Beyond ReLU: How Activations Affect Neural Kernels and Random Wide Networks

David Holzmüller*
INRIA

Max David Schölpple*
University of Stuttgart

Abstract

In recent years, the neural tangent kernel (NTK) and neural network Gaussian process kernel (NNGP) have given theoreticians tractable limiting cases of fully connected neural networks. However, the property of these kernels are poorly understood for activation functions other than powers of the ReLU. Our main contribution is a characterization of the RKHS of these kernels for activation functions whose only non-smoothness is at zero. This extends existing theory to numerous commonly used activation functions such as SELU, ELU, or LeakyReLU. Additionally, we analyze a broad set of special cases such as missing biases, two-layer networks, or polynomial activations. Our results show that a broad class of not infinitely smooth activations generate equivalent RKHSs at different network depths, depending only on the degree of the non-smoothness up to equivalence. On the other hand, the RKHS generated by polynomial activations depends on the network depth. Finally, we derive results for the smoothness of NNGP sample paths, characterizing the smoothness of infinitely wide neural networks at initialization.

1 INTRODUCTION

Despite great efforts, our theoretical understanding of when and why deep learning works remains limited, and much of it is distributed across fragmented case studies that do not integrate into a bigger theory. For example, the influence of activation functions on the training dynamics and generalization behavior of deep learning methods is unclear.

One of the most prominent approaches towards deep learning theory is the study of infinite-width limits of neural networks (Neal, 1996; Jacot et al., 2018; Yang and Hu, 2021). Different parametrizations of neural networks lead to different behavior at infinite width (Yang and Hu, 2021). Here, we will focus on the neural tangent parametrization (NTP), whose infinite-width limit leads to kernel behavior (Jacot et al., 2018). While the kernel regime cannot model all aspects of neural network behavior (Yehudai and Shamir, 2019; Ortiz-Jiménez et al., 2021; Vyas et al., 2022; Wenger et al., 2023), it is one of the models most amenable to theoretical analysis. In particular, at initialization, the function represented by certain infinite-width neural networks is distributed like a Gaussian process with the so-called neural network Gaussian process (NNGP) kernel (Neal, 1996; Daniely et al., 2016; Lee et al., 2018; Matthews et al., 2018). The training then follows the dynamics of kernel gradient descent with the so-called neural tangent kernel (NTK, Jacot et al., 2018; Lee et al., 2019).

To understand the behavior of neural networks in the kernel regime, including training dynamics and generalization, it is essential to understand the properties of the associated kernels themselves. In particular, the structure of the reproducing kernel Hilbert spaces (RKHSs) corresponding to these kernels is central to further theoretical analysis. Here, we limit ourselves to fully connected neural networks, which are sufficient to study the influence of activation functions, and are still practically relevant by themselves (Holzmüller et al., 2024; Gorishniy et al., 2025; Ye et al., 2024; Erickson et al., 2025). While these kernels have been analyzed for (powers of) the ReLU activation and infinitely smooth activation functions (Chen and Xu, 2021; Bietti and Bach, 2021; Vakili et al., 2023), not much is known about their structure for other activation functions.

*Equal contribution.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

Contribution Our main contribution is Theorem 9, illustrated in Remark 10, where we prove a general result analyzing the structure of the RKHS corresponding to the NTK and NNGP kernels on the sphere $\mathbb{S}^d$ of fully connected neural networks. Our proof, of which an overview is given in Appendix A, builds upon the analysis of Bietti and Bach (2021), but generalizes existing results in multiple ways:

- We study a significantly larger class of activations. In particular, we provide exact characterizations for typical activation functions that are infinitely smooth everywhere except at zero. We also provide new results for polynomial activation functions, and we more precisely analyze the conditions under which results for infinitely smooth activation functions hold.
- We study a more general class of fully connected neural networks, covering the cases of normally distributed bias, zero-initialized bias, and no bias. We also study all special cases arising for two-layer networks.
- We analyze all cases for both NNGP and NTK (except discontinuous activation functions, for which the NTK is not well-defined).

As an additional contribution, in Theorem 14, we show how the results on the RKHS translate to the smoothness of functions sampled from the Gaussian process with the NNGP kernel, and hence the smoothness of infinitely wide neural networks at initialization.

After discussing preliminaries (Section 2), we introduce our main results in Section 3 and Section 5. Section 7 discusses some implications of our results on the broader infinite-width theory, such as the (non-)benefit of depth, training dynamics, and generalization. Related work is discussed in Section 8, before Section 9 concludes with a discussion of possible extensions.

2 PRELIMINARIES

We denote the d-dimensional unit sphere as $\mathbb{S}^d \subset \mathbb{R}^{d+1}$, and the square integrable functions over a measure space $(X, \mathcal{A}, \mu)$ as $L_2(X) := L_2(\mu) = \{f \mid \int_X f^2 \, d\mu < \infty\}$. For $s \geq 0$, we use $H^s(X)$ to denote the Sobolev space of order s, which essentially consists of the functions f for which all (fractional) derivatives up to order s are in $L_2(X)$.

Neural kernels For $d \geq 1$, we consider fully connected neural networks (NNs) in neural tangent parametrization (NTP) operating on inputs $\boldsymbol{x} \in \mathbb{R}^{d_0}, d_0 := d + 1$. Typically, neural networks use componentwise-normalized input vectors $\boldsymbol{x} \in \mathbb{R}^{d_0}$ such that $\|\boldsymbol{x}\|_2 \approx \sqrt{d_0}$. Following related theoretical literature, we instead consider inputs on the sphere, $\boldsymbol{x} \in \mathbb{S}^d = \{\boldsymbol{z} \in \mathbb{R}^{d_0} \mid \|\boldsymbol{z}\|_2 = 1\}$, and compensate for this by rescaling the first layer by $\sqrt{d_0}$, which leads to equivalent learning dynamics with gradient descent. We consider the following network architecture:

Definition 1. Let $L \geq 2$ be the number of layers, let $\sigma_w > 0$, and let $\sigma_b, \sigma_i \geq 0$. The network is the function $f_{\boldsymbol{\theta}} : \mathbb{R}^{d_0} \to \mathbb{R}^{d_L}$ given by $f_{\boldsymbol{\theta}}(\boldsymbol{x}^{(0)}) := \boldsymbol{z}^{(L)}$, where

$$\begin{aligned} \boldsymbol{z}^{(1)} &:= \sigma_w \boldsymbol{W}^{(1)} \boldsymbol{x}^{(0)} + \sigma_b \boldsymbol{b}^{(1)} \in \mathbb{R}^{d_1} \\ \boldsymbol{z}^{(l)} &:= \frac{\sigma_w}{\sqrt{d_{l-1}}} \boldsymbol{W}^{(l)} \boldsymbol{x}^{(l-1)} + \sigma_b \boldsymbol{b}^{(l)} \in \mathbb{R}^{d_l}, \\ \boldsymbol{x}^{(l-1)} &:= \varphi(\boldsymbol{z}^{(l-1)}) \in \mathbb{R}^{d_l}. \qquad (l \geq 2) \end{aligned}$$

We assume that all parameters are initialized independently as $\boldsymbol{W}_{jk}^{(l)} \sim \mathcal{N}(0,1)$ and $b_j^{(l)} \sim \mathcal{N}(0, \sigma_i^2)$. ◀

Definition 1 subsumes the three cases of

- bias-free networks by setting $\sigma_b = 0$,
- zero-initialized biases by setting $\sigma_b \neq 0$, $\sigma_i = 0$,
- randomly initialized biases by setting $\sigma_b, \sigma_i \neq 0$.

For $\sigma_i^2 = 1$, the recursive formula for NTK and NNGP kernels are derived in Lee et al. (2019). We derive formulas for the different possible values of $\sigma_i, \sigma_b, \sigma_w$ in Lemma D.2. For the NNGP-kernel, the recursion is given by

$$\begin{aligned} k_1^{\mathrm{NNGP}}(\boldsymbol{x}, \bar{\boldsymbol{x}}) &:= \sigma_b^2 \sigma_i^2 + \sigma_w^2 \langle \boldsymbol{x}, \bar{\boldsymbol{x}} \rangle \\ k_L^{\mathrm{NNGP}}(\boldsymbol{x}, \bar{\boldsymbol{x}}) &= \sigma_b^2 \sigma_i^2 + \sigma_w^2 \mathbb{E}_{(u,v) \sim \boldsymbol{\Sigma}_{L-1}(\boldsymbol{x}, \bar{\boldsymbol{x}})}[\varphi(u)\varphi(v)] \\ \boldsymbol{\Sigma}_L(\boldsymbol{x}, \bar{\boldsymbol{x}}) &= \begin{pmatrix} k_L^{\mathrm{NNGP}}(\boldsymbol{x}, \boldsymbol{x}) & k_L^{\mathrm{NNGP}}(\boldsymbol{x}, \bar{\boldsymbol{x}}) \\ k_L^{\mathrm{NNGP}}(\bar{\boldsymbol{x}}, \boldsymbol{x}) & k_L^{\mathrm{NNGP}}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{x}}) \end{pmatrix} , \end{aligned}$$

For the NTK, we refer to Lemma D.2.

To analyze NTK and NNGP kernels, we restrict them to the sphere $\mathbb{S}^d$, where they are dot-product kernels due to the rotation-invariance induced by the Gaussian initialization of the weight matrices $\boldsymbol{W}^{(l)}$.

Dot-product kernels on the sphere Let $\kappa : [-1, 1] \to \mathbb{R}$ such that $k := k_{\kappa,d} : \mathbb{S}^d \times \mathbb{S}^d \to \mathbb{R}, (\boldsymbol{x}, \boldsymbol{x}') \mapsto \kappa(\langle \boldsymbol{x}, \boldsymbol{x}' \rangle)$ is a (positive semidefinite) kernel. Then, k is called a *dot-product kernel.* We leverage the theory of dot-product kernels on spheres to study the RKHS $\mathcal{H}_k$ associated with k (see e.g. Bietti and Bach, 2021; Hubbert et al., 2023; Haas et al., 2023). In particular, these RKHSs can be characterized through the eigenvalues of the associated integral operator $T_k : L_2(\mathbb{S}^d) \to L_2(\mathbb{S}^d)$ given by

$$(T_k f)(\boldsymbol{x}) = \int_{\mathbb{S}^d} k(\boldsymbol{x}, \boldsymbol{x}') f(\boldsymbol{x}') \, \mathrm{d}\boldsymbol{x}' .$$

It can be diagonalized as

$$T_k f = \sum_{l=0}^{\infty} \sum_{i=1}^{N_{l,d}} \mu_l Y_{l,i} \langle Y_{l,i}, f \rangle_{L_2(\mathbb{S}^d)} ,$$

where $\{Y_{l,1},\dots,Y_{l,N_{l,d}}\}$ is an arbitrary orthonormal basis of the space of spherical harmonics (a subset of polynomials) of degree l in $L_2(\mathbb{S}^d)$, cf. Müller (2006), but the details are not important in the following. The eigenfunctions $Y_{l,i}$ are the same for all dot-product kernels but the eigenvalues $\mu_l = \mu_l(k) = \mu_l(\kappa, d)$ can be different and characterize the RKHS $\mathcal{H}_k$ associated with k:

$$\mathcal{H}_k = \left\{ \sum_{l=0}^{\infty} \sqrt{\mu_l} \sum_{i=1}^{N_{l,d}} a_{l,i} Y_{l,i} \;\middle|\; \sum_{l=0}^{\infty} \sum_{i=1}^{N_{l,d}} a_{l,i}^2 < \infty \right\} .$$

Its inner product is

$$\left\langle \sum_{l=0}^{\infty} \sqrt{\mu_l} \sum_{i=1}^{N_{l,d}} a_{l,i} Y_{l,i}, \sum_{l=0}^{\infty} \sqrt{\mu_l} \sum_{i=1}^{N_{l,d}} b_{l,i} Y_{l,i} \right\rangle_{\mathcal{H}_k} = \sum_{l=0}^{\infty} \sum_{i=1}^{N_{l,d}} a_{l,i} b_{l,i} \ .$$

Hence, the RKHS as a set is characterized by 1) the asymptotic behavior of the eigenvalues and 2) the set of indices for which the eigenvalues are nonzero. By Proposition 3 in Schölpple and Steinwart (2025), if two RKHSs $\mathcal{H}_1, \mathcal{H}_2$ are equal as sets, then their norms are equivalent. In this case, we call $\mathcal{H}_1, \mathcal{H}_2$ *equivalent*, and write $\mathcal{H}_1 \cong \mathcal{H}_2$. Polynomial eigenvalue decays lead to RKHSs equivalent to Sobolev spaces:

Lemma 2 (Sobolev spaces on the sphere)**.** *Let $d \in \mathbb{N}$ and $s > d/2$. The RKHS $\mathcal{H}_k$ corresponding to the dot-product kernel k on $\mathbb{S}^d$ is equivalent to the Sobolev space $H^s(\mathbb{S}^d)$ if and only if there exist $c, C > 0$ with $c(l+1)^{-2s} \leq \mu_l \leq C(l+1)^{-2s}$ for all $l \geq 0$.*

Proof. Depending on the definition of $H^s(\mathbb{S}^d)$, this follows either directly or from classical theory, see Hubbert et al. (2023) and Lemma B.1 in Haas et al. (2023). □

Sobolev spaces on the sphere consist of those functions f that locally, when composed with a smooth chart $\varphi : U \subseteq \mathbb{R}^d \to V \subseteq \mathbb{S}^d$, are Sobolev functions on $\mathbb{R}^d$, e.g. Haas et al. (2023), p. 34.

3 THE STRUCTURE OF NEURAL KERNELS

Before we introduce our main theorem, we need to impose some assumptions on the involved activation functions.

Definition 3 (Functions with polynomially bounded derivatives)**.** Let $I \subseteq \mathbb{R}$ be an interval. We define the set $\mathcal{S}^{(\infty)}(I)$ to contain all functions $\varphi_I \in C^\infty(I)$ with polynomially bounded derivatives. In other words: For all $m \in \mathbb{N}_0$, there exist $a_m, b_m, q_m > 0$ such that for all $x \in I$, $|\varphi_I^{(m)}(x)| \leq a_m |x|^{q_m} + b_m$. ◀

Polynomial boundedness for $m \in \{0,1\}$ is required to get the NTK formulas from (Yang, 2020, Theorem 7.2, Box 1); for higher derivatives we only need $\varphi_I^{(m)} \in L_2(\mathcal{N}(0,\sigma^2))$ for all $\sigma^2 > 0$.

Assumption 4 (Activation function)**.** We assume that $\varphi : \mathbb{R} \to \mathbb{R}$ is of the form[1]

$$\varphi(x) = \begin{cases} \varphi_+(x) & , x > 0 \\ \varphi_-(x) & , x < 0 \\ \frac{1}{2}(\lim_{x \searrow 0} \varphi_+(x) + \lim_{x \nearrow 0} \varphi_-(x)) & , x = 0 \end{cases}$$

with $\varphi_- \in \mathcal{S}^{(\infty)}((-\infty, 0))$ and $\varphi_+ \in \mathcal{S}^{(\infty)}((0,\infty))$ and that φ is not the zero function. ◀

Proposition 5 (All common activation function satisfy Assumption 4)**.**

(a) If f is constructed by addition, multiplication, and composition of polynomials, sigmoid, tanh, softplus, sin, cos, RBF *and* Φ, *then $f \in \mathcal{S}^{(\infty)}(\mathbb{R})$. Moreover, the functions $g(x) = \exp(ax)$ are in $\mathcal{S}^{(\infty)}((-\infty, 0])$ for every $a \geq 0$.*

(b) Assumption 4 is satisfied for all activation functions from Table 1. It is also satisfied for all $\phi \in \mathcal{S}^{(\infty)}(\mathbb{R})$.

We prove Proposition 5 in Appendix E.

Definition 6 (Smoothness of an activation function)**.** For an activation function φ as in Assumption 4, we define its smoothness as

$$\operatorname{smoothness}(\varphi) := \inf\{m \in \mathbb{N}_0 \mid \lim_{t\searrow 0} \varphi^{(m)}(t) \neq \lim_{t \nearrow 0} \varphi^{(m)}(t)\} \in \mathbb{N}_0 \cup \{\infty\}. \quad ◀$$

Example 7 (Smoothness of the ReLU activation)**.** The ReLU activation function $\varphi(x) = \max\{0, x\}$ fulfills Assumption 4 and decomposes as $\varphi_+(x) = x \ , x > 0$, $\varphi_-(x) = 0 \ , x < 0$. Hence we have

$$\lim_{t\searrow 0} \varphi(t) = 0 = \lim_{t\nearrow 0} \varphi(t)$$
$$\text{but} \qquad \lim_{t\searrow 0} \varphi'(t) = 1 \neq 0 = \lim_{t\nearrow 0} \varphi'(t) \ ,$$

implying smoothness(ReLU) $= 1$. ◀

Remark 8 (Location of the non-smoothness)**.** We conjecture that our analysis could be extended to the case of non-smoothnesses at $b \neq 0$ with similar results by extending a part of our proofs, cf. Remark A.7. ◀

[1] The definition at zero ensures that φ is continuous when φ_+ and φ_- allow, and that its even/odd parts are zero everywhere whenever they are zero almost everywhere.

Table 1: Activation functions and their smoothness (Definition 6). Here, $\mathcal{P}_m$ means that the function is a polynomial of degree m. $\mathcal{P}_{-\infty}$ means that the function is zero. For a more extensive overview, we refer to Dubey et al. (2022). In the typical cases i) of Theorem 9, a smoothness $s \in (1, \infty)$ of φ yields $\mathcal{H}_{k_L^{\mathrm{NNGP}}} \cong H^{d/2+s+1/2}(\mathbb{S}^d)$ and $\mathcal{H}_{k_L^{\mathrm{NTK}}} \cong H^{d/2+s-1/2}$.

Activation	Formula		Smoothness s		
	$x < 0$	$x \geq 0$	φ	φ_{even}	φ_{odd}
ReLU	0	x	1	1	$\mathcal{P}_1$
LeakyReLU ($\varepsilon \neq -1$)	$-\varepsilon x$	x	1	1	$\mathcal{P}_1$
SELU	$\lambda\alpha(e^x - 1)$	λx	1	1	2
ELU ($\alpha \neq 1$)	$\alpha(e^x - 1)$	x	1	1	2
ELU ($\alpha = 1$), CELU	$\alpha(e^{x/\alpha} - 1)$	x	2	3	2
RePU, power m even	0	x^m	m	$\mathcal{P}_m$	m
RePU, power m odd	0	x^m	m	m	$\mathcal{P}_m$
Heaviside	$\frac{1}{2}\mathbb{1}_{\{0\}}(x) + \mathbb{1}_{\mathbb{R}_{>0}}(x)$		0	$\mathcal{P}_0$	0
tanh	$(e^x - e^{-x})/(e^x + e^{-x})$		∞	$\mathcal{P}_{-\infty}$	∞
Sigmoid	$1/(1 + e^{-x})$		∞	$\mathcal{P}_0$	∞
GeLU	$\frac{1}{2}x(1 + \mathrm{erf}(x/\sqrt{2}))$		∞	∞	$\mathcal{P}_1$
SiLU	$x/(1 + e^{-x})$		∞	∞	∞
RBF	$\exp(-x^2)$		∞	∞	$\mathcal{P}_{-\infty}(\mathbb{S}^d)$.

Our main result, Theorem 9, contains multiple cases. The "regular" cases are discussed afterward in Remark 10. In some special bias-free cases we observe a phenomenon of parity: The even/odd parts of the functions contained in the RKHSs $\mathcal{H}_k$ of the neural kernels depend on the even/odd parts of the activation φ, commonly defined as

$$\varphi_{\mathrm{even}}(x) := \frac{\varphi(x) + \varphi(-x)}{2}, \quad \varphi_{\mathrm{odd}}(x) := \frac{\varphi(x) - \varphi(-x)}{2},$$

fulfilling $\varphi = \varphi_{\mathrm{even}} + \varphi_{\mathrm{odd}}$. For example, $\mathrm{ReLU}_{\mathrm{even}}(x) = \frac{1}{2}|x|$ and $\mathrm{ReLU}_{\mathrm{odd}}(x) = \frac{1}{2}x$. In the bias-free case ($\sigma_b^2 = 0$), the function f_φ represented by a two-layer network with activation φ can be written as $f_\varphi = f_{\varphi_{\mathrm{even}}} + f_{\varphi_{\mathrm{odd}}}$, where $f_{\varphi_{\mathrm{even}}}$ is even and $f_{\varphi_{\mathrm{odd}}}$ is odd. Hence, $(f_{\mathrm{ReLU}})_{\mathrm{odd}} = f_{\mathrm{ReLU}_{\mathrm{odd}}}$ is a linear function, which explains the results of Bietti and Mairal (2019) where the odd eigenvalues satisfy $\mu_1 > 0$, $\mu_3 = \mu_5 = \ldots = 0$. A much more general version of this calculation is performed in Proposition D.4.

By $\mathcal{P}_m$ we denote the polynomials of degree at most m. We additionally define

$$\mathcal{F}_{\mathrm{even}} := \{f : \mathbb{S}^d \to \mathbb{R} \mid f \text{ is even}\}, \quad \mathcal{F}_{\mathrm{odd}} := \{f : \mathbb{S}^d \to \mathbb{R} \mid f \text{ is odd}\}.$$

Theorem 9 (Main result, summary below)**.** *Let the activation φ fulfill Assumption 4 and let $s :=$* $\mathrm{smoothness}(\varphi)$.

NNGP:

i) Case $\sigma_b^2\sigma_i^2 > 0$ or both $L \geq 3$ and φ is neither even nor odd.

- *a) If $s = 0$, then $\mathcal{H}_{k_L^{\mathrm{NNGP}}} \cong H^{d/2+2^{1-L}}(\mathbb{S}^d)$.*
- *b) If $1 \leq s < \infty$, then $\mathcal{H}_{k_L^{\mathrm{NNGP}}} \cong H^{d/2+s+1/2}(\mathbb{S}^d)$.*
- *c) If $s = \infty$ and φ is not a polynomial, then $\mathcal{H}_{k_L^{\mathrm{NNGP}}} \subset H^t(\mathbb{S}^d)$ for all $t \in \mathbb{R}$ and $\mathcal{H}_{k_L^{\mathrm{NNGP}}}$ contains all polynomials.*
- *d) If $s = \infty$ and φ is a polynomial of degree m, then $\mathcal{H}_{k_L^{\mathrm{NNGP}}} \cong \mathcal{P}_{m^{L-1}}$.*

ii) Case $\sigma_b^2\sigma_i^2 = 0$ and φ is even or odd.

- *a) If $s = 0$, then $\mathcal{H}_{k_L^{\mathrm{NNGP}}} \cong H^{d/2+2^{1-L}}(\mathbb{S}^d) \cap \mathcal{F}_{even/odd}$.*
- *b) If $1 \leq s < \infty$, then $\mathcal{H}_{k_L^{\mathrm{NNGP}}} \cong H^{d/2+s+1/2}(\mathbb{S}^d) \cap \mathcal{F}_{even/odd}$.*
- *c) If $s = \infty$ and φ is not a polynomial, then $\mathcal{H}_{k_L^{\mathrm{NNGP}}} \subset H^t(\mathbb{S}^d) \cap \mathcal{F}_{even/odd}$ for all $t \in \mathbb{R}$ and $H_{k_L^{\mathrm{NNGP}}}$ contains all even/odd polynomials.*
- *d) If $s = \infty$ and φ is a polynomial of degree m, then $\mathcal{H}_{k_L^{\mathrm{NNGP}}} \cong \mathcal{P}_{m^{L-1}} \cap \mathcal{F}_{even/odd}$.*

iii) Case $\sigma_b^2\sigma_i^2 = 0$ and $L = 2$. We have

$$\mathcal{H}_{k_L^{\mathrm{NNGP}}} \cong \mathcal{H}_{k^{\mathrm{NNGP}}_{\varphi_{\mathrm{even}},L}} \oplus \mathcal{H}_{k^{\mathrm{NNGP}}_{\varphi_{\mathrm{odd}},L}} \tag{1}$$

where the RKHSs $\mathcal{H}_{k^{\mathrm{NNGP}}_{\varphi_{\mathrm{even}},L}}$ and $\mathcal{H}_{k^{\mathrm{NNGP}}_{\varphi_{\mathrm{odd}},L}}$ of the even/odd activation functions φ_{even} and φ_{odd} can be found by Case ii) with corresponding $s :=$ $\mathrm{smoothness}(\varphi_{\mathrm{even}})$ *respectively $s :=$* $\mathrm{smoothness}(\varphi_{\mathrm{odd}})$.

NTK:

i) Case $\sigma_b^2 > 0$ or both $L \geq 3$ and φ is neither even nor odd.

- *a) If $1 \leq s < \infty$, then $\mathcal{H}_{k_L^{\mathrm{NTK}}} \cong H^{d/2+s-1/2}(\mathbb{S}^d)$.*
- *b) If $s = \infty$ and φ is not a polynomial, then $\mathcal{H}_{k_L^{\mathrm{NTK}}} \subset H^t(\mathbb{S}^d)$ for all $t \in \mathbb{R}$ and $\mathcal{H}_{k_L^{\mathrm{NTK}}}$ contains all polynomials.*
- *c) If $s = \infty$ and φ is a polynomial of degree m, then $\mathcal{H}_{k_L^{\mathrm{NTK}}} \cong \mathcal{P}_{m^{L-1}}$.*

ii) Case $\sigma_b^2 = 0$ and φ is even or odd.

- *a) If $1 \leq s < \infty$, then $\mathcal{H}_{k_L^{\mathrm{NTK}}} \cong H^{d/2+s-1/2}(\mathbb{S}^d) \cap \mathcal{F}_{even/odd}$.*
- *b) If $s = \infty$ and φ is not a polynomial, then $\mathcal{H}_{k_L^{\mathrm{NTK}}} \subset H^t(\mathbb{S}^d) \cap \mathcal{F}_{even/odd}$ for all $t \in \mathbb{R}$ and $H_{k_L^{\mathrm{NTK}}}$ contains all even/odd polynomials.*
- *c) If $s = \infty$ and φ is a polynomial of degree m, then $\mathcal{H}_{k_L^{\mathrm{NTK}}} = \mathcal{P}_{m^{L-1}} \cap \mathcal{F}_{even/odd}$.*

iii) Case $\sigma_b^2 = 0$ and $L = 2$. We have

$$\mathcal{H}_{k_L^{\mathrm{NTK}}} \cong \mathcal{H}_{k^{\mathrm{NTK}}_{\varphi_{\mathrm{even}},L}} \oplus \mathcal{H}_{k^{\mathrm{NTK}}_{\varphi_{\mathrm{odd}},L}} \tag{2}$$

where the RKHSs $\mathcal{H}_{k^{\mathrm{NTK}}_{\varphi_{\mathrm{even}},L}}$ and $\mathcal{H}_{k^{\mathrm{NTK}}_{\varphi_{\mathrm{odd}},L}}$ of the even/odd activation functions φ_{even} and φ_{odd} and can be found by Case ii) with correspond-

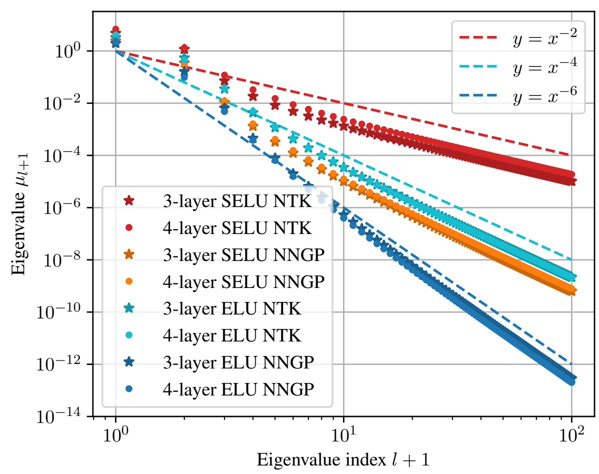

Figure 1: **Eigenvalues μ_l of different neural kernels on $\mathbb{S}^2$.** We use $\sigma_w = \sigma_b = \sigma_i = 1$. We use a custom method to numerically compute dual activations, described in Appendix G.

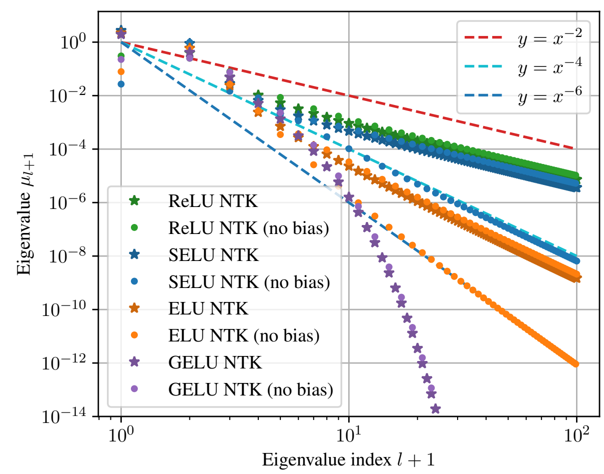

Figure 2: **Eigenvalues μ_l of different two-layer NTKs on $\mathbb{S}^2$.** We use $\sigma_w = \sigma_b = \sigma_i = 1$ for the case with biases and set $\sigma_b = \sigma_i = 0$ for the no-bias case. We use a custom method to numerically compute dual activations, described in Appendix G.

ing $s :=$ smoothness(φ_{even}) *respectively* $s :=$ smoothness(φ_{odd}).

Note that $H^t(\mathbb{S}^d) \cap \mathcal{F}_{\text{even}}$ and $H^t(\mathbb{S}^d) \cap \mathcal{F}_{\text{odd}}$ are closed subspaces of $H^t(\mathbb{S}^d)$ for any t. We equip them with the restricted norm. If in the sub-cases iii) $\varphi_{\text{even}} = 0$ or $\varphi_{\text{odd}} = 0$ occurs, the RKHS corresponding to that even/odd part in Eq. (1) *respectively Eq.* (2) *is $\mathcal{P}_{-\infty} = \{0\}$.*

We show a proof sketch in Section 4, a more detailed proof overview in Appendix A, and the proof for deriving this version of the theorem in Appendix D.4. The smoothness of many common activation functions, as well as the smoothness of their even/odd parts, can be found in Table 1. Note that the results of Theorem 9 can be expressed in terms of the eigenvalues μ_l of the integral operator instead of the RKHSs as discussed in Section 2; the eigenvalue-based formulation can be found in Theorem D.10.

Remark 10 (General takeaway)**.** Theorem 9 is most easy to understand for the case where

- the NN has $L \geq 2$ layers, contains biases ($\sigma_b^2 > 0$), and in the NNGP case the biases are initialized with nonzero variance ($\sigma_i^2 > 0$), or
- the NN has $L \geq 3$ layers and the activation function φ is neither even nor odd.

In this case, the decay of the eigenvalues μ_l and the corresponding RKHS depend on the smoothness s of the activation function φ as follows:

- If the activation function is discontinuous ($s = 0$), the RKHS of the NNGP kernel is equivalent to a Sobolev space of smoothness $d/2 + 2^{1-L}$, while the NTK is not defined.
- If the activation function has finite smoothness $1 \leq s < \infty$, the RKHSs are equivalent to Sobolev spaces of order $d/2 + s - 1/2$ for the NTK and $d/2 + s + 1/2$ for the NNGP. This applies to ReLU, LeakyReLU, ELU ($\alpha \neq 1$) and SELU with $s = 1$ and to CELU with $s = 2$. Figure 1 shows that the corresponding eigenvalue decays $\mu_l = d + 2s - 1$ for the NTK and $\mu_l = d + 2s + 1$ (cf. Lemma 2) are attained in numerical experiments. Rectified power unit activations $\max\{0, x\}^s$ have smoothness $s \in \mathbb{N}$, yielding a family of Sobolev spaces. Our results show that this is due to the behavior at zero and that the superlinear growth of these functions for $s \geq 2$ is not necessary.
- If the activation function is infinitely smooth ($s = \infty$) but not a polynomial, all μ_l are nonzero but decay faster than any inverse polynomial. Hence, the RKHSs are contained in all Sobolev spaces, but the kernels are universal. This applies for example to the GELU, SiLU/Swish, Mish, softplus, sigmoid, and tanh activation functions (though tanh is an odd function, so the conclusion only holds with biases).
- If the activation function is a polynomial, for NTK and NNGP only finitely many eigenvalues μ_l are nonzero and the RKHSs only contain polynomials, whose maximum degree grows exponentially with the depth of the network. ◀

Example 11 (Special cases)**.** As mentioned above, two-layer networks without bias ($\sigma_b^2 = 0$ for NTK or $\sigma_b^2\sigma_i^2 = 0$ for NNGP) can have special behavior. Following Table 1, we obtain for CELU the NTK RKHS $(H^{d/2+5/2} \cap \mathcal{F}_{\text{even}}) \oplus (H^{d/2+3/2} \cap \mathcal{F}_{\text{odd}})$, so the even parts of functions are smoother than the odd parts of functions. For ELU ($\alpha \neq 1$) and SELU, we obtain $(H^{d/2+1/2} \cap \mathcal{F}_{\text{even}}) \oplus (H^{d/2+3/2} \cap \mathcal{F}_{\text{odd}})$. For ReLU and LeakyReLU, the odd parts are just linear functions. For RePU, one of the two parts are polynomials up to a certain degree. Figure 2 shows the corresponding behavior of the eigenvalues μ_l for the two-layer NTKs for different activations: In the bias-free case, the odd SELU eigenvalues decay faster than the even ones, while the even ELU ($\alpha = 1$) eigenvalues decay faster than the odd ones. For bias-free ReLU and GELU, the eigenvalues μ_l for odd $l \geq 3$ are zero.

The other special case is for even or odd activations and arbitrarily deep networks without bias. For example, tanh is an odd activation, hence the corresponding NNGP and NTK RKHSs will only contain odd functions. ◀

Remark 12 (NTKs and NNGPs of non-smooth activations are equivalent to Matérn kernels)**.** Chen and Xu (2021) and Bietti and Bach (2021) showed that on $\mathbb{S}^d$, the RKHS of the deep ReLU NTK is equivalent to that of the Laplace kernel. Theorem 1 in Vakili et al. (2021) further shows that the RKHSs of deep RePU NTKs and NNGPs are equivalent to the ones of Matérn kernels, which are equivalent to Sobolev spaces. Specifically, for the Matérn kernel k_ν of order $\nu > 0$ on the sphere $\mathbb{S}^d$, it holds that $\mathcal{H}_{k_\nu} \cong H^{d/2+\nu}(\mathbb{S}^d)$ (see also Proposition 5.2 (c) in Hubbert et al. 2023). For other not-infinitely-smooth activations, our Theorem 9 shows that their RKHSs are generally also equivalent to Sobolev spaces, and hence the corresponding kernels are equivalent to Matérn kernels. ◀

Section 7 discusses implications of our results, for example, on the equivalence of RKHSs for different network depths.

4 PROOF SKETCH FOR THE MAIN THEOREM

Here, we provide a short overview of our main theorem. A more detailed overview can be found in Appendix A. We obtain Theorem 9 by case distinction from Theorem D.10, the central technical result of this paper, which yields the eigenvalue asymptotics of the neural kernel k in dependence of the smoothness of the activation function φ. Those eigenvalue asymptotics then directly yield the equivalent Sobolev space $\mathcal{H}_k \cong H^s(\mathbb{S}^d)$ by Lemma 2. Theorem D.10 is the major theoretical effort of this paper. It leverages Theorem A.3, a variant of Bietti and Bach (2021, Theorem 7), which reduces the question of the eigenvalue decay to an investigation of the asymptotic behavior at $\{\pm 1\}$ of the function $\kappa : [-1, 1] \to \mathbb{R}$ satisfying

$$k(x, x') = \kappa(\langle x, x' \rangle) \ .$$

This function κ is best described with *dual activation functions* (Definition A.4, Appendix C). The dual activation function $\widehat{\varphi} : [-1, 1] \to \mathbb{R}$ has been introduced in Daniely et al. (2016) and is given by a quadratic form: $\widehat{\varphi} = b(\varphi, \varphi)$, where

$$b(f, g)(t) = \mathbb{E}_{(u,v)\sim\mathcal{N}(0,\Sigma_t)}[f(u)g(v)], \quad \Sigma_t = \begin{pmatrix} 1 & t \\ t & 1 \end{pmatrix}$$

describes the infinite-width behavior of the activation in a single hidden layer. Recursively, the function κ_l^{NNGP} corresponding to the NNGP kernel of an l-layer network for $l \geq 2$ is then given by

$$\begin{aligned} \kappa_1^{\text{NNGP}}(t) &:= \sigma_b^2 \sigma_i^2 + \sigma_w^2 t \\ \kappa_l^{\text{NNGP}}(t) &:= \sigma_b^2 \sigma_i^2 + \sigma_w^2 \widehat{\varphi_{\cdot\sqrt{\alpha_{l-1}}}}(\kappa_{l-1}^{\text{NNGP}}(t)/\alpha_{l-1}), \quad (3) \end{aligned}$$

where $\varphi_{\cdot\sqrt{\alpha_{l-1}}}$ means that φ rescales its input, which is explained in Definition A.4 and Lemma A.5, but is not important as it does not change the smoothness at zero. This recursion alongside with the corresponding formula for the NTK kernel κ_l^{NTK} can be found in Lemma A.5. The analysis of the asymptotic behavior of κ_l^{NNGP} is built on the analysis of the asymptotic behavior of the dual activation $\widehat{\varphi}$ at $\{\pm 1\}$. The following theorem is hence a central cornerstone for the analysis of dual activations that may be useful beyond the specific structure of fully-connected neural networks.

Theorem 13 (Theorem A.8 reformulated)**.** *Let φ be an activation function of finite smoothness s. Then, there exists a constant $b_s > 0$ depending only on s such that for $\tau \in \{\pm 1\}$ and $t \in (0, 2)$*

$$\begin{aligned} \widehat{\varphi}(\tau(1-t)) = &\big(\varphi^{(s)}(0+) - \varphi^{(s)}(0-)\big)(-\tau)^{s+1} b_s t^{s+1/2} \\ &+ p_\tau(t) + q_\tau(t) \ , \end{aligned}$$

where $\varphi^{(s)}(0+), \varphi^{(s)}(0-)$ are right- and left-sided limits, $p_\tau(t)$ is a polynomial and $q_\tau(t)$ fulfills $q_\tau^{(n)}(t) = o(t^{(s-n)})$ for all $n \in \mathbb{N}_0$.

We prove Theorem 13 by decomposing

$$\varphi(x) = \sum_{k=s}^{K-1} (\varphi^{(s)}(0+) - \varphi^{(s)}(0-)) s_k(x) + r(x) \ ,$$

where $s_k(x) := \frac{1}{2k!} \operatorname{sgn}(x) x^k$ are analytically tractable functions of smoothness k and r is a remainder term of smoothness K. The analysis of Bietti and Bach (2021) is similar to computing $\widehat{s}_1 = b(s_1, s_1)$, and we introduce multiple new arguments to study higher-order terms $b(s_k, s_k)$, and also mix-terms $b(s_k, s_m)$, $b(s_k, r)$, and $b(r,r)$. While Bietti and Bach (2021) use arguments from Chen and Xu (2021) based on complex analysis to control the regularity of derivatives for q_τ, we circumvent this part through direct computations since it is not obvious whether functions like $b(r,r)$ are analytic.

While Theorem 13 is only for dual activations and not the final kernels, it already provides the correct functional form “non-integer power plus remainder terms” needed for applying the main theorem of Bietti and Bach (2021), which says that the eigenvalue decay depends on the smallest non-integer power. In Appendix B, we develop a calculus that investigates how this functional form is preserved for sums, producs, and compositions of functions. To obtain our main technical result, Theorem D.10, we apply this calculus in all cases together with Proposition D.4 for handling the bias-free special cases and a separate computation for polynomial activations in Appendix D.5.

5 SMOOTHNESS OF NNGP SAMPLE PATHS

The following result connects RKHSs of kernels on the sphere to the smoothness of the paths of their Gaussian process.

Theorem 14. *Let k be a dot-product kernel on $\mathbb{S}^d$ whose RKHS is equivalent to a Sobolev space $H^{d+\alpha}(\mathbb{S}^d)$, $\alpha > 0$. Let X be a Gaussian process on $\mathbb{S}^d$ with zero mean and covariance kernel k.*

i) For any $\varepsilon \geq 0$ we have $P(Y \in H^{d/2+\alpha+\varepsilon}(\mathbb{S}^d)) = 0$ for any version Y of X.
ii) For any $0 < \varepsilon < \alpha$, there exists a version Y of X with $\mathbb{P}\left(Y \in H^{d/2+\alpha-\varepsilon}\right) = 1$.

The requirement $\varepsilon < \alpha$ in Theorem 14 ii) stems from the underlying theory, which requires the investigated spaces to be RKHSs. For Sobolev spaces $H^s(\mathbb{S}^d)$ is equivalent to $s > d/2$. Theorem 14 is proven in Appendix F.

Example 15 (Application to infinite-width NNs)**.** Consider a network with biases, that is $\sigma_b^2\sigma_i^2 > 0$, and an activation of smoothness $1 \leq s < \infty$. From Theorem 9 we know $\mathcal{H}_{k_L^{\mathrm{NTK}}} \cong H^{d/2+s-1/2}(\mathbb{S}^d)$ and $\mathcal{H}_{k_L^{\mathrm{NNGP}}} \cong H^{d/2+s+1/2}(\mathbb{S}^d)$. If $s+1/2 > d/2$, by Theorem 14, the NNGP sample paths are in $H^{s+1/2-\varepsilon}(\mathbb{S}^d)$ but not $H^{s+1/2+\varepsilon}(\mathbb{S}^d)$ for any $\varepsilon > 0$. We conjecture that randomly initialized finite-width networks are only in $H^s(\mathbb{S}^d)$ but not $H^{s+\varepsilon}(\mathbb{S}^d)$, so the infinite-width limit would gain an extra half-order of smoothness. ◀

Example 16 (ReLU sample paths)**.** For the ReLU activation, in the case with biases ($\sigma_b^2\sigma_i^2 > 0$), we know from Theorem 9 as well as Bietti and Bach (2021) that $\mathcal{H}_{k_L^{\mathrm{NNGP}}} \cong H^{d/2+3/2}(\mathbb{S}^d)$. We have $H^{d/2+3/2}(\mathbb{S}^d) = H^{d+\alpha}(\mathbb{S}^d)$ with $\alpha = 3/2 - d/2$. Therefore, by Theorem 14, the sample paths of the ReLU NNGP essentially have the smoothness $d/2 + \alpha = 3/2$, assuming $\alpha > 0$ or, equivalently, $d \in \{1, 2\}$. ◀

This phenomenon is reminiscent of how, in a suitable small-step limit, a random walk of increasingly small step size converges to the Brownian motion, which essentially has paths of smoothness 1/2:

Remark 17 (Intuition: analogy to random walks)**.** Let $g_n : [0,1] \to \mathbb{R}$ be the random walk of n steps, defined as $g_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} Z_i$ with i.i.d. coin tosses Z_i. Then g_n has n non-smoothnesses of degree $s = 0$ which are increasingly dense in $[0,1]$, and converges weakly to a standard Brownian motion on $[0,1]$. The paths of a Brownian motion are essentially of (Hölder) smoothness 1/2; they gain half an order of smoothness. Intuitively, one might think of this that as the non-smoothnesses become dense, they also become smaller and in this sense less severe, however covering the whole interval.

This is similar to what can be observed in a two-layer neural network: Consider a shallow network with bias $\sigma_b^2\sigma_i^2 > 0$ and the Heaviside activation function $\varphi(x) = \mathbb{1}_{\mathbb{R}_{\geq 0}}(x)$ on the one-dimensional sphere $\mathbb{S}^1$. The network of width n has at most n non-smoothnesses, and as n increases, these non-smoothnesses will be increasingly dense in $\mathbb{S}^1$. So, in analogy to the random walk, the finite networks are functions of smoothness 0 with increasingly small distances between the non-smoothnesses, and the jumps in these points shrink as n grows. The NNGP-RKHS of the infinite-width network is $\mathcal{H}_{k_L^{\mathrm{NNGP}}} = H^{3/2}(\mathbb{S}^1)$ by Theorem 9. The path smoothness theorem 14 shows that the sample paths of the NNGP, that is the infinite width networks at initialization, are essentially of smoothness 1/2. We observe the same effect: the paths gain half an order of smoothness compared to the finite networks.

The same intuition applies for activations of higher smoothness $s \geq 1$.

For the NTK, essentially the same intuition can be envoked. Since the NTK includes derivatives of the activation function, the path paths are one level rougher, they have smoothness $s - 1/2$. Note that the Heaviside activation does not allow an NTK, since it is not weakly differentiable. ◀

6 POSSIBLE EXTENSIONS

While this work considers neural networks consisting only of feedforward-layers, we briefly discuss possible extensions to more complex architectures, by sketching the required steps to include layer normalization and residual layers. In a similar style, adapting our argumentation to many other architectures might be possible.

Our core strategy is to write the kernels as sums, products, and compositions of dual activations and linear functions. Our results from Appendix C prove the boundary behavior of dual activations, and Appendix B introduces a calculus for how this boundary behavior behaves under sums, products, and compositions. Hence, as long as other kernels can be written in the same form, our tools should be applicable to them, and in this case one can proceed similarly as we did to obtain Theorem 9.

Note that a network with complex architecture remains (in distribution, in the case of finite width) rotationally invariant if the first operation in the network is a linear layer with Gaussian initialization.

Residual layers Residual layers fit well into the framework we considered: Similar recursive fromulas as presented in Lemma D.2 can be derived for a network containing residual layers using Appendix E of Yang (2020). Then, it may be possible to imitate the analysis done in the further course of the Appendix D for this recursion formula, which would work similar to the analysis of the NNGP-kernel.

Layer normalization For simplicity, let us consider RMSNorm (Zhang and Sennrich, 2019), which is popular in Transformers, without parameters. Let the l-th post-activation layer for $l \geq 2$ be normalized to a unit vector as $\tilde{\boldsymbol{x}}^{(l-1)} := \varphi(\boldsymbol{z}^{(l)})/\left\|\varphi(\boldsymbol{z}^{(l)})\right\|_2$. In a recursive decomposition of the kernel as in Lemma D.2, this leads to a linear rescaling of the kernel at each layer. In a standard feedforward-network as in Definition 1, the norm of the post activation vector *without layer normalization* converges almost surely to a constant value, namely $\sqrt{k_l^{\mathrm{NNGP}}(\boldsymbol{x}, \boldsymbol{x})}$, by Lemma D.2. Our analysis shows that the scale of that value does not influence the RKHSs associated to the neural network – as an illustrative example, the RKHSs do not change when changing the activation function from φ to $\lambda\varphi$ for $\lambda \neq 0$.

7 IMPLICATIONS

Our results can be used to obtain a more complete understanding of neural networks in the kernel regime. Below, we list some non-exhaustive ways in which our results could complement existing theory.

Deep vs. shallow networks By showing that the RKHSs of deep and shallow ReLU NTKs are equivalent, Bietti and Bach (2021) concluded that the NTK regime is not sufficient to model the benefits of depth for practical neural networks. Our results show that this holds for a much larger class of not infinitely smooth activation functions: Whenever $s \in [1, \infty)$, all depths ≥ 2 (or ≥ 3 in the bias-free case) yield equivalent RKHSs for the NTK as well as for the NNGP. On the other hand, our results show that the RKHSs are not equivalent for polynomial activations, while the situation for other infinitely smooth activations remains unclear, as more precise results are only known in the two-layer case (Murray et al., 2023).

Training dynamics Regarding theoretical modeling, the dynamics of gradient flow are difficult to study for ReLU activations since their derivative is discontinuous at zero. The use of smoother but not infinitely smooth activations such as CELU can enable theoretical studies of gradient flow in a setting where the RKHS is well-known, without limiting results to powers of ReLU as in previous work (Vakili et al., 2023). For example, the gradient flow analysis of Bowman and Montufar (2022) yields finite-time bounds for the deviation from the infinite-width training trajectory but requires activation functions smoother than ReLU. The use of semi-smooth activation functions might also be interesting to the NTK analysis of physics-informed neural networks (Wang et al., 2022), since their training involves taking higher-order derivatives of neural networks. The faster eigenvalue decay for smoother activations can lead to slower training dynamics (e.g., Raskutti et al., 2014; Cao et al., 2019). Compared to the pure kernel case, gradient flow on infinite-width NNs in the kernel regime also has to learn to remove the random initial function (Lee et al., 2019). Thanks to Theorem 14, we now know the smoothness of this function, which makes it possible to derive convergence rates for the regression of this function with the NTK. To this end, we refer to the proof of Theorem G.5 in Haas et al. (2023) for convergence rates with regularization in the noisy case, and to Theorem 3.3 in Le Gia et al. (2006) for an approach towards convergence rates of interpolation.

Generalization Our main result allows the application of generalization results to more activation functions. For example, the inconsistency results of Haas et al. (2023) for overfitting with NTKs and NNGPs apply whenever the RKHS is a Sobolev space, and hence by our results, they apply to a wide range of finitely smooth activation functions. Our results also suggest that the spectral bias of neural networks (Rahaman et al., 2019) can be influenced by changing the smoothness of the activation function. Finally, while Simon et al. (2022) show that a large class of kernels can be represented as NTKs and NNGPs, our results help to elucidate the properties of activation functions that realize these kernels and consider networks of any depth.

8 RELATED WORK

Neural kernels Multiple results have shown that in typical settings, infinite-width neural networks behave like Gaussian processes at initialization (Neal, 1996; Daniely et al., 2016; Lee et al., 2018; Matthews et al., 2018). The associated covariance function is known as the neural network Gaussian process (NNGP) kernel or random features kernel. Jacot et al. (2018) discovered that the training dynamics of such infinite-width neural networks can be described by a different kernel, the so-called neural tangent kernel (NTK). The same observation has been made by Lee et al. (2019). Follow-up work has generalized the scope of these results (e.g., Arora et al., 2019; Yang, 2019b, 2020). These works do not yet give deep insights into the nature of the NTK and NNGP kernels.

Smoothness of neural kernels Belkin et al. (2018) noted similarities in generalization behavior between the Laplace kernel and neural networks. A follow-up line of work managed to characterize the RKHS of different NTKs on $\mathbb{S}^d$. Bietti and Mairal (2019) derived the structure of the RKHS of the two-layer ReLU NTK without bias ($\sigma_b^2 = 0$). Basri et al. (2019) analyzed two-layer ReLU NTKs with and without bias. Geifman et al. (2020) showed that the two-layer ReLU NTK with bias is equivalent to the Laplace kernel, and showed that the RKHS of the deep ReLU NTK is at least as large as the RKHS of the shallow one. The equivalence of these RKHSs was then shown by Chen and Xu (2021) for $\sigma_b^2 > 0, \sigma_i = 0, L \geq 2$. Simultaneously, Bietti and Bach (2021) characterized the structure of the deep ReLU NTK and NNGP without bias ($\sigma_b^2 = 0, L \geq 3$) and also showed some results for step activations as well as infinitely differentiable activations. Vakili et al. (2023) extended this analysis to RePU (integer powers of the ReLU activation). Finally, Haas et al. (2023) formally established the connection of these RKHSs to Sobolev spaces on the sphere. Murray et al. (2023) prove further spectral properties of the NTK, including a characterization of the RKHS of two-layer NNs for tanh and RBF activation functions that is more precise than our result for these activations. Scetbon and Harchaoui (2021) prove some results about spectral properties of dot-product kernels, with weak results concerning the NTK. However, to the best of our knowledge, we are the first to provide a characterization of the RKHS for many non-smooth activation functions like LeakyReLU, SELU, or ELU.

The NTK has also been analyzed for other architectures such as residual networks (Belfer et al., 2024) and convolutional residual networks (Barzilai et al., 2023). Dandi and Jacot (2021) investigate the contributions of individual layers towards the properties of the NTK. Simon et al. (2022) show that a large class of dot-product kernels on the sphere can be realized as NTK or NNGP kernels through suitable activation functions. While most results analyze the NTK on $\mathbb{S}^d$, Lai et al. (2023) analyze the eigenvalue decay on $\mathbb{R}$ and Li et al. (2024) leverage known results on $\mathbb{S}^d$ to obtain decay rates for $\mathbb{R}^{d+1}$. In terms of activation functions, these results do not go beyond the known results for $\mathbb{S}^d$.

Spectral properties Panigrahi et al. (2020) investigate minimum eigenvalues of NTK matrices and find a dependence on the smoothness of the activation function. Nguyen et al. (2021) and Karhadkar et al. (2024) provide more bounds for minimum eigenvalues. These results are related to the structure of the RKHS through kernel matrix concentration inequalities but do not reveal the full structure of the RKHS.

9 CONCLUSION

We have shown general results for the structure of the RKHS for fully connected neural networks with different activation functions, as well as results for the sample path smoothness of randomly initialized wide neural networks. We refer the reader to Appendix A for a short overview of the central techniques and objects of our main proof.

Possible extensions Our work offers several possibilities for extensions. First, following Li et al. (2024), an extension of our results from $x \in \mathbb{S}^d$ to $x \in \mathbb{R}^{d+1}$ could be possible. Second, one could study different activation functions in different layers; we strongly conjecture that the behavior of the NTK and NNGP will be determined by the least smooth activation function. Third, while we study non-smoothness in zero, it might be possible to apply our proof technique to functions that are non-smooth in other points—the main obstacle is to find reference functions s_k with the given non-smoothness that are similarly amenable to analysis as the functions s_k we use in the appendix. Fourth, a more precise characterization of the RKHS for infinitely smooth non-polynomial activations is still open but might require stronger tools (e.g., Minh et al., 2006; Azevedo and Menegatto, 2014; Murray et al., 2023). Finally, our results might be extensible to analyses for residual networks (e.g., Belfer et al., 2024; Tirer et al., 2022), convolutional networks (e.g., Geifman et al., 2022; Barzilai et al., 2023), or more general architectures including transformers (Yang, 2020).

ACKNOWLEDGMENTS

The authors thank Ingo Steinwart, Tizian Wenzel, Jens Wirth, and Alberto Bietti for valuable discussions.

Max Schölpple was funded by the Deutsche Forschungsgemeinschaft (DFG) in the project STE 1074/5-1, within the DFG priority programm SPP 2298 "Theoretical Foundations of Deep Learning", and by the International Max Planck Research School for Intelligent Systems (IMPRS-IS).

References

Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, Russ R. Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, 2019.

D. Azevedo and V.A. Menegatto. Sharp estimates for eigenvalues of integral operators generated by dot product kernels on the sphere. *Journal of Approximation Theory*, 177:57–68, 2014. ISSN 0021-9045.

Daniel Barzilai, Amnon Geifman, Meirav Galun, and Ronen Basri. A kernel perspective of skip connections in convolutional networks. In *International Conference on Learning Representations*, 2023.

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

Yuval Belfer, Amnon Geifman, Meirav Galun, and Ronen Basri. Spectral analysis of the neural tangent kernel for deep residual networks. *Journal of Machine Learning Research*, 25(184):1–49, 2024.

Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, 2018.

Alberto Bietti and Francis Bach. Deep equals shallow for ReLU networks in kernel regimes. In *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*, 2019.

M. S. Birman and M. Z. Solomjak. *Spectral Theory of Self-Adjoint Operators in Hilbert Space*. Springer, Dordrecht, 1987.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pages 1024–1034. PMLR, 2020.

Benjamin Bowman and Guido F Montufar. Spectral bias outside the training set for deep networks in the kernel regime. *Advances in Neural Information Processing Systems*, 2022.

Johann S. Brauchart and Josef Dick. A characterization of Sobolev spaces on the sphere and an extension of Stolarsky's invariance principle to arbitrary smoothness. *Constructive approximation*, 38 (3):397–445, 2013.

Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan Gu. Towards understanding the spectral bias of deep learning. *arXiv:1912.01198*, 2019.

Lin Chen and Sheng Xu. Deep neural tangent kernel and laplace kernel have the same RKHS. In *International Conference on Learning Representations*, 2021.

Yatin Dandi and Arthur Jacot. Understanding layer-wise contributions in deep neural networks through spectral analysis. *arXiv:2111.03972*, 2021.

Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. *Advances in Neural Information Processing Systems*, 2016.

Shiv Ram Dubey, Satish Kumar Singh, and Bidyut Baran Chaudhuri. Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing*, 503:92–108, 2022.

Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmüller, Prateek Mutalik Desai, David Salinas, and Frank Hutter. TabArena: A living benchmark for machine learning on tabular data. In *Advances in Neural Information Processing Systems*, 2025.

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*, 2020.

Amnon Geifman, Meirav Galun, David Jacobs, and Basri Ronen. On the spectral bias of convolutional neural tangent and Gaussian process kernels. *Advances in Neural Information Processing Systems*, 2022.

Yury Gorishniy, Akim Kotelnikov, and Artem Babenko. TabM: Advancing tabular deep learning with parameter-efficient ensembling. In *International Conference on Learning Representations*, 2025.

Moritz Haas, David Holzmüller, Ulrike Luxburg, and Ingo Steinwart. Mind the spikes: Benign overfitting of kernels and neural networks in fixed dimension. In *Advances in Neural Information Processing Systems*, 2023.

David Holzmüller, Léo Grinsztajn, and Ingo Steinwart. Better by default: Strong pre-tuned MLPs and boosted trees on tabular data. *Advances in Neural Information Processing Systems*, 2024.

Simon Hubbert, Quôc Thông Lê Gia, and Tanya M. Morton. *Spherical Radial Basis Functions, Theory and Applications*. Springer Briefs in Mathematics, 2015.

Simon Hubbert, Emilio Porcu, Chris J Oates, and Mark Girolami. Sobolev spaces, kernels and discrepancies over hyperspheres. *Transactions on Machine Learning Research*, 2023.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018.

Kedar Karhadkar, Michael Murray, and Guido F Montufar. Bounds for the smallest eigenvalue of the NTK for arbitrary spherical data of arbitrary dimension. *Advances in Neural Information Processing Systems*, 37:138197–138249, 2024.

Jianfa Lai, Manyun Xu, Rui Chen, and Qian Lin. Generalization ability of wide neural networks on $\mathbb{R}$. *arXiv:2302.05933*, 2023.

Quoc Thong Le Gia, Francis J Narcowich, Joseph D Ward, and Holger Wendland. Continuous and discrete least-squares approximation by radial basis functions on spheres. *Journal of Approximation Theory*, 143(1):124–133, 2006.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, 2019.

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

Milan N. Lukić and Jay H. Beder. Stochastic processes with sample paths in reproducing kernel Hilbert spaces. *Transactions of the American Mathematical Society*, 353:3945–3969, 2001.

Alexander G de G Matthews, Jiri Hron, Mark Rowland, Richard E Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018.

Ha Quang Minh, Partha Niyogi, and Yuan Yao. Mercer's theorem, feature maps, and smoothing. In *International Conference on Computational Learning Theory*. Springer, 2006.

Claus Müller. *Spherical Harmonics*. Springer Berlin Heidelberg, 2006.

Michael Murray, Hui Jin, Benjamin Bowman, and Guido Montufar. Characterizing the spectrum of the NTK via a power series expansion. In *The Eleventh International Conference on Learning Representations*, 2023.

Radford M Neal. Priors for infinite networks. *Bayesian learning for neural networks*, pages 29–53, 1996.

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

Guillermo Ortiz-Jiménez, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. What can linearized neural networks actually say about generalization? *Advances in Neural Information Processing Systems*, 2021.

Abhishek Panigrahi, Abhishek Shetty, and Navin Goyal. Effect of activation functions on the training of overparametrized neural nets. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=rkgfdeBYvH`.

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 *International Conference on Machine Learning*, 2019.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *Journal of Machine Learning Research*, 15(1):335–366, 2014.

Meyer Scetbon and Zaid Harchaoui. A spectral analysis of dot-product kernels. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Max Schölpple and Ingo Steinwart. Which spaces can be embedded in reproducing kernel Hilbert spaces? *Constructive Approximation*, 2025.

James Benjamin Simon, Sajant Anand, and Mike Deweese. Reverse engineering the neural tangent kernel. In *International Conference on Machine Learning*, 2022.

Ingo Steinwart. When does a Gaussian process have its paths in a reproducing kernel Hilbert space? *arXiv:2407.11898*, 2024.

Tom Tirer, Joan Bruna, and Raja Giryes. Kernel-based smoothness analysis of residual networks. In *Mathematical and Scientific Machine Learning*, 2022.

Sattar Vakili, Michael Bromberg, Jezabel Garcia, Dashan Shiu, and Alberto Bernacchia. Uniform generalization bounds for overparameterized neural networks. *arXiv:2109.06099*, 2021.

Sattar Vakili, Michael Bromberg, Jezabel Garcia, Dashan Shiu, and Alberto Bernacchia. Information gain and uniform generalization bounds for neural kernel models. In *IEEE International Symposium on Information Theory*, 2023.

Nikhil Vyas, Yamini Bansal, and Preetum Nakkiran. Limitations of the NTK for understanding generalization in deep learning. *arXiv:2206.10012*, 2022.

Sifan Wang, Xinling Yu, and Paris Perdikaris. When and why PINNs fail to train: A neural tangent kernel perspective. *Journal of Computational Physics*, 449:110768, 2022.

Jonathan Wenger, Felix Dangel, and Agustinus Kristiadi. On the disconnect between theory and practice of neural networks: Limits of the NTK perspective. *arXiv:2310.00137*, 2023.

Andreas Winkelbauer. Moments and absolute moments of the normal distribution. *arXiv:1209.4340*, 2012.

Greg Yang. Tensor programs i: Wide feedforward or recurrent neural networks of any architecture are gaussian processes. *arXiv:1910.12478*, 2019a.

Greg Yang. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *arXiv:1902.04760*, 2019b.

Greg Yang. Tensor programs II: Neural tangent kernel for any architecture. *arXiv:2006.14548*, 2020.

Greg Yang and Edward J Hu. Tensor programs IV: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, 2021.

Han-Jia Ye, Si-Yang Liu, Hao-Run Cai, Qi-Le Zhou, and De-Chuan Zhan. A closer look at deep learning on tabular data. *arXiv:2407.00956*, 2024.

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

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, 2019.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Not Applicable]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] This paper is dedicated to theoretical results.
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

Appendix Contents.

A OVERVIEW AND NOTATION

Appendix A – D are dedicated to the proof of the main theorem, Theorem 9. They build on each other and should be read in order. Appendix E and Appendix F contain smaller proofs that are independent of the other appendices. Appendix G derives a way to numerically compute dual activations for our experiments.

Analysis through boundary behavior. As mentioned before, the NTK and NNGP kernels on the sphere are dot-product kernels, i.e., they are of the form $k(x, x') = \kappa(\langle x, x'\rangle)$. To obtain eigenvalue decays for k, we build on the results of Bietti and Bach (2021), which requires to study the behavior of $\kappa(t)$ for $t \to 1$ and $t \to -1$. To unify these analyses, we let $\tau \in \{-1, 1\}$ and study the behavior of $\kappa_\tau : (0, 2) \to \mathbb{R}, t \mapsto \kappa(\tau(1-t))$ for $t \searrow 0$. We first introduce function classes with certain boundary behavior that will be central in simplifying computations later on:

Definition A.1 (Function classes with controlled boundary behavior)**.** Let $\alpha, \beta, \gamma \in \mathbb{R}$. We define sets of functions $\mathcal{P}_{\alpha,\beta}$, $\mathcal{R}_\gamma$, and $\mathcal{Q}_{\alpha,\beta}$ as follows:

(a) **Sums of high-enough powers:** We define $\mathcal{P}_{\alpha,\beta}$ as the set of functions $f : (0, 2) \to \mathbb{R}$ of the form

$$f(t) = \sum_{i=1}^{n} a_i t^{\alpha_i} \quad \text{with} \quad n \in \mathbb{N}_0, a_i, \alpha_i \in \mathbb{R}, \begin{cases} \alpha_i > \alpha & \text{for all } i \text{ with } \alpha_i \in \mathbb{Z} \\ \alpha_i > \beta & \text{for all } i \text{ with } \alpha_i \notin \mathbb{Z} \end{cases}$$

Note the use of $>$, which will be convenient later.

(b) **Negligible remainder functions:** We define $\mathcal{R}_\gamma$ as the set of C^∞-functions $f : (0, 2) \to \mathbb{R}$ such that for all $m \in \mathbb{N}_0$, the m-th derivative $f^{(m)}$ satisfies

$$|f^{(m)}(t)| \leq O(t^{\gamma - m}) \quad \text{for } t \searrow 0 .$$

Intuitively, $\mathcal{R}_\gamma$ contains functions that are at least as benign as $t \mapsto t^\gamma$, in terms of the behavior of all derivatives for $t \to 0$.

(c) **Sums of powers plus negligible remainder:** We define $\mathcal{Q}_{\alpha,\beta} := \bigcap_{\gamma\in\mathbb{R}}(\mathcal{P}_{\alpha,\beta} + \mathcal{R}_\gamma)$. In other words, $\mathcal{Q}_{\alpha,\beta}$ contains those functions that are sums of powers plus an arbitrarily "high-order" remainder. Note for $f \in \mathcal{Q}_{\alpha,\beta}$ that in general, the number n of terms in the sum of powers when writing f as $f \in \mathcal{P}_{\alpha,\beta} + \mathcal{R}_\gamma$ will depend on the imposed order γ of the remainder.

We will mostly use $\mathcal{Q}_{\alpha,\beta}$ in the paper. Instead of $f \in \mathcal{Q}_{\alpha,\beta}$, we also write $f(t) = \mathcal{Q}_{\alpha,\beta}(t)$ analogous to common O-notation. We often track coefficients of leading terms separately, e.g., writing $g(t) = at^\alpha + bt^\beta + \mathcal{Q}_{\alpha,\beta}(t)$. ◀

Furthermore, we require refined adapted asymptotic notation for sequences, as we need the asympototic behavior of a sequence (a_n) but also need to know wether there are elements $a_n = 0$, which is crucial to compare Sobolev spaces based on eigenvalues, cf. Lemma 2.

Definition A.2 (Asymptotic notation)**.** For functions κ, we use standard asymptotic notation like $\kappa(t) = O(t^{1/2})$ for $t \searrow 0$. For sequences $(a_n)_{n\in I}, (b_n)_{n\in I} \subseteq \mathbb{R}_{\geq 0}$ indexed by $I \subseteq \mathbb{N}_0$, we use an index "$\forall n$" for asymptotic notation to denote that it should hold for all n and not only for almost all n, and to indicate the variable n that the constant is independent of. Specifically,

$$\begin{aligned} a_n = O_{\forall n}(b_n) &:\Leftrightarrow \exists C > 0 : \forall n \in I : a_n \leq C b_n \\ a_n = \Omega_{\forall n}(b_n) &:\Leftrightarrow b_n = O_{\forall n}(a_n) \\ a_n = \Theta_{\forall n}(b_n) &:\Leftrightarrow a_n = O_{\forall n}(b_n) \text{ and } b_n = O_{\forall n}(a_n) \\ a_n = o_{\forall n}(b_n) &:\Leftrightarrow a_n = O_{\forall n}(b_n) \text{ and } \forall C > 0 : \exists n_0 \in \mathbb{N}_0 : \forall n \geq n_0 : a_n \leq C b_n . \end{aligned}$$

Using this notation, when we write $\mu_l = \Theta_{\forall l}((l + 1)^{-d-2s})$, it implies that all μ_l are positive. We write $l + 1$ and not l to make the right-hand side well-defined for all $l \in \mathbb{N}_0$. The constant C implied in this notation can depend on d and s. ◀

A key element of our analysis is the following result, formulated using the previously defined boundary function classes:

Theorem A.3 (Adaptation of Theorem 7 in arXiv v4 of Bietti and Bach 2021)**.** *Let $\kappa : [-1, 1] \to \mathbb{R}$ be a function that is smooth on $(-1, 1)$ such that $k_{\kappa,d}(\boldsymbol{x}, \boldsymbol{x}') = \kappa(\langle \boldsymbol{x}, \boldsymbol{x}'\rangle)$ is a positive semi-definite kernel on all* *spheres* $\mathbb{S}^d, d \in \mathbb{N}_{\geq 1}$*. Suppose that there exists* $0 < \beta \in \mathbb{R} \setminus \mathbb{Z}$ *and* $b_{-1}, b_1 \in \mathbb{R}$ *such that for* $\tau \in \{-1, 1\}$,

$$\kappa(\tau(1-t)) = b_\tau t^\beta + \mathcal{Q}_{-1,\beta}(t) \ .$$

Then, for a given dimension $d \in \mathbb{N}_{\geq 1}$*, the eigenvalues* $\mu_l = \mu_l(\kappa, d)$ *as defined in Section 2 satisfy:*

(a) *For* $l \in \mathbb{N}_0$ *even, if* $b_{-1} \neq -b_1$*, then* $\mu_l = \Theta_{\forall l}((l+1)^{-d-2\beta})$.
(b) *For* $l \in \mathbb{N}_0$ *even, if* $b_{-1} = -b_1$*, then* $\mu_l = o_{\forall l}((l+1)^{-d-2\beta})$.
(c) *For* $l \in \mathbb{N}_0$ *odd, if* $b_{-1} \neq b_1$*, then* $\mu_l = \Theta_{\forall l}((l+1)^{-d-2\beta})$.
(d) *For* $l \in \mathbb{N}_0$ *odd, if* $b_{-1} = b_1$*, then* $\mu_l = o_{\forall l}((l+1)^{-d-2\beta})$.

This leads to the following strategy, detailed below, for analyzing the eigenvalues of neural kernels on the sphere:

(1) Write neural kernels on the sphere as a composition of dual activation functions.
(2) Analyze the boundary behavior of dual activation functions.
(3) Analyze the boundary behavior for sums, products, and compositions of functions, which will be done in Proposition B.2.
(4) Analyze the behavior of even and odd functions to obtain stronger results for the special cases (b) and (d) in Theorem A.3.
(5) Assemble everything.

Dual activations and neural kernels. To apply the theorem above, similar to Bietti and Bach (2021), we derive a convenient expression for κ using *dual activation functions* introduced by Daniely et al. (2016).

Definition A.4 (Dual and rescaled activation functions)**.** For a function $\varphi \in L_2(\mathcal{N}(0,1))$, we follow Daniely et al. (2016) and define the dual activation

$$\widehat{\varphi} : [-1,1] \to \mathbb{R}, t \mapsto \mathbb{E}_{(u,v)\sim\mathcal{N}(0,\Sigma_t)}[\varphi(u)\varphi(v)], \qquad \Sigma_t := \begin{pmatrix} 1 & t \\ t & 1 \end{pmatrix} \ .$$

Moreover, for a function $\varphi : \mathbb{R} \to \mathbb{R}$ and $a \in \mathbb{R}$, we define the rescaled function

$$\varphi_{\cdot a} : \mathbb{R} \to \mathbb{R}, \varphi_{\cdot a}(x) := \varphi(ax) \ . \qquad ◀$$

Our following result, derived in Appendix D.1, shows that the NTK and NNGP kernels restricted to the sphere can be expressed using dual rescaled activations:

Lemma A.5 (Neural kernels on the unit sphere)**.** *Let the activation function* $\varphi : \mathbb{R} \to \mathbb{R}$ *fulfill Assumption 4. Consider a neural network* $f_\theta : \mathbb{R}^{d+1} \to \mathbb{R}$ *initialized as in Definition 1. For* $l \geq 2$ *and* $t \in [-1,1]$ *we recursively define*

$$\begin{aligned}
\alpha_1 &:= \sigma_b^2\sigma_i^2 + \sigma_w^2 \\
\alpha_l &:= \sigma_b^2\sigma_i^2 + \sigma_w^2\widehat{\varphi_{\cdot\sqrt{\alpha_{l-1}}}}(1) \\
\kappa_1^{\mathrm{NNGP}}(t) &:= \sigma_b^2\sigma_i^2 + \sigma_w^2 t \\
\kappa_l^{\mathrm{NNGP}}(t) &:= \sigma_b^2\sigma_i^2 + \sigma_w^2\widehat{\varphi_{\cdot\sqrt{\alpha_{l-1}}}}(\kappa_{l-1}^{\mathrm{NNGP}}(t)/\alpha_{l-1}) \\
\kappa_1^{\mathrm{NTK}}(t) &:= \sigma_b^2(1-\sigma_i^2) + \kappa_1^{\mathrm{NNGP}}(t) \\
\kappa_l^{\mathrm{NTK}}(t) &:= \sigma_b^2(1-\sigma_i^2) + \kappa_l^{\mathrm{NNGP}}(t) + \sigma_w^2\kappa_{l-1}^{\mathrm{NTK}}(t)\widehat{(\varphi')_{\cdot\sqrt{\alpha_{l-1}}}}(\kappa_{l-1}^{\mathrm{NNGP}}(t)/\alpha_{l-1}) \ .
\end{aligned}$$

Then for all $l \geq 1, \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{S}^d$ *we have* $\alpha_l > 0$ *and*

$$\begin{aligned}
k_l^{\mathrm{NNGP}}(\boldsymbol{x}, \bar{\boldsymbol{x}}) &= \kappa_l^{\mathrm{NNGP}}(\langle \boldsymbol{x}, \bar{\boldsymbol{x}} \rangle) \\
k_l^{\mathrm{NTK}}(\boldsymbol{x}, \bar{\boldsymbol{x}}) &= \kappa_l^{\mathrm{NTK}}(\langle \boldsymbol{x}, \bar{\boldsymbol{x}} \rangle) \ .
\end{aligned}$$

Boundary behavior of dual activations. To study the behavior of dual activations at ± 1, we first need a definition:

Definition A.6 (Reference activations)**.** For $k \in \mathbb{N}_0$, we define the activation functions $s_k : \mathbb{R} \to \mathbb{R}$ by

$$s_k(x) := \frac{1}{2k!}\operatorname{sgn}(x)x^k \ , \qquad \operatorname{sgn}(x) := \begin{cases} -1 & , x < 0 \\ 0 & , x = 0 \\ 1 & , x > 0 \ . \end{cases}$$

Let φ be an activation function as in Assumption 4. For $k \in \mathbb{N}_0$, we define the coefficients

$$\Delta_k(\varphi) := \varphi^{(k)}(0+) - \varphi^{(k)}(0-) ,$$

where $\varphi^{(k)}(0+)$ is the right-sided limit of the k-th derivative of φ at zero, and $\varphi^{(k)}(0-)$ is the left-sided limit. ◀

The motivation behind Definition A.6 is given in Lemma C.15, which says that we can decompose

$$\varphi = \sum_{k=0}^{m-1} \Delta_k(\varphi) s_k + \varphi_m , \tag{A.1}$$

where φ_m is m times pseudo-differentiable (see Definition C.1). This decomposition is central to our proof.

Remark A.7. The reason why our analysis is restricted to a non-smoothness at the origin is that the "reference activations" $s_k(x) = \frac{1}{2k!}\operatorname{sgn}(x)x^k$ introduced in Definition A.6 have their only non-smoothness in the origin. For these reference activations, we can study their dual activation through analytic means (e.g., Lemmata C.6 and C.7). Generalizing our results to non-smoothness in another point b would require finding a similar family of reference functions that have their non-smoothness in b while still allowing similar calculations for the dual activation. This is non-trivial, but if it can be done, the rest of our analysis should still apply in the same way. ◀

As in Daniely et al. (2016), we use h_n to denote the n-th normalized probabilist's Hermite polynomial, which means that h_n is a polynomial of degree n and $(h_n)_{n\in\mathbb{N}_0}$ forms an orthonormal basis of the Hilbert space $L_2(\mathcal{N}(0,1))$. For a function $f \in L_2(\mathcal{N}(0,1))$ and $n \in \mathbb{N}_0$, we define its n-th Hermite coefficient

$$a_n(f) := \langle f, h_n \rangle_{L_2(\mathcal{N}(0,1))} ,$$

such that we have the following Hermite expansion in $L_2(\mathcal{N}(0,1))$:

$$f = \sum_{n=0}^{\infty} a_n(f) h_n .$$

Daniely et al. (2016) show that the Hermite series of φ is related to the Maclaurin series of $\widehat{\varphi}$ via

$$\widehat{\varphi}(t) = \sum_{n=0}^{\infty} a_n(\varphi)^2 t^n . \tag{A.2}$$

This also demonstrates that the "dualization" of activation functions should be thought of as a (function-valued) quadratic form and not dualization. Using our decomposition in Eq. (A.1), we therefore obtain

$$\begin{aligned}\widehat{\varphi}(t) &= \sum_{i=0}^{m}\sum_{j=0}^{m}\sum_{n=0}^{\infty} \Delta_i(\varphi)\Delta_j(\varphi) a_n(s_i) a_n(s_j) t^n \\ &\quad + \sum_{i=0}^{m}\sum_{n=0}^{\infty} 2\Delta_i(\varphi) a_n(s_i) a_n(\varphi_m) t^n + \sum_{n=0}^{\infty} a_n(\varphi_m)^2 t^n .\end{aligned} \tag{A.3}$$

We will treat the first term involving $a_n(s_i)a_n(s_j)$ analytically, while using the smoothness of φ_m to obtain fast decay rates for $a_n(\varphi_m)$, which will imply that the behavior at the boundary is dominated by the first term. In total, we obtain the following result at the end of Appendix C.5:

Theorem A.8 (Boundary behavior of dual activations)**.** *Let φ be an activation function as in Assumption 4 and let $m \in \mathbb{N}_0$ such that $\Delta_k(\varphi) = 0$ for all $k < m$ (see Definition A.6). Then, there exists $b_m > 0$ depending only on m such that for $\tau \in \{\pm 1\}$,*

$$\widehat{\varphi}(\tau(1-t)) = \Delta_m(\varphi)^2 (-\tau)^{m+1} b_m t^{m+1/2} + \mathcal{Q}_{-1,m+1/2}(t)$$

holds.

To deal with the derivative of the activation in the NTK, we can use the convenient formula $\widehat{\varphi'} = \widehat{\varphi}'$ from Daniely et al. (2016), of which a generalized formulation (for pseudo-derivatives, see Definition C.1) is proved in Lemma C.2.

Even and odd functions. Let $I \subseteq \mathbb{R}$ be symmetric around zero and let $f : I \to \mathbb{R}$ be a function. We say that f is even if $f(x) = f(-x)$ for all $x \in I$, and that f is odd if $f(x) = -f(-x)$ for all $x \in I$. Every function $f : I \to \mathbb{R}$ can be decomposed in its even and odd part

$$f_{\text{even}}(x) = \frac{f(x) + f(-x)}{2}, \qquad f_{\text{odd}}(x) = \frac{f(x) - f(-x)}{2} \ .$$

We then have $f = f_{\text{even}} + f_{\text{odd}}$, f_{even} is even, and f_{odd} is odd.

For $n \in \mathbb{Z}$, we define

$$\text{even}(n) \coloneqq \begin{cases} 1 & , n \text{ is even} \\ 0 & , n \text{ is odd,} \end{cases} \qquad \text{odd}(n) \coloneqq \begin{cases} 0 & , n \text{ is even} \\ 1 & , n \text{ is odd,} \end{cases}$$

The Hermite polynomials h_n are even for even n and odd for odd n.

We show in Proposition D.4 that two-layer bias-free neural kernels satisfy an even-odd decomposition, and deeper bias-free neural kernels can be even/odd whenever the activation function is even/odd.

B BOUNDARY BEHAVIOR

In the following, we will prove some rules for algebraic manipulations with the function classes $\mathcal{P}_{\alpha,\beta}, \mathcal{R}_\gamma, \mathcal{Q}_{\alpha,\beta}$ from Definition A.1. While we mostly care about $\mathcal{Q}_{\alpha,\beta}$ later, we will first prove rules for $\mathcal{P}_{\alpha,\beta}$ and $\mathcal{R}_\gamma$ as a simpler intermediate step, before providing rules for $\mathcal{Q}_{\alpha,\beta}$ in Proposition B.2, which are summarized in Table B.1:

Lemma B.1 (Rules for $\mathcal{P}_{\alpha,\beta}$ and $\mathcal{R}_\gamma$)**.**

(a) For $\alpha \in \mathbb{R}$, we have $\mathcal{P}_{\alpha,\alpha} \subseteq \mathcal{R}_\alpha$.
(b) Let $\gamma_1, \gamma_2 \in \mathbb{R}$. Let $f_1, f_2 : (0,2) \to \mathbb{R}$ with $f_i \in \mathcal{R}_{\gamma_i}$. Then, $f_1 \cdot f_2 \in \mathcal{R}_{\gamma_1+\gamma_2}(t)$.
(c) If $f_i \in \mathcal{P}_{\alpha_i,\beta_i}$, $\alpha_i, \beta_i \in \mathbb{R}$, $\alpha_i \le \beta_i$, then

$$f_1 \cdot f_2 \in \mathcal{P}_{\alpha_1+\alpha_2, \min\{\alpha_1+\beta_2, \beta_1+\alpha_2\}} \ .$$

(d) Let $A, B > 0$, $g_1 : (0,2) \to (0,2)$, $g_2 : (0,2) \to \mathbb{R}$. Suppose that $g_2 \in \mathcal{R}_{\gamma_2}, \gamma_2 \in \mathbb{R}$ and that $g_1(t) = at^\alpha + \mathcal{R}_{\gamma_1}(t)$, $\gamma_1 > \alpha > 0$ and $a > 0$. Then,

$$g_2 \circ g_1 \in \mathcal{R}_{\alpha\gamma_2} \ .$$

(e) Let $J \subseteq \mathbb{R}$ be an interval, let $g_1 : (0,2) \to J$, and let $g_2 \in C^\infty(J)$ with $0 \in J$. Moreover, suppose that $g_1(t) = at^\alpha + bt^\beta + \mathcal{P}_{\alpha,\beta}(t) + \mathcal{R}_\gamma(t)$ with $\lim_{t \searrow 0} g_1(t) = 0$, $0 \le \alpha \le \beta \le \gamma$, $\alpha \in \mathbb{N}_0$, $\beta, \gamma \in \mathbb{R}$ and $a, b \in \mathbb{R}$. Then,

$$g_2(g_1(t)) = g_2(0) + g_2'(0)\alpha a t^\alpha + g_2'(0)\beta b t^\beta + \mathcal{P}_{\alpha,\beta}(t) + \mathcal{R}_\gamma(t) \ .$$

(f) Let $g_1 : (0,2) \to (0,\infty)$ such that $g_1(t) = at^\alpha + \mathcal{P}_{\alpha,\alpha}(t) + \mathcal{R}_\gamma(t)$ for $a, \alpha > 0$ and $\gamma > \alpha$. Then, for $\delta > 0$,

$$g_1(t)^\delta = a^\delta t^{\alpha\delta} + \mathcal{P}_{\alpha\delta,\alpha\delta}(t) + \mathcal{R}_{\gamma-\alpha+\alpha\delta}(t) \ .$$

Proof.

(a) This is straightforward.
(b) For $n \in \mathbb{N}_0$, we have

$$\begin{aligned} |(f_1 \cdot f_2)^{(n)}(t)| &= \left| \sum_{k=0}^n \binom{n}{k} f_1^{(k)}(t) f_2^{(n-k)}(t) \right| \\ &\le \sum_{k=0}^n O(t^{\gamma_1 - k}) O(t^{\gamma_2-(n-k)}) = O(t^{\gamma_1+\gamma_2-n}) \ . \end{aligned}$$

(c) Consider integers $\alpha_i' > \alpha_i$ and non-integers $\beta_i' > \beta_i$ such that $t^{\alpha_i'} = \mathcal{P}_{\alpha_i,\beta_i}(t)$ and $t^{\beta_i'} = \mathcal{P}_{\alpha_i,\beta_i}(t)$. We now investigate all possible products of such terms:

- The term $t^{\alpha_1'+\alpha_2'}$ has integer power $\alpha_1' + \alpha_2' > \alpha_1 + \alpha_2$.
- The term $t^{\alpha_1'+\beta_2'}$ has non-integer power $\alpha_1' + \beta_2' > \alpha_1 + \beta_2$.
- The term $t^{\beta_1'+\alpha_2'}$ has non-integer power $\beta_1' + \alpha_2' > \beta_1 + \alpha_2$.
- The term $t^{\beta_1'+\beta_2'}$ may have integer or non-integer power. In any case, due to the assumption $\alpha_i \leq \beta_i$, we have

$$\beta_1' + \beta_2' > \beta_1 + \beta_2 \geq \min\{\alpha_1 + \beta_2, \beta_1 + \alpha_2\} \geq \alpha_1 + \alpha_2 \ .$$

(d) Let $g_2 \in \mathcal{R}_{\gamma_2}$. We show

$$|(g_2 \circ g_1)^{(n)}(t)| = O(t^{\gamma_2 \alpha - n})$$

by induction on $n \in \mathbb{N}_0$. Since $0 < \alpha < \gamma_1$, we have $g_1(t) = \Theta(t^\alpha)$ for $t \to 0$ and $\lim_{t\to 0} g_1(t) = 0$. Therefore,

$$g_2(g_1(t)) = O(g_1(t)^{\gamma_2}) = O(t^{\gamma_2 \alpha}) \ .$$

Here, we used $g_1(t) = O(t^\alpha)$ in the case $\gamma_2 > 0$ and $g_1(t) = \Omega(t^\alpha)$ in the case $\gamma_2 < 0$.
For the induction step $n \to n+1$, we use $g_2' \in \mathcal{R}_{\gamma_2 - 1}$ to obtain

$$\begin{aligned}
|(g_2 \circ g_1)^{(n+1)}(t)| &= \left| \frac{\mathrm{d}^n}{\mathrm{d}t^n} (g_2' \circ g_1)(t) \cdot g_1'(t) \right| \\
&= \left| \sum_{k=0}^{n} \binom{n}{k} (g_2' \circ g_1)^{(k)}(t) g_1^{(1+(n-k))}(t) \right| \\
&\leq \sum_{k=0}^{n} O(t^{(\gamma_2 - 1)\alpha - k}) O(t^{(\alpha - 1 - (n-k))}) \\
&= O(t^{\gamma_2 \alpha - (n+1)}) \ .
\end{aligned}$$

(e) We decompose $g_2 = p_2 + r_2$, where p_2 is the degree-$(N-1)$ Taylor polynomial of g_2 around 0 for some $N \geq 2$ to be defined later. Then,

$$p_2(g_1(t)) = g_2(0) + g_2'(0) g_1(t) + \sum_{k=2}^{N-1} \frac{g_2^{(k)}(0)}{k!} g_1(t)^k \ .$$

Since we required $\lim_{t \searrow 0} g_1(t) = 0$, g_1 contains no degree-zero polynomial terms. Therefore, it follows from (a)–(c) that the powers $g_1(t)^k$ for $k \geq 2$ satisfy $g_1(t)^k = \mathcal{P}_{\alpha,\beta}(t) + \mathcal{R}_\gamma(t)$, which shows that

$$p_2(g_1(t)) = g_2(0) + g_2'(0) a t^\alpha + g_2'(0) b t^\beta + \mathcal{P}_{\alpha,\beta}(t) + \mathcal{R}_\gamma(t) \ .$$

Since $\lim_{t \searrow 0} g_1(t) = 0$, we know that $g_1 \in \mathcal{R}_{\tilde{\beta}}$ for some $\tilde{\beta} > 0$. For the remainder function r_2, we could show that $r_2 \in \mathcal{R}_N$ but cannot directly apply (d) since it requires a lower bound of the form $g_1(t) = \Omega(t^{\tilde{\beta}})$, which does not necessarily hold if $a = b = 0$. However, here, the lower bound is not needed: Since we know that $r_2 \in C^\infty$, all derivatives of r_2 are bounded around zero, which is not the case for all $\mathcal{R}_N$ functions. Hence, here is a variant of the proof of (d) for the current setting:
In order to prove that $r_2 \circ g_1 \in \mathcal{R}_{N\tilde{\beta}}$, we want to prove the following statement by induction on n:

> For all $n \in \mathbb{N}_0$, if $h_2 \in C^\infty(J)$ and $M \in \mathbb{N}_0$ such that $h_2(0) = h_2'(0) = \ldots = h_2^{(M-1)}(0) = 0$, then $|(h_2 \circ g_1)^{(n)}(t)| = O(t^{M\tilde{\beta} - n})$.

By choosing $M := N$ large enough such that $M\tilde{\beta} \geq \gamma$ and setting $h_2 := r_2$, we can then conclude that $r_2 \circ g_1 \in \mathcal{R}_\gamma$.
Base case: Let $n = 0$. By applying Taylor's theorem, we find that

$$\begin{aligned}
|h_2(g_1(t))| &= \left| \sum_{k=0}^{M} \frac{h_2^{(k)}(0)}{k!} g_1(t)^k + o(|g_1(t)|^M) \right| \leq \left| \frac{h_2^{(N)}(0)}{N!} g_1(t)^N \right| + o(|g_1(t)|^N) \\
&= O(|g_1(t)|^N) = O(t^{N\tilde{\beta}}).
\end{aligned}$$

Table B.1: Summary of the boundary computation rules in Proposition B.2.

Rule	Assumptions
$\mathcal{Q}_{\alpha_2,\beta_2}(t) = \mathcal{Q}_{\alpha_1,\beta_1}(t)$	$\alpha_1 \leq \alpha_2, \beta_1 \leq \beta_2$
$\mathcal{Q}_{\alpha_1,\beta_1}(t) + \mathcal{Q}_{\alpha_2,\beta_2}(t) = \mathcal{Q}_{\min\{\alpha_1,\alpha_2\},\min\{\beta_1,\beta_2\}}(t)$	—
$\mathcal{Q}_{\alpha_1,\beta_1}(t) \cdot \mathcal{Q}_{\alpha_2,\beta_2}(t) = \mathcal{Q}_{\alpha_1+\alpha_2,\min\{\alpha_1+\beta_2,\alpha_2+\beta_1\}}(t)$	$\alpha_i \leq \beta_i$
$g_2(at^\alpha + bt^\beta + \mathcal{Q}_{\alpha,\beta}(t)) = g_2(0) + g_2'(0)at^\alpha + g_2'(0)bt^\beta + \mathcal{Q}_{\alpha,\beta}(t)$	$g_2 \in C^\infty$ (also at 0), $\alpha \in \mathbb{N}_0$, $\beta \geq \alpha$, $a = 0$ if $\alpha = 0$, $b = 0$ if $\beta = 0$
$(at^\alpha + \mathcal{Q}_{\alpha,\alpha}(t))^\delta = a^\delta t^{\alpha\delta} + \mathcal{Q}_{\alpha\delta,\alpha\delta}(t)$	$a, \alpha, \delta > 0$, $at^\alpha + \mathcal{Q}_{\alpha,\alpha}(t) > 0$
$\mathcal{Q}_{\delta,\delta}(at^\alpha + \mathcal{Q}_{\alpha,\alpha}(t)) = \mathcal{Q}_{\alpha\delta,\alpha\delta}(t)$	$a, \alpha, \delta > 0$, $at^\alpha + \mathcal{Q}_{\alpha,\alpha}(t) \in (0, 2)$

Induction step $n \to n+1$: We apply the induction hypothesis to $\tilde{h}_2 := h_2' \in C^\infty(J)$, which satisfies the derivative condition for $\tilde{M} := \max\{0, M-1\}$, and obtain

$$\left|(h_2' \circ g_1)^{(k)}(t)\right| = O(t^{\tilde{N}\tilde{\beta}-k}) \leq O(t^{(N-1)\tilde{\beta}-k})$$

for all $0 \leq k \leq n$. Hence,

$$\begin{aligned}
|(h_2 \circ g_1)^{(n+1)}(t)| &= \left|\frac{\mathrm{d}^n}{\mathrm{d}t^n}(h_2' \circ g_1)(t) \cdot g_1'(t)\right| \\
&= \left|\sum_{k=0}^{n} \binom{n}{k} (h_2' \circ g_1)^{(k)}(t) \cdot g_1^{(1+n-k)}(t)\right| \\
&\leq \sum_{k=0}^{n} O(t^{(M-1)\tilde{\beta}-k}) O(t^{\tilde{\beta}-1-n+k}) \\
&= O(t^{M\tilde{\beta}-(n+1)}) ,
\end{aligned}$$

which completes the induction step.

(f) We obtain

$$\begin{aligned}
g_1(t)^\delta &= (at^\alpha + \mathcal{P}_{\alpha,\alpha}(t) + \mathcal{R}_\gamma(t))^\delta \\
&= (at^\alpha \cdot (1 + t^{-\alpha}\mathcal{P}_{\alpha,\alpha}(t) + t^{-\alpha}\mathcal{R}_\gamma(t)))^\delta \\
&\overset{\text{(b), (c)}}{=} a^\delta t^{\alpha\delta}(1 + \mathcal{P}_{0,0}(t) + \mathcal{R}_{\gamma-\alpha}(t))^\delta \\
&\overset{\text{(e)}}{=} a^\delta t^{\alpha\delta}(1 + \mathcal{P}_{0,0}(t) + \mathcal{R}_{\gamma-\alpha}(t)) \\
&= a^\delta t^{\alpha\delta} + \mathcal{P}_{\alpha\delta,\alpha\delta}(t) + \mathcal{R}_{\gamma-\alpha+\alpha\delta}(t) .
\end{aligned}$$

□

We now obtain the following rules for $\mathcal{Q}_{\alpha,\beta}$, summarized in Table B.1:

Proposition B.2 (Rules for $\mathcal{Q}_{\alpha,\beta}$). *Let $J \subseteq \mathbb{R}$ be an interval. Let $f_1, f_2 : [0,2] \to \mathbb{R}$, $g_1 : [0,2] \to J$ and $g_2 : J \to \mathbb{R}$ with*

$$\begin{aligned}
f_1(t) &= \mathcal{Q}_{\alpha_1,\beta_1}(t), \\
f_2(t) &= \mathcal{Q}_{\alpha_2,\beta_2}(t)
\end{aligned}$$

for some $\alpha_i, \beta_i \in \mathbb{R}$.

*(a) **Inclusion**: If $\alpha_1 \leq \alpha_2$ and $\beta_1 \leq \beta_2$, then $\mathcal{Q}_{\alpha_2,\beta_2} \subseteq \mathcal{Q}_{\alpha_1,\beta_1}$.*

*(b) **Sum:** We have*

$$f_1(t) + f_2(t) = \mathcal{Q}_{\min\{\alpha_1,\alpha_2\},\min\{\beta_1,\beta_2\}}(t) .$$

*(c) **Product:** If $\alpha_i \leq \beta_i$, we have*

$$f_1(t) \cdot f_2(t) = \mathcal{Q}_{\alpha_1+\alpha_2,\min\{\alpha_1+\beta_2,\alpha_2+\beta_1\}}(t) .$$

(d) ***Composition with C^∞:*** *Suppose that $g_1(t) = at^\alpha + bt^\beta + \mathcal{Q}_{\alpha,\beta}(t)$ with $g_1(0) = 0$ for some $\alpha \in \mathbb{N}_0$ and $\beta \geq \alpha$. Additionally, suppose that $g_2 \in C^\infty(J)$ with $0 \in J$. Then,*

$$g_2(g_1(t)) = g_2(0) + g_2'(0)at^\alpha + g_2'(0)bt^\beta + \mathcal{Q}_{\alpha,\beta}(t) \ .$$

(e) ***Composition with power:*** *Suppose that $J \subseteq (0, \infty)$ with $g_1(t) = at^\alpha + \mathcal{Q}_{\alpha,\alpha}(t)$ for $a, \alpha > 0$. Then, for $\delta > 0$,*

$$g_1(t)^\delta = a^\delta t^{\alpha\delta} + \mathcal{Q}_{\alpha\delta,\alpha\delta}(t) \ .$$

(f) ***Composition with $\mathcal{Q}_{\delta,\delta}$:*** *Suppose that $J = (0, 2)$, $g_1(t) = at^\alpha + \mathcal{Q}_{\alpha,\alpha}(t)$ with $a, \alpha > 0$, and $g_2(t) = \mathcal{Q}_{\delta,\delta}(t)$ with $\delta \geq 0$. Then,*

$$g_2(g_1(t)) = \mathcal{Q}_{\alpha\delta,\alpha\delta}(t) \ .$$

Proof.

(a) Follows from the definition.
(b) For given $\gamma \in \mathbb{R}$, write $f_i = p_i + r_i$ with $p_i \in \mathcal{P}_{\alpha_i,\beta_i}$ and $r_i \in \mathcal{R}_\gamma$. Then, obviously $p_1 + p_2 \in \mathcal{P}_{\min\{\alpha_1,\alpha_2\},\min\{\beta_1,\beta_2\}}$ and $r_1 + r_2 \in \mathcal{R}_\gamma$, which shows the claim.
(c) Using a decomposition as in the proof of (b), we write $f_1 f_2 = p_1 p_2 + (p_1 r_2 + r_1 p_2 + r_1 r_2)$. Here, it follows from Lemma B.1 (c) that $p_1 p_2 \in \mathcal{P}_{\alpha_1+\alpha_2,\min\{\alpha_1+\beta_2,\alpha_2+\beta_1\}}$. Moreover, using Lemma B.1 (a) we have $p_i \in \mathcal{R}_{\alpha_i}$. Using Lemma B.1 (b), we obtain

$$p_1 r_2 + r_1 p_2 + r_1 r_2 \in \mathcal{R}_{\gamma+\min\{\alpha_1,\alpha_2\}} \ ,$$

where $\gamma \in \mathbb{R}$ was arbitrary. This shows the claim.
(d) Follows from Lemma B.1 (e).
(e) Follows from Lemma B.1 (f).
(f) Consider $\gamma \in \mathbb{R}$ to be chosen later. We can then write $g_2(t) = p_2(t) + r_2(t)$ with $p_2 \in \mathcal{P}_{\delta,\delta}$ and $r_2 \in \mathcal{R}_\gamma$. Since p_2 is a sum of powers, (a) and (e) show that $p_2(g_1(t)) = \mathcal{Q}_{\alpha\delta,\alpha\delta}(t)$. For r_2, we use Lemma B.1 (a), (d) to conclude that for some $\varepsilon > 0$,

$$r_2(g_1(t)) = r_2(at^\alpha + \mathcal{R}_{\alpha+\varepsilon}(t)) = \mathcal{R}_{\alpha\gamma}(t) \ .$$

Since γ was arbitrary, it can be chosen such that $\alpha\gamma$ can be arbitrarily large (since $\alpha > 0$). This completes the proof. □

C DUAL ACTIVATIONS

C.1 General properties

Definition C.1 (Pseudo-derivative)**.** Let $f, g : \mathbb{R} \to \mathbb{R}$. We call g a *pseudo-derivative* of f if g is Lebesgue integrable on compact intervals and

$$f(x) = f(0) + \int_0^x g(t) \,\mathrm{d}t$$

for all $x \in \mathbb{R}$. ◀

If $f : \mathbb{R} \to \mathbb{R}$ is continuously differentiable, then f' is a pseudo-derivative of f. As a non-differentiable example, the Heaviside theta function $\mathbb{1}_{(0,\infty)}(x)$ is a pseudo-derivative of the ReLU function. From basic Lebesgue integration theory, it follows that pseudo-derivatives are unique up to null sets. Note that any *continuous* activation function $\varphi : \mathbb{R} \to \mathbb{R}$ fulfilling Assumption 4 is pseudo-differentiable.

The following lemma generalizes the differentiation part of Lemma 11 by Daniely et al. (2016) to pseudo-differentiable functions.

Lemma C.2 (Properties of pseudo-derivatives)**.** *Let $g \in \mathcal{L}_2(\mathcal{N}(0, 1))$ be a pseudo-derivative of $f : \mathbb{R} \to \mathbb{R}$. Then,*

(a) $f \in \mathcal{L}_2(\mathcal{N}(0, 1))$,

(b) for $n \geq 1$, $a_n(f) = n^{-1/2} a_{n-1}(g)$,
(c) the Hermite expansion $f = \sum_{n=0}^{\infty} a_n(f) h_n$ converges pointwise,
(d) $\hat{f}$ is differentiable on $[-1, 1]$ with $\hat{f}' = \hat{g}$ (cf. Definition A.4).

Proof. Let ϕ be the p.d.f. of the normal distribution $\mathcal{N}(0, 1)$. Without loss of generality, let $x \geq 0$. Then,

$$\begin{aligned}
f(x) &= f(0) + \int_0^x g(t)\,\mathrm{d}t = f(0) + \left\langle \frac{\mathbb{1}_{[0,x]}}{\phi}, g \right\rangle_{L_2(\mathcal{N}(0,1))} \\
&= f(0) + \sum_{n=0}^{\infty} a_n(g) \left\langle \frac{\mathbb{1}_{[0,x]}}{\phi}, h_n \right\rangle_{L_2(\mathcal{N}(0,1))} \\
&= f(0) + \sum_{n=0}^{\infty} a_n(g) \int_0^x h_n(t)\,\mathrm{d}t \\
&= f(0) + \sum_{n=0}^{\infty} a_n(g) \int_0^x (n+1)^{-1/2} h'_{n+1}(t)\,\mathrm{d}t \\
&= f(0) + \sum_{n=0}^{\infty} (n+1)^{-1/2} a_n(g) [h_{n+1}(x) - h_{n+1}(0)] \\
&= f(0) + \sum_{n=1}^{\infty} n^{-1/2} a_{n-1}(g) [h_n(x) - h_n(0)] \\
&= \left[f(0) - \sum_{n=1}^{\infty} n^{-1/2} a_{n-1}(g) h_n(0) \right] h_0(x) + \sum_{n=1}^{\infty} n^{-1/2} a_{n-1}(g) h_n(x) \;. \qquad \text{(C.1)}
\end{aligned}$$

In the last step, we used $h_0(x) = 1$ for all x. Moreover, splitting the series in the last step is valid since the extracted series is absolutely summable:

$$\begin{aligned}
\sum_{n=1}^{\infty} |n^{-1/2} a_{n-1}(g) h_n(0)| &= \sum_{n=1}^{\infty} n^{-1/2} |a_{n-1}(g)| \operatorname{even}(n) \frac{(n-1)!!}{\sqrt{n!}} \\
&\overset{\text{Lemma C.8}}{\leq} \sum_{n=1}^{\infty} |a_{n-1}(g)| O(n^{-3/4}) \\
&= \left\langle (|a_{n-1}(g)|)_{n \geq 1}, O(n^{-3/4}) \right\rangle_{\ell_2(\mathbb{N}_{\geq 1})} \\
&\overset{\text{Cauchy-Schwarz}}{<} \infty \;.
\end{aligned}$$

Here, we have used that $(a_n(g))_{n \geq 1} \in \ell_2(\mathbb{N}_{\geq 1})$ since $g \in \mathcal{L}_2(\mathcal{N}(0, 1))$. An analogous argument shows that (C.1) also equals $f(x)$ for $x < 0$. Since $|n^{-1/2} a_{n-1}(g)| \leq |a_{n-1}(g)|$, we know that the series representation (C.1) not only converges pointwise but also in $L_2(\mathcal{N}(0, 1))$, and these limits must be identical. This shows (a), (b) and (c). In order to show (d), we compute

$$\hat{f}(t) = \sum_{n=0}^{\infty} a_n(f)^2 t^n \overset{\text{(b)}}{=} a_0(f)^2 + \sum_{n=1}^{\infty} n^{-1} a_{n-1}(g)^2 t^n$$

for $t \in [-1, 1]$, which implies

$$\hat{f}'(t) = \sum_{n=1}^{\infty} n^{-1} a_{n-1}(g)^2 n t^{n-1} = \sum_{n=0}^{\infty} a_n(g)^2 t^n = \hat{g}(t) \;.$$

This also holds at the boundary $t \in \{-1, 1\}$: Thanks to Abel's theorem, $\hat{f}$ and $\hat{g}$ are continuous on $[-1, 1]$. Without loss of generality, let $t = 1$. The mean value theorem of integration yields $\hat{f}(1) - \hat{f}(1-h) = \int_{1-h}^{1} \hat{f}'(t)\,\mathrm{d}t = hf'(\xi_h)$ for some $\xi_h \in (1-h, 1)$, thus

$$\hat{f}'(1) = \lim_{h \searrow 0} \hat{f}'(\xi_h) = \lim_{u \nearrow 1} \hat{f}'(u) = \lim_{u \nearrow 1} \hat{g}(u) = \hat{g}(1) \;. \qquad \square$$

While the dualization $\varphi \mapsto \hat{\varphi}$ from Definition A.4 is a quadratic and not a linear mapping, it still interchanges nicely with the even-odd decomposition (since even and odd functions are orthogonal w.r.t. the corresponding "inner product"):

Lemma C.3 (General properties of dual activations)**.** *Let $\varphi \in \mathcal{L}_2(\mathcal{N}(0,1))$. Then,*

(a) $\varphi_{\text{even}}, \varphi_{\text{odd}} \in \mathcal{L}_2(\mathcal{N}(0,1))$ *and*

$$(\hat{\varphi})_{\text{even}} = \widehat{\varphi_{\text{even}}}, \qquad (\hat{\varphi})_{\text{odd}} = \widehat{\varphi_{\text{odd}}} \ .$$

(b) $\widehat{\varphi}$, $\widehat{\varphi_{\text{even}}}$ *and* $\widehat{\varphi_{\text{odd}}}$ *are nonnegative and increasing on* $[0,1]$*. Moreover, if φ is not almost surely constant, then $\widehat{\varphi}$ is strictly increasing on* $[0,1]$*.*

(c) We have

$$|\widehat{\varphi}(t)| \le \widehat{\varphi}(|t|) \le \widehat{\varphi}(1), \tag{C.2}$$

for all $t \in [-1,1]$. If φ is not almost surely constant, then $|\widehat{\varphi}(t)| < \widehat{\varphi}(1)$ for $t \in (-1,1)$. Moreover, $|\widehat{\varphi}(-1)| = \widehat{\varphi}(1)$ if and only if φ is almost surely even or almost surely odd.

(d) We have $\widehat{\varphi}|_{(-1,1)} \in C^\infty((-1,1))$*.*

Proof.

(a) It is easy to see that $\varphi_{\text{even}}, \varphi_{\text{odd}} \in \mathcal{L}_2(\mathcal{N}(0,1))$. As mentioned in Appendix A, the Hermite polynomials are even for even n and odd for odd n. Hence,

$$\varphi_{\text{even}} = \sum_{n=0}^{\infty} \operatorname{even}(n) a_n(\varphi) h_n$$

$$\varphi_{\text{odd}} = \sum_{n=0}^{\infty} \operatorname{odd}(n) a_n(\varphi) h_n \ .$$

We also know that the dual activation satisfies $\hat{\varphi}(t) = \sum_{n=0}^{\infty} a_n(\varphi)^2 t^n$. Since the monomials t^n are also even for even n and odd for odd n, we obtain

$$(\hat{\varphi})_{\text{even}}(t) = \sum_{n=0}^{\infty} \operatorname{even}(n) a_n(\varphi)^2 t^n = \widehat{\varphi_{\text{even}}}(t)$$

$$(\hat{\varphi})_{\text{odd}}(t) = \sum_{n=0}^{\infty} \operatorname{odd}(n) a_n(\varphi)^2 t^n = \widehat{\varphi_{\text{odd}}}(t) \ . \tag{C.3}$$

Alternatively, this statement can also be proven directly using the definition of the dual activation.

(b) It follows from the Hermite expansions in Item (a) that $\widehat{\varphi_{\text{even}}}$ and $\widehat{\varphi_{\text{odd}}}$ are nonnegative and increasing on $[0,1]$, and therefore this also holds for $\widehat{\varphi}$. If φ is not almost surely constant, we have $a_n(\varphi) \neq 0$ for some $n \ge 1$. Therefore, the Hermite expansion $\widehat{\varphi}(t) = \sum_{n=0}^{\infty} a_n(\varphi)^2 t^n$ implies that $\widehat{\varphi}$ is strictly increasing on $[0,1]$.

(c) For $t \in [0,1]$, the statement follows from (b). For $t \in [-1,0)$, (b) implies

$$\begin{aligned} |\widehat{\varphi}(t)| &= |\widehat{\varphi_{\text{even}}}(-t) - \widehat{\varphi_{\text{odd}}}(-t)| \\ &\le |\widehat{\varphi_{\text{even}}}(-t)| + |\widehat{\varphi_{\text{odd}}}(-t)| = \widehat{\varphi_{\text{even}}}(-t) + \widehat{\varphi_{\text{odd}}}(-t) = \widehat{\varphi}(-t) = \widehat{\varphi}(|t|) \le \widehat{\varphi}(1) \ . \end{aligned}$$

If φ is not almost surely constant, then $\widehat{\varphi}$ is strictly increasing on $[0,1]$ by (b) and the previous inequality implies $|\widehat{\varphi}(t)| < \widehat{\varphi}(1)$ for $t \in (-1,1)$. Moreover, the inequality $|\widehat{\varphi_{\text{even}}}(-t) - \widehat{\varphi_{\text{odd}}}(-t)| \le |\widehat{\varphi_{\text{even}}}(-t)| + |\widehat{\varphi_{\text{odd}}}(-t)|$ is sharp iff $\widehat{\varphi_{\text{even}}}(-1) = 0$ or $\widehat{\varphi_{\text{odd}}}(-1) = 0$, which is the case iff φ is almost surely even or almost surely odd.

(d) By Eq. (A.2) the dual activation $\widehat{\varphi}$ is a convergent power series on $[-1,1]$. □

C.2 Dominating terms

Definition C.4 (Double factorial)**.** Following Daniely et al. (2016), we define the double factorial for $n \in \mathbb{Z}$ as

$$n!! := \begin{cases} 1 & , n \le 0 \\ n \cdot (n-2) \cdots 4 \cdot 2 & , n > 0 \text{ even} \\ n \cdot (n-2) \cdots 3 \cdot 1 & , n > 0 \text{ odd.} \end{cases}$$ ◀

Proposition C.5. *For all $k \in \mathbb{N}_0$, the reference activation s_k from Definition A.6 satisfies $s_k \in L_2(\mathcal{N}(0,1))$ and s_k is a pseudo-derivative of s_{k+1}. Moreover, for all $n \in \mathbb{N}_0$, the Hermite coefficients satisfy*

$$a_n(s_k) = \operatorname{odd}(n-k) \frac{(-1)^{\frac{\max(1,n-k)-1}{2}} (n-k-2)!!}{(k-n)!! \sqrt{2\pi n!}}$$

Proof. The first two statements are straightforward to show, hence we only show the formula for the Hermite coefficients. We first note that for $X \sim \mathcal{N}(0,1)$, we have for odd k the central absolute moments following (see e.g. Winkelbauer, 2012):

$$\begin{aligned}\mathbb{E}[|X|^k] &= 2^{k/2}\Gamma\left(\frac{k+1}{2}\right)\pi^{-1/2} = 2^{k/2}((k-1)/2)!\pi^{-1/2} = 2^{1/2}(k-1)!!\pi^{-1/2} \\ &= \sqrt{\frac{2}{\pi}}(k-1)!! \;.\end{aligned}$$

This allows us to show the statement for $n = 0$:

$$\begin{aligned}a_0(s_k) = \langle h_0, s_k \rangle_{\mathcal{L}_2(\mathcal{N}(0,1))} &= \int_{\mathbb{R}} s_k(x) \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \,\mathrm{d}x = \operatorname{odd}(k) \frac{1}{2k!} \int_{\mathbb{R}} |x^k| \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \,\mathrm{d}x \\ &= \operatorname{odd}(k) \frac{(k-1)!!}{k!\sqrt{2\pi}} = \operatorname{odd}(k) \frac{1}{k!!\sqrt{2\pi}} = \operatorname{odd}(0-k) \frac{(-1)^{\frac{\max(1,0-k)-1}{2}} (0-k-2)!!}{(k-0)!!\sqrt{2\pi 0!}} \;,\end{aligned}$$

Finally, we show the formula for all n, k via induction on k. (We do not use induction on n.) For the base case $k = 0$, we note that $h_0 \equiv 1$ and $s_0 = 2^{-1/2}\sqrt{2}\mathbb{1}_{(0,\infty)} - \frac{1}{2}h_0$ almost everywhere. Daniely et al. (2016) show in Section 8 that

$$a_n(\sqrt{2}\mathbb{1}_{(0,\infty)}) = \begin{cases} 2^{-1/2} & , n = 0 \\ \frac{(-1)^{\frac{n-1}{2}}(n-2)!!}{\sqrt{\pi n!}} & , n \text{ odd} \\ 0 & , 2 \leq n \text{ even.} \end{cases}$$

Therefore, we obtain for $n \geq 1$:

$$\begin{aligned}a_n(s_0) = 2^{-1/2}a_n(\sqrt{2}\mathbb{1}_{(0,\infty)}) &= \operatorname{odd}(n) \frac{(-1)^{\frac{n-1}{2}}(n-2)!!}{\sqrt{2\pi n!}} \\ &= \operatorname{odd}(n-0) \frac{(-1)^{\frac{\max(1,n-0)-1}{2}}(n-0-2)!!}{(0-n)!!\sqrt{2\pi n!}} \;,\end{aligned}$$

which completes the base case $k = 0$ since the case $n = 0$ has already been treated above.

For the induction $k \to k+1$, we note that s_k is a pseudo-derivative of s_{k+1} and use Lemma C.2 to obtain for all $n \geq 0$:

$$\begin{aligned}a_{n+1}(s_{k+1}) = (n+1)^{-1/2}a_n(s_k) &= \operatorname{odd}(n-k) \frac{(-1)^{\frac{\max(1,n-k)-1}{2}}(n-k-2)!!}{(k-n)!!\sqrt{2\pi(n+1)!}} \\ &= \operatorname{odd}((n+1)-(k+1)) \frac{(-1)^{\frac{\max(1,(n+1)-(k+1))-1}{2}}((n+1)-(k+1)-2)!!}{((k+1)-(n+1))!!\sqrt{2\pi(n+1)!}} \;,\end{aligned}$$

which completes the induction. □

Lemma C.6. *We have $\widehat{s_0}(t) = \frac{1}{4} - \frac{1}{2\pi}\arccos(t)$.*

Proof. By Section 8 in Daniely et al. (2016), the function $f := \sqrt{2}\mathbb{1}_{(0,\infty)}$ has the dual activation $\hat{f}(t) = 1 - \frac{1}{\pi}\arccos(t)$. Hence, $g := \mathbb{1}_{(0,\infty)}$ has the dual activation

$$\hat{g}(t) = \frac{1}{2} - \frac{1}{2\pi}\arccos(t) \;.$$

Now, since $s_0 = g - 1/2$ almost everywhere, we have $a_n(s_0) = a_n(g)$ for all $n \neq 0$, and hence

$$\widehat{s_0}(t) = \sum_{n=0}^{\infty} a_n(s_0)^2 t^n = C + \sum_{n=0}^{\infty} a_n(g)^2 t^n = C + \frac{1}{2} - \frac{1}{2\pi}\arccos(t)$$

for some constant $C \in \mathbb{R}$. Since s_0 is odd, $\widehat{s_0}$ must be odd by Lemma C.3, which yields $C = -1/4$. □

Lemma C.7. *For all $k \in \mathbb{N}_0$, there exists $b_k > 0$ such that for $\tau \in \{\pm 1\}$,*

$$\widehat{s_k}(\tau(1-t)) = (-\tau)^{k+1} b_k t^{k+1/2} + \mathcal{Q}_{-1,k+1/2}(t) \ .$$

Proof. **Step 1: Analysis of s_0.** First, consider the case $\tau = 1$. We have

$$\frac{\mathrm{d}}{\mathrm{d}t}\arccos(1-t) = \frac{1}{\sqrt{1-(1-t)^2}} = t^{-1/2}(2-t)^{-1/2} \ .$$

It is easily seen using induction that

$$\frac{\mathrm{d}^n}{\mathrm{d}t^n}(2-t)^{-1/2} = \frac{(2n-1)!!}{2^n}(2-t)^{-\frac{2n+1}{2}} \ .$$

Since $t \mapsto (2-t)^{-1/2}$ is analytic in a neighborhood of $t = 0$, it is equal to its Taylor expansion:

$$(2-t)^{-1/2} = \sum_{n=0}^{\infty} \frac{(2n-1)!!}{2^{2n+\frac{1}{2}} n!} t^n \ .$$

Since all coefficients are positive, we can apply the monotone convergence theorem to obtain

$$\begin{aligned}
\arccos(1-t) &= \arccos(1-0) + \int_0^t \frac{\mathrm{d}}{\mathrm{d}u}\arccos(1-u)\,\mathrm{d}u \\
&= \int_0^t \sum_{n=0}^{\infty} \frac{(2n-1)!!}{2^{2n+\frac{1}{2}} n!} u^{n-1/2}\,\mathrm{d}u \\
&= \sum_{n=0}^{\infty} \int_0^t \frac{(2n-1)!!}{2^{2n+\frac{1}{2}} n!} u^{n-1/2}\,\mathrm{d}u \\
&= \sum_{n=0}^{\infty} \frac{(2n-1)!!}{2^{2n+\frac{1}{2}} n!} \left[\frac{1}{n+1/2} u^{n+1/2}\right]_0^t \\
&= \sum_{n=0}^{\infty} \frac{(2n-1)!!}{2^{2n+\frac{1}{2}} n!(n+1/2)} t^{n+1/2} \ .
\end{aligned}$$

For arbitrary $1 \leq N \in \mathbb{N}_0$, this yields using Lemma C.6, the identity

$$\begin{aligned}
\widehat{s_0}(1-t) &= \frac{1}{4} - \sum_{n=0}^{\infty} \frac{(2n-1)!!}{2^{2n+\frac{3}{2}} n!(n+1/2)\pi} t^{n+1/2} \\
&= \frac{1}{4} - \frac{1}{\pi\sqrt{2}} t^{1/2} + \underbrace{\sum_{n=1}^{N-1} \frac{(2n-1)!!}{2^{2n+\frac{3}{2}} n!(n+1/2)\pi} t^{n+1/2}}_{=\mathcal{P}_{0,1/2}(t)} \\
&\quad + \underbrace{t^{N+1/2}}_{=\mathcal{R}_{N+1/2}(t)} \underbrace{\sum_{n=N}^{\infty} \frac{(2n-1)!!}{2^{2n+\frac{3}{2}} n!(n+1/2)\pi} t^{n-N}}_{=\mathcal{R}_0(t)} .
\end{aligned}$$

Step 2: Induction on k. We now show the lemma via induction on $k \in \mathbb{N}_0$. For arbitrary $\gamma \in \mathbb{R}$, by choosing N with $N + 1/2 \geq \gamma$, we obtain using Lemma B.1:

$$\widehat{s_0}(1-t) = (-1)^1 b_0 t^{0+1/2} + \mathcal{P}_{-1,1/2}(t) + \mathcal{R}_\gamma(t) \ .$$

For the induction step $k \to k+1$, we note that since s_k is a pseudo-derivative of s_{k+1} (see Definition C.1), we have $\widehat{s_{k+1}}' = \widehat{s'_{k+1}} = \widehat{s_k}$ by Lemma C.2, which yields

$$\begin{aligned}\widehat{s_{k+1}}(1-t) &= \widehat{s_{k+1}}(1) - \int_0^t \widehat{s_{k+1}}'(1-u)\,\mathrm{d}u \\ &= \widehat{s_{k+1}}(1) - \int_0^t \widehat{s_k}(1-u)\,\mathrm{d}u \\ &= \widehat{s_{k+1}}(1) - \int_0^t (c_k + (-1)^{k+1} b_k u^{k+1/2} + \mathcal{P}_{0,k+1/2}(u) + \mathcal{R}_\gamma(u))\,\mathrm{d}u \\ &= c_{k+1} + (-1)^{k+2} b_{k+1} t^{(k+1)+1/2} + \mathcal{P}_{0,(k+1)+1/2}(t) + \mathcal{R}_{\gamma+1}(t)\end{aligned}$$

for suitable constants $c_{k+1}, b_{k+1} > 0$. This shows the claim for $\tau = 1$. For $\tau = -1$, we note that s_k is odd for even k and even for odd k. By Lemma C.3, the same holds for $\widehat{s_k}$. This shows

$$\begin{aligned}\widehat{s_k}(-1+t) &= (-1)^{k+1}\widehat{s_k}(1-t) \\ &= b_k t^{k+1/2} + \mathcal{P}_{-1,k+1/2}(t) + \mathcal{R}_\gamma(t)\ .\end{aligned}$$

□

We will also need the asymptotic decay of $a_n(s_k)$, which will be studied in the following two lemmas.

Lemma C.8. *For $p \in \mathbb{N}_0$ and odd $m \in \mathbb{N}_0$, we have*

$$\frac{m!!}{\sqrt{(m+p)!}} = \Theta_{\forall m}(m^{1/4-p/2})\ .$$

Proof. For even $n = 2k \geq 2$, we have $n!! = n(n-2)\cdots 2 = (2\cdot k)(2\cdot(k-1))\cdots(2\cdot 1) = 2^k k! = 2^{n/2}(n/2)!$. Using Stirling's formula, we obtain

$$\frac{(n-1)!!}{\sqrt{n!}} = \frac{n!}{n!!\sqrt{n!}} = \frac{\sqrt{n!}}{2^{n/2}(n/2)!} \sim \frac{(2\pi n)^{1/4} n^{n/2} e^{-n/2}}{2^{n/2}(\pi n)^{1/2}(n/2)^{n/2}e^{-n/2}} = \Theta_{\forall n}(n^{-1/4})\ .$$

By setting $n \coloneqq m+1$, we obtain

$$\begin{aligned}\frac{m!!}{\sqrt{(m+p)!}} &= \Theta_{\forall m}(m^{1/2-p/2})\frac{m!!}{\sqrt{(m+1)!}} = \Theta_{\forall m}(m^{1/2-p/2})\Theta_{\forall m}(m^{-1/4}) \\ &= \Theta_{\forall m}(m^{1/4-p/2})\ .\end{aligned}$$

□

Lemma C.9. *We have $|a_n(s_k)| = \Theta_{\forall n}(\mathrm{odd}(n-k)(n+1)^{-3/4-k/2})$.*

Proof. Since $(k-n)!! = 1$ for $n \geq k$ by our definition of the double factorial in Definition C.4, we obtain

$$\begin{aligned}|a_n(s_k)| &\overset{\text{Proposition C.5}}{=} \Theta_{\forall n}\left(\mathrm{odd}(n-k)\frac{(n-k-2)!!}{\sqrt{2\pi n!}}\right) \\ &\overset{\text{Lemma C.8}}{=} \Theta_{\forall n}(\mathrm{odd}(n-k)(n+1)^{-3/4-k/2})\ .\end{aligned}$$

□

C.3 Mix terms

Now, we want to investigate some of the mix-terms arising in the decomposition in Eq. (A.3). For convenience, we will give them a new name:

Definition C.10. For $i, j \in \mathbb{N}_0$, define $f_{i,j} : [-1,1] \to \mathbb{R}$ by

$$f_{i,j}(t) \coloneqq \sum_{n=0}^{\infty} a_n(s_i)a_n(s_j)t^n\ .$$

Note that $f_{i,j} \equiv 0$ for odd $i-j$ and $f_{i,j} = f_{j,i}$. Moreover, $f_{i,i} = \widehat{s_i}$. ◀

Lemma C.11 (Recursive characterization of mix-terms)**.** *Let $k, l \in \mathbb{N}_0$. Then, there exists a polynomial $p_{k,k+2l}$ such that for $t \in [-1, 1]$,*

$$\begin{aligned} f_{k,k+2l}(t) &= (k+2) f_{k+2,(k+2)+2(l-1)}(t) - t f_{k+1,(k+1)+2(l-1)}(t) + p_{k,k+2l}(t) \\ f_{k,k+2l+1}(t) &= 0 \ . \end{aligned}$$

Proof. Since $a_n(s_k) = 0$ for even $n - k$ and $a_n(s_{k+2l+1}) = 0$ for even $n - (k + 2l + 1)$, we have

$$f_{k,k+2l+1}(t) = \sum_{n=0}^{\infty} a_n(s_k) a_n(s_{k+2l+1}) t^n = 0 \ .$$

For the other formula, we use Proposition C.5 and obtain

$$\begin{aligned} a_n(s_k) a_n(s_{k+2l}) = &\operatorname{odd}(n-k) \operatorname{odd}(n-(k+2l)) (-1)^{\frac{\max(1,n-k)-1}{2} + \frac{\max(1,n-k-2l)-1}{2}} \\ &\cdot \frac{(n-k-2)!!(n-k-2l-2)!!}{(k-n)!!(k+2l-n)!!2\pi n!} \ . \end{aligned}$$

Here, $\operatorname{odd}(n-k) \operatorname{odd}(n-(k+2l)) = \operatorname{odd}(n-k)$. Moreover, for odd $n - k$ and $n \geq k + 2l$, we have

$$(-1)^{\frac{\max(1,n-k)-1}{2} + \frac{\max(1,n-k-2l)-1}{2}} = (-1)^{n-k-l-1} = (-1)^{-l} = (-1)^l \ .$$

In the following, we write

$$f \simeq_{\text{pol}} g$$

if $f - g$ is a polynomial.

The terms $(k - n)!!$ and $(k + 2l - n)!!$ are equal to one for $n \geq k + 2l$. Therefore, we obtain for all but finitely many n

$$a_n(s_k) a_n(s_{k+2l}) = \operatorname{odd}(n-k)(-1)^l \frac{(n-k-2)!!(n-k-2l-2)!!}{2\pi n!} \ .$$

By writing $(n - k - 2)!! = -(k + 2)(n - k - 4)!! + n(n - k - 4)!!$, we obtain:

$$\begin{aligned} &\sum_{n=0}^{\infty} a_n(s_k) a_n(s_{k+2l}) t^n \\ \simeq_{\text{pol}}\ & -(k+2) \sum_{n=0}^{\infty} \operatorname{odd}(n-k)(-1)^l \frac{(n-k-4)!!(n-k-2l-2)!!}{2\pi n!} t^n \\ &+ \sum_{n=0}^{\infty} n \operatorname{odd}(n-k)(-1)^l \frac{(n-k-4)!!(n-k-2l-2)!!}{2\pi n!} t^n \\ \simeq_{\text{pol}}\ & (k+2) \sum_{n=0}^{\infty} \operatorname{odd}(n-(k+2))(-1)^{l-1} \frac{(n-(k+2)-2)!!(n-(k+2)-2(l-1)-2)!!}{2\pi n!} t^n \\ &- t \cdot \sum_{n=1}^{\infty} \operatorname{odd}((n-1)-(k-1))(-1)^{l-1} \frac{((n-1)-k-3)!!((n-1)-k-2l-1)!!}{2\pi (n-1)!} t^{n-1} \\ =\ & (k+2) \sum_{n=0}^{\infty} \operatorname{odd}(n-(k+2))(-1)^{l-1} \frac{(n-(k+2)-2)!!(n-(k+2)-2(l-1)-2)!!}{2\pi n!} t^n \\ &- t \cdot \sum_{n=0}^{\infty} \operatorname{odd}(n-(k+1))(-1)^{l-1} \frac{(n-(k+1)-2)!!(n-(k+1)-2(l-1)-2)!!}{2\pi n!} t^n \\ \simeq_{\text{pol}}\ & (k+2) \sum_{n=0}^{\infty} a_n(s_{k+2}) a_n(s_{(k+2)+2(l-1)}) t^n - t \cdot \sum_{n=0}^{\infty} a_n(s_{k+1}) a_n(s_{(k+1)+2(l-1)}) t^n \ , \end{aligned}$$

which completes the proof. □

Proposition C.12 (Boundary behavior of mix terms)**.** *Let $i, j \in \mathbb{N}_0$. Then, for any $\tau \in \{\pm 1\}$ and $t \in (0, 2)$,*

$$f_{i,j}(\tau(1-t)) = \mathcal{Q}_{-1,(i+j)/2}(t) \ .$$

Proof. Recall Lemma C.11. Since $f_{i,j} = f_{j,i}$, we can assume $i \leq j$ without loss of generality. Moreover, for odd $j - i$, the statement is trivial since $f_{i,j} \equiv 0$. For the remaining cases where $j - i \geq 0$ is even, it suffices to prove the following statement by induction on l:

For all $l \in \mathbb{N}_0$: For all $k \in \mathbb{N}_0$:

$$f_{k,k+2l} = \mathcal{Q}_{-1,k+l}(t) \ .$$

For $l = 0$, this follows from Lemma C.7 since $f_{k,k} = \widehat{s_k}$. For the induction step, we use Lemma C.11 to obtain for all $k \in \mathbb{N}_0$:

$$\begin{aligned} & f_{k,k+2l}(\tau(1-t)) \\ &= \mathcal{Q}_{-1,\infty}(t) + (k+2) f_{k+2,k+2+2(l-1)}(\tau(1-t)) - \tau(1-t) f_{k+1,k+1+2(l-1)}(\tau(1-t)) \\ &= \mathcal{Q}_{-1,\infty}(t) + \mathcal{Q}_{-1,k+l+1}(t) - \tau(1-t)(\mathcal{Q}_{-1,k+l}(t)) \\ &= \mathcal{Q}_{-1,k+l}(t) \ . \end{aligned}$$

Here, we used that for a polynomial p, the map $t \mapsto p(\tau(1-t))$ is also a polynomial and thus in $\mathcal{Q}_{-1,\infty}(t)$. ☐

C.4 Results for smooth terms

In the following, we analyze the decay of Hermite coefficients for smooth functions, which can then be used to establish the smoothness of certain components of dual activations. For smooth f, we could show the smoothness of $\widehat{f}$ by using $\widehat{f}' = \widehat{f'}$, but this approach does not work directly for mix-terms, and hence the intermediate step via coefficient decay is helpful.

Lemma C.13. *If $f : \mathbb{R} \to \mathbb{R}$ has a m-fold pseudo-derivative $f^{(m)}$ with $f^{(m)} \in \mathcal{L}_2(\mathcal{N}(0,1))$, then $|a_n(f)| < o_m((n+1)^{-m/2})$.*

Proof. By Lemma C.2, we obtain for $n \geq m$:

$$a_n(f) = [n(n-1) \cdot \ldots \cdot (n-m+1)]^{-1/2} \, a_{n-m}(f^{(m)}) \ .$$

Since $a_{n-m}(f^{(m)}) < o_{\forall n}(1)$, we obtain $|a_n(f)| < o_{\forall n}((n+1)^{-m/2})$. ☐

Lemma C.14 (Differentiability of power series)**.** *Let $f : [-1, 1] \to \mathbb{R}, x \mapsto \sum_{n=0}^{\infty} b_n x^n$ with $|b_n| = O_{\forall n}((n+1)^{-(k+1+\varepsilon)})$ for some $k \in \mathbb{N}_0$ and $\varepsilon > 0$. Then, $f \in C^k([-1,1])$.*

Proof. We prove by induction on k that $f \in C^k([-1, 1])$ and

$$f^{(k)}(t) = \sum_{n=k}^{\infty} n \cdot \ldots \cdot (n-k+1) b_n t^{n-k}$$

for $t \in [-1, 1]$. For $k = 0$, we have $\sum_{n=0}^{\infty} |b_n| < \infty$ and the result follows from Abel's theorem on power series. Now, let the statement hold for $k - 1 \geq 0$. We know from the case $k = 0$ that

$$g : [-1, 1] \to \mathbb{R}, t \mapsto \sum_{n=1}^{\infty} n b_n t^{n-1}$$

as well as f are continuous. By elementary analysis, $f' = g$ on $(-1, 1)$. Moreover,

$$f(1) - f(1-h) = \lim_{h' \searrow 0} f(1-h') - f(1-h) = \lim_{h' \searrow 0} \int_{1-h}^{1-h'} g(x) \, \mathrm{d}x = \int_{1-h}^{1} g(x) \, \mathrm{d}x = h g(\xi_h)$$

for a suitable $\xi_h \in [1-h, 1]$ by the mean value theorem of integration. Since g is continuous, it follows that f is differentiable in 1 with $f'(1) = g(1)$. An analogous calculation can be applied for $t = -1$. By applying the induction hypothesis to $g = f'$, the proof is completed. ☐

C.5 General activation functions

Now, we want to obtain the asymptotic boundary behavior in the sense of Appendix B for general activation functions φ as in Assumption 4. To this end, we use a decomposition into reference functions and smooth remainders:

Lemma C.15. *Let φ be an activation function as in Assumption 4 and let $m \in \mathbb{N}_0$. Then, using $\Delta_k(\varphi)$ as defined in Definition A.6,*

$$\varphi_m \coloneqq \varphi - \sum_{k=0}^{m-1} \Delta_k(\varphi) s_k$$

is m times pseudo-differentiable and the m-fold pseudo-derivative $\varphi_m^{(m)}$ is in $L_2(\mathcal{N}(0,1))$.

Proof. Note that Δ_m is a linear operator with $\Delta_m(s_k) = \delta_{mk}$. Hence, for $l \leq m-1$, we have

$$\Delta_l(\varphi_m) = \Delta_l(\varphi) - \sum_{k=0}^{m-1} \Delta_k(\varphi)\Delta_l(s_k) = \Delta_l(\varphi) - \Delta_l(\varphi) = 0 \ .$$

We now prove by induction on l with $0 \leq l \leq m$ that $g_l : \mathbb{R} \to \mathbb{R}$ defined by

$$g_0 \coloneqq \varphi_m, g_l(x) \coloneqq \begin{cases} \varphi_m^{(l)}(x) & , x \neq 0 \\ \varphi_m^{(l)}(0-) & , x = 0 \end{cases} \quad (l \geq 1)$$

is a l-fold pseudo-derivative of φ_m. For $l = 0$, we have $g_0 = \varphi_m$ by definition, hence g_0 is a 0-fold pseudo-derivative of φ_m. For the induction step $l \to l+1$, we observe that $l \leq m-1$, hence $\Delta_l(\varphi_m) = 0$, which implies $\varphi_m^{(l)}(0-) = \varphi_m^{(l)}(0+)$ and therefore, g_l is continuous. Moreover, g_l is continuously differentiable on $(0, x)$ and $(-x, 0)$, and g_l' is bounded on these intervals. Thus,

$$\begin{aligned} g_l(x) &= g_l(0) + \int_0^x g_l'(t)\,\mathrm{d}t = g_l(0) + \int_0^x g_{l+1}(t)\,\mathrm{d}t \\ g_l(-x) &= g_l(0) + \int_0^{-x} g_l'(t)\,\mathrm{d}t = g_l(0) + \int_0^{-x} g_{l+1}(t)\,\mathrm{d}t \ , \end{aligned}$$

which shows that g_{l+1} is a pseudo-derivative of g_l.

It remains to show that $g_m \in L_2(\mathcal{N}(0,1))$. For $x \neq 0$, we have $s_k^{(m)}(x) = 0$ for $k < m$ and therefore $g_m(x) = \varphi^{(m)}(x)$. By Assumption 4, $\varphi|_{(0,\infty)} \in \mathcal{S}^\infty((0,\infty))$ and $\varphi|_{(-\infty,0)} \in \mathcal{S}^\infty((-\infty,0))$, which implies the desired integrability for $\varphi^{(m)}$, hence $g_m \in L_2(\mathcal{N}(0,1))$. □

Theorem A.8 (Boundary behavior of dual activations)**.** *Let φ be an activation function as in Assumption 4 and let $m \in \mathbb{N}_0$ such that $\Delta_k(\varphi) = 0$ for all $k < m$ (see Definition A.6). Then, there exists $b_m > 0$ depending only on m such that for $\tau \in \{\pm 1\}$,*

$$\widehat{\varphi}(\tau(1-t)) = \Delta_m(\varphi)^2(-\tau)^{m+1} b_m t^{m+1/2} + \mathcal{Q}_{-1,m+1/2}(t)$$

holds.

Proof. **Step 0: Proof strategy** Fix $\tau \in \{-1, 1\}$. For an integer $M_1 > m$ to be defined later, we decompose φ using Lemma C.15 as

$$\varphi = g_1 + g_2, \qquad g_1 \coloneqq \sum_{k=m}^{M_1-1} \Delta_k(\varphi) s_k, \qquad g_2 \coloneqq \varphi_{M_1} \ .$$

Here, Lemma C.15 tells us that g_2 is M_1 times pseudo-differentiable and $g_2^{(M_1)} \in L_2(\mathcal{N}(0,1))$. The basic strategy is as follows: To verify $f(t) = \mathcal{Q}_{-1,m+1/2}(t)$ for a suitable function, we need to show that for all $\gamma \in \mathbb{R}$, there exists a decomposition $f = q + r$ with $q \in \mathcal{P}_{-1,m+1/2}$ and $r \in \mathcal{R}_\gamma$. We will select M_1 depending on γ to construct such a decomposition. To show $r \in \mathcal{R}_\gamma$, we need to show $|r^{(N)}(t)| = O(t^{\gamma-N})$ for all $N \in \mathbb{N}_0$. To this end, we will decompose φ again with an order $M_2 > M_1$ that depends on N (but does not change r).

Step 1: Defining the decomposition of $\widehat{\varphi}$. Using Definition C.10, we obtain

$$\begin{aligned}\widehat{g_1}(t) &= \sum_{n=0}^{\infty}\left(\sum_{k=m}^{M_1-1}\Delta_k(\varphi)a_n(s_k)\right)^2 t^n\\ &= \sum_{k_1=m}^{M_1-1}\sum_{k_2=m}^{M_1-1}\Delta_{k_1}(\varphi)\Delta_{k_2}(\varphi)\sum_{n=0}^{\infty}a_n(s_{k_1})a_n(s_{k_2})t^n\\ &= \sum_{k_1=m}^{M_1-1}\sum_{k_2=m}^{M_1-1}\Delta_{k_1}(\varphi)\Delta_{k_2}(\varphi)f_{k_1,k_2}(t).\end{aligned}$$

For $k_1 = k_2 = m$, we have

$$f_{m,m}(\tau(1-t)) = \widehat{s_m}(\tau(1-t)) \overset{\text{Lemma C.7}}{=} (-1)^{m+1}\Delta_m(\varphi)^2\tau^{m+1}b_m t^{m+1/2} + \mathcal{Q}_{-1,m+1/2}(t)\ .$$

For $k_1 > m$ or $k_2 > m$, Proposition C.12 yields

$$f_{k_1,k_2}(\tau(1-t)) = \mathcal{Q}_{-1,(k_1+k_2)/2}(t) = \mathcal{Q}_{-1,m+1/2}(t).$$

Therefore,

$$\widehat{g_1}(\tau(1-t)) = \Delta_m(\varphi)^2(-1)^{m+1}\tau^{m+1}b_m t^{m+1/2} + \mathcal{Q}_{-1,m+1/2}(t)$$

for some constant $b_m > 0$. Moreover, for $M_1 \geq 2+m$, we have $M_1/2 \geq 3/4 + m/2$. Using Lemma C.9 and Lemma C.13, we obtain

$$\begin{aligned}\widehat{\varphi}(t) - \widehat{g_1}(t) &= \sum_{n=0}^{\infty}\left(a_n(g_2) + \sum_{k=m}^{M_1-1}\Delta_k(\varphi)a_n(s_k)\right)^2 t^n - \sum_{n=0}^{\infty}\left(\sum_{k=m}^{M_1-1}\Delta_k(\varphi)a_n(s_k)\right)^2 t^n\\ &= \sum_{n=0}^{\infty}a_n(g_2)\left(a_n(g_2) + 2\sum_{k=m}^{M_1-1}\Delta_k(\varphi)a_n(s_k)\right)t^n\\ &= \sum_{n=0}^{\infty}o((n+1)^{-M_1/2})O((n+1)^{-3/4-m/2})t^n\\ &= \sum_{n=0}^{\infty}o((n+1)^{-3/4-m/2-M_1/2})t^n\ .\end{aligned}$$

Let $\gamma \in \mathbb{R}$ be arbitrary. We choose M_1 large enough such that

$$3/4 + m/2 + M_1/2 > \lceil\gamma\rceil + 1 \tag{C.4}$$

Then, Lemma C.14 yields

$$h := \widehat{\varphi} - \widehat{g_1} \in C^{\lceil\gamma\rceil}([-1,1])\ .$$

We now define the Taylor polynomials

$$p_\tau(t) := \sum_{k=0}^{\lceil\gamma\rceil-1}\frac{\frac{\mathrm{d}^k}{\mathrm{d}u^k}h(\tau(1-u))|_{u=0}}{k!}t^k = \mathcal{Q}_{-1,\infty}(t)$$

and the rest terms

$$r_\tau(t) := h(\tau(1-t)) - p_\tau(t)\ . \tag{C.5}$$

It remains to show that $r_\tau(t) = \mathcal{R}_\gamma(t)$, which will then yield

$$\widehat{\varphi}(\tau(1-t)) = \widehat{g_1}(\tau(1-t)) + p_\tau(t) + r_\tau(t)$$

$$= -\Delta_m(\varphi)^2\tau^{m+1}(-1)^{m+1}b_m t^{m+1/2} + \mathcal{P}_{-1,m+1/2}(t) + \mathcal{R}_\gamma(t)$$

for arbitrary $\gamma \in \mathbb{R}$.

Step 2: Analyzing the rest term. Let $\tau \in \{-1, 1\}$ and $N \in \mathbb{N}_0$ be arbitrary. We need to show that $|r_\tau^{(N)}(t)| = O(t^{\gamma-N})$, where r_τ is defined in Eq. (C.5). For an integer $M_2 > M_1$ yet to be specified, we use Lemma C.15 again to decompose

$$g_2 = \varphi_{M_2} + \sum_{k=M_1}^{M_2-1} \Delta_k(\varphi)s_k \ .$$

With the index set

$$\mathcal{I} \coloneqq \{m, m+1, \dots, M_2-1\}^2 \setminus \{m, m+1, \dots, M_1-1\}^2 \ ,$$

we then obtain similar to the calculation above

$$\begin{aligned}
h(t) &= h_1(t) + h_2(t) \\
h_1(t) &\coloneqq \sum_{(k_1,k_2)\in\mathcal{I}} \Delta_{k_1}(\varphi)\Delta_{k_2}(\varphi)f_{k_1,k_2}(t) \\
h_2(t) &\coloneqq \sum_{n=0}^{\infty} a_n(\varphi_{M_2})\left(a_n(\varphi_{M_2}) + 2\sum_{k=m}^{M_2-1}\Delta_k(\varphi)a_n(s_k)\right)t^n \\
&= \sum_{n=0}^{\infty} o((n+1)^{-M_2/2})O((n+1)^{-3/4-m/2})t^n = \sum_{n=0}^{\infty} o((n+1)^{-3/4-m/2-M_2/2})t^n \ .
\end{aligned}$$

We now choose M_2 sufficiently large such that $3/4+m/2+M_2/2 > \tilde{N}+1$, where $\tilde{N} \coloneqq \max\{N, \lceil\gamma\rceil\}$. Lemma C.14 yields

$$h_2 \in C^{\tilde{N}}([-1, 1]) \ .$$

Since (C.4) implies $(m + M_1)/2 \geq \gamma$, we obtain from Proposition C.12

$$h_1(\tau(1-t)) = \mathcal{P}_{-1,(m+M_1)/2}(t) + \mathcal{R}_\gamma(t) = \mathcal{P}_{-1,\infty}(t) + \mathcal{R}_\gamma(t) \ .$$

In other words, we can find a polynomial $p_{1,\tau}$ of degree $\leq \lceil\gamma\rceil - 1$ and a function $r_{1,\tau} \in \mathcal{R}_\gamma$ such that

$$h_1(\tau(1-t)) = p_{1,\tau}(t) + r_{1,\tau}(t) \ .$$

We therefore investigate

$$\begin{aligned}
r_{2,\tau}(t) &\coloneqq r_\tau(t) - r_{1,\tau}(t) = (h(\tau(1-t)) - p_\tau(t)) - (h_1(\tau(1-t)) - p_{1,\tau}(t)) \\
&= h_2(\tau(1-t)) - p_\tau(t) + p_{1,\tau}(t) \ ,
\end{aligned}$$

which satisfies $r_{2,\tau} \in C^{\tilde{N}}([-1, 1])$ We now distinguish two cases:

- If $N \geq \gamma$, we simply use the continuity of $r_{2,\tau}^{(N)}$ in 0 to obtain
$$|r_{2,\tau}^{(N)}(t)| = O(1) = O(t^{\gamma-N}) \ .$$
- If $N < \gamma$, we proceed differently. For $0 \leq n \leq \lceil\gamma\rceil - 1$, we have
$$r_{2,\tau}^{(n)}(0) = r_{2,\tau}^{(n)}(0+) = r_\tau^{(n)}(0+) - r_{1,\tau}^{(n)}(0+) = 0 - 0 = 0 \ ,$$
where we used that $r_\tau^{(n)}(0+) = 0$ by construction and that $r_{1,\tau}^{(n)}(0+) = 0$ thanks to $r_{1,\tau} \in \mathcal{R}_\gamma$ and $n < \gamma$. Thus, Taylor's theorem with Peano's form of the remainder yields
$$\begin{aligned}
|r_{2,\tau}^{(N)}(t)| &= \left|\sum_{k=0}^{\lceil\gamma\rceil-N} \frac{r_{2,\tau}^{(N+k)}(0)}{k!}t^k + o(t^{\lceil\gamma\rceil-N})\right| \\
&= O(t^{\lceil\gamma\rceil-N}) = O(t^{\gamma-N}) \ .
\end{aligned}$$

Since $r_{1,\tau} \in \mathcal{R}_\gamma$, we obtain

$$|r_\tau^{(N)}(t)| \leq |r_{1,\tau}^{(N)}(t)| + |r_{2,\tau}^{(N)}(t)| \leq O(t^{\gamma-N}) + O(t^{\gamma-N}) = O(t^{\gamma-N}) \ ,$$

which completes the proof. □

D NEURAL KERNELS

D.1 Analytical formulas

Recall the network's architecture in Definition 1.

For the subsequent consideration concerning kernels, we compare the network behavior for two inputs $\boldsymbol{x}, \bar{\boldsymbol{x}}$. All terms $\bullet$ correspond to the inputs $\boldsymbol{x}$, all terms $\bar{\bullet}$ to the input $\bar{\boldsymbol{x}}$.

Definition D.1 (Neural kernels)**.** Consider a network of depth $L \geq 2$ and output dimension $d_L = 1$. Let $d = d_1 = \cdots = d_{L-1}$. Define the *neural network Gaussian process*-kernel $k_L^{\mathrm{NNGP}} : \mathbb{R}^{d_0} \times \mathbb{R}^{d_0} \to \mathbb{R}$ as

$$k_L^{\mathrm{NNGP}}(\boldsymbol{x}, \bar{\boldsymbol{x}}) \coloneqq \lim_{d\to\infty} \operatorname{Cov}\left(\boldsymbol{z}^{(L)}, \bar{\boldsymbol{z}}^{(L)}\right)$$

and the neural tangent kernel $k_L^{\mathrm{NTK}} : \mathbb{R}^{d_0} \times \mathbb{R}^{d_0} \to \mathbb{R}$ as

$$k_L^{\mathrm{NTK}}(\boldsymbol{x}, \bar{\boldsymbol{x}}) \coloneqq \lim_{d\to\infty} \left\langle \nabla_{\boldsymbol{\theta}} \boldsymbol{z}^{(L)}, \nabla_{\boldsymbol{\theta}} \bar{\boldsymbol{z}}^{(L)} \right\rangle ,$$

where $\nabla_{\boldsymbol{\theta}} \boldsymbol{z}^{(L)}$ denotes derivation of the output $\boldsymbol{z}^{(L)}$ by all parameters $\boldsymbol{\theta}$. ◀

Lemma D.2. *Let the activation function $\varphi : \mathbb{R} \to \mathbb{R}$ fulfill Assumption 4. Then, the NNGP and NTK kernels introduced above converge almost surely and they are given by*

$$\begin{aligned}
k_1^{\mathrm{NNGP}}(\boldsymbol{x}, \bar{\boldsymbol{x}}) &\coloneqq \sigma_b^2 \sigma_i^2 + \sigma_w^2 \langle \boldsymbol{x}, \bar{\boldsymbol{x}} \rangle \\
k_L^{\mathrm{NNGP}}(\boldsymbol{x}, \bar{\boldsymbol{x}}) &= \sigma_b^2 \sigma_i^2 + \sigma_w^2 \mathbb{E}_{(u,v)\sim \boldsymbol{\Sigma}_{L-1}(\boldsymbol{x},\bar{\boldsymbol{x}})}[\varphi(u)\varphi(v)] \\
k_L^{\mathrm{NTK}}(\boldsymbol{x}, \bar{\boldsymbol{x}}) &= \sigma_b^2(1-\sigma_i^2) + k_L^{\mathrm{NNGP}}(\boldsymbol{x}, \bar{\boldsymbol{x}}) + \sigma_w^2 \mathbb{E}_{(u,v)\sim \mathcal{N}(0,\boldsymbol{\Sigma}_{L-1}(\boldsymbol{x},\bar{\boldsymbol{x}}))}[\varphi'(u)\varphi'(v)] k_{L-1}^{\mathrm{NTK}}(\boldsymbol{x}, \bar{\boldsymbol{x}}) \\
\boldsymbol{\Sigma}_L(\boldsymbol{x}, \bar{\boldsymbol{x}}) &= \begin{pmatrix} k_L^{\mathrm{NNGP}}(\boldsymbol{x}, \boldsymbol{x}) & k_L^{\mathrm{NNGP}}(\boldsymbol{x}, \bar{\boldsymbol{x}}) \\ k_L^{\mathrm{NNGP}}(\bar{\boldsymbol{x}}, \boldsymbol{x}) & k_L^{\mathrm{NNGP}}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{x}}) \end{pmatrix}
\end{aligned}$$

where we define $k_1^{\mathrm{NTK}}(\boldsymbol{x}, \bar{\boldsymbol{x}}) \coloneqq \sigma_b^2(1-\sigma_i^2) + k_1^{\mathrm{NNGP}}(\boldsymbol{x}, \bar{\boldsymbol{x}})$.

Proof. Denote $\tilde{d}_l \coloneqq d$ if $l \in \{1, \dots, L-1\}$ and $\tilde{d}_l = 1$ otherwise. Consider the inputs $\boldsymbol{x}, \bar{\boldsymbol{x}} \in \mathbb{R}^{d_0}$. For a fixed hidden layer width $d \in \mathbb{N}$, the finite version of the NTK kernel is given by

$$\begin{aligned}
&\left\langle \nabla_{\boldsymbol{\theta}} \boldsymbol{z}^{(L)}, \nabla_{\boldsymbol{\theta}} \bar{\boldsymbol{z}}^{(L)} \right\rangle = \sum_{l=1}^{L} \left\langle \nabla_{\boldsymbol{W}^{(l)}} \boldsymbol{z}^{(L)}, \nabla_{\boldsymbol{W}^{(l)}} \bar{\boldsymbol{z}}^{(L)} \right\rangle + \sum_{l=1}^{L} \left\langle \nabla_{\boldsymbol{b}^{(l)}} \boldsymbol{z}^{(L)}, \nabla_{\boldsymbol{b}^{(l)}} \bar{\boldsymbol{z}}^{(L)} \right\rangle , \\
&\nabla_{\boldsymbol{b}^{(l)}} \boldsymbol{z}^{(L)} = \sigma_b \nabla_{\boldsymbol{z}^{(l)}} \boldsymbol{z}^{(L)} , \\
&\left\langle \nabla_{\boldsymbol{W}^{(l)}} \boldsymbol{z}^{(L)}, \nabla_{\boldsymbol{W}^{(l)}} \bar{\boldsymbol{z}}^{(L)} \right\rangle = \frac{\sigma_w^2}{\tilde{d}_{l-1}} \left\langle \nabla_{\boldsymbol{z}^{(l)}} \boldsymbol{z}^{(L)} \boldsymbol{x}^{(l-1)\top}, \nabla_{\bar{\boldsymbol{z}}^{(l)}} \bar{\boldsymbol{z}}^{(L)} \bar{\boldsymbol{x}}^{(l-1)\top} \right\rangle ,
\end{aligned} \tag{D.1}$$

where inner products involving matrices are Frobenius inner products. Using the outer product structure we can rewrite the weight-matrix gradient as

$$\begin{aligned}
&\left\langle \nabla_{\boldsymbol{W}^{(l)}} \boldsymbol{z}^{(L)}, \nabla_{\boldsymbol{W}^{(l)}} \bar{\boldsymbol{z}}^{(L)} \right\rangle \\
&= \sigma_w^2 \left\langle \nabla_{\boldsymbol{z}^{(l)}} \boldsymbol{z}^{(L)}, \nabla_{\bar{\boldsymbol{z}}^{(l)}} \bar{\boldsymbol{z}}^{(L)} \right\rangle \frac{1}{\tilde{d}_{l-1}} \left\langle \boldsymbol{x}^{(l-1)}, \bar{\boldsymbol{x}}^{(l-1)} \right\rangle \\
&= \frac{\sigma_w^2}{d_l} \left\langle \sqrt{d_l} \nabla_{\boldsymbol{z}^{(l)}} \boldsymbol{z}^{(L)}, \sqrt{d_l} \nabla_{\bar{\boldsymbol{z}}^{(l)}} \bar{\boldsymbol{z}}^{(L)} \right\rangle \frac{1}{\tilde{d}_{l-1}} \left\langle \boldsymbol{x}^{(l-1)}, \bar{\boldsymbol{x}}^{(l-1)} \right\rangle .
\end{aligned} \tag{D.2}$$

The occurring backpropagation terms can be recursively unrolled as

$$\begin{aligned}
&\nabla_{\boldsymbol{z}^{(L)}} \boldsymbol{z}^{(l)} = \frac{\sigma_w}{\sqrt{d_l}} \boldsymbol{W}^{(l+1)\top} \nabla_{\boldsymbol{z}^{(l+1)}} \boldsymbol{z}^{(L)} \odot \varphi'(\boldsymbol{z}^{(l)}) && 1 \leq l \leq L-1 , \\
&\nabla_{\boldsymbol{z}^{(L)}} \boldsymbol{z}^{(L)} = 1,
\end{aligned}$$

where $\odot$ denotes component-wise multiplication.

In order to rigorously calculate the NNGP and the NTK, we use the simplified $\texttt{NETSOR}^\top$ program from Section 7 in Yang (2020). To this end, define the set of d-dimensional initial vectors $\mathcal{V} := \{\sigma_w \boldsymbol{W}^{(1)}\boldsymbol{x}, \sigma_w \boldsymbol{W}^{(1)}\bar{\boldsymbol{x}}, \boldsymbol{W}^{(L)\top}, \sigma_i\sigma_b\boldsymbol{b}^{(1)}, \sigma_i\sigma_b\boldsymbol{b}^{(2)}, \ldots, \sigma_i\sigma_b\boldsymbol{b}^{(L-1)}\}$. For $1 \leq i \leq d$, the process $(v_i)_{v\in\mathcal{V}}$ is a centered Gaussian process distributed as $(Z^v)_{v\in\mathcal{V}}$, where[2] $\mathrm{Cov}(Z^{\boldsymbol{W}^{(1)}\boldsymbol{y}}, Z^{\boldsymbol{W}^{(1)}\bar{\boldsymbol{y}}}) = \langle \boldsymbol{y}, \bar{\boldsymbol{y}}\rangle$ for $\boldsymbol{y}, \bar{\boldsymbol{y}} \in \{\boldsymbol{x}, \bar{\boldsymbol{x}}\}$, $\mathrm{Cov}(Z^{\boldsymbol{b}^{(l)}}, Z^v) = \mathbb{1}_{\{\boldsymbol{b}^{(l)}\}}(v)$ and $\mathrm{Cov}\left(Z^{\boldsymbol{W}^{(L)\top}}, v\right) = \mathbb{1}_{\{\boldsymbol{W}^{(L)\top}\}}(v)$. Define the set of $\mathbb{R}^{d\times d}$ random matrices $\mathcal{W} := \{\tilde{\boldsymbol{W}}^{(2)}, \ldots, \tilde{\boldsymbol{W}}^{(L-1)}\}$ by $\tilde{\boldsymbol{W}}^{(i)}_{jk} \sim \mathcal{N}(0, \sigma_w^2/d)$. Now we recursively define the vectors

$$\begin{aligned}
\boldsymbol{h}^{(1)} &:= \sigma_w \boldsymbol{W}^{(1)}\boldsymbol{x}, & \boldsymbol{h}^{(l)} &:= \tilde{\boldsymbol{W}}^{(l)}\boldsymbol{x}^{(l-1)} \\
\boldsymbol{z}^{(l)} &:= \boldsymbol{h}^{(l)} + \sigma_b\sigma_i\boldsymbol{b}^{(l)}, & \boldsymbol{x}^{(l)} &:= \varphi(\boldsymbol{z}^{(l)}), \\
\mathrm{d}\boldsymbol{z}^{(L-1)} &:= \boldsymbol{W}^{(L)\top} \odot \varphi'(\boldsymbol{z}^{(L-1)}), & \mathrm{d}\boldsymbol{h}^{(l)} &:= \tilde{\boldsymbol{W}}^{(l+1)\top}\, \mathrm{d}\boldsymbol{z}^{(l+1)}, \\
\mathrm{d}\boldsymbol{z}^{(l)} &:= \mathrm{d}\boldsymbol{h}^{(l)} \odot \varphi'(\boldsymbol{z}^{(l)}), &&
\end{aligned} \tag{D.3}$$

where the index l has the range $1 \leq l \leq L-1$ for $\boldsymbol{x}^{(l)}$, $1 \leq l \leq L-2$ for $\mathrm{d}\boldsymbol{h}^{(l)}$, $\mathrm{d}\boldsymbol{z}^{(l)}$ and $2 \leq l \leq L-1$ for $\boldsymbol{h}^{(l)}, \boldsymbol{z}^{(l)}$. Also note $\mathrm{d}\boldsymbol{z}^{(l)} = \frac{1}{\sqrt{d}}\nabla_{\boldsymbol{z}^{(l)}}\boldsymbol{z}^{(L)}$ for $1 \leq l \leq L-1$ due to the „missing normalization"in the definition of $\mathrm{d}\boldsymbol{z}^{(L-1)}$. The individual vectors in Eq. (D.3) are given by a non-linear operation or a matrix operation as in (Yang, 2020, Box 1 p. 7). This enables applying Theorem 7.2 and Box 1 from Yang (2020), which states that limits of the kernels

$$\begin{aligned}
B^{(l)}(\boldsymbol{x}, \bar{\boldsymbol{x}}) &:= \lim_{d\to\infty} \frac{1}{d}\left\langle \boldsymbol{z}^{(l)}, \bar{\boldsymbol{z}}^{(l)}\right\rangle, \\
C^{(l)}(\boldsymbol{x}, \bar{\boldsymbol{x}}) &:= \lim_{d\to\infty} \frac{1}{d}\left\langle \boldsymbol{x}^{(l)}, \bar{\boldsymbol{x}}^{(l)}\right\rangle, \\
D^{(l,L)}(\boldsymbol{x}, \bar{\boldsymbol{x}}) &:= \lim_{d\to\infty} \frac{1}{d}\left\langle \mathrm{d}\boldsymbol{z}^{(l)}, \mathrm{d}\bar{\boldsymbol{z}}^{(l)}\right\rangle, \\
E^{(l)}(\boldsymbol{x}, \bar{\boldsymbol{x}}) &:= \lim_{d\to\infty} \frac{1}{d}\left\langle \varphi'(\boldsymbol{z}^{(l)}), \varphi'(\bar{\boldsymbol{z}}^{(l)})\right\rangle.
\end{aligned}$$

exist almost surely and furthermore yields recursive formulas for those limits. Note that the backpropagation terms $D^{(l,L)}$ depend on the network's depth L. By the independence of the parameters in the network we have

$$\begin{aligned}
k_L^{\mathrm{NNGP}}(\boldsymbol{x}, \bar{\boldsymbol{x}}) &= \lim_{d\to\infty} \mathrm{Cov}\left(\boldsymbol{z}^{(L)}, \bar{\boldsymbol{z}}^{(L)}\right) \\
&= \lim_{d\to\infty} \frac{\sigma_w^2}{d}\mathbb{E}\big[\boldsymbol{x}^{(L-1)\top}\boldsymbol{W}^{(L)^\top}\boldsymbol{W}^{(L)}\bar{\boldsymbol{x}}^{(L-1)}\big] + \sigma_i^2\sigma_b^2\mathbb{E}\big[(\boldsymbol{b}^{(L)})^2\big] \\
&= \lim_{d\to\infty} \sigma_w^2\mathbb{E}\frac{1}{d}\left\langle \boldsymbol{x}^{(L-1)}, \bar{\boldsymbol{x}}^{(L-1)}\right\rangle + \sigma_i^2\sigma_b^2 \\
&= \sigma_w^2 C^{(L-1)}(\boldsymbol{x}, \bar{\boldsymbol{x}}) + \sigma_i^2\sigma_b^2 .
\end{aligned}$$

Here, the convergence of the expectation follows from Yang (2019a, Proposition G.4). By Eqs. (D.1) and (D.2) the NTK-kernel can be expressed as

$$k_L^{\mathrm{NTK}}(\boldsymbol{x}, \bar{\boldsymbol{x}}) = \sigma_w^2 \sum_{l=1}^{L} D^{(l)}(\boldsymbol{x}, \bar{\boldsymbol{x}})C^{(l-1)}(\boldsymbol{x}, \bar{\boldsymbol{x}}) + \sigma_b^2 \sum_{l=1}^{L} D^{(l)}(\boldsymbol{x}, \bar{\boldsymbol{x}}) .$$

For a kernel k denote $\Sigma_k(\boldsymbol{y}, \bar{\boldsymbol{y}}) := \begin{pmatrix} k(\boldsymbol{y},\boldsymbol{y}) & k(\boldsymbol{y},\bar{\boldsymbol{y}}) \\ k(\boldsymbol{y},\bar{\boldsymbol{y}}) & k(\bar{\boldsymbol{y}},\bar{\boldsymbol{y}}) \end{pmatrix}$. Using Eq. (D.3) and Yang (2020), the above kernels are recursively given by

$$\begin{aligned}
&B^{(l)}(\boldsymbol{x}, \bar{\boldsymbol{x}}) = \sigma_w^2 C^{l-1}(\boldsymbol{x}, \bar{\boldsymbol{x}}) + \sigma_b^2\sigma_i^2, \\
&C^{(l)}(\boldsymbol{x}, \bar{\boldsymbol{x}}) = \mathbb{E}_{(u,v)\sim\mathcal{N}(0,\Sigma_{B^{(l)}}(\boldsymbol{x},\bar{\boldsymbol{x}}))}[\varphi(u)\varphi(v)], \quad (l \geq 1) \\
&C^{(0)}(\boldsymbol{x}, \bar{\boldsymbol{x}}) = \langle \boldsymbol{x}, \bar{\boldsymbol{x}}\rangle,
\end{aligned} \tag{D.4}$$

[2]Note that Z^v is a real-valued random variable for $v \in \mathcal{V}$.

$$D^{(l,L)}(\boldsymbol{x},\bar{\boldsymbol{x}}) = \sigma_w^2 D^{(l+1)}(\boldsymbol{x},\bar{\boldsymbol{x}})E^{(l)}(\boldsymbol{x},\bar{\boldsymbol{x}}), \qquad l \le L-1,$$
$$D^{(L,L)}(\boldsymbol{x},\bar{\boldsymbol{x}}) = 1,$$
$$E^{(l)}(\boldsymbol{x},\bar{\boldsymbol{x}}) = \mathbb{E}_{(u,v)\sim\mathcal{N}(0,\Sigma_{B^{(l)}}(\boldsymbol{x},\bar{\boldsymbol{x}}))}[\varphi'(u)\varphi'(v)],$$

and a recursive formula for $D^{(l,L)}$ over the depth L follows as

$$D^{(l,L+1)} = \sigma_w^2 E^{(l)} \cdot \ldots \cdot \sigma_w^2 E^{(L)} = \sigma_w^2 E^{(L)} D^{(l,L)} \ .$$

The recursive formula for $k_L^{\mathrm{NNGP}}(\boldsymbol{x},\bar{\boldsymbol{x}})$ can directly be derived from Eq. (D.4). We investigate $k_L^{\mathrm{NTK}}(\boldsymbol{x},\bar{\boldsymbol{x}})$ by induction. For $L=1$ there is nothing to show. Assume that the claimed formula holds for $L \in \mathbb{N}$. Then we have

$$\begin{aligned}
k_{L+1}^{\mathrm{NTK}}(\boldsymbol{x},\bar{\boldsymbol{x}}) &= \sigma_w^2 \sum_{l=1}^{L+1} D^{(l,L+1)}(\boldsymbol{x},\bar{\boldsymbol{x}})C^{(l-1)}(\boldsymbol{x},\bar{\boldsymbol{x}})x + \sigma_b^2 \sum_{l=1}^{L} D^{(l,L+1)}(\boldsymbol{x},\bar{\boldsymbol{x}}) \\
&= \sigma_w^2 E^{(L)}(\boldsymbol{x},\bar{\boldsymbol{x}})\left(\sigma_w^2 \sum_{l=1}^{L} D^{(l,L)}(\boldsymbol{x},\bar{\boldsymbol{x}})C^{(l-1)}(\boldsymbol{x},\bar{\boldsymbol{x}}) + \sigma_b^2 \sum_{l=1}^{L} D^{(l,L)}(\boldsymbol{x},\bar{\boldsymbol{x}})\right) \\
&\quad + \sigma_w^2 C^{(L)}(\boldsymbol{x},\bar{\boldsymbol{x}}) + \sigma_b^2 \\
&= \sigma_b^2(1-\sigma_i^2) + k_{L+1}^{\mathrm{NNGP}}(\boldsymbol{x},\bar{\boldsymbol{x}}) + \sigma_w^2 \mathbb{E}_{(u,v)\sim\mathcal{N}(0,\boldsymbol{\Sigma}_L)}[\varphi'(u)\varphi'(v)]k_L^{\mathrm{NTK}}(x,x').
\end{aligned}$$

□

A dot-product kernel k on the sphere $\mathbb{S}^d$ is conveniently described by a function $\kappa : [-1,1] \to \mathbb{R}$ by defining $\kappa(\langle x,y\rangle) := k(x,y)$. We translate the NNGP- and NTK-recursion of Lemma D.2 to a recursion for the corresponding functions κ describing the restriction of these kernels to the sphere.

Lemma A.5 (Neural kernels on the unit sphere)**.** *Let the activation function $\varphi : \mathbb{R} \to \mathbb{R}$ fulfill Assumption 4. Consider a neural network $f_\theta : \mathbb{R}^{d+1} \to \mathbb{R}$ initialized as in Definition 1. For $l \ge 2$ and $t \in [-1,1]$ we recursively define*

$$\begin{aligned}
\alpha_1 &:= \sigma_b^2\sigma_i^2 + \sigma_w^2 \\
\alpha_l &:= \sigma_b^2\sigma_i^2 + \sigma_w^2 \widehat{\varphi_{\cdot\sqrt{\alpha_{l-1}}}}(1) \\
\kappa_1^{\mathrm{NNGP}}(t) &:= \sigma_b^2\sigma_i^2 + \sigma_w^2 t \\
\kappa_l^{\mathrm{NNGP}}(t) &:= \sigma_b^2\sigma_i^2 + \sigma_w^2 \widehat{\varphi_{\cdot\sqrt{\alpha_{l-1}}}}(\kappa_{l-1}^{\mathrm{NNGP}}(t)/\alpha_{l-1}) \\
\kappa_1^{\mathrm{NTK}}(t) &:= \sigma_b^2(1-\sigma_i^2) + \kappa_1^{\mathrm{NNGP}}(t) \\
\kappa_l^{\mathrm{NTK}}(t) &:= \sigma_b^2(1-\sigma_i^2) + \kappa_l^{\mathrm{NNGP}}(t) + \sigma_w^2 \kappa_{l-1}^{\mathrm{NTK}}(t)\widehat{(\varphi')_{\cdot\sqrt{\alpha_{l-1}}}}(\kappa_{l-1}^{\mathrm{NNGP}}(t)/\alpha_{l-1}) \ .
\end{aligned}$$

Then for all $l \ge 1, \boldsymbol{x},\boldsymbol{x}' \in \mathbb{S}^d$ we have $\alpha_l > 0$ and

$$k_l^{\mathrm{NNGP}}(\boldsymbol{x},\bar{\boldsymbol{x}}) = \kappa_l^{\mathrm{NNGP}}(\langle \boldsymbol{x},\bar{\boldsymbol{x}}\rangle)$$
$$k_l^{\mathrm{NTK}}(\boldsymbol{x},\bar{\boldsymbol{x}}) = \kappa_l^{\mathrm{NTK}}(\langle \boldsymbol{x},\bar{\boldsymbol{x}}\rangle) \ .$$

Proof. As $\sigma_w^2 > 0$ and φ is not almost surely equal to zero, $\alpha_l > 0$ follows for all l. By a simple induction we see that $\alpha_l = k_l^{\mathrm{NNGP}}(1)$ holds for all $l \ge 1$. The identities for the NNGP can be shown by induction on l as well. The induction base $k_1^{\mathrm{NNGP}}(\boldsymbol{x},\bar{\boldsymbol{x}}) = \kappa_1^{\mathrm{NNGP}}(\langle \boldsymbol{x},\bar{\boldsymbol{x}}\rangle)$ is straightforward. For the induction step Lemma D.2 yields

$$k_l^{\mathrm{NNGP}}(\boldsymbol{x},\bar{\boldsymbol{x}}) = \sigma_b^2\sigma_i^2 + \sigma_w^2 \mathbb{E}_{(u,v)\sim\mathcal{N}(0,\boldsymbol{\Sigma}_{l-1}(\boldsymbol{x},\bar{\boldsymbol{x}}))}[\varphi(u)\varphi(v)],$$
$$\boldsymbol{\Sigma}_{l-1}(\boldsymbol{x},\bar{\boldsymbol{x}}) = \begin{pmatrix} k_{l-1}^{\mathrm{NNGP}}(\boldsymbol{x},\boldsymbol{x}) & k_{l-1}^{\mathrm{NNGP}}(\boldsymbol{x},\boldsymbol{x}') \\ k_{l-1}^{\mathrm{NNGP}}(\boldsymbol{x},\boldsymbol{x}') & k_{l-1}^{\mathrm{NNGP}}(\boldsymbol{x}',\boldsymbol{x}') \end{pmatrix} = \alpha_{l-1} \cdot \begin{pmatrix} 1 & \kappa_{l-1}^{\mathrm{NNGP}}(\langle \boldsymbol{x},\bar{\boldsymbol{x}}\rangle)/\alpha_{l-1} \\ \kappa_{l-1}^{\mathrm{NNGP}}(\langle \boldsymbol{x},\bar{\boldsymbol{x}}\rangle)/\alpha_{l-1} & 1 \end{pmatrix}$$

and the identity $k_l^{\mathrm{NNGP}}(\boldsymbol{x},\bar{\boldsymbol{x}}) = \kappa_l^{\mathrm{NNGP}}(\langle \boldsymbol{x},\bar{\boldsymbol{x}}\rangle)$ follows directly from the definition of the dual activation (A.4). A similar argument shows $k_l^{\mathrm{NTK}}(\boldsymbol{x},\bar{\boldsymbol{x}}) = \kappa_l^{\mathrm{NTK}}(\langle \boldsymbol{x},\bar{\boldsymbol{x}}\rangle)$. □

D.2 Even and odd parts of neural kernels

Lemma D.3 (Even/odd algebra)**.** *Let $f, g : \mathbb{R} \to \mathbb{R}$. Then,*

(a) $(f+g)_{\text{even}} = f_{\text{even}} + g_{\text{even}}$ *and* $(f+g)_{\text{odd}} = f_{\text{odd}} + g_{\text{odd}}$.
(b) $(f \cdot g)_{\text{even}} = f_{\text{even}} \cdot g_{\text{even}} + f_{\text{odd}} \cdot g_{\text{odd}}$ *and* $(f \cdot g)_{\text{odd}} = f_{\text{even}} \cdot g_{\text{odd}} + f_{\text{odd}} \cdot g_{\text{even}}$.
(c) If g *is odd,* $(f \circ g)_{\text{even}} = f_{\text{even}} \circ g$ *and* $(f \circ g)_{\text{odd}} = f_{\text{odd}} \circ g$*. If* g *is even,* $f \circ g$ *is even.*
(d) If f' *is a weak derivative of* f*, then* $(f')_{\text{even}}$ *is a weak derivative of* f_{odd} *and* $(f')_{\text{odd}}$ *is a weak derivative of* f_{even}.

Proof. Statements (a) – (c) are straightforward to prove. For (d), we obtain for $x \in \mathbb{R}$:

$$\begin{aligned}\int_0^x (f')_{\text{even}}(t)\,\mathrm{d}t &= \int_0^x \frac{f'(t)+f'(-t)}{2}\,\mathrm{d}t \\ &= \frac{1}{2}\left(\int_0^x f'(t)\,\mathrm{d}t + \int_0^x f'(-t)\,\mathrm{d}t\right) \\ &= \frac{1}{2}\left(\int_0^x f'(t)\,\mathrm{d}t - \int_0^{-x} f'(t)\,\mathrm{d}t\right) \\ &= \frac{1}{2}((f(x)-f(0)) - (f(-x)-f(0))) \\ &= f_{\text{odd}}(x)\ .\end{aligned}$$

This shows that $(f')_{\text{even}}$ is a weak derivative of f_{odd}, and a similar computation shows that $(f')_{\text{odd}}$ is a weak derivative of f_{even}. □

Proposition D.4 (Special cases for NNGP/NTK with even/odd functions)**.** *Let the activation function* φ *fulfill Assumption 4.*

(a) Let $\sigma_b^2\sigma_i^2 = 0$*. If* φ *is even/odd, then* κ_l^{NNGP} *is even/odd for all* $l \geq 2$.
(b) Let $\sigma_b^2\sigma_i^2 = 0$*. Then,* $\left(\kappa_{2,\varphi}^{\text{NNGP}}\right)_{\text{even}} = \kappa_{2,\varphi_{\text{even}}}^{\text{NNGP}}$ *and* $\left(\kappa_{2,\varphi}^{\text{NNGP}}\right)_{\text{odd}} = \kappa_{2,\varphi_{\text{odd}}}^{\text{NNGP}}$.
(c) Let $\sigma_b^2 = 0$*. If* φ *is even/odd, then* κ_l^{NTK} *is even/odd for all* $l \geq 2$.
(d) Let $\sigma_b^2 = 0$*. Then,* $\left(\kappa_{2,\varphi}^{\text{NTK}}\right)_{\text{even}} = \kappa_{2,\varphi_{\text{even}}}^{\text{NTK}}$ *and* $\left(\kappa_{2,\varphi}^{\text{NTK}}\right)_{\text{odd}} = \kappa_{2,\varphi_{\text{odd}}}^{\text{NTK}}$.

Here we denote the activation function in the index to clarify the network architecture the kernels belong to.

Proof.

(a) Since $\sigma_b^2\sigma_i^2 = 0$, κ_1^{NNGP} is odd. If φ is even/odd, then so is $\varphi_{\cdot\sqrt{\alpha_l}}$ and by Lemma C.3 also $\widehat{\varphi_{\cdot\sqrt{\alpha_l}}}$. The claim then follows by induction using Lemma D.3 (c).
(b) We have $\kappa_{2,\varphi}^{\text{NNGP}}(t) = \sigma_w^2\widehat{\varphi_{\cdot\sqrt{\alpha_1}}}(t)$. Note that α_1 does not depend on φ. By Lemma C.3, we obtain

$$\left(\kappa_{2,\varphi}^{\text{NNGP}}\right)_{\text{even}}(t) = \sigma_w^2\widehat{\left(\varphi_{\cdot\sqrt{\alpha_1}}\right)_{\text{even}}}(t) = \sigma_w^2\widehat{(\varphi_{\text{even}})_{\cdot\sqrt{\alpha_1}}}(t) = \kappa_{2,\varphi_{\text{even}}}^{\text{NNGP}}(t)$$

and similar for the odd part.
(c) Since $\sigma_b^2 = 0$, κ_1^{NTK} is odd. Moreover,

$$\kappa_l^{\text{NTK}}(t) = \sigma_w^2\left(\kappa_l^{\text{NNGP}}(t) + \kappa_{l-1}^{\text{NTK}}(t)\cdot\widehat{(\varphi')_{\cdot\sqrt{\alpha_{l-1}}}}(\kappa_{l-1}^{\text{NNGP}}(t)/\alpha_{l-1})\right)$$

holds and we obtain the claim for $l = 2$. We use Lemma D.3 and Lemma C.3 to obtain that $\widehat{(\varphi')_{\sqrt{\alpha_l}}}$ is odd/even if φ is even/odd, and the claim holds for $l \geq 3$ by induction since $\kappa_l^{\text{NNGP}}(t)$ is even/odd by (a).
(d) Just as in (c), this follows from Lemma C.3, Lemma D.3 and (b). □

D.3 Adapting the main theorem from Bietti and Bach (2021)

Here, we derive Theorem A.3, an adapted version of the main theorem of Bietti and Bach (2021) that is closer to our notation. First, we restate the original theorem with minor adaptations (using τ, expanding at $t = 0$ instead of $t \in \{-1, 1\}$, rewriting derivatives of powers, using $\mathbb{S}^d$ instead of $\mathbb{S}^{d-1}$). Here, the notation $a_n \sim b_n$ means $\lim_{n\to\infty} a_n/b_n = 1$.

Theorem D.5 (Theorem 7 in Bietti and Bach (2021), arXiv version 4)**.** *Let* $k : [-1, 1] \to \mathbb{R}$ *be a function that is* C^∞ *on* $(-1, 1)$ *and has for* $\tau \in \{1, -1\}$ *the following expansion for* $t \searrow 0$:

$$k(\tau(1-t)) = p_\tau(t) + \sum_{j=1}^{r} c_{j,\tau}t^{\nu_j} + O(t^{\nu_1+1+\varepsilon}), \tag{D.5}$$

where p_τ are polynomials and $0 < \nu_1 < \cdots < \nu_r$ are not integers and $0 < \varepsilon < \nu_2 - \nu_1$. Assume further that the derivatives $k^{(s)}$ have for any $s \in \mathbb{N}_0$ the following expressions for $t \searrow 0$, where $(t^{\nu_j})^{(s)}$ is the s-th derivative of $t \mapsto t^{\nu_j}$:

$$k^{(s)}(\tau(1-t)) = p_{s,\tau}(t) + \sum_{j=1}^{r} c_{j,\tau}(t^{\nu_j})^{(s)} + O\big(t^{\nu_1+1+\varepsilon-s}\big) \tag{D.6}$$

for some polynomials $p_{s,\tau}$. Then, for $d \geq 1$, the eigenvalues $\mu_k = \mu_k(\kappa, d)$ as defined in Section 2 satisfy, for an absolute constant $C(d, \nu_1)$,

- *For k even, if $c_{1,1} \neq -c_{1,-1}$: $\mu_k \sim (c_{1,1} + c_{1,-1})C(d,\nu_1)k^{-d-2\nu_1}$;*
- *For k even, if $c_{1,1} = -c_{1,-1}$: $\mu_k = o(k^{-d-2\nu_1})$;*
- *For k odd, if $c_{1,1} \neq c_{1,-1}$: $\mu_k \sim (c_{1,1} - c_{1,-1})C(d,\nu_1)k^{-\tilde{d}-2\nu_1}$;*
- *For k odd, if $c_{1,1} = c_{1,-1}$: $\mu_k = o(k^{-d-2\nu_1})$.*

We reformulate this in terms of $\mathcal{Q}$, see Definition A.1. This perspective allows conveniently handling the $O(t^{\nu_1+1+\varepsilon})$-term and its derivatives. For greater precision, we use the following variant of Lemma B.2 in Haas et al. (2023) to investigate the sign of eigenvalues μ_i:

Lemma D.6 (Guaranteeing strictly positive eigenvalues for kernels on spheres)**.** *Let $\kappa : [-1,1] \to \mathbb{R}$ be continuous, let $d \geq 1$ and consider the radial kernels*

$$k_d : \mathbb{S}^d \times \mathbb{S}^d \to \mathbb{R}, k_d(x,y) := \kappa(\langle x,y\rangle)$$
$$k_{d+2} : \mathbb{S}^{d+2} \times \mathbb{S}^{d+2} \to \mathbb{R}, k_{d+2}(x,y) := \kappa(\langle x,y\rangle).$$

Suppose that k_{d+2} is a kernel. Then, k_d is a kernel. If an eigenvalue $\mu_{\hat{l}}$ of k_d fulfills $\mu_{\hat{l}} > 0$ and $\hat{l}$ is even/odd, then for all even/odd $l \leq \hat{l}$

$$\mu_l > 0 \tag{D.7}$$

follows. Especially, if Eq. (D.7) holds for infinitely many even/odd l, then it holds for all even/odd l.

The only difference between Lemma B.2 in Haas et al. (2023) and the above lemma is that the eigenvalues are described in greater detail here, the proof remains the same.

Theorem A.3 (Adaptation of Theorem 7 in arXiv v4 of Bietti and Bach 2021)**.** *Let $\kappa : [-1,1] \to \mathbb{R}$ be a function that is smooth on $(-1,1)$ such that $k_{\kappa,d}(\boldsymbol{x},\boldsymbol{x}') = \kappa(\langle \boldsymbol{x},\boldsymbol{x}'\rangle)$ is a positive semi-definite kernel on all spheres $\mathbb{S}^d, d \in \mathbb{N}_{\geq 1}$. Suppose that there exists $0 < \beta \in \mathbb{R} \setminus \mathbb{Z}$ and $b_{-1}, b_1 \in \mathbb{R}$ such that for $\tau \in \{-1,1\}$,*

$$\kappa(\tau(1-t)) = b_\tau t^\beta + \mathcal{Q}_{-1,\beta}(t)\ .$$

Then, for a given dimension $d \in \mathbb{N}_{\geq 1}$, the eigenvalues $\mu_l = \mu_l(\kappa, d)$ as defined in Section 2 satisfy:

(a) For $l \in \mathbb{N}_0$ even, if $b_{-1} \neq -b_1$, then $\mu_l = \Theta_{\forall l}((l+1)^{-d-2\beta})$.
(b) For $l \in \mathbb{N}_0$ even, if $b_{-1} = -b_1$, then $\mu_l = o_{\forall l}((l+1)^{-d-2\beta})$.
(c) For $l \in \mathbb{N}_0$ odd, if $b_{-1} \neq b_1$, then $\mu_l = \Theta_{\forall l}((l+1)^{-d-2\beta})$.
(d) For $l \in \mathbb{N}_0$ odd, if $b_{-1} = b_1$, then $\mu_l = o_{\forall l}((l+1)^{-d-2\beta})$.

Proof. Set $\gamma := \beta + 2$. Fix $\tau \in \{-1,1\}$. Then, we can write $\kappa(\tau(1-t)) = b_\tau t^\beta + p(t) + r(t)$ for some $p \in \mathcal{P}_{-1,\beta}$ and $r \in \mathcal{R}_\gamma(t)$. To apply Theorem D.5, we can rewrite $b_\tau t^\beta = p_\tau(t) + \sum_{j=1}^{r} c_{j,\tau} t^{\nu_j}$ with $\nu_1 = \beta$ and $c_{1,\tau} = b_\tau$. The rest term $r(t)$ is covered by the $O(t^{\nu_1+1+\varepsilon})$ term in Theorem D.5 by setting $\varepsilon := \frac{1}{2}\min\{1, \nu_2 - \nu_1\}$, and the expressions for the derivatives in Theorem D.5 are satisfied by the definition of $\mathcal{R}_\gamma$.

Hence, we can apply Theorem D.5 to obtain the correct asymptotic decay rates, and it remains to show that in the cases (a) and (c), all eigenvalues μ_l are strictly positive for all $l \in \mathbb{N}_0$. This claim follows from the following considerations:

- $C(d,\nu_1) \neq 0$: While this is not shown in Bietti and Bach (2021), it follows by considering $\tilde{\kappa}(t) := (1-t)^{\nu_1}$: This choice of κ satisfies the assumptions of Theorem D.5 with $c_{1,1} = 1 \neq 0 = c_{1,-1}$, hence the cases (a) or respectively (c) apply.

Suppose $C(d, \nu_1) = 0$, then we obtain $\mu_k \sim 0$ and hence only finitely many μ_k are nonzero. Following Appendix A in Bietti and Bach (2021), $\tilde{\kappa}$ can be expressed as

$$\tilde{\kappa}(t) = \sum_{l=0}^{\infty} \mu_l N_{d,l} P_l(t) \ ,$$

where the P_l are Legendre polynomials of degree l for the dimension $d+1$. Hence, $\tilde{\kappa}$ would be a polynomial, contradicting the definition $\tilde{\kappa}(t) = (1-t)^{\nu_1}$.

- Since we assumed $k_{\kappa,d}$ to be a (positive semi-definite) kernel, we cannot have negative eigenvalues, hence $(c_{1,1} + c_{1,-1})C(d, \nu_1) > 0$ or respectively $(c_{1,1} - c_{1,-1})C(d, \nu_1) > 0$ follows in the cases (a) and (c).
- The notation $\sim$ from Theorem D.5 allows a finite number of μ_l to be zero. However, as there are infinitely many both even and odd indices for which μ_l is strictly positive, Lemma D.6 shows that $\mu_l > 0$ holds for all $l \in \mathbb{N}_0$.

□

D.4 Eigenvalue decay

Some preliminary constructions useful for both the NNGP and the NTK kernel are first established here, stressing the perspective of layerwise computation in the neural network.

We will study $\kappa_L^{\mathrm{NNGP}}, \kappa_L^{\mathrm{NTK}} : [-1, 1] \to \mathbb{R}$ as compositions of functions which break the kernels given by depth L down to a "composition" of kernels of depth 1.

Recall the recursive form of the NNGP-kernel on the sphere given in Lemma A.5 and the dual activation function $\widehat{\varphi}$ in Definition A.4.

Notation 7 (NNGP- and NTK related terms)**.** Recursively define the functions $g_l, G_l, H_l : [-1, 1] \to [-1, 1]$ by

$$\begin{aligned}
\alpha_l &:= \kappa_{l,\varphi}^{\mathrm{NNGP}}(1), \\
g_1(t) &:= \frac{\sigma_b^2 \sigma_i^2 + \sigma_w^2 t}{\alpha_1}, \\
g_l(t) &:= \frac{\sigma_b^2 \sigma_i^2 + \sigma_w^2 \widehat{\varphi_{\cdot \sqrt{\alpha_{l-1}}}}(t)}{\alpha_l}, \\
G_l(t) &:= g_l \circ \cdots \circ g_1(t), \\
H_1(t) &:= \sigma_b^2 + \sigma_w^2 t, \\
H_l(t) &:= \sigma_b^2(1 - \sigma_i^2) + \alpha_l G_l(t) + \sigma_w^2 \widehat{\varphi'_{\cdot \sqrt{\alpha_{l-1}}}}(G_{l-1}(t)) H_{l-1}(t) \ .
\end{aligned}$$

◀

By the recursive formulation of $\kappa_l^{\mathrm{NNGP}}, \kappa_l^{\mathrm{NTK}}$ in Lemma A.5 we observe for all $t \in [-1, 1]$

$$\begin{aligned}
\alpha_l G_l(t) &= \kappa_l^{\mathrm{NNGP}}(t), \\
H_l(t) &= \kappa_l^{\mathrm{NTK}}(t) \ .
\end{aligned}$$

Lemma D.8. *Under the assumptions of Lemma A.5, we have for any $l \geq 1$*

$$\max_{t \in [-1,1]} |g_l(t)| = g_l(1) = 1 \ .$$

Furthermore,

$$G_l[-1, 1) \subset (-1, 1)$$

holds for all $l \geq 1$ when $\sigma_b^2 \sigma_i^2 > 0$, and for all $l \geq 2$ when φ is neither even nor odd.

Proof. Follows from the definition of g_l and Lemma C.3. □

Lemma D.9 (Behavior at 1)**.** *Let the activation function φ fulfill Assumption 4 and let $m := \inf\{k \in \mathbb{N}_0 \mid \varphi^{(k)}(0+) \neq \varphi^{(k)}(0-)\}$ be its smoothness, see Definition 6. Let $l \geq 2$.*

1. *Let* $m = 0$*. Then we have for* $t \searrow 0$

$$g_l(1-t) = 1 - \Big(c_l t^{1/2} + \mathcal{Q}_{0,1/2}(t)\Big),$$
$$G_l(1-t) = 1 - \Big(C_l t^{2^{1-l}} + \mathcal{Q}_{0,2^{1-l}}(t)\Big),$$

for constants $c_l, C_l > 0$*.*

2. *Let* $m \in \mathbb{N}_{\geq 1}$*. Then we have for* $t \searrow 0$

$$g_l(1-t) = 1 - \Big(b_l t + (-1)^m c_l t^{m+1/2} + \mathcal{Q}_{1,m+1/2}(t)\Big),$$
$$G_l(1-t) = 1 - \Big(B_l t + (-1)^m C_l t^{m+1/2} + \mathcal{Q}_{1,m+1/2}(t)\Big),$$
$$H_l(1-t) = A_l - \Big((-1)^{m+1} D_l t^{m-1/2} + \mathcal{Q}_{0,m-1/2}(t)\Big)$$

for constants $b_l = g_l'(1), c_l, A_l, B_l, C_l, D_l, > 0$*. Furthermore, we have for* $l \geq 3$ *the recursive relation* $C_l \geq g_l'(1) C_{l-1}$*.*

3. *Let* $m = \infty$*. Then we have for any* $\tau \in \{\pm 1\}, q \in \mathbb{R}$ *and* $t \searrow 0$

$$g_l(\tau(1-t)) = \mathcal{Q}_{-1,q}(t),$$
$$G_l(\tau(1-t)) = \mathcal{Q}_{-1,q}(t),$$
$$H_l(\tau(1-t)) = \mathcal{Q}_{-1,q}(t).$$

Proof.

1. The formula for g_l follows from Theorem A.8 and Lemma D.8. We obtain the formula for G_l via induction, where in the base case for $l = 1$ by definition $G_1(1-t) = 1 - c_1 t$ with $c_1 > 0$ holds. For $l = 2$ we have

$$G_2(1-t) = g_2(G_1(1-t)) = g_2(1 - (1 - G_1(1-t))) = 1 - c_2 c_1^{1/2} t^{1/2} + \mathcal{Q}_{0,1/2}(c_1 t)$$

as claimed. For $l \geq 3$, note that the identities $\mathcal{Q}_{0,2^{2-l}}(t) = \mathcal{Q}_{2^{2-l},2^{2-l}}(t)$ and $\mathcal{Q}_{0,1/2}(t) = \mathcal{Q}_{1/2,1/2}(t)$ hold by Definition A.1. From Proposition B.2 (e,f) we obtain

$$\begin{aligned}
G_l(1-t) &= g_l(G_{l-1}(1-t)) = g_l(1 - (1 - G_{l-1}(1-t))) \\
&= 1 - c_l \Big(C_{l-1} t^{2^{2-l}} + \mathcal{Q}_{2^{2-l},2^{2-l}}(t)\Big)^{1/2} + \mathcal{Q}_{1/2,1/2}\Big(C_{l-1} t^{2^{2-l}} + \mathcal{Q}_{2^{2-l},2^{2-l}}(t)\Big) \\
&= 1 - \Big(C_l t^{2^{1-l}} + \mathcal{Q}_{2^{1-l},2^{1-l}}(t)\Big)
\end{aligned}$$

and the claim follows from the identity $\mathcal{Q}_{2^{1-l},2^{1-l}}(t) = \mathcal{Q}_{0,2^{1-l}}(t)$.

2. Define $\beta \coloneqq m + 1/2$. From Theorem A.8 we obtain

$$g_l(1-t) = (-1)^{m+1} c_l t^{\beta} + \mathcal{Q}_{-1,\beta}(t)$$

with $c_l > 0$. Lemma D.8 shows $g_l(0) = 1$ and furthermore we observe

$$\widehat{\varphi \cdot \sqrt{\alpha_{l-1}}}'(1) = \mathbb{E}_{u \sim \mathcal{N}(0,\alpha_{l-1})}[\varphi'(u)^2] > 0$$

as φ is not constant, since $1 \leq m < \infty$. From Lemma D.8 and using $b_l \coloneqq g_l'(1) > 0$, it follows that

$$g_l(1-t) = 1 - \Big(b_l t + (-1)^m c_l t^{m+1/2} + \mathcal{Q}_{1,m+1/2}(t)\Big)\ .$$

Note that for the NNGP term G_l we have to argue more carefully to obtain the strictly positive factor $B_l > 0$ for the linear term, while this is not required for the NTK term H_l where we can rely on the more precisely calculated NNGP term. We show the formula for G_l via induction. For $l = 2$ it follows directly from $G_2 = g_2 \circ g_1$ using $g_1(1-t) = 1 - c_1 t$. In the induction step, let $l \geq 3$ and define

$$\bar{G}_{l-1}(t) \coloneqq 1 - G_{l-1}(t) = B_{l-1} t + (-1)^m C_{l-1} t^{\beta} + \mathcal{Q}_{1,\beta}(t)\ .$$

We rewrite

$$\begin{aligned} G_l(1-t) &= g_l(G_{l-1}(1-t)) = g_l(1-\bar{G}_{l-1}(t)) \\ &= 1 - \left(b_l\bar{G}_{l-1}(t) + (-1)^m c_l\big(\bar{G}_{l-1}(t)\big)^\beta + \mathcal{Q}_{1,\beta}\big(\bar{G}_{l-1}(t)\big)\right) \end{aligned}$$

and investigate the single summands. The linear summand straightforwardly yields

$$b_l\bar{G}_{l-1}(t) = b_lB_{l-1}t + (-1)^m b_lC_{l-1}t^\beta + \mathcal{Q}_{1,\beta}(t).$$

The power summand can be handled with Proposition B.2 (e), where we write $\bar{G}_{l-1} = b_lB_{l-1}t + \mathcal{Q}_{1,1}(t)$ and consequently we have

$$\begin{aligned} (\bar{G}_{l-1}(t))^\beta &= (b_lB_{l-1})^\beta t^\beta + \mathcal{Q}_{\beta,\beta}(t) \\ &= (b_lB_{l-1})^\beta t^\beta + \mathcal{Q}_{1,\beta}(t)\ . \end{aligned}$$

In order to calculate $\mathcal{Q}_{1,\beta}(\bar{G}_{l-1}(t))$ we decompose $\mathcal{Q}_{1,\beta}(t) = p(t) + q(t)$, where $p(t)$ is a polynomial with integer powers greater than 1 and $q(t) \in \mathcal{Q}_{\beta,\beta}(t)$. Using Proposition B.2 (d) we have, as $p(0) = p'(0) = 0$ holds,

$$p(\bar{G}_{L-1}(t)) = p(B_{l-1}t + (-1)^m C_{l-1}t^\beta + \mathcal{Q}_{1,\beta}(t)) = \mathcal{Q}_{1,\beta}(t)\ .$$

Proposition B.2 (f) yields $q(t) = \mathcal{Q}_{\beta,\beta}(t) = \mathcal{Q}_{1,\beta}(t)$ and we obtain $\mathcal{Q}_{1,\beta}\left(\bar{G}_{l-1}(t)\right) = \mathcal{Q}_{1,\beta}(t)$. Altogether, we see that

$$G_l(1-t) = 1 - \big(b_lB_{l-1}t + (-1)^m(b_lC_{l-1}(b_lB_{l-1})^\beta)t^\beta + \mathcal{Q}_{1,\beta}(t)\big)$$

holds as desired, and furthermore we have

$$C_l \coloneqq b_lC_{l-1}(b_lB_{l-1})^\beta \geq b_lC_{l-1} = g_l'(1)C_{l-1}\ .$$

We move on to calculate H_l for $l \geq 2$. An elementary argument yields $\kappa^{\mathrm{NTK}}(1) > 0$ as a consequence of $\sigma_w > 0$, hence we have $A_l > 0$ for all $l \geq 2$. Again, the induction is based on the recursive construction in Notation 7, where we handle the term arising from $\widehat{\varphi'_{\cdot\sqrt{\alpha_l}}}$ using Theorem A.8. Since φ' has smoothness $m-1$ we obtain

$$g_l'(1-t) = \widehat{\varphi'_{\cdot\sqrt{\alpha_l}}}(1-t) = a_l(-1)^m t^{\beta-1} + \mathcal{Q}_{-1,\beta-1}(t)$$

for some $a_l > 0$. For $l = 2$, the claimed form of H_l now follows straightforwardly. For the induction step, assume $l \geq 3$. Then we have

$$\widehat{\varphi'_{\cdot\sqrt{\alpha_{l-1}}}}(G_{l-1}(1-t)) = \bar{c}_l + a_l(-1)^m\big(\bar{G}_{l-1}(t)\big)^{\beta-1} + \mathcal{Q}_{0,\beta-1}\big(\bar{G}_{l-1}(t)\big)$$

where $\bar{c}_l \geq 0$ holds since we have $\widehat{\varphi'_{\cdot\sqrt{\alpha_{l-1}}}}(G_{l-1}(1-0)) = \widehat{\varphi'_{\cdot\sqrt{\alpha_{l-1}}}}(1) \geq 0$. A similar argument as above shows

$$\begin{aligned} \mathcal{Q}_{-1,\beta-1}\big(\bar{G}_{l-1}(t)\big) &= \mathcal{Q}_{-1,\beta-1}(t), \\ \big(\bar{G}_{l-1}(t)\big)^{\beta-1} &= B_{l-1}^{\beta-1}t^{\beta-1} + \mathcal{Q}_{-1,\beta-1}(t)\ . \end{aligned}$$

We define $\bar{D}_l \coloneqq a_lB_{l-1}^{\beta-1}$ and conclude

$$\begin{aligned} &\widehat{\varphi'_{\cdot\sqrt{\alpha_{l-1}}}}(G_{l-1}(1-t))H_{l-1}(1-t) \\ &= \big(\bar{c}_l + (-1)^m\bar{D}_lt^{\beta-1} + \mathcal{Q}_{0,\beta-1}(t)\big)\big(A_{l-1} - \big((-1)^{m+1}D_{l-1}t^{\beta-1} + \mathcal{Q}_{0,\beta-1}(t)\big)\big) \\ &= A_l - \big((-1)^{m+1}D_lt^{\beta-1} + \mathcal{Q}_{0,\beta-1}(t)\big), \end{aligned}$$

where $D_l > 0$, and the claim follows.

3. Choose a natural number $m \geq q - 1/2$ and apply Theorem A.8 to obtain

$$g_l(\tau(1-t)) = \mathcal{Q}_{-1,m+1/2}(t) = \mathcal{Q}_{-1,q}(t).$$

Using Proposition B.2 c), we obtain the claim by induction. □

Theorem D.10. *Let the assumptions of Lemma A.5 be satisfied. Choose a number of layers $L \geq 2$ and a parity $r \in \{0,1\}$ of the eigenvalues to be analyzed. Define $\varphi^{[0]} := \varphi_{\text{even}}$ and $\varphi^{[1]} := \varphi_{\text{odd}}$.*

(NNGP) Define the simplified activation $\tilde{\varphi}$ by

$$\tilde{\varphi} := \begin{cases} \varphi^{[r]} & , \text{ if } \sigma_b^2\sigma_i^2 = 0 \text{ and } (L = 2 \text{ or } \varphi \text{ is even or odd}) \\ \varphi & , \text{ otherwise,} \end{cases}$$

and let $s :=$ smoothness$(\tilde{\varphi})$. Note that in all cases, the kernels k_L^{NNGP} and k_L^{NTK} refer to the original *activation φ, while the smoothness parameter s is determined by the simplified activation $\tilde{\varphi}$.*

(1.1) If $s = 0$, then $\mu_{2t+r,d}(k_L^{\text{NNGP}}) = \Theta_{\forall t}\Big((2t+r+1)^{-d-2^{2-L}}\Big)$.

(1.2) If $1 \leq s < \infty$, then $\mu_{2t+r,d}(k_L^{\text{NNGP}}) = \Theta_{\forall t}\left((2t+r+1)^{-d-2s-1}\right)$.

(1.3) If $s = \infty$ and $\tilde{\varphi}$ is not a polynomial, then $\mu_{2t+r,d}(k_L^{\text{NNGP}}) > 0$ for all t and $\mu_{2t+r,d}(k_L^{\text{NNGP}}) = o_{\forall t}((2t+r+1)^{-q})$ for all $q > 0$.

(1.4) If $\tilde{\varphi}$ is a polynomial, refer to the polynomial case below.

(NTK) Define the simplified activation $\tilde{\varphi}$ by

$$\tilde{\varphi} := \begin{cases} \varphi^{[r]} & , \text{ if } \sigma_b^2 = 0 \text{ and } (L = 2 \text{ or } \varphi \text{ is even or odd}) \\ \varphi & , \text{ otherwise,} \end{cases}$$

and let $s :=$ smoothness$(\tilde{\varphi})$.

(2.1) If $1 \leq s < \infty$, then $\mu_{2t+r,d}(k_L^{\text{NTK}}) = \Theta_{\forall t}\left((2t+r+1)^{-d-2s+1}\right)$.

(2.2) If $s = \infty$ and $\tilde{\varphi}$ is not a polynomial, then $\mu_{2t+r,d}(k_L^{\text{NTK}}) > 0$ for all t and $\mu_{2t+r,d}(k_L^{\text{NTK}}) < o_{\forall t}((2t+r+1)^{-q})$ for all $q > 0$.

(2.3) If $\tilde{\varphi}$ is a polynomial, refer to the polynomial case below.

Polynomial case:

(3.1) Let the activation $\tilde{\varphi}(t) = \sum_{i\geq 0} \lambda_i t^i$ be a nonzero polynomial. Define the even/odd degree of $\tilde{\varphi}$ as the degree of $\tilde{\varphi}(t) + \tilde{\varphi}(-t)$ or of $\tilde{\varphi}(t) - \tilde{\varphi}(-t)$. For the NNGP, define $\sigma := \sigma_b^2\sigma_i^2$ and for the NTK define $\sigma := \sigma_b^2$ respectively.

Define N_{even}, N_{odd} according to the following table. By $\deg_{ev}^{Her}(\tilde{\varphi}), \deg_{od}^{Her}(\tilde{\varphi})$ we denote the even/odd degree, see Definition D.11, of $\tilde{\varphi}$ displayed in the Hermite basis, which is possible as φ fulfills Assumption 4.

	N_{even}	N_{odd}
$\sigma^2 > 0,\ \deg_{ev}^{Her}(\tilde{\varphi}) > \deg_{od}^{Her}(\tilde{\varphi})$	$\deg_{ev}^{Her}(\tilde{\varphi})^{L-1}$	$\deg_{ev}^{Her}(\tilde{\varphi})^{L-1} - 1$
$\sigma^2 > 0,\ \deg_{ev}^{Her}(\tilde{\varphi}) < \deg_{od}^{Her}(\tilde{\varphi})$	$\deg_{od}^{Her}(\tilde{\varphi})^{L-1} - 1$	$\deg_{od}^{Her}(\tilde{\varphi})^{L-1}$
$\sigma^2 = 0, \deg_{ev}^{Her}(\tilde{\varphi}) > \deg_{od}^{Her}(\tilde{\varphi})$	$\deg_{ev}^{Her}(\tilde{\varphi})^{L-1}$	$\deg_{ev}^{Her}(\tilde{\varphi})^{L-1} - \deg_{ev}^{Her}(\tilde{\varphi}) + \deg_{od}^{Her}(\tilde{\varphi})$
$\sigma^2 = 0, \deg_{ev}^{Her}(\tilde{\varphi}) < \deg_{od}^{Her}(\tilde{\varphi})$	$\deg_{od}^{Her}(\tilde{\varphi})^{L-1} - \deg_{od}^{Her}(\tilde{\varphi}) + \deg_{ev}^{Her}(\tilde{\varphi})$	$\deg_{od}^{Her}(\tilde{\varphi})^{L-1}$

Then,

$$\mu_{2t+r,d}(k_L^{\text{NNGP}}/k_L^{\text{NTK}}) > 0 \text{ if and only if } \begin{cases} 2t+r \leq N_{even} & , r \text{ even,} \\ 2t+r \leq N_{odd} & , r \text{ odd.} \end{cases}$$

(3.2) If $\tilde{\varphi} = 0$, then $\mu_{2t+r,d}(k_L^{\text{NNGP}}) > 0$ if and only if $2t+r = 0$ and $\sigma_b^2\sigma_i^2 > 0$, and $\mu_{2t+r,d}(k_L^{\text{NTK}}) > 0$ if and only if $2t+r = 0$ and $\sigma_b^2 > 0$.

The definition of $\tilde{\varphi}$ may look complicated. However, in contrast to φ, it prevents that $c_{1,1} = -(-1)^r c_{1,-1}$ in Theorem A.3, which then would not provide an exact decay rate.

Proof. For a bounded radial kernel $k(x,y) := \kappa(\langle x,y\rangle)$ on $\mathbb{S}^d$ we define $k_{\text{even}}(x,y) := \kappa_{\text{even}}(\langle x,y\rangle), k_{\text{odd}}(x,y) := \kappa_{\text{odd}}(\langle x,y\rangle)$. The spherical harmonics of degree l are even if l is even and odd if l is odd. Hence, we obtain

$$\mu_{2t,d}(k) = \mu_{2t,d}(k_{\text{even}}),$$

$$\mu_{2t+1,d}(k) = \mu_{2t+1,d}(k_{\text{odd}})$$

for all $t \in \mathbb{N}_0$. For the NNGP- and NTK-kernel, Proposition D.4 shows that in the cases where $\tilde{\varphi} \neq \varphi$ holds we have

$$\begin{aligned}\kappa^{\text{NNGP}}_{L,\varphi_{\text{even}}} &= \left(\kappa^{\text{NNGP}}_{L,\varphi}\right)_{\text{even}}, \\ \kappa^{\text{NTK}}_{L,\varphi_{\text{even}}} &= \left(\kappa^{\text{NTK}}_{L,\varphi}\right)_{\text{even}}, \\ \kappa^{\text{NNGP}}_{L,\varphi_{\text{odd}}} &= \left(\kappa^{\text{NNGP}}_{L,\varphi}\right)_{\text{odd}}, \\ \kappa^{\text{NTK}}_{L,\varphi_{\text{odd}}} &= \left(\kappa^{\text{NTK}}_{L,\varphi}\right)_{\text{odd}},\end{aligned}$$

hence we conclude

$$\begin{aligned}\mu_{2t+r}\left(k^{\text{NNGP}}_{L,\tilde{\varphi}}\right) &= \mu_{2t+r}\left(k^{\text{NNGP}}_{L,\varphi}\right), \\ \mu_{2t+r}\left(k^{\text{NTK}}_{L,\tilde{\varphi}}\right) &= \mu_{2t+r}\left(k^{\text{NTK}}_{L,\varphi}\right),\end{aligned}$$

and it suffices to investigate the kernels induced by $\tilde{\varphi}$. In the following, g_l, G_l and H_l from Notation 7 refer to the functions induced by the activation function $\tilde{\varphi}$.

NNGP:

(1.1) Lemma D.9 yields

$$G_L(1-t) = C_{L,+}t^{2^{1-L}} + \mathcal{Q}_{-1,2^{1-L}}(t)$$

for $t \searrow 0$, where $|C_{L,+}| > 0$. By Theorem A.3, it suffices to show

$$\begin{aligned}G_L(-(1-t)) &= C_{L,-}t^{2^{1-L}} + \mathcal{Q}_{-1,2^{1-L}}(t), \\ C_{L,-} &\neq \begin{cases} -C_{L,+} & \text{, if } r = 0 \\ C_{L+} & \text{, if } r = 1\ . \end{cases}\end{aligned} \tag{D.8}$$

(a) Suppose $\sigma_i^2\sigma_b^2 > 0$. By Lemma D.8 we have $g_1([-1,1)) \subset (-1,1)$ and $G_l([-1,1)) \subset (-1,1)$ for $l \geq 2$. As g_1 is a polynomial, $g_1(-(1-t)) = \mathcal{Q}_{-1,2^{1-L}}(t)$ holds. By Lemma C.3 (d) we have $g_l|_{(-1,1)} \in C^\infty((-1,1))$ for all $l \geq 2$, which allows leveraging Proposition B.2 (d) to obtain $G_l(-(1-t)) = g_l(G_{l-1}(-(1-t))) = \mathcal{Q}_{-1,2^{1-L}}(t)$ and Eq. (D.8) follows.

(b) Suppose $\sigma_b^2\sigma_i^2 = 0$, $\tilde{\varphi}$ be neither even nor odd and $L \geq 3$. Then, $g_1(t) = t$ and $G_2(t) = \widehat{\varphi}(t)/\widehat{\varphi}(1)$ hold and from Theorem A.8 we obtain

$$G_2(-(1-t)) = C_{2,-}t^{1/2} + \mathcal{Q}_{-1,1/2}(t).$$

Lemma D.8 yields $G_l([-1,1)) \subset (-1,1)$ for $l \geq 2$ and we furthermore have $g_l|_{(-1,1)} \in C^\infty(-1,1)$ by Lemma C.3 (d). Recursively applying Proposition B.2 (d) now yields

$$G_L(-(1-t)) = g_L\left(G_{L-1}(-(1-t))\right) = Ct^{1/2} + \mathcal{Q}_{-1,1/2}(t) = \mathcal{Q}_{-1,2^{1-L}}(t)$$

and we have Eq. (D.8).

(c) Let $\sigma_b^2\sigma_i^2 = 0$ and let $\tilde{\varphi}$ be odd. Then, $r = 1$ follows by construction. Proposition D.4 shows that G_l is odd for all $l \in \mathbb{N}$ and we obtain

$$G_L(-(1-t)) = -G_L(1-t) = -\left(C_{L,+}t^{2^{1-L}} + \mathcal{Q}_{-1,2^{1-L}}(t)\right)$$

follows. As desired, we have $C_{L,+} \neq C_{L,-} = -C_{L,+}$.

(d) The case $\sigma_b^2\sigma_i^2 = 0$ and $\tilde{\varphi}$ even is not possible since even activation functions $\tilde{\varphi}$ cannot have smoothness $s = 0$.

(1.2) Define $\beta \coloneqq s + 1/2$. Lemma D.9 yields

$$G_L(1-t) = C_{L,+}t^\beta + \mathcal{Q}_{-1,\beta}(t)$$

for $t \searrow 0$, where $|C_{L,+}| > 0$. By Theorem A.3 it suffices to show

$$\begin{aligned}G_L(-(1-t)) &= C_{L,-}t^\beta + \mathcal{Q}_{-1,\beta}(t)\ , \\ C_{L,-} &\neq \begin{cases} -C_{L,+} & \text{, if } r = 0 \\ C_{L+} & \text{, if } r = 1\ . \end{cases}\end{aligned} \tag{D.9}$$

(a) Let $\sigma_b^2\sigma_i^2 > 0$. By Lemma D.8 we have $g_1([-1,1)) \subset (-1,1)$ and $G_i([-1,1)) \subset (-1,1)$ for $i \geq 2$. As g_1 is a polynomial, $g_1(-(1-t)) = \mathcal{Q}_{-1,\beta}(t)$ holds. Lemma C.3 (d) shows $g_i|_{(-1,1)} \in C^\infty((-1,1))$ for all $i \geq 2$ which allows leveraging Proposition B.2 (d) to obtain $G_L(-(1-t)) = g_L(G_{L-1}(-(1-t))) = \mathcal{Q}_{-1,\beta}(t)$ by induction. Eq. (D.9) follows.

(b) Let $\sigma_b^2\sigma_i^2 = 0$, $\tilde{\varphi}$ be neither even nor odd and $L \geq 3$, that is, $\tilde{\varphi} = \varphi$. Then, $g_1(t) = t$ and $G_2(t) = \widehat{\varphi}(t)/\widehat{\varphi}(1)$ hold and from Theorem A.8 we obtain

$$G_2(1-t) = C_{2,+}t^\beta + \mathcal{Q}_{-1,\beta}(t),$$
$$G_2(-(1-t)) = C_{2,-}t^\beta + \mathcal{Q}_{-1,\beta}(t),$$

where $|C_{2,+}| = |C_{2,-}| \neq 0$. Recursively applying Lemma D.9 furthermore yields

$$|C_{L,+}| \geq g'_L(1) \cdot \ldots \cdot g'_3(1) \cdot |C_{2,+}|.$$

Lemma D.8 yields $G_l([-1,1)) \subset (-1,1)$ for $l \geq 2$ and we have $g_l|_{(-1,1)} \in C^\infty(-1,1)$ by Lemma C.3. Iteratively applying Proposition B.2 (d) yields

$$G_L(-(1-t)) = g'_L(G_{L-1}(-1)) \cdot \ldots \cdot g'_3(G_2(-1)) \cdot C_{2,-}t^\beta + \mathcal{Q}_{-1,\beta}(t).$$

Note that we have $g'_l(t) = \widehat{\varphi'_{\cdot\sqrt{\alpha_{l-1}}}(t)}/\alpha_l$ and by Lemma C.3 (c) we have

$$|g'_l(t)| < |g'_l(1)|$$

for all $t \in (-1,1)$. Altogether we conclude

$$G_L(1-t) = C_{L,+}t^\beta + \mathcal{Q}_{-1,\beta}(t),$$
$$G_L(-(1-t)) = C_{L,-}t^\beta + \mathcal{Q}_{-1,\beta}(t),$$

where $|C_{L,+}| > |C_{L,-}|$ holds and we obtain Eq. (D.9).

(c) Let $\sigma_b^2\sigma_i^2 = 0$ and let $\tilde{\varphi}$ be odd, so we have $r = 1$. By Proposition D.4, G_l is odd for all $l \in \mathbb{N}$ and

$$G_L(-(1-t)) = -G_L(1-t) = -\big(C_{L,+}t^\beta + \mathcal{Q}_{-1,\beta}(t)\big)$$

follows. We obtain Eq. (D.9) as $C_{L,+} \neq C_{L,-} = -C_{L,+}$ holds.

(d) In the case $\sigma_b^2\sigma_i^2 = 0$ and $\tilde{\varphi}$ even Proposition D.4 yields G_l even and we argue as in (c).

(1.3) For $q > 0$ choose $\tilde{q} > 0, \tilde{q} \notin \mathbb{Z}$ such that $-d - 2\tilde{q} \leq -q$. By Lemma D.9 we have for $\tau \in \{1,-1\}$:

$$G_L(\tau(1-t)) = \mathcal{Q}_{-1,\tilde{q}}(t)$$

and Theorem A.8 yields $\mu_{2t+r,d}(k_L^{\mathrm{NNGP}}) = o_{\forall t}((2t+r+1)^{-d-2\tilde{q}}) = o_{\forall t}((2t+r+1)^{-q})$.
We proceed to show that all Eigenvalues $\mu_{2t+1,d}(k_L^{\mathrm{NNGP}})$ are strictly positive. As $\tilde{\varphi}$ is not a polynomial, its Hermite series representation has infinitely many nonzero coefficients; that is, we have $\deg^{\mathrm{Her}}(\tilde{\varphi}) = \infty$. See Appendix D.5 for more details. We apply Lemma D.13.
Assume even parity $r = 0$. In the case $\deg_{\mathrm{ev}}(\tilde{\varphi}) = \infty$, the claim follows directly. In the case $\deg_{\mathrm{ev}}(\tilde{\varphi}) < \deg_{\mathrm{od}}(\tilde{\varphi}) = \infty$, we note that $\tilde{\varphi} = \varphi$ follows and investigate sub-cases. If $\sigma_b^2\sigma_i^2 > 0$, then Lemma D.13 directly yields $\deg_{\mathrm{ev}}(G_l) = \infty$ as desired. If $\sigma_b^2\sigma_i^2 = 0$ holds, then the definition of $\tilde{\varphi}$ implies that $\deg_{\mathrm{ev}}(\tilde{\varphi}) \geq 0$ and $L \geq 3$ hold, and again the claim follows directly from Lemma D.13. The case of odd parity $r = 1$ is proven analogously.

NTK:

The argumentation for the NTK-kernel is more technical as for the NNGP-kernel, as by Notation 7 H_l follows the recursion

$$H_1(t) = \sigma_b^2 + \sigma_w^2 t \ ,$$
$$H_l(t) = \sigma_b^2(1-\sigma_i^2) + \alpha_l G_l(t) + \sigma_w^2 \widehat{\varphi'_{\cdot\sqrt{\alpha_{l-1}}}}(G_{l-1}(t))H_{l-1}(t) \ , \quad l \geq 2 \ , \tag{D.10}$$

which involves more terms in a complicated fashion.

(2.1) Define $\gamma := s - 1/2$. Lemma D.9 yields

$$H_L(1-t) = C_{L,+}t^\gamma + \mathcal{Q}_{-1,\gamma}(t)$$

for $t \searrow 0$, where $|C_{L,+}| > 0$. By Theorem A.3 we need to show

$$\begin{aligned} H_L(-(1-t)) &= C_{L,-}t^\gamma + \mathcal{Q}_{-1,\gamma}(t) \ , \\ C_{L,-} &\neq \begin{cases} -C_{L,+} & \text{, if } r = 0 \\ C_{L+} & \text{, if } r = 1 \ . \end{cases} \end{aligned} \tag{D.11}$$

In case (1.2) we saw for the NNGP-kernel that

$$G_l(\tau(1-t)) = c_{l,\tau}t^{\gamma+1} + \mathcal{Q}_{-1,\gamma+1}(t) = \mathcal{Q}_{-1,\gamma}(t) \tag{D.12}$$

holds for $\tau = \pm 1,\ l \geq 1$. Hence, it suffices to show

$$\begin{aligned} \sigma_w^2 \widehat{\varphi'_{\cdot\sqrt{\alpha_{L-1}}}}(G_{L-1}(-(1-t)))H_{L-1}(-(1-t)) &= C_{L,-}t^\gamma + \mathcal{Q}_{-1,\gamma}(t) \ , \\ C_{L,-} &\neq \begin{cases} -C_{L,+} & \text{, if } r = 0 \\ C_{L+} & \text{, if } r = 1 \ . \end{cases} \end{aligned} \tag{D.13}$$

in order to obtain Eq. (D.11).

(a) Let $\sigma_b^2 > 0$. This is the most difficult sub-case. Without loss of generality assume $\sigma_w^2 = 1$. When analyzing $H_l(\tau(1-t))$, Eq. (D.10) the constants $C_{l,\tau}$ of t^γ stem from the summand

$$\widehat{\varphi'_{\cdot\sqrt{\alpha_{l-1}}}}(G_{l-1}(\tau(1-t)))H_{l-1}(\tau(1-t)) = C_{l,\tau}t^\gamma + \mathcal{Q}_{-1,\gamma}(t) \tag{D.14}$$

as the other summands in Eq. (D.10) are in $\mathcal{Q}_{-1,\gamma}(t)$ by Eq. (D.12).
We show by induction in $l \geq 2$ that the following statements hold:

- $\operatorname{sgn} C_{l,+} = (-1)^{s+1}$,
- $|C_{l,+}| > |C_{l,-}|$.

The first point is a technical tool for the induction, ensuring that the coefficients of t^γ for $\tau = 1$ suffer no annihilation because of different signs of its summands: The coefficients of t^γ are the sum of the product of the t^γ-coefficient of one factor with the constant coefficient of the other factor in Eq. (D.14). The second point then proves this sub-case of the theorem.
By Theorem A.8 we have

$$\widehat{\varphi'_{\cdot\sqrt{\alpha_l}}}(\tau(1-t)) = (-\tau)^{s+1}d_l t^\gamma + \mathcal{Q}_{-1,\gamma}(t) \ , \tag{D.15}$$

where $d_l > 0$, and furthermore $\widehat{\varphi'_{\cdot\sqrt{\alpha_l}}}(\cdot)$ is smooth on $(-1,1)$ by Lemma C.3 (d).
For the induction base $l = 2$ we have $G_1(\tau(1-t)) = \tau(1-t), H_1(\tau(1-t)) = \sigma_b^2 + \tau - \tau t$ which yields

$$\begin{aligned} \widehat{\varphi'_{\cdot\sqrt{\alpha_1}}}(G_1(\tau(1-t)))H_1(\tau(1-t)) &= ((-\tau)^{s+1}d_2 t^\gamma + \mathcal{Q}_{-1,\gamma}(t))(\sigma_b^2 + \tau - \tau t) \\ &= (-\tau)^{s+1}d_2(\sigma_b^2 + \tau)t^\gamma + \mathcal{Q}_{-1,\gamma}(t) \ , \end{aligned}$$

i.e. $C_{2,+} = (-1)^{s+1}d_2(\sigma_b^2 + 1)$ and $C_{2,-} = d_2(\sigma_b^2 - 1)$ and the claim holds for $l = 2$ since $\sigma_b^2 > 0$.
For the induction step, let the claim hold for $l \geq 2$. First we show the auxiliary statement

$$H_r(1) > |H_r(t)| \quad , t \in [-1,1),\ r \in \mathbb{N} \ , \tag{D.16}$$

which is in the spirit of Lemma D.8, by induction in $r \geq 1$. In the base case $r = 1$, we have $H_1(t) = \sigma_b^2 + t$ and the claim holds. In the induction step, let the claim hold for $r \geq 1$. We use Notation 7 and obtain

$$H_{r+1}(t) = \sigma_b^2 + \widehat{\varphi_{\cdot\sqrt{\alpha_r}}}(G_r(t)) + \widehat{\varphi'_{\cdot\sqrt{\alpha_r}}}(G_r(t))H_r(t) \ .$$

Now, for any $t \in [-1,1]$ we have $G_r(1) = 1, G_r(t) \in [-1,1]$ by Lemma D.8 and by Lemma C.3 we obtain

$$\widehat{\varphi_{\cdot\sqrt{\alpha_r}}}(1) \geq \left|\widehat{\varphi_{\cdot\sqrt{\alpha_r}}}(t)\right|$$

$$\widehat{\varphi'_{\cdot\sqrt{\alpha_r}}}(1) \geq \left|\widehat{\varphi'_{\cdot\sqrt{\alpha_r}}}(t)\right| \tag{D.17}$$

and Eq. (D.16) follows.
By a Taylor expansion we have

$$G_l(\tau(1-t)) = G_l(\tau) - \tau t G_l'(\tau) + \mathcal{Q}_{1,\gamma}(t) \tag{D.18}$$

as $G_l(\tau(1-t)) = \mathcal{Q}_{-1,\gamma}(t)$ by Eq. (D.12) and furthermore $G_l(t) \in [-1,1], G_l(1) = 1$ by Lemma D.8.
For $\tau = 1$ we have by Eq. (D.18) and Eq. (D.15)

$$\widehat{\varphi'_{\cdot\sqrt{\alpha_l}}}(G_l(1-t)) = (-1)^{s+1} d_l G_l'(1) t^\gamma + \mathcal{Q}_{-1,\gamma}(t) \ .$$

$$\widehat{\varphi'_{\cdot\sqrt{\alpha_l}}}(G_l(1-t)) H_l(1-t) = \underbrace{\left((-1)^{s+1} d_l G_l'(1) H_l(1) + \widehat{\varphi'_{\cdot\sqrt{\alpha_l}}}(G_l(1)) C_{l,+}\right)}_{=C_{l+1,+}} t^\gamma + \mathcal{Q}_{-1,\gamma}(t)$$

where furthermore

$$G_l'(1) \geq |G_l'(t)| \quad G_l'(1) > 0 \tag{D.19}$$

hold for all $t \in [-1,1]$ by Lemma C.3 and Notation 7. We observe that $\operatorname{sgn} C_{l,+} = (-1)^{s+1}$ holds as claimed, that is, $C_{l+1,+}$ suffers no annihilation from different signs of its summands.
For $\tau = -1$ the possible cases are $G_l(-1) \in (-1,1)$, for which we obtain $\widehat{\varphi'_{\cdot\sqrt{\alpha_l}}}(G_l(-(1-t))) = \mathcal{Q}_{-1,\gamma}(t)$ by Proposition B.2 (d) and Eq. (D.12), and $|G_l(-1)| = 1$, for which we obtain

$$\widehat{\varphi'_{\cdot\sqrt{\alpha_l}}}(G_l(-(1-t))) = \tilde{a}_{l,\tau} d_l G_l'(-1) t^\gamma + \mathcal{Q}_{-1,\gamma}(t) \ ,$$

$\tilde{a}_{l,\tau} = \pm 1$ and $|G_l'(-1)| \leq G_l'(1)$ by Eq. (D.19). Hence, we have

$$\begin{aligned} \left|\widehat{\varphi'_{\cdot\sqrt{\alpha_l}}}(G_l(-(1-t)))\right| &= a_{l,-} t^\gamma + \mathcal{Q}_{-1,\gamma}(t) \\ |a_{l,-}| &\leq |d_l G_l'(1)| \end{aligned} \tag{D.20}$$

This yields

$$\widehat{\varphi'_{\cdot\sqrt{\alpha_l}}}(G_l(-(1-t))) H_l(-(1-t)) = \underbrace{\left(a_{l,-} H_l(-1) + \widehat{\varphi'_{\cdot\sqrt{\alpha_l}}}(G_l(-1)) C_{l,-}\right)}_{=C_{l+1,-}} t^\gamma + \mathcal{Q}_{-1,\gamma}(t)$$

and comparing the coefficients $C_{l+1,+}$ and $C_{l+1,-}$ we finally obtain

$$\begin{aligned} |C_{l+1,+}| &= |d_l G_l'(1) H_l(1)| + \left|\widehat{\varphi'_{\cdot\sqrt{\alpha_l}}}(G_l(1)) C_{l,+}\right| \\ &> |a_{l,-} H_l(-1)| + \left|\widehat{\varphi'_{\cdot\sqrt{\alpha_l}}}(G_l(-1)) C_{l,-}\right| \\ &\geq |C_{l+1,-}| \end{aligned}$$

using Eq. (D.16) and Eq. (D.20) to compare the first summand and the induction hypothesis and Eq. (D.17) for the second summand.

(b) Suppose $\sigma_b^2 = 0$, $\tilde{\varphi}$ neither even nor odd and $L \geq 3$, that is $\tilde{\varphi} = \varphi$.
Without loss of generality assume $\sigma_w^2 = 1$ for simpler notation. We show $|C_{2,+}| \geq |C_{2,-}|$, for $l \geq 3$ we show that $|C_{l-1,+}| \geq |C_{l-1,-}|$ implies $|C_{l,+}| > |C_{l,-}|$. That yields Eq. (D.11). Furthermore we show that $\operatorname{sgn}(C_{l,+}) = (-1)^s$ holds, which is required for the corresponding recursion.
As $G_1(t) = t$, we have for $\tau = \pm 1$ by Theorem A.8

$$\widehat{\varphi'_{\cdot\sqrt{\alpha_{l-1}}}}(G_1(\tau(1-t))) = (-\tau)^s C_2 t^\gamma + \mathcal{Q}_{-1,\gamma}(t) \ , \quad t \searrow 0 \ ,$$

where $C_2 > 0$. Now, since we have $H_1(\tau(1-t)) = \tau(1-t)$ we obtain from Eq. (D.10)

$$H_2(\tau(1-t)) = \tau(-\tau)^s C_2 t^\gamma + \mathcal{Q}_{-1,\gamma}(t)$$

and the claim holds for $l = 2$. Now, we take a close look at $C_{l,+}$ for $l \geq 3$. Recalling $G_l(1-t) = G_{l-1}(0) - tG'_{l-1}(1) + \mathcal{Q}_{1,\gamma}(t)$ from the considerations done for the NNGP above, we see that for

$$H_l(1-t) = \sigma_b^2(1-\sigma_i^2) + \alpha_l G_l(t) + \widehat{\varphi'_{\cdot\sqrt{\alpha_{l-1}}}}(G_{l-1}(1-t))H_{l-1}(1-t)$$

the terms contributing to the coefficient of t^γ only stem from the product of $\widehat{\varphi'_{\cdot\sqrt{\alpha_{l-1}}}}(G_{l-1}(1-t))$ and $H_{l-1}(1-t)$ and by Theorem A.8 we have

$$\widehat{\varphi'_{\cdot\sqrt{\alpha_{l-1}}}}(G_{l-1}(1-t)) = (-1)^s G'_{l-1}(1)c_l t^\gamma + \widehat{\varphi'_{\cdot\sqrt{\alpha_{l-1}}}}(1) + \mathcal{Q}_{0,\gamma}(t)$$

for a constant $c_l > 0$. As $G'_{l-1}(1) > 0$ by Lemma D.8, the sign of the coefficient is $(-1)^s$. Together with

$$H_{l-1}(1-t) = H_{l-1}(1) + C_{l-1,+}t^\gamma + \mathcal{Q}_{0,\gamma}(t)$$

we obtain

$$\begin{aligned}
&\widehat{\varphi'_{\cdot\sqrt{\alpha_{l-1}}}}(G_{l-1}(1-t))H_{l-1}(1-t) \\
&= \left((-1)^s G'_{l-1}(1)c_l t^\gamma + \widehat{\varphi'_{\cdot\sqrt{\alpha_{l-1}}}}(1) + \mathcal{Q}_{0,\gamma}(t)\right) \cdot (H_{l-1}(1) + C_{l-1,+}t^\gamma + \mathcal{Q}_{0,\gamma}(t)) \\
&= \underbrace{\left((-1)^s G'_{l-1}(1)c_l H_{l-1}(1) + \widehat{\varphi'_{\cdot\sqrt{\alpha_{l-1}}}}(1)C_{l-1,+}\right)}_{=C_{l,+}} t^\gamma + \mathcal{Q}_{-1,\gamma}(t) \ .
\end{aligned}$$

We observe that as $G'_{l-1}(1), c_l, H_{l-1}(1), \widehat{\varphi'_{\cdot\sqrt{\alpha_{l-1}}}}(1) > 0$ holds we indeed have $\operatorname{sgn}(C_{l,+}) = (-1)^s$ and observe

$$|C_{l,+}| > \widehat{\varphi'_{\cdot\sqrt{\alpha_{l-1}}}}(1)|C_{l-1,+}| \ . \tag{D.21}$$

Investigating $C_{l,-}$ we note that as $G_{l-1}([-1,1)) \subseteq (-1,1)$ holds by Lemma D.8, Lemma C.3 (d) yields $\widehat{\varphi'_{\cdot\sqrt{\alpha_{l-1}}}}|_{(-1,1)} \in C^\infty((-1,1))$ and by Proposition B.2 (e) we have

$$\widehat{\varphi'_{\cdot\sqrt{\alpha_{l-1}}}}(G_{l-1}(-(1-t))) = \widehat{\varphi'_{\cdot\sqrt{\alpha_{l-1}}}}(G_{l-1}(-1)) + \mathcal{Q}_{0,\gamma}(t) \ ,$$

since $G_{l-1}(-(1-t)) = \mathcal{Q}_{-1,\gamma}(t)$ holds. As before, coefficients of t^γ can only stem from the product of $\widehat{\varphi'_{\cdot\sqrt{\alpha_{l-1}}}}(G_{l-1}(-(1-t)))$ and $H_{l-1}(-(1-t))$ and where we have

$$\begin{aligned}
&\widehat{\varphi'_{\cdot\sqrt{\alpha_{l-1}}}}(G_{l-1}(-(1-t)))H_{l-1}(-(1-t)) \\
&= \left(\widehat{\varphi'_{\cdot\sqrt{\alpha_{l-1}}}}(G_{l-1}(-1)) + \mathcal{Q}_{0,\gamma}(t)\right) \cdot (H_{l-1}(1) + C_{l-1,-}t^\gamma + \mathcal{Q}_{0,\gamma}(t)) \\
&= \underbrace{\widehat{\varphi'_{\cdot\sqrt{\alpha_{l-1}}}}(G_{l-1}(-1))C_{l-1,-}}_{=C_{l,-}} t^\gamma + \mathcal{Q}_{-1,\gamma}(t) \ .
\end{aligned}$$

Lemma C.3 (c) yields

$$\widehat{\varphi'_{\cdot\sqrt{\alpha_{l-1}}}}(G_{l-1}(-1)) < \widehat{\varphi'_{\cdot\sqrt{\alpha_{l-1}}}}(G_{l-1}(1))$$

as $G_{l-1}(-1) \in (-1,1)$ and we obtain from Eq. (D.21) as desired

$$\begin{aligned}
|C_{l,-}| &= \left|\widehat{\varphi'_{\cdot\sqrt{\alpha_{l-1}}}}(G_{l-1}(-1))C_{l-1,-}\right| < \left|\widehat{\varphi'_{\cdot\sqrt{\alpha_{l-1}}}}(G_{l-1}(1))C_{l-1,+}\right| \\
&< |C_{l,+}| \ .
\end{aligned}$$

(c) Let $\sigma_b^2 = 0$ and let $\tilde{\varphi}$ be odd, so we have $r = 1$. By Proposition D.4 H_l is odd for all $l \in \mathbb{N}$ and

$$H_L(-(1-t)) = -H_L(1-t) = -(C_{L,+}t^\gamma + \mathcal{Q}_{-1,\gamma}(t))$$

follows. We obtain Eq. (D.11) as $C_{L,+} \neq C_{L,-} = -C_{L,+}$ holds.

(d) In the case $\sigma_b^2 = 0$ and $\tilde{\varphi}$ even Proposition D.4 yields H_L even and we argue as above.

(2.2), (2.3) The arguments work just as in the NNGP cases (1.3) or respectively (1.4).

Polynomial case:

(3.1) This is the statement of Lemma D.13.

(3.2) Elementary. □

Finally, we are able to prove Theorem 9.

Proof of Theorem 9. Lemma D.2 yields the convergence of the NNGP- and NTK-kernels. Theorem D.10 combined with Lemma 2 and a case distinction then yields the claim. The sub-cases i) and ii) are straightforward, in the sub-cases iii) we use Theorem D.10 with the parity $r = 0$ to obtain the eigenvalues of even index and Theorem D.10 with the parity $r = 1$ to obtain the eigenvalues of odd index, and then combine these results to obtain Eqs. (1) and (2), where we use that the spherical harmonics of even/odd degree are even/odd. □

D.5 A close look at the even/odd degree of G_l and H_l

In this subsection we develop the methods required to deal with smooth activations in Theorem 9, where the only difficulty not covered by Theorem A.3 is investigating which eigenvalues of the NNGP-kernel and NTK-kernel are strictly positive.

Definition D.11. Let $p(t) = \sum_{n\geq 0} \lambda_n t^n, q(t) = \sum_{n\geq 0} \eta_n t^n$ be *power series* which converge on an interval $I \subset \mathbb{R}$. We define the degree of p as $\deg(p) := \sup(\{n \in \mathbb{N}_0 \mid \lambda_n \neq 0\} \cup \{-\infty\}) \in \mathbb{N}_0 \cup \{-\infty, \infty\}$ and the even/odd degree of p as

$$\deg_{\text{ev}}(p) := \deg(t \mapsto p(t) + p(-t)) \ ,$$
$$\deg_{\text{od}}(p) := \deg(t \mapsto p(t) - p(-t))$$

and say

$$p =_{\deg} q$$

if and only if $\deg_{\text{ev}}(p) = \deg_{\text{ev}}(q)$ and $\deg_{\text{od}}(p) = \deg_{\text{od}}(q)$. That is to say, we compare only the highest even/odd degrees for this notion of equality.

Let $f \in \mathcal{L}_2(\mathcal{N}(0,1))$ be a function. Then f equals its convergent *Hermite* representation $t \mapsto \sum_{n\geq 0} a_n(f)h_n(t)$ almost everywhere, and we define $\deg^{\text{Her}}(f), \deg^{\text{Her}}_{\text{ev}}(f), \deg^{\text{Her}}_{\text{od}}(f)$ and $f =^{\text{Her}}_{\deg} g$ based on the Hermite coefficients $(a_n)_{n\geq 0}$ analogously to the above definitions.

Furthermore we also use the notation $f =^{\text{Pow,Her}}_{\deg} g$ to denote that the even/odd degree of f and g coincide when f is a power series and and g is displayed in the Hermite basis. ◀

Note that $\deg_{\text{ev}}(p) = -\infty$ if p is odd and respectively $\deg_{\text{od}}(p) = -\infty$ if p is even, resembling the common convention for the degree of polynomials.

Lemma D.12 (Odd/even degree of power series with non-negative coefficients)**.** *Let $f, g, \tilde{f}, \tilde{g}$ be power series with non-negative coefficients converging on an interval I, such that the products and compositions below are absolutely convergent on $\bar{I}$. We use the conventions $\infty - \infty := -\infty, 0 \cdot \infty := 0, a \vee b := \max\{a, b\}$.*

1. *(a) The sum $g + f$ fulfills*

$$\deg_{ev}(g+f) = \deg_{ev}(g) \vee \deg_{ev}(f) \,,$$
$$\deg_{od}(g+f) = \deg_{od}(g) \vee \deg_{od}(f) \ .$$

(b) The product $g \cdot f$ fulfills

$$\deg_{ev}(g \cdot f) = (\deg_{ev}(g) + \deg_{ev}(f)) \vee (\deg_{od}(g) + \deg_{od}(f)) \,,$$
$$\deg_{od}(g \cdot f) = (\deg_{ev}(g) + \deg_{od}(f)) \vee (\deg_{od}(g) + \deg_{ev}(f)) \ .$$

(c) Let $f \neq 0$. The composition $g \circ f$ fulfills

$$\begin{aligned}\deg_{ev}(g \circ f) &= (\deg_{ev}(g)\deg_{ev}(f)) \vee (\deg_{ev}(g)\deg_{od}(f)) \\ &\vee\ (\deg_{od}(g)\deg_{ev}(f)) \vee ((\deg_{od}(g)-1)\deg_{od}(f)+\deg_{ev}(f)) \\ \deg_{od}(g \circ f) &= (\deg_{od}(g)\deg_{od}(f)) \vee ((\deg_{od}(g)-1)\deg_{ev}(f)+\deg_{od}(f)) \\ &\vee\ ((\deg_{ev}(g)-1)\deg_{od}(f)+\deg_{ev}(f)) \\ &\vee\ ((\deg_{ev}(g)-1)\deg_{ev}(f)+\deg_{od}(f))\ .\end{aligned}$$

2. Let $\tilde{f} =_{\deg} f, \tilde{g} =_{\deg} g$. Then we have

$$\begin{aligned} g + f &=_{\deg} \tilde{g} + \tilde{f}\ , \\ g \cdot f &=_{\deg} \tilde{g} \cdot \tilde{f}\ , \\ g \circ f &=_{\deg} \tilde{g} \circ \tilde{f}\ . \end{aligned}$$

Proof. 1. The requirement that f, g have nonzero coefficients is crucial as it prevents the elimination of coefficients.

We only prove the even-statement in iii) as this is the most difficult statement. The other claims are similar but easier to prove.

As a consequence of the absolute convergence we observe that $g \circ f = \lim_{n\to\infty} g_n \circ f_n$ and $\deg_{\text{ev}}(g \circ f) = \lim_{n\to\infty} \deg_{\text{ev}}(g_n \circ f_n), \deg_{\text{od}}(g \circ f) = \lim_{n\to\infty} \deg_{\text{od}}(g_n \circ f_n)$ holds, where $f_n(t) = \sum_{i=0}^n \lambda_i t^i, g_n(t) = \sum_{j=0}^n \mu_j t^j$. Hence, we can assume without loss of generality that $\deg f, \deg g < \infty$ holds and obtain

$$\begin{aligned} g \circ f(t) &= \sum_{i\geq 0} \mu_i \Big(\sum_{j\geq 0} \lambda_j t^j \Big)^i = \sum_{i \geq 0} \mu_i \Big(\sum_{j_1,\ldots,j_i \geq 0} \lambda_{j_1} \cdot \ldots \cdot \lambda_{j_i} t^{j_1+\cdots+j_i} \Big) \\ &= \sum_{i\geq 0} \sum_{j_1,\ldots,j_i\geq 0} \mu_i \lambda_{j_1} \cdot \ldots \cdot \lambda_{j_i} t^{j_1+\cdots+j_i}\ . \end{aligned}$$

We note that all occurring coefficients are non-negative, as they are products of the non-negative coefficients λ_j, μ_i. In the above expressions, there are multiple summands contributing to a power t^r, however, the coefficient of t^r in the power series $g \circ f$ is nonzero if and only if any of the summands $\mu_i \lambda_{j_1} \cdot \ldots \cdot \lambda_{j_i}$ with $j_1 + \cdots + j_i = r$ is nonzero. Define

$$\begin{aligned} I(f) &\coloneqq \{i \in \mathbb{N}_0 \mid \lambda_i \neq 0\} \cup \{-\infty\}, \\ I(g) &\coloneqq \{i \in \mathbb{N}_0 \mid \mu_i \neq 0\} \cup \{-\infty\} \end{aligned}$$

and it is immediately clear that the coefficient $\mu_i \lambda_{j_1} \cdot \ldots \cdot \lambda_{j_i}$ is nonzero if and only if $i \in I(g) \setminus \{-\infty\}, j_1, \ldots, j_i \in I(f) \setminus \{-\infty\}$. The reason we add $\{-\infty\}$ to $I(f), I(g)$ is that it allows dealing with cases such as $f = 0$ or f odd in a simple manner. In order to investigate $\deg_{\text{ev}}(g \circ f)$ we stepwise simplify:

$$\begin{aligned} \deg_{\text{ev}}(g \circ f) &= \max_{i \in I(g)} \max_{\substack{j_1,\ldots,j_i \in I(f) \\ j_1+\cdots+j_i \text{ even}}} j_1 + \cdots + j_i \\ &= \max_{i\in I(g)} \max_{\substack{\tilde{i} \leq i \\ \tilde{i} \text{ even}}} (i - \tilde{i}) \deg_{\text{ev}}(f) + \tilde{i} \deg_{\text{od}}(f) && \text{(D.22)} \\ &= \max\Big\{ \max_{\substack{\tilde{i} \leq \deg_{\text{ev}}(g) \\ \tilde{i} \text{ even}}} (\deg_{\text{ev}}(g) - \tilde{i}) \deg_{\text{ev}}(f) + \tilde{i} \deg_{\text{od}}(f), && \text{(D.23)} \\ &\qquad \max_{\substack{\tilde{i} \leq \deg_{\text{od}}(g) \\ \tilde{i} \text{ even}}} (\deg_{\text{od}}(g) - \tilde{i}) \deg_{\text{ev}}(f) + \tilde{i} \deg_{\text{od}}(f) \Big\} \\ &= \max\Big\{ \max\{ \deg_{\text{ev}}(g) \deg_{\text{ev}}(f), \deg_{\text{ev}}(g) \deg_{\text{od}}(f)\}, && \text{(D.24)} \\ &\qquad \max\{ \deg_{\text{od}}(g) \deg_{\text{ev}}(f), \deg_{\text{ev}}(f) + (\deg_{\text{od}}(g) - 1) \deg_{\text{od}}(f)\}\Big\}\ . \end{aligned}$$

Here, Eq. (D.22) is obtained by observing

$$j_1 + \cdots + j_i \leq \deg_{\text{ev}}(f) + j_2 + \cdots + j_i \quad \text{, if } j_1 \text{ is even or respectively}$$

$$j_1 + \cdots + j_i \le \deg_{\text{od}}(f) + j_2 + \cdots + j_i \quad \text{, if } j_1 \text{ is odd ,}$$

where the sum on the right hand side remains even. Similarly, Eq. (D.23) follows by observing for $m \in \mathbb{N}_0$ the inequality

$$(i - \tilde{i}) \deg_{\text{ev}}(f) + \tilde{i} \deg_{\text{od}}(f) \le \max \big\{ (2m + i - (2m + \tilde{i})) \deg_{\text{ev}}(f) + (2m + \tilde{i}) \deg_{\text{od}}(f), \\ (2m + i - \tilde{i}) \deg_{\text{ev}}(f) + \tilde{i} \deg_{\text{od}}(f) \big\}$$

as by assumption $\deg_{\text{ev}}(f) \ge 0$ or $\deg_{\text{od}}(f) \ge 1$ hold. Finally, Eq. (D.24) follows by observing that extremal behavior in Eq. (D.23) is obtained by the highest or the lowest valid $\tilde{i}$. The equation for $\deg_{\text{od}}(g \circ f)$ follows analogously.

2. Follows from 1. □

In the following lemma, we investigate the even/odd degree of κ_L^{NNGP} and κ_L^{NTK} and then deploy Lemma D.6 to characterize which eigenvalues of $k_L^{\text{NNGP}}, k_L^{\text{NTK}}$ are strictly positive.

Lemma D.13. *Let the activation function φ fulfill Assumption 4 and let $L \ge 2$. We use the conventions $\infty^0 := 1$ and $\infty - \infty := -\infty$. Then, the nonzero eigenvalues of the NNGP-kernel k_L^{NNGP} and the NTK-kernel k_L^{NTK} defined in Notation 7 fulfill*

$$\mu_{2k+r,d}(k_L^{\text{NNGP}}) > 0 \text{ if and only if } \begin{cases} 2k + r \le \deg_{ev}(G_L) & , r = 0, \\ 2k + r \le \deg_{od}(G_L) & , r = 1, \end{cases}$$

$$\mu_{2k+r,d}(k_L^{\text{NTK}}) > 0 \text{ if and only if } \begin{cases} 2k + r \le \deg_{ev}(H_L) & , r = 0, \\ 2k + r \le \deg_{od}(H_L) & , r = 1 \;. \end{cases}$$

For the NNGP-term G_L define $\sigma^2 := \sigma_i^2 \sigma_b^2$, and for the NTK-term H_L define $\sigma^2 := \sigma_b^2$. That is, the relevant row of the table is different for G_L or respectively for H_L in the case $\sigma_i^2 = 0,\ \sigma_b^2 > 0$.

The even/odd degree of G_L or respectively H_L is given by

	$\deg_{ev}(G_L), \deg_{ev}(H_L)$	$\deg_{od}(G_L), \deg_{od}(H_L)$
$\sigma^2 > 0,\ \deg_{ev}^{Her}(\varphi) > \deg_{od}^{Her}(\varphi)$	$\deg_{ev}^{Her}(\varphi)^{L-1}$	$\deg_{ev}^{Her}(\varphi)^{L-1} - 1$
$\sigma^2 > 0,\ \deg_{ev}^{Her}(\varphi) < \deg_{od}^{Her}(\varphi)$	$\deg_{od}^{Her}(\varphi)^{L-1} - 1$	$\deg_{od}^{Her}(\varphi)^{L-1}$
$\sigma^2 = 0, \deg_{ev}^{Her}(\varphi) > \deg_{od}^{Her}(\varphi)$	$\deg_{ev}^{Her}(\varphi)^{L-1}$	$\deg_{ev}^{Her}(\varphi)\left(\deg_{ev}^{Her}(\varphi)^{L-2} - 1\right) + \deg_{od}^{Her}(\varphi)$
$\sigma^2 = 0, \deg_{ev}^{Her}(\varphi) < \deg_{od}^{Her}(\varphi)$	$\deg_{od}^{Her}(\varphi)\left(\deg_{od}^{Her}(\varphi)^{L-2} - 1\right) + \deg_{ev}^{Her}(\varphi)$	$\deg_{od}^{Her}(\varphi)^{L-1}$

Proof. Let $\kappa : [-1, 1] \to \mathbb{R}, \kappa(t) = \sum_{j \ge 0} \lambda_i t^j$ be a power series. Let k be the corresponding radial kernel on $\mathbb{S}^d$ given by $k(x, y) := \kappa(\langle x, y \rangle)$ and let $(\mu_i)_{i \ge 0}$ be the decreasing series of eigenvalues of k counted by algebraic multiplicity.

Reduction to polynomial degrees of κ: We show that the largest even/odd index i corresponding to a nonzero eigenvalue $\mu_i \ne 0$ of k is the same as the largest even/odd index j corresponding to a nonzero coefficient λ_j of the power series representation of κ. By convention, if there are infinitely many such indices, we say that the largest one equals ∞ and we say that the largest even/odd nonzero index equals $-\infty$ if all corresponding eigenvalues or respectively coefficients are equal to zero.

- Hubbert et al. (2015) show that the eigenvalues μ_i have the same sign as the Legendre coefficients $b_k(\kappa)$ obtained by displaying κ by Legendre polynomials, which form an orthonormal basis of $L_2([-1, 1])$.
- The bases of the Legendre polynomials $(p_i)_{i \ge 0}$, the Hermite polynomials $(h_i)_{i \ge 0}$ and the monomials $(t \mapsto t^i)_{i \ge 0}$ share the properties that the i-th element is a polynomial of degree i and is it even/odd if i is even/odd. Hence, when we can display κ as a sum in those bases, the largest even/odd index i of nonzero coefficients coincide.
- Hence, we see that indeed the largest even/odd index of a nonzero eigenvalue of k corresponds to the largest even/odd index of a nonzero coefficient of the power series representation of κ.

Hence, in order to obtain the largest even/odd index of a nonzero eigenvalue of k_L^{NNGP} or respectively k_L^{NTK} it suffices to determine the largest even/odd index of the nonzero coefficients of the power series representation of G_L or respectively H_L. Given a nonzero eigenvalue with even/odd index, Lemma D.6 shows that all eigenvalues with smaller even/odd index are strictly positive and we can conclude the claim. Hence, all that remains is to show that G_L and H_L can be represented as a power series and that the even/odd degrees of G_l and H_l are as the table claims.

Simplifying Notation: By Assumption 4 we have $\varphi \in \mathcal{L}_2(\mathcal{N}(0,1))$, enabling us to display φ in the Hermite basis as $\varphi(t) = \sum_{i\geq 0} \eta_i h_i(t)$. Eq. (A.2) shows

$$\widehat{\varphi}(t) = \sum_{i\geq 0} \eta_i^2 t^i \ . \tag{D.25}$$

Essentially, we obtain for any $\alpha > 0$

$$\widehat{\varphi_{\cdot\sqrt{\alpha}}} =_{\deg} \widehat{\varphi} =_{\deg}^{\mathrm{Pow,Her}} \varphi, \quad \widehat{\varphi'_{\cdot\sqrt{\alpha}}} =_{\deg} \widehat{\varphi'} =_{\deg}^{\mathrm{Pow,Her}} \varphi' \ .$$

This observation allows us to work with $\widehat{\varphi}$ or respectively $\widehat{\varphi'}$ instead of the rescaled versions occurring in G_l and H_l, simplifying the notation. By Eq. (D.25) all coefficients of the power series of $\widehat{\varphi_{\cdot\sqrt{\alpha}}}, \widehat{\varphi'_{\cdot\sqrt{\alpha}}}$ are non-negative and hence the coefficients of the power series G_l, H_l are non-negative as well, enabling the use of Lemma D.12 in the following.

Calculating G_l: We directly obtain

$$\begin{aligned}
\deg(G_l) &= \deg(g_l \circ \cdots \circ g_1) = \deg(\widehat{\varphi})^{l-1}, \quad l \geq 2, \\
g_1(t) &=_{\deg} t + \sigma_b^2 \sigma_i^2, \\
g_l(t) &=_{\deg} \widehat{\varphi}(t) + \sigma_b^2 \sigma_i^2, \quad l \geq 2 \ .
\end{aligned}$$

If $\deg_{\mathrm{ev}}(\widehat{\varphi}) = \deg_{\mathrm{od}}(\widehat{\varphi})$ holds, then $\deg_{\mathrm{ev}}(\widehat{\varphi}) = \deg_{\mathrm{od}}(\widehat{\varphi}) = \infty$ follows and the claim is trivial.

Assume $\deg_{\mathrm{ev}}(\widehat{\varphi}) > \deg_{\mathrm{od}}(\widehat{\varphi})$. We show the claim for $\deg_{\mathrm{od}}(G_l)$ by induction in $l \geq 2$ and directly obtain the induction base

$$\deg_{\mathrm{od}}(G_2) = \begin{cases} \deg_{\mathrm{ev}}(\widehat{\varphi}) - 1 & , \text{ if } \sigma_b^2\sigma_i^2 > 0 \ , \\ \deg_{\mathrm{od}}(\widehat{\varphi}) & , \text{ if } \sigma_b^2\sigma_i^2 > 0 \ . \end{cases}$$

In the induction step, we have by Lemma D.12

$$\begin{aligned}
\deg_{\mathrm{od}}(G_{l+1}) &= ((\deg_{\mathrm{ev}}(\widehat{\varphi}) - 1) \cdot \deg_{\mathrm{ev}}(G_l) + \deg_{\mathrm{od}}(G_l)) \vee ((\deg_{\mathrm{ev}}(\widehat{\varphi}) - 1) \deg_{\mathrm{od}}(G_l) + \deg_{\mathrm{ev}}(G_l)) \\
&\quad \vee (\deg_{\mathrm{od}}(\widehat{\varphi}) \deg_{\mathrm{od}}(G_l)) \vee ((\deg_{\mathrm{od}}(\widehat{\varphi}) - 1) \deg_{\mathrm{ev}}(G_l) + \deg_{\mathrm{od}}(G_l)) \\
&= (\deg_{\mathrm{ev}}(\widehat{\varphi}) - 1) \deg_{\mathrm{ev}}(G_l) + \deg_{\mathrm{od}}(G_l)
\end{aligned}$$

and the claim follows. The case $\deg_{\mathrm{od}}(\widehat{\varphi}) > \deg_{\mathrm{ev}}(\widehat{\varphi})$ is handled analogously.

Calculating H_l:

If $\deg_{\mathrm{ev}}(\widehat{\varphi}) = \deg_{\mathrm{od}}(\widehat{\varphi})$ holds, then $\deg_{\mathrm{ev}}(\widehat{\varphi}) = \deg_{\mathrm{od}}(\widehat{\varphi}) = \infty$ follows and the claim is simple.

By definition we have $H_1(t) =_{\deg} t + \sigma_b^2$. For $l \geq 2$ we simplify $H_l(t)$ as

$$\begin{aligned}
H_l(t) &\coloneqq \sigma_b^2(1-\sigma_i^2) + \alpha_l G_l(t) + \sigma_w^2 \widehat{(\varphi')_{\cdot\sqrt{\alpha_{l-1}}}}(G_{l-1}(t)) H_{l-1}(t) \\
&=_{\deg} \sigma_b^2(1-\sigma_i^2) + \sigma_b^2\sigma_i^2 + \sigma_w^2 \left(\widehat{\varphi} \circ G_{l-1}(t) + \widehat{(\varphi')_{\cdot\sqrt{\alpha_{l-1}}}}(G_{l-1}(t)) H_{l-1}(t) \right) \\
&=_{\deg} G_l(t) + \widehat{\varphi}'(G_{l-1}(t)) H_{l-1}(t) + \sigma_b^2 \ .
\end{aligned}$$

A simple induction yields

$$\deg(H_l) = \deg(G_l) = \deg(\widehat{\varphi})^{l-1}$$

and it remains to calculate $\deg_{\mathrm{ev}}(H_l)$ for $\deg_{\mathrm{od}}(\widehat{\varphi}) > \deg_{\mathrm{ev}}(\widehat{\varphi})$ and $\deg_{\mathrm{od}}(H_l)$ for $\deg_{\mathrm{ev}}(\widehat{\varphi}) > \deg_{\mathrm{od}}(\widehat{\varphi})$.

While we usually have $\left|\deg_{\mathrm{ev}}(\widehat{\varphi}') - \deg_{\mathrm{od}}(\widehat{\varphi}')\right| = 1$, we need some case work to cover the case where one of the degrees is $-\infty$.

Case 1: $\sigma_b^2 > 0$. If $\deg_{\text{ev}}(\widehat{\varphi}) = \deg_{\text{od}}(\widehat{\varphi})$ holds, then $\deg_{\text{ev}}(\widehat{\varphi}) = \deg_{\text{od}}(\widehat{\varphi}) = \infty$ follows and the claim is trivial. Assume $\deg_{\text{ev}}(\widehat{\varphi}) > \deg_{\text{od}}(\widehat{\varphi})$. We show that $\deg_{\text{od}}(H_l) = \deg_{\text{ev}}(\widehat{\varphi}) - 1$ holds for all $l \geq 2$. For $l = 2$ we obtain straightforwardly

$$\deg_{\text{od}}(H_2) = \deg_{\text{ev}}(\widehat{\varphi}) - 1.$$

By induction we obtain

$$\begin{aligned}
\deg_{\text{od}}(H_{l+1}) &\geq \deg_{\text{od}}\left((\widehat{\varphi}' \circ G_l) \cdot H_l\right) \\
&\geq \deg_{\text{ev}}\left(\widehat{\varphi}' \circ G_l\right) + \deg_{\text{od}}(H_l) = (\deg_{\text{ev}}(\widehat{\varphi}) - 1)\deg_{\text{ev}}(\widehat{\varphi})^{l-1} + \deg_{\text{ev}}(\widehat{\varphi})^{l-1} - 1 \\
&= \deg_{\text{ev}}(\widehat{\varphi})^{l} - 1,
\end{aligned}$$

and as we have $\deg(H_{l+1}) = \deg_{\text{ev}}(\widehat{\varphi})^l$, the claim follows. The case $\deg_{\text{od}}(\widehat{\varphi}) > \deg_{\text{ev}}(\widehat{\varphi})$ works analogously.

Case 2: $\sigma_b^2 = 0$. Assume

$$\deg_{\text{od}}(\widehat{\varphi}) > \deg_{\text{ev}}(\widehat{\varphi}).$$

Case 2.1: $\deg_{\text{od}}(\widehat{\varphi}) = \infty$. We want to show

$$\deg_{\text{ev}}(H_L) = \begin{cases} \infty & \text{, if } \deg_{\text{ev}}(\widehat{\varphi}) > -\infty \text{ ,} \\ -\infty & \text{, if } \deg_{\text{ev}}(\widehat{\varphi}) = -\infty \text{ .} \end{cases}$$

In the case $\deg_{\text{ev}}(\widehat{\varphi}) > -\infty$, we have since $\sigma_b^2 = 0$

$$\deg_{\text{ev}}(H_L) = \deg_{\text{ev}}(G_L) \vee \left(\deg_{\text{ev}}\left((\widehat{\varphi}' \circ G_{L-1}) \cdot H_L\right)\right) \geq \deg_{\text{ev}}(G_{l+1}) = \infty$$

by the previous case. In the case $\deg_{\text{ev}}(\widehat{\varphi}) = -\infty$, H_L is odd by Proposition D.4 and hence we have $\deg_{\text{ev}}(H_L) = -\infty$ as desired.

Case 2.2: $\deg_{\text{od}}(\widehat{\varphi}) < \infty$. We show the claim by induction in l. For $l = 2$ we straightforwardly obtain

$$\deg_{\text{ev}}(H_2) = \deg_{\text{ev}}(\widehat{\varphi}).$$

In the induction step, let $l \geq 3$ and let $\deg_{\text{ev}}(H_l) = \deg_{\text{od}}(\widehat{\varphi})(\deg_{\text{od}}(\widehat{\varphi})^{l-2} - 1) + \deg_{\text{ev}}(\widehat{\varphi})$ hold as claimed. Since $\sigma_b^2 = 0$, we have

$$\begin{aligned}
\deg_{\text{ev}}(H_{l+1}) &= \deg_{\text{ev}}(G_{l+1}) \vee \left(\deg_{\text{ev}}\left((\widehat{\varphi}' \circ G_l) \cdot H_l\right)\right) \\
&= \left(\deg_{\text{od}}(\widehat{\varphi})\left(\deg_{\text{od}}(\widehat{\varphi})^{l-1} - 1\right) + \deg_{\text{ev}}(\widehat{\varphi})\right) \vee \left(\deg_{\text{ev}}\left(\widehat{\varphi}' \circ G_l\right) + \deg_{\text{ev}}(H_l)\right) \\
&\quad \vee \left(\deg_{\text{od}}\left(\widehat{\varphi}' \circ G_l\right) + \deg_{\text{od}}(H_l)\right) \text{ .}
\end{aligned}$$

Now we show that the first term is the largest one in order to obtain the claim. Firstly, we bound the second term as

$$\begin{aligned}
\deg_{\text{ev}}\left(\widehat{\varphi}' \circ G_l\right) + \deg_{\text{ev}}(H_l) &\leq \deg\left(\widehat{\varphi}' \circ G_l\right) + \deg_{\text{ev}}(H_l) \\
&= (\deg(\widehat{\varphi}) - 1) \cdot \deg(G_l) + \deg_{\text{ev}}(H_l) \\
&= (\deg(\widehat{\varphi}) - 1) \cdot \deg(\widehat{\varphi})^{l-1} + \deg(\widehat{\varphi})\left(\deg(\widehat{\varphi})^{l-2} - 1\right) + \deg_{\text{ev}}(\widehat{\varphi}) \\
&= \deg(\widehat{\varphi})\left(\deg(\widehat{\varphi})^{l-1} - 1\right) + \deg_{\text{ev}}(\widehat{\varphi}) \text{ .}
\end{aligned}$$

In order to bound the third term, we recall $\deg_{\text{od}}\left(H^l\right) = \deg_{\text{od}}(\widehat{\varphi})^{l-1}$, hence we need to show $\deg_{\text{od}}\left(\widehat{\varphi}' \circ G_l\right) \leq \deg_{\text{od}}(\widehat{\varphi})^l - \deg_{\text{od}}(\widehat{\varphi})^{l-1} + \deg_{\text{ev}}(\widehat{\varphi}) - \deg_{\text{od}}(\widehat{\varphi})$. If $\deg_{\text{od}}(\widehat{\varphi}) = 1$ we can see $\widehat{\varphi}' \equiv c, c \in \mathbb{R}$ and obtain $\deg_{\text{od}}(\widehat{\varphi} \circ G_l) = -\infty$. Assume $\deg_{\text{od}}(\widehat{\varphi}) \geq 3$. In the following we use

$$\deg_{\text{od}}\left(\widehat{\varphi}'\right) \leq \deg_{\text{od}}(\widehat{\varphi}) - 2 < \deg_{\text{od}}(\widehat{\varphi}) - 1 = \deg_{\text{ev}}\left(\widehat{\varphi}'\right) \text{ .}$$

Lemma D.12 (c) allows unrolling $\deg_{\text{od}}\left(\widehat{\varphi}' \circ G_l\right)$ as

$$\begin{aligned}
\deg_{\text{od}}\left(\widehat{\varphi}' \circ G_l\right) = &\left(\deg_{\text{od}}\left(\widehat{\varphi}'\right) \cdot \deg_{\text{od}}\left(G_l\right)\right) \\
&\vee \left((\deg_{\text{od}}\left(\widehat{\varphi}'\right) - 1) \cdot \deg_{\text{ev}}\left(G_l\right) + \deg_{\text{od}}\left(G_l\right)\right) \\
&\vee \left((\deg_{\text{ev}}\left(\widehat{\varphi}'\right) - 1) \cdot \deg_{\text{od}}\left(G_l\right) + \deg_{\text{ev}}\left(G_l\right)\right) \\
&\vee \left((\deg_{\text{ev}}\left(\widehat{\varphi}'\right) - 1) \cdot \deg_{\text{ev}}\left(G_l\right) + \deg_{\text{od}}\left(G_l\right)\right) \\
\leq &\left(\deg_{\text{od}}\left(\widehat{\varphi}\right) - 2\right) \deg_{\text{od}}\left(\widehat{\varphi}\right)^{l-1} \\
&\vee \left(\deg_{\text{od}}\left(\widehat{\varphi}\right) - 3\right) \left(\deg_{\text{od}}\left(\widehat{\varphi}\right)^{l-1} + \deg_{\text{ev}}\left(\widehat{\varphi}\right) - \deg_{\text{od}}\left(\widehat{\varphi}\right)\right) + \deg_{\text{od}}\left(\widehat{\varphi}\right)^{l-1} \\
&\vee \left(\deg_{\text{od}}\left(\widehat{\varphi}\right) - 2\right) \deg_{\text{od}}\left(\widehat{\varphi}\right)^{l-1} + \deg_{\text{od}}\left(\widehat{\varphi}\right)^{l-1} + \deg_{\text{ev}}\left(\widehat{\varphi}\right) - \deg_{\text{od}}\left(\widehat{\varphi}\right) \\
&\vee \left(\deg_{\text{od}}\left(\widehat{\varphi}\right) - 2\right) \left(\deg_{\text{od}}\left(\widehat{\varphi}\right)^{l-1} + \deg_{\text{ev}}\left(\widehat{\varphi}\right) - \deg_{\text{od}}\left(\widehat{\varphi}\right)\right) + \deg_{\text{od}}\left(\widehat{\varphi}\right)^{l-1} \\
\leq &\deg_{\text{od}}\left(\widehat{\varphi}\right)^{l} - \deg_{\text{od}}\left(\widehat{\varphi}\right)^{l-1} + \deg_{\text{ev}}\left(\widehat{\varphi}\right) - \deg_{\text{od}}\left(\widehat{\varphi}\right) \; .
\end{aligned}$$

The case $\deg\left(\widehat{\varphi}\right) = \deg_{\text{ev}}\left(\widehat{\varphi}\right) > \deg_{\text{od}}\left(\widehat{\varphi}\right)$ is handled analogously. □

E POLYNOMIAL BOUNDEDNESS OF DERIVATIVES

Here, we prove Proposition 5. Specifically, we want to find a sufficient criterion for C^∞-functions f such that all of their derivatives $f^{(k)}$ are polynomially bounded. To this end, we leverage that typical activation functions use symbolic expressions whose derivatives can be expressed using the same set of symbols, e.g. $\tanh'(x) = 1 - \tanh(x)^2$. Our goal is to find a set of "base functions" whose derivatives use the same base functions and which are polynomially bounded. We then exploit that the class of polynomially bounded functions is closed under addition, multiplication and composition and therefore also contains the derivatives of the "base functions". Higher-order derivatives are then easily treated using induction. We first formally define the relevant function classes:

Definition E.1. Let $I \subseteq \mathbb{R}$ be an interval. For $m \in \mathbb{N}_0$, let

$$\mathcal{S}^{(m)}(I) := \{f \in C^m(I) \mid \forall 0 \leq k \leq m \; \exists a, b, q > 0 \; \forall x \in I : |f^{(k)}(x)| \leq a|x|^q + b\} \; .$$

We note that the class $\mathcal{S}^{(\infty)}(I)$ from Definition 3 satisfies $\mathcal{S}^{(\infty)}(I) = \bigcap_{m=0}^{\infty} \mathcal{S}^{(m)}(I)$. ◀

Now, we formally define some "base functions":

Definition E.2. Define $\text{sigmoid}, \text{softplus}, \text{RBF} : \mathbb{R} \to \mathbb{R}$ by

$$\begin{aligned}
\text{sigmoid}(x) &:= (1 + e^{-x})^{-1} \\
\text{softplus}(x) &:= \log(1 + e^{-x}) \\
\text{RBF}(x) &:= e^{-x^2} \; .
\end{aligned}$$

Moreover, let Φ be the CDF of the normal distribution $\mathcal{N}(0, 1)$. ◀

Lemma E.3. *Let $I \subseteq \mathbb{R}$ be an interval and let $m \in \mathbb{N}_0 \cup \{\infty\}$. Then,*

(a) If $m \geq 1$, then $\mathcal{S}^{(m)} = \{f \in C^1(I) \mid f \in \mathcal{S}^{(0)}(I)$ and $f' \in \mathcal{S}^{(m-1)}\}(I)$.
(b) $\mathcal{S}^{(m)}(I)$ is closed under addition, multiplication and composition of functions.
(c) $\mathcal{S}^{(m)}(I)$ contains all polynomials.
(d) $\mathcal{S}^{(m)}(I)$ contains $\text{sigmoid}, \tanh, \text{softplus}, \sin, \cos, \text{RBF}$ *and Φ.*

Proof. We prove the statements for $m \in \mathbb{N}_0$, the case $m = \infty$ easily follows.

(a) This statement is straightforward.
(b) For addition, the statement is trivial. Now, let $f, g \in \mathcal{S}^{(m)}$. Choose constants C_1, C_2, q such that $|f(x)|, |g(x)| \leq C_1 + C_2|x|^q$. Then, $|f(x)g(x)| = O(|x|^{2q})$ and $|f(g(x))| \leq C_1 + C_2|g(x)|^q \leq C_1 + C_2|C_1 + C_2|x|^q|^q = O(|x|^{q^2})$. This shows $f \cdot g, f \circ g \in \mathcal{S}^{(0)}$. If $m = 0$, we are done. Otherwise, we obtain

$$(f + g)' = f' + g' \in \mathcal{S}^{(m-1)}t$$

$$(f \cdot g)' = f' \cdot g + f \cdot g' \in \mathcal{S}^{(m-1)}$$
$$(f \circ g)' = (f' \circ g) \cdot g' \in \mathcal{S}^{(m-1)}$$

since $\mathcal{S}^{(m-1)}$ is closed under addition, multiplication and composition by the induction hypothesis. By (b), we have shown that $f \cdot g, f \circ g \in \mathcal{S}^{(m)}$.

(c) This is trivial.

(d) It is not hard to see that all of the mentioned functions are in $\mathcal{S}^{(0)}$. For the induction step, we assume that all of the mentioned functions are in $\mathcal{S}^{(m-1)}$. Since all of these functions are C^∞ functions, their derivatives are in $C^{m-1}(\mathbb{R})$. Moreover, by inspecting the derivatives and using the previously proven facts as well as the induction hypothesis, it is easy to see that the derivatives are in $\mathcal{S}^{(m-1)}(\mathbb{R})$:

$$\begin{aligned}
\text{sigmoid}'(x) &= \text{sigmoid}(x) \cdot (1 - \text{sigmoid}(x)) = \text{sigmoid}(x) \cdot (x^0 + (-x^0) \cdot \text{sigmoid}(x)) \\
\tanh'(x) &= 1 - \tanh(x) \cdot \tanh(x) \\
\text{softplus}'(x) &= \text{sigmoid}(x) \\
\sin'(x) &= \cos(x) \\
\cos'(x) &= -\sin(x) \\
\text{RBF}'(x) &= 2x \cdot \text{RBF}(x) \\
\Phi'(x) &= \frac{1}{\sqrt{2\pi}} e^{-x^2/2} = \frac{1}{\sqrt{2\pi}} \text{RBF}(2^{-1/2} \cdot x) \ .
\end{aligned}$$

Hence, by (b), we have shown that the functions are in $\mathcal{S}^{(m)}$. ☐

Proposition 5 (All common activation function satisfy Assumption 4)**.**

(a) If f is constructed by addition, multiplication, and composition of polynomials, sigmoid*,* tanh*,* softplus*,* sin*,* cos*,* RBF *and* Φ*, then $f \in \mathcal{S}^{(\infty)}(\mathbb{R})$. Moreover, the functions $g(x) = \exp(ax)$ are in $\mathcal{S}^{(\infty)}((-\infty, 0])$ for every $a \geq 0$.*

(b) Assumption 4 is satisfied for all activation functions from Table 1. It is also satisfied for all $\phi \in \mathcal{S}^{(\infty)}(\mathbb{R})$.

Proof.

(a) It follows from elementary calculations that the functions $f(x) = \exp(ax)$ are in $\mathcal{S}^{(\infty)}((-\infty, 0])$ for all $a \geq 0$. For the other functions, this follows from Lemma E.3.

(b) This follows directly from (a). ☐

F SMOOTHNESS OF GP SAMPLE PATHS

To prove Theorem 14, we need the following result, which is a variant of Thm. 7.4 of Lukić and Beder (2001):

Theorem F.1 (Theorem 1.2 in Steinwart 2024)**.** *Let H_1, H_2 be Hilbert spaces on a set T and let X be a centered Gaussian process with covariance kernel k such that H_1 is the RKHS of k. Then the following statements hold true:*

a) If $H_1 \hookrightarrow H_2$ and furthermore this embedding is a Hilbert-Schmidt operator, then there exists a version Y of X with $\mathbb{P}(Y \in H_2) = 1$.

b) Otherwise, for all versions Y of X we have $\mathbb{P}(Y \in H_2) = 0$.

Now, we can prove Theorem 14:

Theorem 14. *Let k be a dot-product kernel on $\mathbb{S}^d$ whose RKHS is equivalent to a Sobolev space $H^{d+\alpha}(\mathbb{S}^d)$, $\alpha > 0$. Let X be a Gaussian process on $\mathbb{S}^d$ with zero mean and covariance kernel k.*

i) For any $\varepsilon \geq 0$ we have $P(Y \in H^{d/2+\alpha+\varepsilon}(\mathbb{S}^d)) = 0$ for any version Y of X.

ii) For any $0 < \varepsilon < \alpha$, there exists a version Y of X with $\mathbb{P}\left(Y \in H^{d/2+\alpha-\varepsilon}\right) = 1$.

Proof. We will apply Theorem F.1 with

$$H_1 := H^{d+\alpha}(\mathbb{S}^d)\,, \qquad \text{and} \qquad H_2 := H^{d/2+\alpha+\varepsilon}(\mathbb{S}^d)\,,$$

where $\varepsilon \in (-\alpha, \infty)$.

Here we first note that in the case $\varepsilon > d/2$ we have $H^{d+\alpha}(\mathbb{S}^d) \not\hookrightarrow H^{d/2+\alpha+\varepsilon}(\mathbb{S}^d)$ and hence *ii)* of Theorem F.1 gives *i)* of Theorem 14 for such ε.

It thus remains to consider $\varepsilon \in (-\alpha, d/2)$. Since in this case the embedding $H^{d+\alpha}(\mathbb{S}^d) \hookrightarrow H^{d/2+\alpha+\varepsilon}(\mathbb{S}^d)$ exists, Theorem F.1 shows that we need to investigate whether this embedding is a Hilbert-Schmidt operator. To this end we apply Brauchart and Dick (2013) where it is shown that the Sobolev space $H^r(\mathbb{S}^d)$ is given by

$$H^r(\mathbb{S}^d) \coloneqq \{f \in L^2(\mathbb{S}^d) \mid \|f\|_{H^r(\mathbb{S}^d)} < \infty\},$$
$$\|f\|_{H^r(\mathbb{S}^d)} \coloneqq \sum_{l=0}^{\infty} \sum_{i=1}^{N_{l,d}} \lambda_{l,r} \hat{f}_{l,i}$$

for a sequence $\lambda_{l,r} = \Theta_{\forall l}(1+l)^{2r}$, where $\hat{f}_{l,i}$ denotes the Laplace-Fourier coefficients given by

$$\hat{f}_{l,i} \coloneqq \int_{\mathbb{S}^d} f(x) Y_{l,i}(x) \, \mathrm{d}\mu_{\mathbb{S}^d}(x) \ .$$

Furthermore $N_{l,d} = \Theta_{\forall l}((l+1)^{d-1})$ holds. Hence, $(\lambda_{l,r}^{-1/2} Y_{l,i})_{l \geq 0, 1 \leq i \leq N_{l,d}}$ forms an ONB of $H^r(\mathbb{S}^d)$ that diagonalizes the embedding operator. The Hilbert-Schmidt norm of the embedding $H^{d+\alpha}(\mathbb{S}^d) \hookrightarrow H^{d/2+\alpha+\varepsilon}(\mathbb{S}^d)$ is now straightforwardly given, cf. Birman and Solomjak (1987, Chapter 11.3), by

$$\left\| H^{d+\alpha}(\mathbb{S}^d) \hookrightarrow H^{d/2+\alpha+\varepsilon}(\mathbb{S}^d) \right\|_{\mathrm{HS}} = \sum_{l=0}^{\infty} \sum_{i=1}^{N_{l,d}} \frac{\lambda_{l,d/2+\alpha+\varepsilon}}{\lambda_{l,d+\alpha}} \ . \tag{F.1}$$

The asymptotics of $\lambda_{l,s}$ and $N_{l,d}$ yield

$$\sum_{i=1}^{N_{l,d}} \frac{\lambda_{l,d/2+\alpha+\varepsilon}}{\lambda_{l,d+\alpha}} = \Theta_{\forall l}(l+1)^{d-1} \frac{\Theta_{\forall l}(l+1)^{d+2\alpha+2\varepsilon}}{\Theta_{\forall l}(l+1)^{2d+2\alpha}} = \Theta_{\forall l}(l+1)^{2\varepsilon-1}$$

and consequently, we see that (F.1) is finite if and only if $\varepsilon < 0$. This shows *ii)* as well as *i)* in the remaining case $0 \leq \varepsilon \leq d/2$. □

G ACTIVATION QUADRATURE

Here, we derive a way to numerically approximate the dual activation for activations f that are smooth everywhere except possibly at zero. The dual activation involve an integral with the term $f(x)f(y)$, and in order to use Gauss-Legendre quadrature, we need to decompose the integration domain into regions where $f(x)f(y)$ is smooth. Our construction is based on separate quadratures for the four quadrants $x, y \geq 0$, $x, -y \geq 0$, $-x, y \geq 0$, and $-x, -y \geq 0$.

Let $f : \mathbb{R} \to \mathbb{R}$ and let $\rho \in [-1, 1]$. Moreover, let $X, Y \sim \mathcal{N}(0, 1)$ be independent and let $\sigma_x \coloneqq \sqrt{\frac{1+\rho}{2}}$, $\sigma_y \coloneqq \sqrt{\frac{1-\rho}{2}}$. Define the (centered) random variables

$$U \coloneqq \sigma_x X + \sigma_y Y$$
$$V \coloneqq \sigma_x X - \sigma_y Y$$

Then,

$$\mathbb{E}[U^2] = \sigma_x^2 + \sigma_y^2 = 1$$
$$\mathbb{E}[UV] = \sigma_x^2 - \sigma_y^2 = \rho$$
$$\mathbb{E}[V^2] = \sigma_x^2 + \sigma_y^2 = 1 \ .$$

Hence, we have

$$\begin{pmatrix} U \\ V \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right) \ .$$

We then want to compute

$$\hat{f}(\rho) = \mathbb{E}[f(U, V)] = \mathbb{E}[f(\sigma_x X + \sigma_y Y, \sigma_x X - \sigma_y Y)] \ .$$

We approximate the integral over a quadrant using the standard normal pdf ϕ as

$$\begin{aligned}
g(\rho) &:= \int_{(0,\infty)^2} f(u, v) \,\mathrm{d}P_{U,V}(u, v) \\
&= \int_0^\infty \int_{-\frac{\sigma_x}{\sigma_y}x}^{\frac{\sigma_x}{\sigma_y}x} f(\sigma_x x + \sigma_y y, \sigma_x x - \sigma_y y)\phi(x)\phi(y) \,\mathrm{d}y \,\mathrm{d}x \\
&\approx \int_0^{c\sigma_y} \int_{-\frac{\sigma_x}{\sigma_y}x}^{\frac{\sigma_x}{\sigma_y}x} f(\sigma_x x + \sigma_y y, \sigma_x x - \sigma_y y)\phi(x)\phi(y) \,\mathrm{d}y \,\mathrm{d}x \\
&\quad + \int_{c\sigma_y}^{c(\sigma_y+1)} \int_{-c\sigma_x}^{c\sigma_x} f(\sigma_x x + \sigma_y y, \sigma_x x - \sigma_y y)\phi(x)\phi(y) \,\mathrm{d}y \,\mathrm{d}x \\
&=: A + B
\end{aligned}$$

The restricted integration domains in the second integral B should not cause a large error since $\phi(x)\phi(y)$ is small for $x \geq c(\sigma_y + 1) \geq c$ and also for $x \geq c\sigma_y, |y| \geq c\sigma_x$ since either σ_y or σ_x are $\geq \frac{1}{\sqrt{2}}$. We can now further simplify A using the substitutions $x =: c\sigma_y x'$, $y =: \frac{\sigma_x}{\sigma_y} x y' = c\sigma_x x' y'$:

$$A = \int_0^1 \int_{-1}^1 c\sigma_y \cdot c\sigma_x x' \cdot f(c\sigma_x\sigma_y x'(1 + y'), c\sigma_x\sigma_y x'(1 - y'))\phi(c\sigma_y x')\phi(c\sigma_x x' y') \,\mathrm{d}y' \,\mathrm{d}x' \ .$$

Now, we simplify B using the substitutions $x =: c\sigma_y + cx''$ and $y =: c\sigma_x y''$:

$$\begin{aligned}
B = \int_0^1 \int_{-1}^1 & c \cdot c\sigma_x \cdot f(c\sigma_x\sigma_y(1 + y'') + c\sigma_x x'', c\sigma_x\sigma_y(1 - y'') + c\sigma_x x'') \\
& \cdot \phi(c\sigma_y + cx'')\phi(c\sigma_x y'') \,\mathrm{d}y'' \,\mathrm{d}x'' \ .
\end{aligned}$$

Both integrals can be approximated using Gauss-Legendre quadrature along both axes.

For the plots in Figure 1 and Figure 2, we use $c = 12$ and a 50×50 grid of Gauss-Legendre points for each 2D integral. To compute the eigenvalues from the NTK, we use the Gegenbauer quadrature implementation from Bordelon et al. (2020) with 1000 quadrature points. We provide code for reproducing the figures at `https://github.com/dholzmueller/beyond_relu`.