). \end{aligned}$$

Using $(1+\alpha^2) \approx 1$ and Lemmas A.19, A.21 and A.22

$$X_\ell = (1+\alpha^2)^{\ell-1}\beta_\ell + K_{\ell-1}(-1+t)\eta_\ell + \tau^2\eta_\ell = (-1+\tau^2)\frac{\sqrt{2}}{\pi}t^{1/2} + \mathcal{O}(t).$$

Using the above and Lemma A.23, we have

$$X_\ell Y_\ell = \left(\left((-1+\tau^2)\frac{\sqrt{2}}{\pi}t^{1/2} + \mathcal{O}(t)\right)\left(1 + (L-\ell)\frac{\sqrt{2}\alpha^2}{\pi}t^{1/2} + \mathcal{O}(t)\right)\right) = (-1+\tau^2)\frac{\sqrt{2}}{\pi}t^{1/2} + \mathcal{O}(t).$$

Consequently,

$$\boldsymbol{r}^{(L)}(-1+t) = C_\tau \sum_{\ell=1}^{L} X_\ell Y_\ell = C_\tau \sum_{\ell=1}^{L}\left((-1+\tau^2)\frac{\sqrt{2}}{\pi}t^{1/2} + \mathcal{O}(t)\right) = C_\tau L(-1+\tau^2)\frac{\sqrt{2}}{\pi}t^{1/2} + o(t^{1/2})$$

Lemma B.4 *In this regime, $c_1 \neq c_{-1}$.*

Proof First, in this regime, we get:

$$c_1 = C_\tau(-\frac{\sqrt{2}}{\pi}L(1+\alpha^2)^{L-2}\left((1+\alpha^2 L)+\tau^2(1+\alpha^2)\right) \approx -C_\tau\frac{\sqrt{2}}{\pi}L\left(1+\tau^2\right)$$

We will show that the division does not equal to 1:

$$\frac{|c_1|}{|c_{-1}|} = \frac{\left|C_\tau\frac{\sqrt{2}}{\pi}L\left(1+\tau^2\right)\right|}{\left|C_\tau L(-1+\tau^2)\frac{\sqrt{2}}{\pi}\right|} = \frac{|1+\tau^2|}{|-1+\tau^2|}$$

This expression $\neq 1$ for any $\tau > 0$, hence $c_1 \neq c_{-1}$. ■

■

B.4 Proof of Theorem 1 from the paper

Theorem B.5 *The eigenvalues λ_k of ResNTK with bias, $\boldsymbol{r}(\mathbf{x},\mathbf{z})$, for $\mathbf{x},\mathbf{z} \in \mathbb{S}^{d-1}$ corresponding to a spherical harmonic eigenfunction $Y_{kl}(\mathbf{x})$ with frequency $k \geq 0$ and phase $1 \leq l \leq N(d,k)$ under the uniform measure decay at the rate of k^{-d} where k denotes frequency.*

Proof By combining Lemmas B.2, B.3 and B.4, the conditions of Theorem A.3 are satisfied, implying the desired decay. ■

Appendix C. Proof of Theorem 2 from the paper

Theorem C.1 *ResNTK $\boldsymbol{r}^{(L)}$ for residual networks with $L \in \mathbb{N}$ layers and the hyperparameter $\alpha = L^{-\gamma}$, $0.5 < \gamma \leq 1$, approaches the 2-layer FC-NTK uniformly in the interval $\mathbf{x}^T\mathbf{z} \in [-1,1]$, where $\mathbf{x},\mathbf{z} \in \mathbb{S}^{d-1}$; that is, let $\epsilon > 0$, $\forall L > c(\epsilon,\gamma)$*

$$|\boldsymbol{r}^{(L)}(\mathbf{x},\mathbf{z}) - \boldsymbol{k}^{(1)}(\mathbf{x},\mathbf{z})| \leq \epsilon.$$

Proof To show uniform convergence on the interval $[-1,1]$ we need to address the singularities of these kernels near the boundaries. Hence, we let $\delta = \frac{\epsilon^2}{4}$, and analyze separately the following intervals

- $\mathbf{x}^T\mathbf{z} \in [-1+\delta, 1-\delta]$
- $\mathbf{x}^T\mathbf{z} \in [1-\delta, 1]$
- $\mathbf{x}^T\mathbf{z} \in [-1, -1+\delta]$

The inner interval, $\mathbf{x}^T\mathbf{z} \in [-1+\delta, 1-\delta]$. We follow the ResNTK notations in Sec. A. We include an additional subscript L to emphasize the dependence of α on L. Let

$$u_{\ell,L} = \frac{K_{\ell,L}}{(1+\alpha^2)^\ell}, \quad u_0 = K_0 = \mathbf{x}^T\mathbf{z}$$

and assume that $-1+\delta < u_0 < 1-\delta$. Following these notations, and using Lemma A.2, we obtain the following relation

$$u_{\ell,L} = \frac{u_{\ell-1,L} + \alpha^2 \kappa_1(u_{\ell-1,L})}{1+\alpha^2}, \tag{24}$$

which implies that

$$u_{\ell,L} - u_{\ell-1,L} = \frac{\alpha^2}{1+\alpha^2}(\kappa_1(u_{\ell-1,L}) - u_{\ell-1,L}). \tag{25}$$

We note that $\kappa_0, \kappa_1 : [-1,1] \to [0,1]$ and $\kappa_1'(s) = \kappa_0(s)$, and therefore, the derivative of the function $\kappa_1(s) - s$ is non-positive, implying that $\kappa_1(s) - s$ is non-increasing. Therefore, the minimal value is attained at $s = 1$ and the maximal value at $s = -1$. Since $\kappa_1(1) - 1 = 0$ and $\kappa_1(-1) + 1 = 1$ this means that $0 \le \kappa_1(s) - s \le 1$. Now, by the relation (25), it is easy to see that $u_{\ell,L} \ge u_{\ell-1,L}$, which means that

$$u_0 \le u_{1,L} \le \ldots \le u_{L-1,L}. \tag{26}$$

In addition, we obtain the following upper bound for $u_{\ell,L} - u_0$

$$u_{\ell,L} - u_0 = \sum_{i=1}^{\ell}(u_{i,L} - u_{i-1,L}) = \frac{\alpha^2}{1+\alpha^2}\sum_{i=1}^{\ell}(\kappa_1(u_{i-1,L}) - u_{i-1,L}) \le \frac{\alpha^2}{1+\alpha^2}(\kappa_1(u_0) - u_0)\ell,$$

where the last inequality uses the observation $u_0 \le u_{i,L}$ and that $\kappa_1(s) - s$ is decreasing. The last inequality is equivalent to

$$u_{\ell,L} \le u_0 + \frac{\alpha^2}{1+\alpha^2}(\kappa_1(u_0) - u_0)\ell. \tag{27}$$

For $\alpha = L^{-\gamma}$, we have $\frac{\alpha^2}{1+\alpha^2} = \frac{1}{1+L^{2\gamma}}$, and since $0 \le \kappa_1(s) - s \le 1$ this inequality implies that

$$u_{L-1,L} \le u_0 + \frac{L}{1+L^{2\gamma}} \le 1 - \delta + L^{1-2\gamma}. \tag{28}$$

Therefore, for $\gamma > 0.5$ and L sufficiently large, this yields a maximal bound $1-\delta'$ over the series (26), with $\delta > \delta' > 0$. In particular, we can let $\delta' = \delta/2$ and then can bound the series from above by $1 - \delta + \delta/2 = 1 - \delta'$ for any L large enough.

Denote by

$$P_{\ell+1,L} = B_{\ell+1,L}(1+\alpha^2)^{-(L-\ell)} = \prod_{i=\ell}^{L-1}\frac{1+\alpha^2\kappa_0(u_{i,L})}{1+\alpha^2},$$

and note that $P_{l+1,L} \in (0,1]$. Since $1 - \frac{1+\alpha^2\kappa_0(u_{i,L})}{1+\alpha^2} = \frac{\alpha^2(1-\kappa_0(u_{i,L}))}{1+\alpha^2}$ and for $a_k \in [0,1]$, $1 - \prod_{k=1}^n(1-a_k) \le \sum_{k=1}^n a_k$ (see Lemma C.2), we obtain

$$\begin{aligned} 1 - P_{\ell+1,L} &= 1 - \prod_{i=\ell}^{L-1}\left(1 - \frac{\alpha^2(1-\kappa_0(u_{i,L}))}{1+\alpha^2}\right) \\ &\le \sum_{i=\ell}^{L-1}\frac{\alpha^2(1-\kappa_0(u_{i,L}))}{1+\alpha^2} = \frac{\alpha^2}{1+\alpha^2}\left(L - \ell - \sum_{i=\ell}^{L-1}\kappa_0(u_{i,L})\right). \end{aligned} \tag{29}$$

Using these notations, ResNTK on the sphere (8) can be written as

$$\boldsymbol{r}^{(L)} = \frac{1}{2L}\sum_{\ell=1}^{L} P_{\ell+1,L}(\kappa_1(u_{\ell-1,L}) + u_{\ell-1,L}\kappa_0(u_{\ell-1,L})). \tag{30}$$

We next bound the distance of each layer from $\kappa_1(u_0) + u_0\kappa_0(u_0)$ from above. In the derivation below we apply several times the mean value theorem, i.e., $\exists\ c \in [a,b]$, such that $\kappa_1(b) - \kappa_1(a) = \kappa_0(c)(b-a) \le \kappa_0(b)(b-a)$. This is valid since the derivative of κ_1 is κ_0. In addition, κ_0 is monotonic increasing, so any $c \in [a,b]$ can be replaced by b.

$$\begin{aligned}
&|P_{\ell+1,L}(\kappa_1(u_{\ell-1,L}) + u_{\ell-1,L}\kappa_0(u_{\ell-1,L})) - (\kappa_1(u_0) + u_0\kappa_0(u_0))| \\
&\le |P_{\ell+1,L}| \cdot |(\kappa_1(u_{\ell-1,L}) + u_{\ell-1,L}\kappa_0(u_{\ell-1,L})) - (\kappa_1(u_0) + u_0\kappa_0(u_0))| + |(\kappa_1(u_0) + u_0\kappa_0(u_0))| \cdot |1 - P_{\ell+1,L}| \\
&\le |\kappa_0(u_{\ell-1,L})(u_{\ell-1,L} - u_0)| + |\kappa_0(u_{\ell-1,L})u_{\ell-1,L} - \kappa_0(u_0)u_0| + |(\kappa_1(u_0) + u_0\kappa_0(u_0))| \cdot |1 - P_{\ell+1,L}|,
\end{aligned}$$

where the last inequality is because $0 < P_{\ell-1,L} \le 1$ and due to the mean value theorem. We next focus on the *first two terms*

$$\begin{aligned}
&|\kappa_0(u_{\ell-1,L})(u_{\ell-1,L} - u_0)| + |\kappa_0(u_{\ell-1,L})u_{\ell-1,L} - \kappa_0(u_0)u_0| \\
&\le |\kappa_0(u_{\ell-1,L})(u_{\ell-1,L} - u_0)| + |\kappa_0(u_{\ell-1,L})u_{\ell-1,L} - \kappa_0(u_{\ell-1,L})u_0 + \kappa_0(u_{\ell-1,L})u_0 - \kappa_0(u_0)u_0| \\
&\le |\kappa_0(u_{\ell-1,L})(u_{\ell-1,L} - u_0)| + |\kappa_0(u_{\ell-1,L})u_{\ell-1,L} - \kappa_0(u_{\ell-1,L})u_0| + |\kappa_0(u_{\ell-1,L})u_0 - \kappa_0(u_0)u_0| \\
&= 2|\kappa_0(u_{\ell-1,L})(u_{\ell-1,L} - u_0)| + |u_0(\kappa_0(u_{\ell-1,L}) - \kappa_0(u_0))| \\
&\le^1 2\kappa_0(u_{\ell-1,L})\frac{\alpha^2}{1+\alpha^2}(\kappa_1(u_0) - u_0)(\ell-1) + |u_0|(u_{l-1,L} - u_0)\kappa_0'(c_{l-1,L}) \\
&= 2\kappa_0(u_{\ell-1,L})\frac{\alpha^2}{1+\alpha^2}(\kappa_1(u_0) - u_0)(\ell-1) + |u_0|(u_{l-1,L} - u_0)\frac{1}{\pi\sqrt{1 - c_{\ell-1,L}^2}} \\
&\le^2 2\kappa_0(u_{\ell-1,L})\frac{\alpha^2}{1+\alpha^2}(\kappa_1(u_0) - u_0)(\ell-1) + \frac{|u_0|(\kappa_1(u_0) - u_0)(\ell-1)}{\pi\sqrt{1 - c_{\ell-1,L}^2}}\frac{\alpha^2}{1+\alpha^2}
\end{aligned}$$

where $\le^1$ is obtained by applying (27) and the mean value theorem for κ_0 with $c_{l-1,L} \in [u_0, u_{l-1,L}]$, and $\le^2$ too is obtained by applying (27).

Third term (29) and the monotonicity of κ_0 yield

$$\begin{aligned}
&|(\kappa_1(u_0) + u_0\kappa_0(u_0))| \cdot |1 - P_{\ell+1,L}| \le |(\kappa_1(u_0) + u_0\kappa_0(u_0))| \cdot \frac{\alpha^2}{1+\alpha^2}(L - \ell - \sum_{i=\ell}^{L-1}\kappa_0(u_{i,L})) \\
&\le |(\kappa_1(u_0) + u_0\kappa_0(u_0))| \cdot \frac{\alpha^2}{1+\alpha^2}(L-\ell)(1 - \kappa_0(u_0))
\end{aligned}$$

To recap, the upper bound for each layer is

$$\begin{aligned}
&|P_{\ell+1,L}(\kappa_1(u_{\ell-1,L}) + u_{\ell-1,L}\kappa_0(u_{\ell-1,L})) - (\kappa_1(u_0) + u_0\kappa_0(u_0))| \\
&\le 2\kappa_0(u_{\ell-1,L})\frac{\alpha^2}{1+\alpha^2}(\kappa_1(u_0) - u_0)(\ell-1) + \frac{|u_0|(\kappa_1(u_0) - u_0)(\ell-1)}{\pi\sqrt{1 - c_{\ell-1,L}^2}}\frac{\alpha^2}{1+\alpha^2} \\
&+ |(\kappa_1(u_0) + u_0\kappa_0(u_0))| \cdot \frac{\alpha^2}{1+\alpha^2}(L-\ell)(1 - \kappa_0(u_0)).
\end{aligned} \tag{31}$$

We would like next to derive a bound for the entire kernel, i.e., to bound from above the following expression

$$
\begin{aligned}
&|\boldsymbol{r}^{(L)}(u_0) - \boldsymbol{k}^{(1)}(u_0)| = \left| \frac{1}{2L} \sum_{\ell=1}^{L} \left\{ P_{\ell+1,L}(\kappa_1(u_{\ell-1,L}) + u_{\ell-1,L}\kappa_0(u_{\ell-1,L})) \right\} - \frac{1}{2}(\kappa_1(u_0) + u_0\kappa_0(u_0)) \right| \\
&= \left| \frac{1}{2L} \sum_{l=1}^{L} \left\{ P_{\ell+1,L}(\kappa_1(u_{\ell-1,L}) + u_{\ell-1,L}\kappa_0(u_{\ell-1,L})) - (\kappa_1(u_0) + u_0\kappa_0(u_0)) \right\} \right| \\
&\leq^3 \frac{1}{2L} \frac{\alpha^2}{1+\alpha^2} \sum_{\ell=1}^{L} \left\{ 2\kappa_0(u_{\ell-1,L})(\kappa_1(u_0) - u_0)(\ell-1) + \frac{|u_0|(\kappa_1(u_0) - u_0)(\ell-1)}{\pi\sqrt{1 - c_{\ell-1,L}^2}} \right. \\
&\left. + |(\kappa_1(u_0) + u_0\kappa_0(u_0))|(L-\ell)(1-\kappa_0(u_0)) \right\} \\
&\leq^4 \frac{1}{2L} \frac{\alpha^2}{1+\alpha^2} \sum_{\ell=1}^{L} \left(2(\kappa_1(u_0) - u_0)(\ell-1) + \frac{|u_0|(\kappa_1(u_0) - u_0)(\ell-1)}{\pi\sqrt{1-(1-\delta')^2}} \right) \\
&+ \frac{1}{2L} \frac{\alpha^2}{1+\alpha^2} |(\kappa_1(u_0) + u_0\kappa_0(u_0))|(1-\kappa_0(u_0)) \frac{L(L-1)}{2} \\
&= \frac{L(L-1)}{2} \frac{1}{2L} \frac{\alpha^2}{1+\alpha^2} \left[2(\kappa_1(u_0) - u_0) + \frac{|u_0|(\kappa_1(u_0) - u_0)}{\pi\sqrt{1-(1-\delta')^2}} + |(\kappa_1(u_0) + u_0\kappa_0(u_0))|(1-\kappa_0(u_0)) \right] \\
&= \frac{L-1}{4} \frac{\alpha^2}{1+\alpha^2} \left[2(\kappa_1(u_0) - u_0) + \frac{|u_0|(\kappa_1(u_0) - u_0)}{\pi\sqrt{1-(1-\delta')^2}} + |(\kappa_1(u_0) + u_0\kappa_0(u_0))|(1-\kappa_0(u_0)) \right],
\end{aligned}
$$

where $\leq^3$ is directly by applying (31), and $\leq^4$ relies on the fact that $0 \leq \kappa_0(s) \leq 1$ and the following argument. We would like to bound from above the term $\frac{1}{\sqrt{1-c_{l-1,L}^2}}$ for $c_{l-1,L} \in [u_0, u_{l-1,L}]$. Since we have

$$-1 + \delta' \leq -1 + \delta \leq u_0 \leq \ldots \leq u_{L-1,L} \leq 1 - \delta \leq 1 - \delta',$$

it follows that $\frac{1}{\sqrt{1-c_{l-1,L}^2}} \leq \frac{1}{\sqrt{1-(1-\delta')^2}}$.

Since for $\alpha = L^{-\gamma}$ we have $\frac{\alpha^2}{1+\alpha^2} = \frac{1}{1+L^{2\gamma}}$ we obtain

$$
\begin{aligned}
&|\boldsymbol{r}^{(L)}(u_0) - \boldsymbol{k}^{(1)}(u_0)| \leq \\
&\frac{L-1}{4} \frac{1}{1+L^{2\gamma}} \left[2(\kappa_1(u_0) - u_0) + \frac{|u_0|(\kappa_1(u_0) - u_0)}{\pi\sqrt{1-(1-\delta')^2}} + |(\kappa_1(u_0) + u_0\kappa_0(u_0))| \cdot (1+\kappa_0(u_0)) \right] \leq \\
&L^{1-2\gamma} \left[2(\kappa_1(u_0) - u_0) + \frac{|u_0|(\kappa_1(u_0) - u_0)}{\pi\sqrt{1-(1-\delta')^2}} + |(\kappa_1(u_0) + u_0\kappa_0(u_0))| \cdot (1+\kappa_0(u_0)) \right] \leq^1 \\
&L^{1-2\gamma} \left[2(2-\delta) + \frac{(1+\delta)(2-\delta)}{\pi\sqrt{1-(1-\frac{\delta}{2})^2}} + (2+\delta) \cdot 2 \right] = \tilde{c}(\delta) L^{1-2\gamma},
\end{aligned}
$$

where $\leq^1$ holds since $\kappa_0, \kappa_1 \in [0,1]$ and in this interval $u_0 \in [-1+\delta, 1-\delta]$ (recall that $\delta' = \frac{\delta}{2}$). Using the relation $\delta = \frac{\epsilon^2}{4}$, we define $c(\epsilon) = \frac{\tilde{c}(\frac{\epsilon^2}{4})}{\epsilon}$. Therefore, given $\epsilon > 0$, $\forall L > c(\epsilon)^{\frac{1}{2\gamma-1}} c(\epsilon, \gamma)$ it holds that $|\boldsymbol{r}^{(L)}(u_0) - \boldsymbol{k}^{(1)}(u_0)| \leq \epsilon$.

The interval on the right, $\mathbf{x}^T\mathbf{z} \in [1-\delta, 1]$. Since the kernels have a singular point at $u_0 = 1$, we will use their asymptotes (20) and (32) with $t \to 0$.

$$\begin{aligned}
\left|\boldsymbol{r}^{(L)}(1-t) - \boldsymbol{k}^{(1)}(1-t)\right| &\leq \left|1 - \frac{1+\alpha^2 L}{\sqrt{2}\pi(1+\alpha^2)}t^{1/2} + o(t^{1/2}) - \left(1 - \frac{1}{\sqrt{2}\pi}t^{1/2} + o(t^{1/2})\right)\right| \\
&= \left|-\frac{1+\alpha^2 L}{\sqrt{2}\pi(1+\alpha^2)}t^{1/2} + \frac{1}{\sqrt{2}\pi}t^{1/2} + o(t^{1/2})\right| \\
&= \left|\frac{(1+\alpha^2) - (1+\alpha^2 L)}{\sqrt{2}\pi(1+\alpha^2)}t^{1/2} + o(t^{1/2})\right| \\
&= \left|\frac{\alpha^2(1-L)}{\sqrt{2}\pi(1+\alpha^2)}t^{1/2} + o(t^{1/2})\right| \\
&\leq \left|\frac{\alpha^2(1-L)}{\sqrt{2}\pi(1+\alpha^2)}\delta^{1/2} + \delta^{1/2}\right| \\
&\leq \delta^{1/2}\left(\frac{\alpha^2(L-1)}{\sqrt{2}\pi(1+\alpha^2)} + 1\right).
\end{aligned}$$

For uniform convergence we require

$$\delta^{1/2}\left(\frac{\alpha^2(L-1)}{\sqrt{2}\pi(1+\alpha^2)} + 1\right) < \epsilon \iff \frac{\alpha^2(L-1)}{1+\alpha^2} \leq \sqrt{2}\pi\frac{\epsilon - \delta^{1/2}}{\delta^{1/2}} =^1 \sqrt{2}\pi \iff$$
$$\frac{L-1}{1+L^{2\gamma}} \leq \sqrt{2}\pi \iff -(1+\sqrt{2}\pi) < L^{2\gamma}(\sqrt{2}\pi - L^{1-2\gamma}),$$

where $=^1$ is obtained by plugging in $\delta = \frac{\epsilon^2}{4}$. Since $L > 0$ (and therefore $L^{2\gamma} > 0$), we only need to make sure that the RHS is positive. This happens $\forall L \geq 1 > (\frac{1}{\sqrt{2}\pi})^{\frac{1}{2\gamma-1}}$. Therefore, it holds that $\forall L \geq 1$, $|\boldsymbol{r}^{(L)}(1-t) - \boldsymbol{k}^{(1)}(1-t)| \leq \epsilon$.

The interval on the left, $\mathbf{x}^T\mathbf{z} \in [-1, -1+\delta]$. Since the kernels are singular at $u_0 = -1$ we again use their Taylor expansions. Note that we are in the vanishing regime, $\alpha^2 L \ll 1$, so for ResNTK we will use Lemma A.24. For FC-NTK we use the expansion from Bietti and Bach (2021).

$$\begin{aligned}
\left|\boldsymbol{r}^{(L)}(-1+t) - \boldsymbol{k}^{(1)}(-1+t)\right| &\leq \left|-\frac{1}{\sqrt{2}\pi}t^{1/2} + o(t^{1/2}) - \left(-\frac{\sqrt{2}}{\pi}t^{1/2} + o(t^{1/2})\right)\right| \\
&= \left|\frac{1}{\sqrt{2}\pi}t^{1/2} + o(t^{1/2})\right| \\
&\leq \frac{1}{\sqrt{2}\pi}\delta^{1/2} + \delta^{1/2} \leq 2\delta^{1/2} =^1 \epsilon,
\end{aligned}$$

where $=^1$ is due to $\delta = \epsilon^2/4$. Combining the proofs for the three intervals the theorem is proven. ■

Lemma C.2 *For $a_k \in [0,1]$, it holds that $1 - \prod_{k=1}^n (1-a_k) \leq \sum_{k=1}^n a_k$.*

Proof By induction. The lemma holds trivially for $k = 1$. Assume the lemma holds for $k \leq n-1$, then

$$\begin{aligned} 1 - \prod_{k=1}^{n}(1-a_k) &= 1 - (1-a_n)\left(\prod_{k=1}^{n-1}(1-a_k)\right) = 1 - \prod_{k=1}^{n-1}(1-a_k) + a_n \prod_{k=1}^{n-1}(1-a_k) \\ &\leq \sum_{k=1}^{n-1} a_k + a_n \prod_{k=1}^{n-1}(1-a_k) \leq \sum_{k=1}^{n} a_k. \end{aligned}$$

■

Appendix D. Steepness of FC-NTK

In this section, we analyze the asymptotic relations between the Laplace and FC-NTK kernels.

Lemma D.1 *Bietti and Bach (2021) With small $t > 0$,*

$$\boldsymbol{k}_{Lap}(1-t) = e^{-c\sqrt{2t}} = 1 - c\sqrt{2t} + \mathcal{O}(t),$$

where $\boldsymbol{k}_{Lap}$ is defined in equation (8) in the paper.

We next prove the Taylor expansion of deep fully connected networks.[1]

Lemma D.2 *With small $t > 0$*

$$\boldsymbol{k}^{(L)}(1-t) = 1 - \frac{L}{\pi\sqrt{2}} t^{1/2} + o(t^{1/2}). \tag{32}$$

Therefore, with $c = \frac{L}{2\pi}$, $\boldsymbol{k}^{(L)}(1-t) - \boldsymbol{k}_{Lap}(1-t) = o(t^{1/2})$.

Proof The proof is by induction on the unnormalized kernel $\tilde{\boldsymbol{k}}^{(\ell)} = (\ell+1)\boldsymbol{k}^{(\ell)}$. With $\ell = 1$:

$$\begin{aligned} \tilde{\boldsymbol{k}}^{(1)}(1-t) &= (1-t)\kappa_0(1-t) + \kappa_1(1-t) = (1-t)\left(1 - \frac{\sqrt{2}}{\pi} t^{1/2} + \mathcal{O}(t^{3/2})\right) + 1 + \mathcal{O}(t) \\ &= 2 - \frac{\sqrt{2}}{\pi} t^{1/2} + o(t^{1/2}). \end{aligned}$$

Note that by the definition of $\tilde{\boldsymbol{k}}^{(\ell)}$

$$\tilde{\boldsymbol{k}}^{(\ell)}(u) = \tilde{\boldsymbol{k}}^{(\ell-1)}(u)\kappa_0(\Sigma^{(\ell-1)}(u)) + \Sigma^{(\ell)}(u).$$

Using

$$\Sigma^{(\ell)}(1-t) = 1 - t + o(t),$$

1. Note that here we fix a slight miscalculation in Bietti and Bach (2021)(Corollary 3) which implied that the coefficient of $t^{1/2}$ is constant with depth.

that was proved in Bietti and Bach (2021). Additionally, using the equation above and Lemma A.10

$$\kappa_0(\Sigma^{(\ell-1)}(1-t)) = \kappa_0(1-t+o(t)) = 1-\frac{\sqrt{2}}{\pi}(t+o(t))^{1/2}+o(t^{1/2}) = 1-\frac{\sqrt{2}}{\pi}t^{1/2}+o(t^{1/2}).$$

Suppose the lemma holds for $j \leq \ell-1$, then

$$\begin{aligned}\tilde{\boldsymbol{k}}^{(\ell)}(1-t) &= \tilde{\boldsymbol{k}}^{(\ell-1)}(1-t)\kappa_0(\Sigma^{(\ell-1)}(1-t)) + \Sigma^{(\ell)}(1-t) \\ &= \ell\left(1-\frac{\ell-1}{\pi\sqrt{2}}t^{1/2}+o(t^{1/2})\right)\left(1-\frac{\sqrt{2}}{\pi}t^{1/2}+o(t^{1/2})\right)+1-t+o(t) \\ &= \ell+1-\frac{\ell(\ell+1)}{\pi\sqrt{2}}t^{1/2}+o(t^{1/2}).\end{aligned}$$

Using $\boldsymbol{k}^{(L)} = \frac{1}{L+1}\tilde{\boldsymbol{k}}^{(L)}$, the first part of the lemma is proven. Finally, using Lemma D.1, the relation to the Laplace kernel is immediate. ■

Hence, with $c = L/(2\pi)$, $\boldsymbol{k}^{(L)}(1-t) - \boldsymbol{k}_{Lap}(1-t) = o(t^{1/2})$, implying that deep FC-NTK becomes steeper near 1.

Appendix E. Implication on generalization

In this section we prove Theorem 3 from the paper.

Theorem E.1 *Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be n i.i.d samples such that $\{\mathbf{x}_i\}_{i=1}^n$ are drawn from the uniform distribution on $\mathbb{S}^{d-1}$ and assuming that $y \in \mathcal{H}_r(\mathbb{S}^{d-1})$. Then,*

1. *There exists L_0 such that $\forall L > L_0$ it holds that with probability at least $1-\delta$, the expected risk of the ResNTK with depth L and $0.5 < \gamma \leq 1$ is upper bounded by*

$$\mathbb{E}(\mathcal{L}(f_{\text{ResNTK}}(\mathbf{x}), y)) \leq O\left(\frac{r^{\frac{3d-2}{2}}}{\sqrt{n}} + \sqrt{\frac{log(1/\delta)}{n}}\right).$$

2. *$\forall \epsilon > 0$ there exists $L_0 = L(\epsilon, n)$, such that $\forall L > L_0$, the NTK predictor for $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ and depth L satisfies almost surely*

$$\mathbb{E}(\mathcal{L}(f_{\text{FC}-\text{NTK}}(\mathbf{x}), y)) = \int_{\mathbb{S}^{d-1}} (f_{\text{FC}-\text{NTK}}(\mathbf{x}) - y(\mathbf{x}))^2 d\mathbf{x} \geq 1 - O(\epsilon).$$

For the proof we consider the following setting. Let $(\mathbf{x}, y) \backsim D$, where D is some underlying distribution over $(\mathbf{x}, y) \in \mathbb{S}^{d-1} \times \mathbb{R}$. The generalization error is defined as the expected risk with expectation taken over new test points sampled from the same distribution D. For a given predictor $f(\mathbf{x})$ and target function $y = y(\mathbf{x})$ we use the truncated ℓ_2 loss

$$L(f(\mathbf{x}), y) = \min(f(\mathbf{x}) - y, c)^2$$

and define the expected risk for $c > 0$ as

$$\mathbb{E}_{(\mathbf{x},y)\backsim D}(L(f(\mathbf{x}),y)) = \int_{\mathbb{S}^{d-1}} L(f(\mathbf{x}),y(\mathbf{x}))d\mathbf{x}.$$

Our analysis assumes that the distribution of $\mathbf{x} \in \mathbb{R}^d$ is uniform on the sphere and that $y(\mathbf{x}) \in \mathcal{H}_r(\mathbb{S}^{d-1})$, where $\mathcal{H}_r(\mathbb{S}^{d-1})$ is the set of band-limited functions on $\mathbb{S}^{d-1}$ with maximal frequency r. For simplicity we also assume that $\mathrm{Var}(y(\mathbf{x})) = 1$ and that $y(\mathbf{x}) = O(1)$. We analyze a case where we draw n i.i.d samples from the distribution, build $f_{NTK}(\mathbf{x}), f_{ResNTK}(\mathbf{x})$ so our bound will depend on the probability of the random draw.

In the following analysis, we show that for any number of samples n and large enough depth L it holds with high probability that $\mathbb{E}_{(\mathbf{x},y)\backsim D}(L(f_{ResNTK}(\mathbf{x}),y)) = O\left(\frac{r^{(3d-2)/2}}{\sqrt{n}}\right)$ while, in contrast, $\mathbb{E}_{(\mathbf{x},y)\sim D}(L(f_{NTK}(\mathbf{x}),y)) \approx 1$.

We begin by citing a fundamental result from Bartlett and Mendelson (2002); Arora et al. (2019a).

Theorem E.2 *Bartlett and Mendelson (2002) Given training data $\{(\mathbf{x}_i,y_i)\}_{i=1}^n$ drawn i.i.d. from a distribution D and a kernel $k(\cdot,\cdot) : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R}$, consider any loss function $l : \mathbb{R} \times \mathbb{R} \to [0,c]$ that is ρ-Lipschitz in the first argument such that $l(y,y) = 0$. With probability at least $1-\delta$, the expected risk of the kernel predictor can be upper bounded by*

$$\mathbb{E}_{(\mathbf{x},y)\backsim D}(l(f_{ResNTK},y)) = O\left(2\rho\frac{\sqrt{\mathbf{y}^T\Theta^{-1}\mathbf{y}\cdot trace(\Theta)}}{n} + c\sqrt{\frac{log(1/\delta)}{n}}\right),$$

where $\Theta_{ij} = k(\mathbf{x}_i,\mathbf{x}_j)$, and $\mathbf{y} = (y_1,..,y_n)^T$.

Next, to prove part 1 of Theorem E.1 we use the following supporting lemmas.

Lemma E.3 *Let $k(\mathbf{x},\mathbf{z}) = \sum_{k\geq 0}\lambda_k\sum_{i=1}^{N(d,k)} Y_{ki}(\mathbf{x})Y_{ki}(\mathbf{z})$ where $\{Y_{ki}(\cdot)\}$ are the spherical harmonics basis. Then:*

1. *The kernel matrix Θ where $\Theta_{ij} = k(\mathbf{x}_i,\mathbf{x}_j)$ can be written as*

$$\Theta = \sum_{k=0}^{\infty}\lambda_k\sum_{i=1}^{N(d,k)} Z_{ki}(X)Z_{ki}(X)^T,$$

   where $Z_{ki}(X) = (Y_{ki}(\mathbf{x}_1),..,Y_{ki}(\mathbf{x}_n))^T \in \mathbb{R}^n$.

2. *For $\bar{\Theta}_{(ki)} = \lambda_k Z_{ki}(X)Z_{ki}(X)^T$ it holds that $\Theta \succcurlyeq \bar{\Theta}_{(ki)}$. Moreover,*

$$\bar{\Theta}^{\dagger}_{(ki)} \succcurlyeq P_{\bar{\Theta}_{(ki)}}\Theta^{-1}P_{\bar{\Theta}_{(ki)}},$$

   where $P_{\bar{\Theta}_{(ki)}} = \bar{\Theta}^{\frac{1}{2}}_{(ki)}\bar{\Theta}^{\dagger}_{(ki)}\bar{\Theta}^{\frac{1}{2}}_{(ki)}$ and $\bar{\Theta}^{\dagger}_{(ki)}$ denotes the pseudo inverse of $\bar{\Theta}_{(ki)}$.

3. *Let $\mathbf{y}_{ki} = (Y_{ki}(\mathbf{x}_1),..,Y_{ki}(\mathbf{x}_n))^T$. Then, $P_{\bar{\Theta}_{(ki)}}\mathbf{y}_{ki} = \mathbf{y}_{ki}$.*

Proof

1. This is straightforward from the definition of the kernel.
2. Note that every term of the form $Z_{ki}(X)Z_{ki}(X)^T$ is a PSD matrix, therefore $\bar{\Theta}_{(ki)}$ is generated from Θ by omitting PSD terms meaning that $\bar{\Theta}_{(ki)} \succcurlyeq \Theta$. Moreover, the inequality
$$\bar{\Theta}^{\dagger}_{(ki)} \succcurlyeq P_{\bar{\Theta}_{(ki)}} \Theta^{-1} P_{\bar{\Theta}_{(ki)}}$$
was proved in Arora et al. (2019a) (Lemma E.1).
3. Since $\mathbf{y}_{ki} = Z_{ki}(X)$ and $P_{\bar{\Theta}_{(ki)}}$ is a projection to the column space of $\Theta_{(ki)}$ it holds that $P_{\bar{\Theta}_{(ki)}} \mathbf{y}_{ki} = \mathbf{y}_{ki}$.

■

Lemma E.4 *Let $r, d \in \mathbb{N}$ be given. Then, $\sum_{k=1}^{r} \sum_{i=0}^{N(d,k)} 1 = O(r^{d-1})$, where $N(d,k)$ is the number of harmonics of degree k in $\mathbb{S}^{d-1}$.*

Proof On $\mathbb{S}^{d-1}$the number of harmonics of degree k are $N(d,k) = \frac{2k+d-2}{k}\binom{k+d-3}{d-2} = O(k^{d-2})$ where the O notation holds for a fixed d. Therefore it is enough to compute $\sum_{k=1}^{r} k^{d-2}$. We estimate this finite sum by lower and upper bounds obtained by integrals on the function $f(\mathbf{x}) = \mathbf{x}^{d-2}$, obtaining
$$\frac{r^{d-1}}{d-1} = \int_0^r \mathbf{x}^{d-2} d\mathbf{x} \leq \sum_1^r k^{d-2} \leq \int_1^{r+1} \mathbf{x}^{d-2} d\mathbf{x} = \frac{(r+1)^{d-1} - 1}{d-1}.$$

■

Proof (Of Thm E.1, part 1) From theorem E.2 we have
$$\mathbb{E}_{(\mathbf{x},y)\backsim D}(L(f_{ResNTK}, y)) = O\left(\frac{\sqrt{\mathbf{y}^T \Theta^{-1} \mathbf{y} \cdot trace(\Theta)}}{n} + \sqrt{\frac{log(1/\delta)}{n}}\right), \tag{33}$$
where $\Theta_{ij} = \boldsymbol{r}^{(L)}(\mathbf{x}_i^T \mathbf{x}_j)$. Since the ResNTK is normalized (i.e $\boldsymbol{r}^{(L)}(\mathbf{x}_i^T \mathbf{x}_i) = 1$), $trace(\Theta) = n$, and (33) becomes:
$$\mathbb{E}_{(\mathbf{x},y)\backsim D}(L(f_{ResNTK}, y)) = O\left(\sqrt{\frac{\mathbf{y}^T \Theta^{-1} \mathbf{y}}{n}} + \sqrt{\frac{log(1/\delta)}{n}}\right).$$
Let $\bar{\mathbf{y}} = \sum_{k=0}^{r} \sum_{i=1}^{N(d,k)} \alpha_{ki} \mathbf{y}_{ki}$ where $\mathbf{y}_{ki} = (Y_{ki}(\mathbf{x}_1), .., Y_{ki}(\mathbf{x}_n))^T$. It holds that
$$\sqrt{\frac{\bar{\mathbf{y}}^T \Theta^{-1} \bar{\mathbf{y}}}{n}} = \sqrt{\frac{\sum_{ki} \sum_{k'i'} \alpha_{ki} \alpha_{k'i'} \mathbf{y}_{ki}^T \Theta^{-1} \mathbf{y}_{k'i'}}{n}} \leq O\left(r^{d-1} \max_{k,i}\left(\frac{|\alpha_{ki}| \sqrt{\mathbf{y}_{ki}^T \Theta^{-1} \mathbf{y}_{ki}}}{\sqrt{n}}\right)\right) \tag{34}$$

where r^{d-1} results from the number of Spherical Harmonics of frequency $\leq r$, which is given by Lemma E.4. Using Lemma E.3 we have that

$$\sqrt{\frac{\mathbf{y}_{ki}^T \Theta^{-1} \mathbf{y}_{ki}}{n}} = \sqrt{\frac{\mathbf{y}_{ki}^T P_{\Theta_{(ki)}} \Theta^{-1} P_{\Theta_{(ki)}} \mathbf{y}_{ki}}{n}} \leq \sqrt{\frac{\mathbf{y}_{ki}^T \bar{\Theta}_{(ki)}^{\dagger} \mathbf{y}_{ki}}{n}},$$

where $\bar{\Theta}_{(ki)}^{\dagger}$ denotes the pseudo-inverse of $\bar{\Theta}_{(ki)}$. Observing that $\bar{\Theta}_{(ki)} = \lambda_k \mathbf{y}_{ki} \mathbf{y}_{ki}^T$, we get that

$$\bar{\Theta}_{(ki)}^{\dagger} = \frac{1}{\|\mathbf{y}_{ki}\|^4 \lambda_k} \mathbf{y}_{ki} \mathbf{y}_{ki}^T.$$

We therefore obtain

$$\sqrt{\frac{\mathbf{y}_{ki}^T \bar{\Theta}_{(ki)}^{\dagger} \mathbf{y}_{ki}}{n}} = \frac{1}{\sqrt{n\lambda_k}}.$$

Plugging results from Theorem 3 in the paper, we have that $\forall k > 0, \lambda_k > 0$ and that $\max_{k \leq r} \frac{1}{\lambda_k} = O(r^d)$. Using (34) we get the final bound of

$$\mathbb{E}_{(\mathbf{x},y)\backsim D}(L(f_{ResNTK}, y)) = O\left(\frac{r^{\frac{3d-2}{2}}}{\sqrt{n}} + \sqrt{\frac{log(1/\delta)}{n}} \right).$$

■

Next, we use the following supporting lemma to prove Theorem E.1, part 2.

Lemma E.5 *Let $\mathbf{x}_1, .., \mathbf{x}_n, \mathbf{x} \in \mathbb{S}^{d-1}$. Fix $\epsilon, \delta > 0$. If $\forall i \neq j,\ |1 - \mathbf{x}_i^T \mathbf{x}_j| > \delta,\ |1 - \mathbf{x}_i^T \mathbf{x}| > \delta$ then there exists a depth $L_0 = L(n, \delta, \epsilon)$ such that for any $L > L_0$ it holds that*

$$|f_{NTK}(\mathbf{x}) - c_n \mathbf{1}^T \mathbf{y}| < \epsilon$$

where $\mathbf{y}_i = y_i$ and $c_n = \left(\frac{1}{3} - \frac{n}{3(3+n)}\right)$.

Proof Recall that given a training set $(\mathbf{x}_1, y_1), .., (\mathbf{x}_n, y_n)$ the kernel regression predictor is of the form

$$f_{NTK}(\mathbf{x}) = K(\mathbf{x}, X) K^{-1}(X, X) \mathbf{y},$$

where the matrix $K(X, X)$ and the vector $K(\mathbf{x}, X)$ are respectively defined as $K(X, X)_{ij} = \boldsymbol{k}^{(L)}(\mathbf{x}_i, \mathbf{x}_j)$ and $K(\mathbf{x}, X)_i = \boldsymbol{k}^{(L)}(\mathbf{x}, \mathbf{x}_i)$. By Theorem 5 in Huang et al. (2020) we have that for $\mathbf{x}, \mathbf{z} \in \mathbb{S}^{d-1}$ such that $|1 - \mathbf{x}^T \mathbf{z}| \geq \delta$ it holds that $|k_{NTK}(\mathbf{x}, \mathbf{z}) - 0.25| < O\left(\frac{polylog(L)}{L}\right)$. Therefore, under the conditions of the lemma, with probability 1,

$$\lim_{L \to \infty} K(X, X) = 0.75 \cdot I + 0.25 \cdot \mathbf{1}\mathbf{1}^T$$
$$\lim_{L \to \infty} K(\mathbf{x}, X) = 0.25 \cdot \mathbf{1}^T.$$

From the continuity of the inverse function together with the Sherman-Morrison formula, we have that

$$\lim_{L\to\infty} K(X,X)^{-1} = \frac{4}{3}\left(I - \frac{\mathbf{1}\mathbf{1}^T}{3+n}\right)$$

$$\begin{aligned}\lim_{L\to\infty} K(\mathbf{x},X)K(X,X)^{-1}\mathbf{y} &= (0.25\cdot\mathbf{1}^T)\frac{4}{3}\left(I - \frac{\mathbf{1}\mathbf{1}^T}{3+n}\right)\mathbf{y} \\ &= \left(\frac{1}{3} - \frac{n}{3(3+n)}\right)\mathbf{1}^T\mathbf{y} = c_n\mathbf{1}^T\mathbf{y}.\end{aligned}$$

We conclude therefore that for a fixed $\epsilon > 0$ and n, there exists L_0 such that for any $L > L_0$,

$$|f_{NTK}(\mathbf{x}) - c_n\mathbf{1}^T\mathbf{y}| < \epsilon.$$

■

Next, we prove theorem E.1 part 2.

Proof (Theorem E.1 part 2) Let $(\mathbf{x}_1, y_1), .., (\mathbf{x}_n, y_n)$ be i.i.d sample of n training points. Let $\delta > 0$ and denote

$$B = \bigcup_{i=1}^{n} \bar{B}(\mathbf{x}_i, \delta),$$

where $\bar{B}(\mathbf{x}_i, \delta) = \{\mathbf{x} : \|\mathbf{x} - \mathbf{x}_i\|^2/2 < \delta\} \cap \mathbb{S}^{d-1}$. From Lemma E.5 we know that for any $\epsilon > 0$ there exists L_0 such that $\forall L > L_0, \forall \mathbf{x} \in \mathbb{S}^{d-1}/B$ it holds that

$$|f_{NTK}(\mathbf{x}) - c_n\mathbf{1}^T\mathbf{y}| < \epsilon.$$

We therefore have

$$\begin{aligned}\mathbb{E}(L(f_{NTK}(\mathbf{x}), y)) &= \int_{\mathbb{S}^{d-1}} (f_{NTK}(\mathbf{x}) - y(\mathbf{x}))^2 d\mathbf{x} \\ &\geq \int_{\mathbb{S}^{d-1}/B} \left(f_{NTK}(\mathbf{x}) - y(\mathbf{x})\right)^2 d\mathbf{x} \geq \int_{\mathbb{S}^{d-1}/B} \left(c_n\mathbf{1}^T\mathbf{y} \pm \epsilon - y(\mathbf{x})\right)^2 d\mathbf{x} \\ &= \int_{\mathbb{S}^{d-1}/B} \left(c_n\mathbf{1}^T\mathbf{y} - y(\mathbf{x})\right)^2 d\mathbf{x} \pm 2\epsilon \int_{\mathbb{S}^{d-1}/B} \left(c_n\mathbf{1}^T\mathbf{y} - y(\mathbf{x})\right) d\mathbf{x} + Area(\mathbb{S}^{d-1})\epsilon^2 \\ &\geq \int_{\mathbb{S}^{d-1}/B} \left(c_n\mathbf{1}^T\mathbf{y} - y(\mathbf{x})\right)^2 d\mathbf{x} - O(\epsilon) \\ &\geq^{(1)} \int_{\mathbb{S}^{d-1}} \left(c_n\mathbf{1}^T\mathbf{y} - y(\mathbf{x})\right)^2 d\mathbf{x} - O(\epsilon) - O(n\delta) \\ &\geq^{(2)} var(y(\mathbf{x})) - O(\epsilon) - O(n\delta) =^{(3)} 1 - O(\epsilon),\end{aligned}$$

where $^{(1)}$ is from adding at most n pieces of area at most δ, $^{(2)}$ is by the fact that $var(z) = min_c\mathbb{E}((z-c)^2)$, and $^{(3)}$ is by choosing L_0 such that $O(n\delta)$ is of order $O(\epsilon)$. ■

References

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

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, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net, 2019b.

Sanjeev Arora, Simon S. Du, Zhiyuan Li, Ruslan Salakhutdinov, Ruosong Wang, and Dingli Yu. Harnessing the power of infinitely wide deep nets on small-data tasks. In *International Conference on Learning Representations*, 2020.

David Balduzzi, Marcus Frean, Lennox Leary, JP Lewis, Kurt Wan-Duo Ma, and Brian McWilliams. The shattered gradients problem: If resnets are the answer, then what is the question? In *International Conference on Machine Learning*, pages 342–350. PMLR, 2017.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Ronen Basri, David W. Jacobs, Yoni Kasten, and Shira Kritchman. The convergence rate of neural networks for learned functions of different frequencies. In *Advances in Neural Information Pro-cessing Systems*, pages 4763–4772, 2019.

Ronen Basri, Meirav Galun, Amnon Geifman, David Jacobs, Yoni Kasten, and Shira Kritchman. Frequency bias in neural networks for input of non-uniform density. In *International Conference on Machine Learning*, pages 685–694. PMLR, 2020.

Alberto Bietti and Francis Bach. Deep equals shallow for relu networks in kernel regimes. *In 9th International Conference on Learning Representations*, 2021.

Alberto Bietti and Julien Mairal. On the inductive bias of neural tangent kernels. In *Advances in Neural Information Processing Systems*, pages 12893–12904, 2019.

Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan Gu. Towards understanding the spectral bias of deep learning. *arXiv preprint arXiv:2009.01198*, 2019.

Lin Chen and Sheng Xu. Deep neural tangent kernel and laplace kernel have the same rkhs. *International Conference on Learning Representations*, 2020.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pages 2937–2947, 2019.

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, pages 342–350. Curran Associates, Inc., 2009.

Amit Daniely and Eran Malach. Learning parities with neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20356–20365. Curran Associates, Inc., 2020.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685. PMLR, 2019.

Zhou Fan and Zhichao Wang. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7710–7721. Curran Associates, Inc., 2020.

Amnon Geifman, Abhay Yadav, Yoni Kasten, Meirav Galun, David Jacobs, and Ronen Basri. On the similarity between the laplace and neural tangent kernels. *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods?, 2020.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, third edition, 1996.

Daniel Greenfeld, Meirav Galun, Ronen Basri, Irad Yavneh, and Ron Kimmel. Learning to optimize multigrid PDE solvers. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 2415–2423, 2019.

Soufiane Hayou, Eugenio Clerico, Bobby He, George Deligiannidis, Arnaud Doucet, and Judith Rousseau. Stable resnet. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1324–1332. PMLR, 13–15 Apr 2021.

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, 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*. Springer International Publishing, 2016b.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1314–1324, 2019.

Kai-xuan Huang, Yuqing Wang, M. Tao, and Tuo Zhao. Why do deep residual networks generalize better than deep feed forward networks? - a neural tangent kernel perspective. *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

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 31*, pages 8571–8580. 2018.

Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.

Jaehoon Lee, Samuel S. Schoenholz, Jeffrey Pennington, Ben Adlam, Lechao Xiao, Roman Novak, and Jascha Sohl-Dickstein. Finite versus infinite neural networks: an empirical study. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33*, 2020.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Advances in neural information processing systems*, pages 6389–6399, 2018.

Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the risk of minimum-norm interpolants and restricted lower isometry of kernels. *arXiv preprint arXiv:1908.10292*, 2019.

Tengyuan Liang, Alexander Rakhlin, et al. Just interpolate: Kernel "ridgeless" regression can generalize. *Annals of Statistics*, 48(3):1329–1347, 2020.

Tianyi Liu, Minshuo Chen, Mo Zhou, Simon S Du, Enlu Zhou, and Tuo Zhao. Towards understanding the importance of shortcut connections in residual networks. In *Advances in neural information processing systems*, pages 7892–7902, 2019.

Eran Malach, Pritish Kamath, Emmanuel Abbe, and Nathan Srebro. Quantifying the benefit of using differentiable learning over tangent kernels, 2021.

Francis J Narcowich, Xinping Sun, and Joseph D Ward. Approximation power of rbfs and their associated sbfs: a connection. *Advances in Computational Mathematics*, 27(1): 107–124, 2007.

Atsushi Nitanda and Taiji Suzuki. Optimal rates for averaged stochastic gradient descent under neural tangent kernel regime. *arXiv preprint arXiv:2006.12297*, 2020.

Nicolò Pagliana, Alessandro Rudi, Ernesto De Vito, and Lorenzo Rosasco. Interpolation and learning with scale dependent kernels. *arXiv preprint arXiv:2006.09984*, 2020.

Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10428–10436, 2020.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5301–5310. PMLR, 2019.

Ana C. Q. Siravenha, Mylena N. F. Reis, Iraquitan Cordeiro, Renan Arthur Tourinho, Bruno D. Gomes, and Schubert R. Carvalho. Residual mlp network for mental fatigue classification in mining workers from brain data. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 407–412, 2019. doi: 10.1109/BRACIS.2019.00078.

Lili Su and Pengkun Yang. On learning over-parameterized neural networks: A functional approximation perspective. *arXiv preprint arXiv:1905.10826*, 2019.

Julián Tachella, Junqi Tang, and Mike Davies. The neural tangent link between cnn denoisers and non-local filters, 2020.

Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2820–2828, 2019.

Tom Tirer, Joan Bruna, and Raja Giryes. Kernel-based smoothness analysis of residual networks. *MSML*, 2021.

Andreas Veit, Michael J Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. *Advances in neural information processing systems*, 29: 550–558, 2016.

Lechao Xiao, Jeffrey Pennington, and Samuel Schoenholz. Disentangling trainability and generalization in deep neural networks. In *International Conference on Machine Learning*, pages 10462–10472. PMLR, 2020.

Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *CoRR*, abs/1901.06523, 2019.

Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Hongyi Zhang, Yann N Dauphin, and Tengyu Ma. Fixup initialization: Residual learning without normalization. *International Conference on Learning Representations*, 2019a.

Huishuai Zhang, Da Yu, Mingyang Yi, Wei Chen, and Tie-yan Liu. Stability and convergence theory for learning resnet: A full characterization. *arXiv preprint arXiv:1903.07120*, 2019b.