Wide and Deep Neural Networks Achieve Optimality for Classification

Adityanarayanan Radhakrishnan[1] Mikhail Belkin [2] Caroline Uhler[1,3]

May 2, 2022

Abstract

While neural networks are used for classification tasks across domains, a long-standing open problem in machine learning is determining whether neural networks trained using standard procedures are optimal for classification, i.e., whether such models minimize the probability of misclassification for arbitrary data distributions. In this work, we identify and construct an explicit set of neural network classifiers that achieve optimality. Since effective neural networks in practice are typically both wide and deep, we analyze infinitely wide networks that are also infinitely deep. In particular, using the recent connection between infinitely wide neural networks and Neural Tangent Kernels, we provide explicit activation functions that can be used to construct networks that achieve optimality. Interestingly, these activation functions are simple and easy to implement, yet differ from commonly used activations such as ReLU or sigmoid. More generally, we create a taxonomy of infinitely wide and deep networks and show that these models implement one of three well-known classifiers depending on the activation function used: (1) 1-nearest neighbor (model predictions are given by the label of the nearest training example); (2) majority vote (model predictions are given by the label of the class with greatest representation in the training set); or (3) singular kernel classifiers (a set of classifiers containing those that achieve optimality). Our results highlight the benefit of using deep networks for classification tasks, in contrast to regression tasks, where excessive depth is harmful.

1 Introduction

Deep learning has produced state-of-the-art results across several application domains including computer vision [14], natural language processing [5], and biology [34]. Despite these empirical successes, our understanding of basic theoretical properties of deep networks is far from satisfactory. In fact, for the fundamental problem of classification it has not been established whether neural networks trained with standard optimization methods can achieve optimality, i.e., whether they minimize the probability of misclassification for arbitrary data distributions (a property referred to as *Bayes optimality* or *consistency* in the statistics literature).

There is a vast literature on the optimality of statistical machine learning methods; in particular, given the modern practice of using models that can interpolate (i.e., fit the training data exactly), recent works analyzed the optimality of interpolating machine learning models including weighted nearest neighbor methods and kernel smoothers (also known as Nadaraya-Watson estimators) [3, 4, 6, 9, 30]. However, little is known about deep neural networks. Classical work [10] analyzing the optimality of neural networks utilizes the results of Cybenko [7] and Hornik [15] to show that the optimal classifier can be approximated by a neural network that is sufficiently wide; i.e., these prior results are concerned with the existence of networks that achieve optimality and do not present computationally feasible algorithms for finding such networks.

[1]Laboratory for Information & Decision Systems, and Institute for Data, Systems, and Society, Massachusetts Institute of Technology

[2]Halıcıoğlu Data Science Institute, University of California, San Diego

[3]Broad Institute of MIT and Harvard

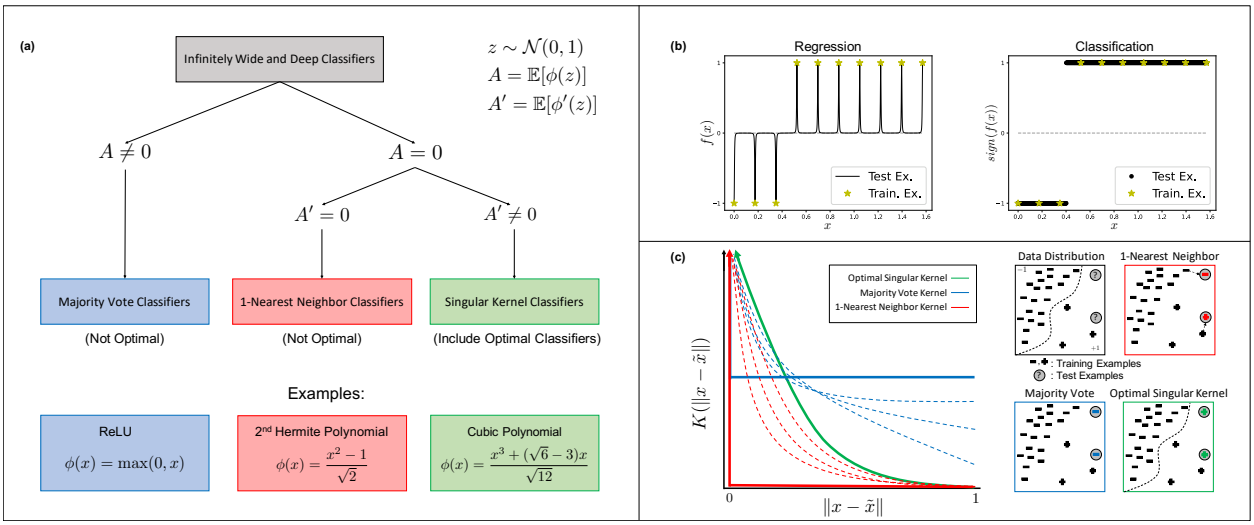

Figure 1: Behavior of infinitely wide and deep neural networks trained with gradient descent. (a) Taxonomy of infinitely wide and deep networks. Depending on the choice of the activation function, $\phi(\cdot)$, these models implement majority vote (blue), 1-nearest neighbor (red), or singular kernel classifiers (green), a subset of which achieve optimality. (b) Regression versus classification using infinitely wide and deep networks. While these models are not effective in the regression setting, since their predictions are near zero almost everywhere, they can achieve optimality for classification, where only the sign of the prediction matters. (c) Illustration of the different behaviors of infinitely wide and deep networks for varying activation functions. Depending on the activation function, infinitely wide and deep networks implement majority vote (blue), 1-nearest neighbor (red), or singular kernel classifiers that can achieve optimality (green). Singular kernels that grow too slowly are akin to majority vote classifiers (dashed blue), whereas those that grow too quickly are akin to weighted nearest neighbor classifiers (dashed red).

By establishing a connection between interpolating kernel smoothers and deep neural networks, we identify and construct an explicit class of neural networks that, when trained with gradient descent, achieve optimality for classification problems. Our results utilize the recent Neural Tangent Kernel (NTK) connection between training wide neural networks and using kernel methods. Several works [17, 20, 22, 23] established conditions under which using a kernel method with the NTK is equivalent to training neural networks, as network width approaches infinity. Given the conceptual simplicity of kernel methods, the NTK has been widely used as a tool for understanding the theoretical properties of neural networks [16, 20, 23, 29, 36]. Since neural networks in practice are often both wide and deep, we consider the natural extension of networks that are both infinitely wide and deep.

In particular, we focus on infinitely wide and deep networks in the classification setting and show that they have markedly different behavior than in the regression setting. Indeed, prior work [13, 16] showed that in the regression setting, infinitely wide and deep neural networks simply predict near-zero values at all test samples and thus, are far from optimal (see Fig. 1b). As a consequence, these models were dismissed as an approach for explaining the strong performance of deep networks in practice. In stark contrast to regression, we show that the sign of the predictor can be informative even when the its numerical output is arbitrarily close to zero (see Fig. 1b for an illustration). In fact, as we show in this work, this is exactly how infinitely wide and deep neural networks can achieve optimal classification accuracy even though the output of the network approaches zero.

To characterize the behavior of infinitely wide and deep classifiers, we establish a taxonomy of such models, and we prove that it includes networks that achieve optimality (see Fig. 1a). More precisely, we prove that infinitely wide and deep neural network classifiers implement one of the following three well-known classifiers depending on the choice of activation function:

1. *1-nearest neighbor (1-NN) classifiers*: the prediction on a new sample is the label of the nearest sample (under Euclidean distance) in the training set.

2. *Majority vote classifiers:* the prediction on a new sample is the label of the class with greater representation in the training set.

3. *Singular kernel classifiers:* the prediction on a new sample is obtained by using the kernel $K(x, \tilde{x}) = \frac{R(\|x-\tilde{x}\|)}{\|x-\tilde{x}\|^\alpha}$ where $\alpha > 0$ is the order of the singularity.[1] As is standard when using kernel smoothers for classification, the prediction, $m(x)$, on a new sample x given training data $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$ is

$$m(x) = \text{sign}\Big(\sum_{i=1}^{n} y^{(i)} K(x^{(i)}, x)\Big). \tag{1}$$

As a corollary of a result in [9] it follows that singular kernel classifiers achieve optimality when α is the dimension of the data, d (see Supplementary Information C). Hence our taxonomy and in particular Theorem 2 of this work provide exact conditions when infinitely wide and deep neural network classifiers achieve optimality for any given data dimension. Notably, we identify a simple class of activation functions that yield singular kernel classifiers with $\alpha = d$, and we thus identify concrete examples of neural networks that achieve optimality. For example, for $d = 2$, the infinitely wide and deep classifier with activation function $\phi(x) = (x^3 + (\sqrt{6} - 3)x)/\sqrt{12}$ achieves optimality. Interestingly, the popular rectified linear unit (ReLU) activation $\phi(x) = \max(x, 0)$ leads to an infinitely wide and deep classifier that implements the majority vote classifier and is thus not optimal. Similarly, the activation function $\phi(x) = (x^2 - 1)/\sqrt{2}$ leads to an infinitely wide and deep classifier that implements the 1-NN classifier and is thus also not optimal.

We note that singular kernels provide a natural transition between 1-NN and majority vote classifiers. Namely, as discussed in [9], for $\alpha > d$, singular kernel classifiers behave akin to weighted nearest neighbor classifiers since $\|x - \tilde{x}\|^\alpha$ is extremely small for $\tilde{x}$ near x. Similarly, for $\alpha < d$, singular kernel classifiers behave akin to majority vote classifiers since $\|x - \tilde{x}\|^\alpha$ is no longer small for $\tilde{x}$ far from x. We visualize this transition between the three classes established in our taxonomy in Fig. 1c.

2 Taxonomy of Infinitely Wide and Deep Neural Networks

In the following, we construct a taxonomy of classifiers implemented by infinitely wide and deep neural networks. Our construction relies on the recent connection between infinitely wide neural networks and kernel methods [17]. In particular, this connection involves utilizing a kernel method known as a kernel machine, which is related to the kernel smoother described in Eq. (1). In contrast to the kernel smoother, a kernel machine with kernel K is given by:

$$\text{sign}\left(y(K_n)^{-1} K(X, x)\right), \tag{2}$$

where $X = [x^{(1)}|x^{(2)}|\dots|x^{(n)}] \in \mathbb{R}^{d\times n}$ denotes the training data, $y = [y^{(1)}, y^{(2)}, \dots y^{(n)}] \in \{-1, 1\}^{1\times n}$ the labels, $K_n \in \mathbb{R}^{n\times n}$ satisfies $(K_n)_{i,j} = K(x^{(i)}, x^{(j)})$ and $K(X, x) \in \mathbb{R}^n$ satisfies $(K(X, x))_i = K(x^{(i)}, x)$. Both kernel methods can be used as prediction schemes for classification [31]. Note that while both algorithms produce predictors with the same functional form, their predictions are generally different. Indeed, understanding the relation between kernel smoothers and kernel machines will be critical to our proof of optimality.

Under certain conditions, training a neural network as width approaches infinity is equivalent[2] to using a kernel machine with a specific kernel known as the Neural Tangent Kernel [17], which is defined below.

Definition 1. *Let $f^{(L)}(x; \mathbf{W})$ denote a fully connected network[3] with L hidden layers with parameters $\mathbf{W}$ operating on data $x \in \mathbb{R}^d$. For $x, \tilde{x} \in \mathbb{R}^d$, the **Neural Tangent Kernel (NTK)** is given by:*

$$K^{(L)}(x, \tilde{x}) = \langle \nabla_{\mathbf{W}} f^{(L)}(x; \mathbf{W}), \nabla_{\mathbf{W}} f^{(L)}(\tilde{x}; \mathbf{W}) \rangle \ .$$

[1] For this order to be well-defined, $R(\cdot)$ is non-negative and satisfies $\inf_{|u|<\epsilon} R(u) > 0$ and $|R(u)| < C$ for some $\epsilon, C > 0$.

[2] This equivalence requires a particular initialization scheme on the weights known as the NTK initialization scheme [17]. Formally, this equivalence holds when an offset term corresponding to the predictions of the neural network at initialization are added to those given by the using a kernel machine with the NTK [17]. Like in prior works (e.g. [1, 13, 16, 19, 29]), we will analyze the NTK without such offset. This model corresponds to averaging the predictions of infinitely many infinite width neural networks [27].

[3] Throughout this work, we consider fully connected networks that have no bias terms.

To work with a simple closed form for the NTK and to avoid symmetries arising from the activation function, we will consider training data with density on $\mathcal{S}_+^d$, where $\mathcal{S}_+^d$ is the intersection of the unit sphere $\mathcal{S}^d$ in $d+1$ dimensions and the non-negative orthant.[4]

In this work, we analyze the behavior of infinitely wide and deep networks by analyzing the kernel machine in Eq. (2), as depth, L, goes to infinity. To perform our analysis, we utilize the recursive formula for the NTK of a deep network originally presented in [17]. Namely, $K^{(L)}$ can be expressed as a function of $K^{(L-1)}$ and the network activation function, $\phi(\cdot)$, yielding a discrete dynamical system indexed by L. The exact formula can be found in Eq. (5), and additional relevant results from prior works that are used in our proofs are referenced in Supplementary Information A.

Remarkably, the properties of the resulting dynamical system as $L \to \infty$ are governed by the mean of $\phi(z)$ and its derivative, $\phi'(z)$, for $z \sim \mathcal{N}(0,1)$. For simplicity, we will assume throughout that $\mathbb{E}[\phi(z)^2] < \infty$ and similarly $\mathbb{E}[\phi'(z)^2] < \infty$, an assumption that holds for many activation functions used in practice including ReLU, leaky ReLU, sigmoid, sinusoids, and polynomials. By defining $A = \mathbb{E}[\phi(z)]$ and $A' = \mathbb{E}[\phi'(z)]$, we break down our analysis into the following three cases:

$$\begin{aligned} &\text{Case 1: } A = 0 \ , \ A' \neq 0 \ , \\ &\text{Case 2: } A = 0 \ , \ A' = 0 \ , \\ &\text{Case 3: } A \neq 0 \ . \end{aligned}$$

Under cases 1 and 2, 0 is the unique fixed point attractor of the recurrence for $K^{(L)}$ and thus $K^{(L)}(x, \tilde{x}) \to 0$ as $L \to \infty$ for $x \neq \tilde{x}$. As a consequence, cases 1 and 2 lead to infinitely wide and deep neural networks that predict 0 almost everywhere. Thus, these networks are far from optimal in the regression setting and were thus dismissed as an approach for explaining the strong performance of deep networks. On the other hand, case 3 yields nonzero values for any pair of examples and thus, prior works that analyzed the regression setting [13, 16] focused on activation functions satisfying case 3.

In stark contrast to the regression setting, we will show that infinitely wide and deep networks with activation functions satisfying case 1 are effective for classification, with a subset achieving optimality. In particular, we will show that networks in case 1 implement singular kernel classifiers while those in case 2 implement 1-NN classifiers. Notably, we will identify conditions and provide explicit examples of activation functions in case 1 that guarantee optimality. We will then show that infinitely wide and deep classifiers with activations satisfying case 3 generally correspond to majority vote classifiers. A summary of our taxonomy is presented in Fig. 1a, and we will now discuss each of the three cases in more depth.

Case 1 ($A = 0, A' \neq 0$) networks implement singular kernel classifiers and can achieve optimality.

We establish conditions on the activation function under which an infinitely wide and deep network implements a singular kernel classifier (Theorem 1). We then utilize results of [9] to show that this set of classifiers contains those that achieve optimality for any given data dimension. Lastly, we will present explicit activation functions that lead to infinitely wide and deep classifiers that achieve optimality. We begin with the following theorem, which establishes conditions under which the infinite depth limit of the NTK is a singular kernel.

Theorem 1. *Let $K^{(L)}$ denote the NTK of a fully connected neural network with L hidden layers and activation function $\phi(\cdot)$. For $z \sim \mathcal{N}(0,1)$, define $A = \mathbb{E}[\phi(z)]$, $A' = \mathbb{E}[\phi'(z)]$, and $B' = \mathbb{E}[\phi'(z)^2]$. If $A = 0$ and $A' \neq 0$, then for $x, \tilde{x} \in \mathcal{S}_+^d$:*

$$\lim_{L \to \infty} \frac{K^{(L)}(x, \tilde{x})}{(A')^{2L}(L+1)} = \frac{R(\|x - \tilde{x}\|)}{\|x - \tilde{x}\|^{\alpha}},$$

where $\alpha = -2\frac{\log(A'^2)}{\log(B')}$ and $R(\cdot)$ is non-negative, bounded from above, and bounded away from 0 around 0.

[4] For example, min-max scaling followed by projection onto the sphere results in the data lying in this region.

The full proof is presented in Supplementary Information B, and we outline its key steps in Section 3. Theorem 2 below characterizes the activation functions for which the infinitely wide and deep network achieves optimality. In particular, we establish the optimality of the classifier, $m_n(\cdot)$, given by taking the limit as $L \to \infty$ of the kernel machine in Eq. (2) with $K = K^{(L)}$, i.e.

$$m_n(x) = \lim_{L\to\infty} \operatorname{sign}\Big(y\big(K_n^{(L)}\big)^{-1} K^{(L)}(X, x)\Big). \tag{3}$$

Theorem 2. *Let m_n denote the classifier in Eq. (3) corresponding to training an infinitely wide and deep network with activation function $\phi(\cdot)$ on n training examples. For $z \sim \mathcal{N}(0,1)$, define $A = \mathbb{E}[\phi(z)]$, $A' = \mathbb{E}[\phi'(z)]$, and $B' = \mathbb{E}[\phi'(z)^2]$. If*

$$A = 0 \quad \textit{and} \quad A' \neq 0 \quad \textit{and} \quad -\frac{\log(A'^2)}{\log(B')} = \frac{d}{2},$$

then this classifier is Bayes optimal.[5]

While the full proof of Theorem 2 is presented in Supplementary Information B and C, we outline its key steps in Section 3. In particular, the proof follows by using Theorem 1 above, proving that m_n is a singular kernel classifier, and then using the results of [9], which establish conditions under which singular kernel estimators achieve optimality. The following corollary (proof in Supplementary Information D) presents a concrete class of activation functions that satisfy the conditions of Theorem 2 for any given data dimension d.

Corollary 1. *Let m_n denote the classifier in Eq. (3) corresponding to training an infinitely wide and deep network with activation function*

$$\phi(x) = \begin{cases} \frac{1}{12\sqrt{70}} h_7(x) + \frac{1}{\sqrt{2}} x & \textit{if } d = 1, \\ \frac{1}{2^{d/4}} \left(\frac{x^3 - 3x}{\sqrt{6}}\right) + \sqrt{1 - \frac{2}{2^{d/2}}} \left(\frac{x^2-1}{\sqrt{2}}\right) + \frac{1}{2^{d/4}} x & \textit{if } d \geq 2, \end{cases}$$

where $h_7(x)$ is the 7^{th} probabilist's Hermite polynomial.[6] *Then the classifier m_n is Bayes optimal.*

We note the remarkable simplicity of the above activation functions yielding infinitely wide and deep networks that achieve optimality. In particular, for $d \geq 2$, these activations are simply cubic polynomials.

Case 2 ($A = 0, A' = 0$) networks implement 1-NN.

We now identify conditions on the activation function under which infinitely wide and deep networks implement the 1-NN classifier.

Theorem 3. *Let m_n denote the classifier in Eq. (3) corresponding to training an infinitely wide and deep network with activation function $\phi(\cdot)$ on n training examples. For $z \sim \mathcal{N}(0,1)$, define $A = \mathbb{E}[\phi(z)]$ and $A' = \mathbb{E}[\phi'(z)]$. If $A = A' = 0$, then $m_n(x)$ implements 1-NN classification on $\mathcal{S}_+^d$.*

The proof of Theorem 3 is provided in Supplementary Information E. The proof strategy is to show that the value of the kernel between a test example and its nearest training example dominates the prediction as $L \to \infty$. In particular, assuming without loss of generality that $x^T x^{(1)} > x^T x^{(j)}$ for $j \in \{2, 3, \ldots, n\}$, we prove that:

$$\lim_{L\to\infty} \frac{K^{(L)}(x, x^{(j)})}{K^{(L)}(x, x^{(1)})} = 0 \ .$$

As a result, after re-scaling by $K^{(L)}(x, x^{(1)})$, we obtain that $m_n(x) = \operatorname{sign}(y^{(1)})$. We note that this proof is analogous to the standard proof that the Gaussian kernel $K(x, \tilde{x}) = \exp\left(-\gamma \|x - \tilde{x}\|^2\right)$ converges to the 1-NN classifier as $\gamma \to \infty$.

[5] Formally, this classifier satisfies that for almost all $x \in \mathcal{S}_+^d$ and for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}_X \left(\left| m_n(x) - \operatorname*{arg\,max}_{\tilde{y}\in\{-1,1\}} \mathbb{P}\left(y = \tilde{y} | x\right) \right| > \epsilon \right) = 0 \ .$$

[6] For $d = 1$, this activation function can be written in closed form as $\frac{x^7 - 21x^5 + 105x^3 + (12\sqrt{35} - 105)x}{12\sqrt{70}}$.

Case 3 ($A \neq 0$) networks implement majority vote classifiers.

We now analyze infinitely wide and deep networks when the activation functon satisfies $\mathbb{E}[\phi(z)] \neq 0$ for $z \sim \mathcal{N}(0,1)$. In this setting, we establish conditions under which the infinitely wide and deep network implements majority vote classification, i.e., the prediction on test samples is simply the label of the class with greatest representation in the training set. More precisely, the following proposition (proof in Supplementary Information F) implies that when the infinite depth NTK is a constant non-zero value for any two non-equal inputs, the resulting classifier is the majority vote classifier.

Proposition 1. *Let m_n denote the classifier in Eq. (3) corresponding to training an infinitely wide and deep network with activation function $\phi(\cdot)$ on n training examples. For any $x, \tilde{x} \in \mathcal{S}_+^d$ with $x \neq \tilde{x}$, if the NTK $K^{(L)}$ satisfies*

$$\lim_{L\to\infty} \frac{K^{(L)}(x,\tilde{x})}{C(L)} = C_1 \quad \textit{and} \quad \lim_{L\to\infty} \frac{K^{(L)}(x,\tilde{x})}{C(L)} \neq \lim_{L\to\infty} \frac{K^{(L)}(x,x)}{C(L)}, \tag{4}$$

with $C_1 > 0$ and $0 < C(L) < \infty$ for any L, then m_n implements the majority vote classifier, i.e.,

$$m_n(x) = \text{sign}\left(\sum_{i=1}^{n} y^{(i)}\right) .$$

We now analyze which activation functions satisfy Eq. (4). As described in [12, 18, 28, 37], under case 3, the value of $B' = \mathbb{E}[\phi'(z)^2]$ for $z \sim \mathcal{N}(0,1)$ determines the fixed point attractors of $K^{(L)}$ as $L \to \infty$. Thus, the infinite depth behavior under case 3 can be broken down into three cases based on the value of B'. Using the terminology from [28], these cases are:

$$(i)\ B' > 1 \text{ (Chaotic Phase)}, \quad (ii)\ B' < 1 \text{ (Ordered Phase)}, \quad (iii)\ B' = 1 \text{ (Edge of Chaos)}.$$

In Lemma 6 in Supplementary Information G, we demonstrate that in the chaotic phase, the resulting infinite depth NTK satisfies the conditions of Proposition 1 and thus implements the majority vote classifier. In Lemma 7 in Supplementary Information G, we similarly show that in the ordered phase the infinite depth NTK also corresponds to the majority vote classifier.[7] The remaining case known as "edge of chaos" has been analyzed in prior works for specific activation functions; for example, the NTK for networks with ReLU activation satisfies Eq. (4) with $C_1 = \frac{1}{4}$ and $C(L) = L+1$ [13, 16]. Hence by Proposition 1, the corresponding infinite depth classifier for ReLU networks corresponds to the majority vote classifier.

3 Outline of Proof Strategy for Theorems 1 and 2

In the following, we outline the proof strategy for our main results. This involves analyzing infinitely wide and deep networks via the limiting NTK kernel given by $K^{(L)}$ as the number of hidden layers $L \to \infty$. As shown in [17], $K^{(L)}$ can be written recursively in terms of $K^{(L-1)}$ and the so-called dual activation function, which was introduced in [8].

Definition 2. *Let $\phi : \mathbb{R} \to \mathbb{R}$ be an activation function satisfying $\mathbb{E}_{x\sim\mathcal{N}(0,1)}[\phi(x)^2] < \infty$. Its **dual activation function** $\check{\phi} : [-1,1] \to \mathbb{R}$ is given by*

$$\check{\phi}(z) = \mathbb{E}_{(u,v)\sim\mathcal{N}(\mathbf{0},\Lambda)}[\phi(u)\phi(v)], \quad \textit{where } \Lambda = \begin{bmatrix} 1 & z \\ z & 1 \end{bmatrix}.$$

While all quantities in our theorems are stated in terms of activation functions, these can be restated in terms of dual activations as follows:

$$A^2 = \check{\phi}(0) \quad \text{and} \quad (A')^2 = \check{\phi}'(0) \quad \text{and} \quad B' = \check{\phi}'(1) .$$

[7] More precisely, we consider the behavior of the infinite depth classifier under ridge-regularization, as the regularization term approaches 0.

Assuming that ϕ is normalized such that $\check{\phi}(1) = 1$,[8] the recursive formula for the NTK of a deep fully connected network for data on the unit sphere was described in [11, 17] in terms of dual activation functions as follows.

Recursive Formula for the NTK. Let $f^{(L)}(x; \mathbf{W})$ denote a fully connected neural network with L hidden layers and activation $\phi(\cdot)$. For $x, \tilde{x} \in \mathcal{S}^d$, let $z = x^T\tilde{x}$. Then $K^{(L)}$ is radial, i.e. $K^{(L)}(x, \tilde{x}) = K^{(L)}(z)$, with

$$K^{(L)}(z) = \check{\phi}^{(L)}(z) + K^{(L-1)}(z)\check{\phi}'\big(\check{\phi}^{(L-1)}(z)\big) \quad \text{and} \quad K^{(0)}(z) = z, \tag{5}$$

where $\check{\phi}^{(L)}(z) = \check{\phi}(\check{\phi}^{(L-1)}(z))$ with $\check{\phi}^{(0)}(z) = \check{\phi}(z)$ and $\check{\phi}'(\cdot)$ denotes the derivative of $\check{\phi}(\cdot)$.

We utilize the dynamical system in Eq. (5) to analyze the behavior of $K^{(L)}(\cdot)$ as $L \to \infty$. Theorem 1 implies that upon normalization by $(L+1)\check{\phi}'(0)^L$, this dynamical system converges to a singular kernel with singularity of order $\alpha = -\log\big(\check{\phi}'(0)\big) / \log\big(\check{\phi}'(1)\big)$. We now present a sketch of the proof of this result.

We first derive the order of the singularity upon iteration of $\check{\phi}$, since as we show in Supplementary Information B, the order of the singularity of the infinite depth NTK is the same as that of the iterated $\check{\phi}$. Since we consider data in $\mathcal{S}^d_+$, $\check{\phi}(\cdot)$ is a function defined on the unit interval $[0, 1]$. Hence, understanding the properties of infinitely wide and deep networks reduces to understanding the properties of iterating a function on the unit interval. To provide intuition around how the iteration of a function on the unit interval can give rise to a function with a singularity, we discuss iterating a piecewise linear function as an illuminating example; see Fig. 2 for a visualization.

Lemma 1. *For $0 < a < 1$ and $b > 1$, let $f : [0, 1] \to \mathbb{R}$ and $c = \frac{b-1}{b-a}$ such that*

$$f(x) = \begin{cases} ax & \text{if } x \in [0, c] \\ 1 - b(1-x) & \text{if } x \in (c, 1] \end{cases}.$$

Then,

$$\lim_{L\to\infty} \frac{f^{(L)}(x)}{a^L} = \frac{R(x)}{(1-x)^{-\log_b a}},$$

where $R(x)$ is non-negative, bounded from above and bounded away from 0 around $x = 1$.

Proof. For any $x \in [0, c]$, we necessarily have:

$$\lim_{L\to\infty} \frac{f^{(L)}(x)}{a^L} = \lim_{L\to\infty} \frac{a^L x}{a^L} = x.$$

Now for fixed $x \in (c, 1)$, since $x = 0$ is an attractive fixed-point of f, let L_0 denote the smallest integer such that $f^{(L_0)}(x) \le c$. Hence, since $f^{(L_0)}(x) \in [0, c]$, we obtain:

$$\lim_{L\to\infty} \frac{f^{(L)}(x)}{a^L} = \lim_{L\to\infty} \frac{f^{(L-L_0)}(f^{(L_0)}(x))}{a^{L-L_0}} \frac{1}{a^{L_0}} = f^{(L_0)}(x) a^{-L_0}. \tag{6}$$

We next solve for L_0 by analyzing the iteration of $g(x) := 1 - b(1-x)$. In particular, we observe that $g^{(L)}(x) = 1 - b^L(1-x)$, and thus L_0 is given by:

$$1 - b^{L_0}(1-x) \le c \implies L_0 = \left\lceil \log_a \left(\frac{1-x}{1-c}\right)^{-\log_b a} \right\rceil \implies a^{-L_0} \in \left[\left(\frac{1-c}{1-x}\right)^{-\log_b a}, \frac{1}{a}\left(\frac{1-c}{1-x}\right)^{-\log_b a} \right].$$

Hence, by Eq. (6) we conclude that for $x \in (c, 1)$, it holds that

$$\lim_{L\to\infty} \frac{f^{(L)}(x)}{a^L} = \frac{R(x)}{(1-x)^{-\log_b a}},$$

where $R(x)$ is non-negative, bounded from above and bounded away from 0 around $x = 1$, which completes the proof. □

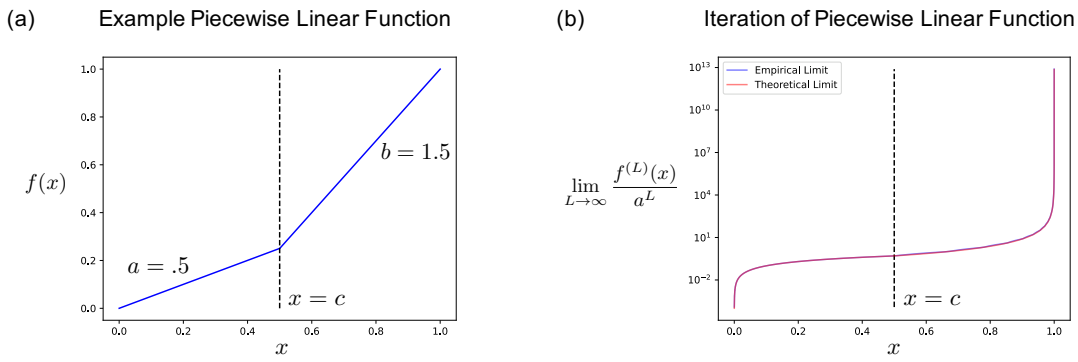

Figure 2: Iteration of a piecewise linear function on a unit interval leads to a function with a singularity at $x = 1$, upon appropriate normalization. (a) We consider the piecewise linear function $f(x)$ given by $1 - b(1 - x)$ on $(c, 1]$ and ax on $[0, c]$, where $a = .5, b = 1.5$ and $c = \frac{b-1}{b-a}$. (b) We observe that upon iterating $f(\cdot)$ numerically to the limit of machine precision, the resulting function strongly agrees with the theoretical limit of Lemma 1 given by a function with singularity of order $-\log_b a \approx 1.7$.

In Supplementary Information B, we extend this analysis to the iteration of dual activations on the unit interval, thereby establishing the order of a singularity obtained by iterating dual activation functions. We then show that this order equals the order of the singularity given by the infinite depth NTK.

Next, we discuss the proof strategy for Theorem 2, which establishes conditions on the activation function under which infinitely wide and deep networks achieve optimality in the classification setting. The proof builds on results in [9] characterizing the optimality of singular kernel smoothers of the form

$$g(x) = \frac{\sum_{i=1}^n y^{(i)} K(x^{(i)},\ x)}{\sum_{i=1}^n K(x^{(i)},\ x)}, \quad \text{where } K(x^{(i)}, x) = \frac{1}{\|x - x^{(i)}\|^\alpha}.$$

In particular, it is shown that if $\alpha = d$, then $g(x)$ achieves optimality. Since Theorem 1 establishes conditions under which the infinite depth NTK implements a singular kernel, to complete the proof we show that infinitely wide and deep classifiers achieve optimality by (1) showing that the classifier m_n implements a singular kernel smoother, and (2) selecting ϕ such that $\alpha = d$ for the corresponding singular kernel.

4 Discussion

In this work, we identified and constructed explicit neural networks that achieve optimality for classification when trained using standard procedures. Furthermore, we provided a taxonomy characterizing the behavior of infinitely wide and deep neural network classifiers. Namely, we showed that these models implement one of the following three well-known types of classifiers: (1) 1-NN (test predictions are given by the label of the nearest training example) ; (2) majority vote (test predictions are given by the label of the class with greatest representation in the training set); or (3) singular kernel classifiers (a set of classifiers containing those that achieve optimality). We conclude by discussing implications of our work and future extensions.

Benefit of Depth in Neural Networks. An emerging trend in machine learning is that larger neural networks capable of interpolating (i.e., perfectly fitting) the training data, can generalize to test data [2, 25, 38]. While the size of neural networks can be increased through width or depth, works such as [2, 25] primarily identified a benefit to increasing network width. Indeed, it remained unclear whether there was any benefit to using extremely deep networks. For example, recent works [26, 35, 36] empirically demonstrated that drastically increasing depth in networks with ReLU or tanh activation could lead to worse performance. In this work, we established a remarkable benefit of very deep networks by proving that they achieve optimality with a careful choice of activation function. In line with previous empirical findings, we proved that deep networks with activations such as ReLU or tanh do not achieve optimality.

[8]Such normalization is always possible for any activation function satisfying $\mathbb{E}[\phi(z)^2] < \infty$ for $z \sim \mathcal{N}(0, 1)$ and has been used in various works before including [11, 12, 13, 16, 21, 36].

Regression versus Classification. Our results demonstrate the benefit of using infinitely wide and deep networks for classification tasks. We note that this in stark contrast to the regression setting, where infinitely deep and wide neural networks are far from optimal, as they simply predict a non-negative constant almost everywhere [13, 16]. Thus, our work provides concrete examples of neural networks that are effective for classification but not regression.

Edge of Chaos Regime. An interesting class of models that are only partially characterized by our taxonomy corresponds to networks with activations in the edge of chaos regime, i.e., when the activation function, $\phi(\cdot)$ satisfies $\mathbb{E}[\phi(z)] \neq 0$ and $\mathbb{E}[\phi'(z)^2] = 1$ for $z \sim \mathcal{N}(0, 1)$. We proved that all activations in this class that have been described so far [13, 16], including the popular ReLU activation, give rise to infinitely wide and deep networks that implement the majority vote classifier. While it appears that all activations in this class lead to the majority vote classifier, it remains open to understand whether there exist other activations in this regime that implement alternative classifiers.

Finite vs. Infinite Neural Networks. In this work, we identified and constructed infinitely wide and deep classifiers that achieve optimality. An important next question is to understand whether interpolating neural networks that are finitely wide and deep can achieve optimality for classification and provide specific activation functions to do so. We also note that Bayes optimality considers the setting when the number of training examples approaches infinity. Another natural next step is to characterize the number of training examples needed for infinitely wide and deep classifiers to reasonably approximate the Bayes optimal classifier. Recent work [24] identified a slow (logarithmic) rate of convergence for singular kernel classifiers, thereby implying that many training examples are needed for these models to be effective in practice. An important open direction of future work is thus to determine not only whether finitely wide and deep networks are optimal for classification but also whether these models require fewer samples to perform well in practice.

Acknowledgements

A.R. and C.U. were partially supported by NSF (DMS-1651995), ONR (N00014-17-1-2147 and N00014-18-1-2765), the MIT-IBM Watson AI Lab, the Eric and Wendy Schmidt Center at the Broad Institute, and a Simons Investigator Award (to C.U.). M.B. acknowledges support from NSF IIS-1815697 and NSF DMS-2031883/Simons Foundation Award 814639.

References

[1] S. Arora, S. S. Du, W. Hu, Z. Li, R. Salakhutdinov, and R. Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, 2019.

[2] M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[3] M. Belkin, D. Hsu, and P. Mitra. Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate. In *Advances in Neural Information Processing Systems*, 2018.

[4] M. Belkin, A. Rakhlin, and A. Tsybakov. Does data interpolation contradict statistical optimality? In *International Conference on Artificial Intelligence and Statistics*, 2018.

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, and D. Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

[6] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[7] G. Cybenko. Approximation by superposition of sigmoidale function. *Mathematics of Control Signal and Systems*, 2:304–314, 01 1989.

[8] A. Daniely, R. F. Frostig, and Y. Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In *Advances in Neural Information Processing Systems*, 2016.

[9] L. Devroye, L. Györfi, and A. Krzyżak. The Hilbert kernel regression estimate. *Journal of Multivariate Analysis*, 65:209–227, 1998.

[10] A. Faragó and G. Lugosi. Strong universal consistency of neural network classifiers. *IEEE Transactions on Information Theory*, 39(4), 1993.

[11] A. Geifman, A. Yadav, Y. Kasten, M. Galun, D. Jacobs, and R. Basri. On the similarity between the Laplace and Neural Tangent Kernels. In *Advances in Neural Information Processing Systems*, 2020.

[12] S. Hayou, A. Doucet, and J. Rousseau. On the impact of the activation function on deep neural networks training. In *International Conference on Machine Learning*, 2019.

[13] S. Hayou, A. Doucet, and J. Rousseau. Mean-field behaviour of Neural Tangent Kernel for deep neural networks. *arXiv:1905.13654*, 2021.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*, 2016.

[15] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1989.

[16] K. Huang, Y. Wang, M. Tao, and T. Zhao. Why do deep residual networks generalize better than deep feedforward networks? — A Neural Tangent Kernel perspective. In *Advances in Neural Information Processing Systems*, 2020.

[17] A. Jacot, F. Gabriel, and C. Hongler. Neural Tangent Kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018.

[18] J. Lee, Y. Bahri, R. Novak, S. Schoenholz, J. Pennington, and J. Sohl-Dickstein. Deep neural networks as Gaussian processes. In *International Conference on Learning Representations*, 2017.

[19] J. Lee, S. Schoenholz, J. Pennington, B. Adlam, L. Xiao, R. Novak, and J. Shol-Dickstein. Finite versus infinite neural networks: an empirical study. In *Advances in Neural Information Processing Systems*, 2020.

[20] J. Lee, L. Xiao, S. Schoenholz, Y. Bahri, J. Sohl-Dickstein, and J. Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, 2019.

[21] T. Liang and H. Tran-Bach. Mehler's formula, branching process, and compositional kernels of deep neural networks. *Journal of the American Statistical Association*, pages 1–35, 11 2020.

[22] C. Liu, L. Zhu, and M. Belkin. On the linearity of large non-linear models: when and why the tangent kernel is constant. In *Advances in Neural Information Processing Systems*, 2020.

[23] C. Liu, L. Zhu, and M. Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 01 2022.

[24] P. P. Mitra and C. Sire. Parameter-free statistically consistent interpolation: Dimension-independent convergence rates for Hilbert kernel regression. *arXiv:2106.03354*, 2021.

[25] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference in Learning Representations*, 2020.

[26] E. Nichani, A. Radhakrishnan, and C. Uhler. Increasing depth leads to U-shaped test risk in over-parameterized convolutional networks. In *International Conference on Machine Learning Workshop on Over-parameterization: Pitfalls and Opportunities*, 2021.

[27] R. Novak, L. Xiao, J. Hron, J. Lee, A. A. Alemi, J. Sohl-Dickstein, and S. S. Schoenholz. Neural Tangents: Fast and easy infinite neural networks in Python. In *International Conference on Learning Representations*, 2020.

[28] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances in Neural Information Processing Systems*, 2016.

[29] A. Radhakrishnan, G. Stefanakis, M. Belkin, and C. Uhler. Simple, fast, and flexible framework for matrix completion with infinite width neural networks. *Proceedings of the National Academy of Sciences*, 119(16):e2115064119, 2022.

[30] A. Rakhlin and X. Zhai. Consistency of interpolation with Laplace kernels is a high-dimensional phenomenon. In *Conference on Learning Theory*, 2019.

[31] B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, 2002.

[32] E. M. Stein and R. Shakarchi. *Complex Analysis*, volume II. Princeton University Press, 2003.

[33] S. Strogatz. *Nonlinear Dynamics and Chaos*, volume 2. Westview Press, 2015.

[34] K. Tunyasuvunakool, J. Adler, Z. Wu, T. Green, M. Zielinski, A. Žídek, A. Bridgland, A. Cowie, C. Meyer, A. Laydon, S. Velankar, G. Kleywegt, A. Bateman, R. Evans, A. Pritzel, M. Figurnov, O. Ronneberger, R. Bates, S. Kohl, and D. Hassabis. Highly accurate protein structure prediction for the human proteome. *Nature*, 596:1–9, 2021.

[35] L. Xiao, Y. Bahri, J. Sohl-Dickstein, S. Schoenholz, and J. Pennington. Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks. In *International Conference on Machine Learning*, 2018.

[36] L. Xiao, J. Pennington, and S. Schoenholz. Disentangling trainability and generalization in deep learning. In *International Conference on Machine Learning*, 2019.

[37] G. Yang and S. Schoenholz. Mean field residual networks: On the edge of chaos. In *Advances in Neural Information Processing Systems*, 2017.

[38] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Supplementary Information

A Preliminaries on NTK and Dual Activations

In this section, we briefly review properties of dual activations that we will use to prove our main results. In order to analyze the behavior of the iterated dual activation, we reference the following result of [8], which implies that the dual activation is analytic around 0 on the interval $[-1, 1]$.

Analyticity of Dual Activations. Let $\phi(\cdot)$ be an activation function such that $\mathbb{E}_{x\sim\mathcal{N}(0,1)}[\phi(x)^2] = 1$, and let $\check{\phi} : [-1, 1] \to \mathbb{R}$ denote the dual activation. Then, for $z \in [-1, 1]$,

$$\check{\phi}(z) = \sum_{i=0}^{\infty} a_i z^i, \tag{7}$$

where $a_i \geq 0$ for all $i \in \mathbb{N}$.

As proven in [8, Lemma 11], several key properties are implied by Eq. (7). Those utilized in this work are: (1) $\check{\phi}$ is increasing on $[0, 1]$, and (2) non-negativity of $\check{\phi}(\cdot)$ on $[0, 1]$. Eq. (7) also implies the following property of dual activations that we will use to construct our taxonomy of infinitely wide and deep neural network classifiers.

Lemma 2. *Let $\check{\phi} : [-1, 1] \to \mathbb{R}$ be a dual activation such that $\check{\phi}(0) = 0$, $\check{\phi}(1) = 1$, and $\check{\phi}(z) \neq z$. Then, $0 \leq \check{\phi}'(0) < 1$.*

Proof. By Eq. (7), we need only show that $0 \leq a_1 < 1$. Since $\check{\phi}(1) = 1$, we obtain that $\sum_{i=1}^{\infty} a_i = 1$. Since $a_i \geq 0$ for all $i \in \mathbb{N}$, we conclude that $0 \leq a_1 \leq 1$. Now if $a_1 = 1$, then $a_i = 0$ for $i \geq 2$, which implies that $\check{\phi}(z) = z$. Hence, we conclude that $0 \leq a_1 < 1$, which completes the proof. □

B Proofs of Theorem 1 and Theorem 2

We first prove Theorem 1, which is expressed below in terms of the dual activation function.

Theorem. *Let $K^{(L)}$ denote the NTK of a fully connected neural network with L hidden layers and activation function $\phi(\cdot)$. For $x, \tilde{x} \in \mathcal{S}_+^d$, let $z = x^T\tilde{x}$. If the dual activation function $\check{\phi}(\cdot)$ satisfies*

1) $\check{\phi}(0) = 0$, $\check{\phi}(1) = 1$,

2) $0 < \check{\phi}'(0) < 1$ and $\check{\phi}'(1) < \infty$,

then:

$$\lim_{L\to\infty} \frac{K^{(L)}(x, \tilde{x})}{\check{\phi}'(0)^L(L + 1)} = \frac{R(x^T\tilde{x})}{\|x - \tilde{x}\|^\alpha} ,$$

where $\alpha = -2\frac{\log(\check{\phi}'(0))}{\log(\check{\phi}'(1))}$ and $R(u) \geq 0$ is bounded for $u \in [0, 1]$ and bounded away from 0 around $u = 1$.

In order to prove this theorem, we first prove that the iterated, normalized NTK converges to a singular kernel without explicitly identifying the order of the singularity.

Lemma 3. *Let $K^{(L)}$ denote the NTK of a depth L fully connected network with normalized activation function ϕ. Assuming $\check{\phi}$ satisfies the conditions of Theorem 1, then for any $x, \tilde{x} \in \mathcal{S}_+^d$ it holds that*

$$\lim_{L\to\infty} \frac{K^{(L)}(x, \tilde{x})}{a_1^L(L + 1)} = \psi(x^T\tilde{x}),$$

where $\psi : [0, 1] \to \mathbb{R}$ can be written as a power series with non-negative coefficients with a singularity at 1.

Proof. We utilize the form of the NTK given in [1] and utilize the radial form of the kernel in Eq. (5). Namely, for $z \in [0,1]$, we have:

$$K^{(L)}(z) = \sum_{i=0}^{L} \check{\phi}^{(i)}(z) \prod_{j=i}^{L-1} \check{\phi}'\left(\check{\phi}^{(j)}(z)\right), \tag{8}$$

where $\check{\phi}^{(i)}$ denotes the iteration of $\check{\phi}$ i times. By Eq. (7) and since $\check{\phi}(0) = 0$, we have that $\check{\phi}(z) = \sum_{i=1}^{\infty} a_i z^i$ for all $z \in [0,1]$.[9] Now, we bound $\check{\phi}$ by quadratic functions in z and bound $\check{\phi}'$ by linear functions in z. In particular, using the conditions $\check{\phi}(1) = 1$ and $\check{\phi}'(1) = C < \infty$, we obtain the upper bounds:

$$\check{\phi}(z) = a_1\left(z + \sum_{i=2}^{\infty} \frac{a_i}{a_1} z^i\right) \leq a_1\left(z + \sum_{i=2}^{\infty} \frac{a_i}{a_1} z^2\right) = a_1\left(z + \left(\frac{1}{a_1} - 1\right) z^2\right),$$
$$\check{\phi}'(z) = a_1\left(1 + \sum_{i=2}^{\infty} i\frac{a_i}{a_1} z^{i-1}\right) \leq a_1\left(1 + \sum_{i=2}^{\infty} \frac{a_i}{a_1} iz\right) = a_1\left(1 + \left(\frac{C}{a_1} - 1\right) z\right).$$

Similarly, we obtain the lower bounds:

$$\check{\phi}(z) = a_1\left(z + \sum_{i=2}^{\infty} \frac{a_i}{a_1} z^i\right) \geq a_1\left(z + \frac{a_2}{a_1} z^2\right),$$
$$\check{\phi}'(z) = a_1\left(1 + \sum_{i=2}^{\infty} i\frac{a_i}{a_1} z^{i-1}\right) \geq a_1\left(1 + \frac{2a_2}{a_1} z\right) \geq a_1\left(1 + \frac{a_2}{a_1} z\right).$$

Now, substituting the above lower and upper bounds into the recursion for $\check{\phi}^{(i)}$, we obtain

$$a_1^i z \prod_{j=0}^{i-1}\left(1 + \frac{a_2}{a_1}\check{\phi}^{(j)}(z)\right) \leq \check{\phi}^{(i)}(z) \leq a_1^i z \prod_{j=0}^{i-1}\left(1 + \left(\frac{1}{a_1} - 1\right)\check{\phi}^{(j)}(z)\right). \tag{9}$$

Lastly, since $C \geq 1$, substituting Eq.(9) and the bounds on $\check{\phi}'$ into Eq. (8) for $K^{(L)}$, we obtain

$$(L+1)a_1^L z \prod_{j=0}^{L-1}\left(1 + \frac{a_2}{a_1}\check{\phi}^{(j)}(z)\right) \leq K^{(L)}(z) \leq (L+1)a_1^L z \prod_{j=0}^{L-1}\left(1 + \left(\frac{C}{a_1} - 1\right)\check{\phi}^{(j)}(z)\right).$$

Hence, to prove that $\psi(z) := \lim_{L\to\infty} \frac{K^{(L)}(z)}{a_1^L(L+1)}$ is finite for $z \in [0,1)$, we need to show that

$$\prod_{j=0}^{\infty}\left(1 + \tilde{C}\check{\phi}^{(j)}(z)\right) < \infty$$

for all $z \in [0,1)$ and any constant $\tilde{C}$. By the Cauchy criterion [32, Ch.5], the above infinite product converges if and only if the following sum converges:

$$\sum_{j=0}^{\infty} \tilde{C}\check{\phi}^{(j)}(z) < \infty.$$

This sum converges by the ratio test. In particular,

$$\lim_{j\to\infty} \frac{\check{\phi}^{(j)}(z)}{\check{\phi}^{(j-1)}(z)} = \lim_{z\to 0} \frac{\check{\phi}(z)}{z} = a_1 < 1,$$

where we used the contractive mapping theorem [33] to establish the first equality, since 0 is a fixed point attractor of $\check{\phi}$. As a consequence, $\psi(z) < \infty$ for $z \in [0,1)$. Now according to Eq. (8), $\psi(z)$ can be written as

[9] Note that the sum starts from a_1 since $\check{\phi}(0) = 0 \implies a_0 = 0$.

a convergent power series with non-negative coefficients for $z \in [0,1)$. To establish the singularity of $\psi(z)$ at $z = 1$, we show that for any constant $R > 0$, there exists z_0 such that $\psi(z) > R$ for $z > z_0$. In particular, note that for any fixed L_0,

$$\lim_{L\to\infty} \frac{K^{(L)}(z)}{a_1^L(L+1)} = \psi(z) \geq z \prod_{j=0}^{L_0-1} \left(1 + \frac{a_2}{a_1}\check{\phi}^{(j)}(z)\right).$$

The right-hand side is a continuous function with maximum value $\left(1 + \frac{a_2}{a_1}\right)^L$. Hence, by selecting L_0 such that $\left(1 + \frac{a_2}{a_1}\right)^{L_0} > R$, we can then pick z_0 such that $\psi(z) > R$ for all $z > z_0$. Hence, we conclude that

$$\lim_{L\to\infty} \frac{K^{(L)}(z)}{a_1^L(L+1)} = \psi(z),$$

where $\psi(z)$ can be written as a convergent power series with non-negative coefficients on $[0,1)$ with a singularity at $z = 1$, which completes the proof. □

We will now prove Theorem 1 by establishing the order of the singularity of ψ from Lemma 3. To characterize the order of this singularity, we will generally characterize the order of the singularity arising from iterating functions on the interval $[0,1]$. In particular, we begin by establishing the order of the singularity of the normalized iteration of a function that is linear around $x = 1$.

Lemma 4. *Let*

$$f(z) = \begin{cases} g(z) & \text{if } z \in [0,d] \\ 1 - b(1-z) & \text{if } z \in (d,1] \end{cases},$$

with $d < 1$ such that $f(z)$ is strictly monotonically increasing and $g(z)$ can be written as a convergent power series with non-negative coefficients with $g(0) = 0$, $g'(0) = a < 1$, and $b > 1$. Then for $z \in (d,1]$, it holds that

$$\lim_{L\to\infty} \frac{f^{(L)}(z)}{a^L} = \frac{R(z)}{(1-z)^{-\log_b a}},$$

where $R(z)$ is non-negative for $z \in [0,1]$, bounded from above, and bounded away from 0 around $z = 1$.

Proof. We first visualize the curve $f(z)$ in Fig. 3a. For any $z \in (d,1]$, let $L_0(x)$ denote the smallest number of iterations until $f^{(L)}(z) = z' \leq d$. Then for $z \in (d,1)$, we have that

$$\lim_{L\to\infty} \frac{f^{(L)}(z)}{a^L} = \lim_{L\to\infty} \frac{f^{(L-L_0(z))}(z')}{a^{L-L_0(z)}} a^{-L_0(z)}.$$

Now by the proof of Lemma 3, we know that

$$\lim_{L\to\infty} \frac{f^{(L-L_0(z))}(z')}{a^{L-L_0(z)}} = \tilde{R}(z') \ ,$$

with $\tilde{R}(z') \geq z'$. Thus, we need only analyze the term $a^{-L_0(z)}$ to determine the pole order. In particular, we have that $L_0(z)$ is the least integer that satisfies:

$$1 - b^{L_0(z)}(1-z) \leq d.$$

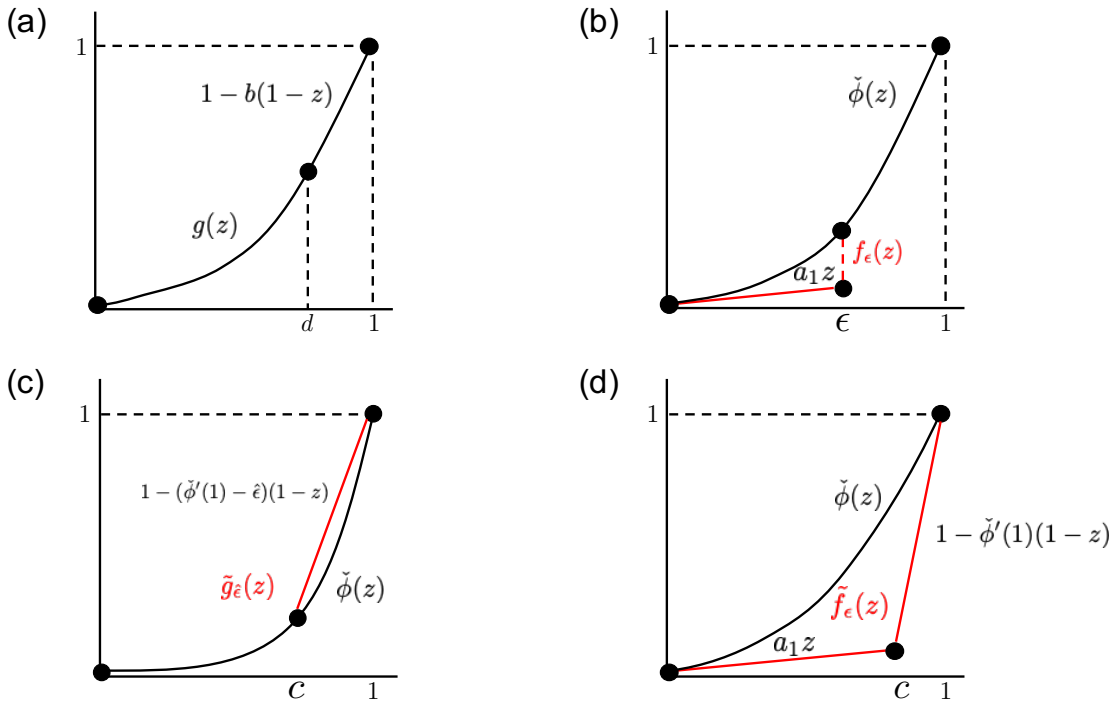

Figure 3: A visualization of the four functions bounding $\check{\phi}(z)$ that are used to prove Theorem 1.

Hence, $L_0(z)$ is given by:

$$\begin{aligned} L_0(z) &= \left\lceil \log_b \left(\frac{1-d}{1-z} \right) \right\rceil \\ &= \left\lceil \log_a \left(\frac{1-d}{1-z} \right)^{\frac{1}{\log_a b}} \right\rceil \\ &= \left\lceil \log_a \left(\frac{1-z}{1-d} \right)^{-\log_b a} \right\rceil \\ &\in \left[\log_a \left(\frac{1-z}{1-d} \right)^{-\log_b a}, \quad \log_a \left(\frac{1-z}{1-d} \right)^{-\log_b a} + 1 \right]. \end{aligned}$$

As a consequence,

$$a^{-L_0(z)} \in \left[\left(\frac{1-d}{1-z} \right)^{-\log_b a}, \frac{1}{a} \left(\frac{1-d}{1-z} \right)^{-\log_b a} \right].$$

Thus we conclude that for $z \in (d, 1)$:

$$\lim_{L \to \infty} \frac{f^{(L)}(z)}{a^L} = \frac{R(z)}{(1-z)^{-\log_b a}},$$

where $R(z)$ is non-negative for $z \in [0, 1]$, bounded from above, and bounded away from 0 around $z = 1$, which concludes the proof. □

We will now utilize Lemma 4 to prove Theorem 1.

Proof. Let $\check{\phi}(z) = \sum_{i=1}^{\infty} a_i z^i$. We will lower bound the dual activation $\check{\phi}$ and its derivative by the piecewise functions:

$$f_\epsilon(z) = \begin{cases} a_1 z & \text{if } z \in [0, \epsilon) \\ \check{\phi}(z) & \text{if } z \in [\epsilon, 1] \end{cases} \quad \text{and} \quad h_\epsilon(z) = \begin{cases} a_1 & \text{if } z \in [0, \epsilon) \\ \check{\phi}'(z) & \text{if } z \in [\epsilon, 1] \end{cases}.$$

The function $f_\epsilon(z)$ is visualized in Fig. 3b. Now consider the function $k_\epsilon^{(L)}(z)$ defined as follows:

$$k_\epsilon^{(L)}(z) = k_\epsilon^{(L-1)}(z) h_\epsilon(f_\epsilon^{(L-1)}(z)) + f_\epsilon^{(L)}(z).$$

By definition, we have that $K^{(L)}(z) \geq k_\epsilon^{(L)}(z)$ for all $z \in [0, 1]$. We will now show that for any $\tilde{\epsilon}$, we can select k_ϵ such that

$$\lim_{L \to \infty} \frac{K^{(L)}(z) - k_\epsilon^{(L)}(z)}{a_1^L (L+1)} < \tilde{\epsilon}. \tag{10}$$

To prove Eq. (10), we first consider the updates for $L > L_0$ where L_0 is the largest integer such that $k^{(L_0)}(z) = K^{(L_0)}(z)$, $h_\epsilon(f_\epsilon^{(L_0+1)}(z)) = a_1$, and $f_\epsilon^{(L_0)}(z) = \check{\phi}^{(L_0)}(z)$. We will first prove inductively that for $T \in \mathbb{N}$:

$$\check{\phi}^{(L_0+T)}(z) - f_\epsilon^{(L_0+T)}(z) \leq C_1(z) \left(\sum_{i=0}^{T-1} a_1^{2L_0+T-1+i} \right), \tag{11}$$

where $C_1(z)$ is a term independent of T. We begin with the base case of $T = 1$. Namely, we have for $z \in (\epsilon, 1)$:

$$\begin{aligned}
\check{\phi}^{(L_0+1)}(z) - f_\epsilon^{(L_0+1)}(z) &\leq \sum_{i=1}^{\infty} a_i \left(\check{\phi}^{(L_0)}(z) \right)^i - a_1 f_\epsilon^{(L_0)}(z) \\
&= \sum_{i=2}^{\infty} a_i \left(\check{\phi}^{(L_0)}(z) \right)^i \qquad \left(\text{since } f_\epsilon^{(L_0)}(z) = \check{\phi}^{(L_0)}(z) \right) \\
&\leq a_1^{2L_0} \tilde{C}_1(z) (1 - a_1) \qquad \left(\text{where } \tilde{C}_1(z) = \left(\lim_{L \to \infty} \frac{\check{\phi}^{(L)}(z)}{a_1^L} \right)^2 \right) \\
&= C_1(z) a_1^{2L_0},
\end{aligned}$$

which concludes the base case. Now assume the statement is true for $T = T_0$. Then for $T = T_0 + 1$, we have:

$$\begin{aligned}
\check{\phi}^{(L_0+T_0+1)}(z) - f_\epsilon^{(L_0+T_0+1)}(z) &= \sum_{i=1}^{\infty} a_i \left(\check{\phi}^{(L_0+T_0)}(z) \right)^i - a_1 f_\epsilon^{(L_0+T_0)}(z) \\
&\leq C_1(z) \left(\sum_{i=0}^{T_0-1} a_1^{2L_0+T_0+i} \right) + \sum_{i=2}^{\infty} a_i \left(\check{\phi}^{(L_0+T_0)}(z) \right)^i \\
&\leq C_1(z) \left(\sum_{i=0}^{T_0-1} a_1^{2L_0+T_0+i} \right) + C_1(z) a_1^{2L_0+2T_0} \\
&= C_1(z) \left(\sum_{i=0}^{T_0} a_1^{2L_0+T_0+i} \right),
\end{aligned}$$

which concludes the proof by induction. We will next prove inductively that for $T \in \mathbb{N}$:

$$K^{(L_0+T)}(z) - k_\epsilon^{(L_0+T)}(z) \leq C_1(z) \left(\sum_{i=0}^{T-1} (T-i) a_1^{2L_0+T-1+i} \right) + C_2(z) \left(\sum_{i=0}^{T-2} (L_0+i+2) a_1^{2L_0+T+i} \right), \tag{12}$$

where $C_1(z), C_2(z)$ are terms independent of T. We begin with the base case of $T = 1$. Namely, we have for $z \in (\epsilon, 1)$:

$$\begin{aligned} K^{(L_0+1)}(z) - k_\epsilon^{(L_0+1)}(z) &\leq [K^{(L_0)}(z)\check{\phi}'(\check{\phi}^{(L_0)}(z)) - k_\epsilon^{(L_0)}(z)h_\epsilon(f_\epsilon^{(L_0)}(z))] + [\check{\phi}^{(L_0+1)}(z) - f_\epsilon^{(L_0+1)}(z)] \\ &= \check{\phi}^{(L_0+1)}(z) - f_\epsilon^{(L_0+1)}(z) \\ &\leq C_1(z)a_1^{2L_0} \quad \text{(by Eq. (11))}\,, \end{aligned}$$

which concludes the base case. Now, assume that Eq. (12) holds for $T = T_0$. Then for $T = T_0 + 1$, we have:

$$\begin{aligned} K^{(L_0+T_0+1)}(z) - k_\epsilon^{(L_0+T_0+1)}(z) &\leq [K^{(L_0+T_0)}(z)\check{\phi}'(\check{\phi}^{(L_0+T_0)}(z)) - k_\epsilon^{(L_0+T_0)}(z)h_\epsilon(f_\epsilon^{(L_0+T_0)}(z))] \\ &\quad + [\check{\phi}^{(L_0+T_0+1)}(z) - f_\epsilon^{(L_0+T_0+1)}(z)] \\ &= \left[K^{(L_0+T_0)}(z)\sum_{i=1}^{\infty} ia_i(\check{\phi}^{(L_0+T_0)}(z))^{i-1} - a_1 k_\epsilon^{(L_0+T_0)}(z)\right] \\ &\quad + \left[\check{\phi}^{(L_0+T_0+1)}(z) - f_\epsilon^{(L_0+T_0+1)}(z)\right]. \end{aligned}$$

Next we simplify each term in brackets via the inductive hypothesis. Let

$$S_1 = \left[K^{(L_0+T_0)}(z)\sum_{i=1}^{\infty} ia_i(\check{\phi}^{(L_0+T_0)}(z))^{i-1} - a_1 k_\epsilon^{(L_0+T_0)}(z)\right].$$

Then, given $\check{\phi}'(1) = C < \infty$, for

$$C_2(z) = (C - a_1)\lim_{L\to\infty}\frac{K^{(L)}(z)}{a_1^L(L+1)}\lim_{L\to\infty}\frac{\check{\phi}^{(L)}(z)}{a_1^L}\,,$$

which is finite by Lemma 3, we have:

$$\begin{aligned} S_1 &\leq C_1(z)\left(\sum_{i=0}^{T_0-1}(T_0 - i)a_1^{2L_0+T_0+i}\right) + C_2(z)\left(\sum_{i=0}^{T_0-2}(L_0 + i + 2)a_1^{2L_0+T_0+1+i}\right) \\ &\quad + K^{(L_0+T_0)}(z)\sum_{i=2}^{\infty} ia_i(\check{\phi}^{(L_0+T_0)}(z))^{i-1} \\ &\leq C_1(z)\left(\sum_{i=0}^{T_0-1}(T_0 - i)a_1^{2L_0+T_0+i}\right) + C_2(z)\left(\sum_{i=0}^{T_0-2}(L_0 + i + 2)a_1^{2L_0+T_0+1+i}\right) \\ &\quad + C_2(z)a_1^{L_0+T_0}(L_0 + T_0 + 1)a_1^{L_0+T_0} \\ &\leq C_1(z)\left(\sum_{i=0}^{T_0-1}(T_0 - i)a_1^{2L_0+T_0+i}\right) + C_2(z)\left(\sum_{i=0}^{T_0-1}(L_0 + i + 2)a_1^{2L_0+T_0+1+i}\right). \end{aligned}$$

Next, let:

$$S_2 = [\check{\phi}^{(L_0+T_0+1)}(z) - f_\epsilon^{(L_0+T_0+1)}(z)].$$

Then, we have by Eq. (11) that

$$S_2 \leq C_1(z)\left(\sum_{i=0}^{T_0} a_1^{2L_0+T_0+i}\right).$$

Therefore, combining the bounds on S_1, S_2, we conclude that

$$\begin{aligned}
K^{(L_0+T_0+1)}(z) - k_\epsilon^{(L_0+T_0+1)}(z) &\leq S_1 + S_2 \\
&\leq C_1(z)\left(\sum_{i=0}^{T_0} a_1^{2L_0+T_0+i}\right) + C_1(z)\left(\sum_{i=0}^{T_0-1}(T_0-i)a_1^{2L_0+T_0+i}\right) \\
&\quad + C_2(z)\left(\sum_{i=0}^{T_0-1}(L_0+i+2)a_1^{2L_0+T_0+1+i}\right) \\
&= C_1(z)\left(\sum_{i=0}^{T_0}(T_0+1-i)a_1^{2L_0+T_0+i}\right) + C_2(z)\left(\sum_{i=0}^{T_0-1}(L_0+i+2)a_1^{2L_0+T_0+1+i}\right),
\end{aligned}$$

which concludes the proof by induction and establishes Eq. (12). Next, Eq. (12) implies:

$$\begin{aligned}
\frac{K^{(L_0+T)}(z) - k_\epsilon^{(L_0+T)}(z)}{a_1^{L_0+T}(L_0+T+1)} &\leq C_1(z)\left(\sum_{i=0}^{T-1}\frac{T-i}{T+L_0+1}a_1^{L_0-1+i}\right) + C_2(z)\left(\sum_{i=0}^{T-2}\frac{L_0+i+2}{T+L_0+1}a_1^{L_0+i}\right) \\
&\leq \left(C_1(z)a_1^{L_0-1} + C_2(z)a_1^{L_0}\right)\frac{1}{1-a_1}\ .
\end{aligned}$$

Hence, since the right-hand side does not depend on T, we conclude that

$$\lim_{L\to\infty}\frac{K^{(L)}(z) - k_\epsilon^{(L)}(z)}{a_1^L(L+1)} \leq \left(C_1(z)a_1^{L_0-1} + C_2(z)a_1^{L_0}\right)\frac{1}{1-a_1}.$$

Lastly, note that by selecting ϵ small enough, we can make $L_0(z)$ arbitrarily large. Hence, for any fixed $z \in [0,1]$, we conclude that

$$\lim_{\epsilon\to 0}\lim_{L\to\infty}\frac{K^{(L)}(z) - k_\epsilon^{(L)}(z)}{a_1^L(L+1)} = 0,$$

and as a consequence that

$$\begin{aligned}
\lim_{L\to\infty}\frac{K^{(L)}(z)}{a_1^L(L+1)} &= \lim_{\epsilon\to 0}\lim_{L\to\infty}\frac{K^{(L)}(z) - k_\epsilon^{(L)}(z)}{a_1^L(L+1)} + \lim_{\epsilon\to 0}\lim_{L\to\infty}\frac{k_\epsilon^{(L)}(z)}{a_1^L(L+1)} \\
&= \lim_{\epsilon\to 0}\lim_{L\to\infty}\frac{k_\epsilon^{(L)}(z)}{a_1^L(L+1)}. \qquad (13)
\end{aligned}$$

By uniformly bounding the right-hand side over ϵ, we will establish an upper bound on the pole order for the iterated, normalized NTK. To do this, we first show that the iterated, normalized k_ϵ and f_ϵ are equal for $z \in (\epsilon, 1)$. Let $\alpha(z) = k_\epsilon^{(L_0(z))}(z)$ and $\beta(z) = f_\epsilon^{(L_0(z))}(z)$ for $z \in (\epsilon, 1)$. We prove by induction for $T > 0$ that

$$k_\epsilon^{(L_0(z)+T)}(z) = a_1^T[\alpha(z) + T\beta(z)]. \qquad (14)$$

The base case for $T = 1$ follows by

$$k_\epsilon^{(L_0(z)+1)}(z) = a_1[\alpha(z)] + a_1\beta(z).$$

Proceeding inductively, we assume that Eq. (14) holds for time T. Then at time $T+1$, we have

$$\begin{aligned}
k_\epsilon^{(L_0(z)+T+1)}(z) &= a_1 k_\epsilon^{L_0(z)+T} + f_\epsilon^{L_0+T+1}(z) \\
&= a_1^{T+1}[\alpha(z) + T\beta(z)] + a_1^{T+1}\beta(z) \\
&= a_1^{T+1}[\alpha(z) + (T+1)\beta(z)],
\end{aligned}$$

which concludes the proof by induction. Thus, we obtain that

$$\begin{aligned}\lim_{L\to\infty}\frac{k_\epsilon^{(L)}(z)}{a_1^L(L+1)} &= \lim_{L\to\infty}\frac{k_\epsilon^{(L-L_0(z))}(\alpha(z))}{a_1^{L-L_0(z)}(L-L_0(z)+1)}\left(\frac{L-L_0(z)+1}{L+1}\right)a_1^{-L_0(z)}\\ &= \lim_{T\to\infty}\frac{a_1^T[\alpha(z)+T\beta(z)]}{a_1^T(T+1)}a_1^{-L_0(z)}\\ &= \beta(z)a_1^{-L_0(z)}\\ &= \lim_{L\to\infty}\frac{f_\epsilon^{(L)}(z)}{a_1^L}.\end{aligned}$$

Next, we will uniformly bound the iterated, normalized f_ϵ. In particular, since $\check{\phi}\geq f_\epsilon$ and the two functions have the same normalizing constant, we obtain

$$\lim_{L\to\infty}\frac{f_\epsilon^{(L)}(z)}{a_1^L}\leq\lim_{L\to\infty}\frac{\check{\phi}^{(L)}(z)}{a_1^L}.$$

Now, we have that for any $\hat{\epsilon}$, $\check{\phi}$ is upper bounded by the function:

$$\tilde{g}_{\hat{\epsilon}}(z)=\begin{cases}\check{\phi}(z) & \text{if } z\in[0,c)\\ 1-(\check{\phi}'(1)-\hat{\epsilon})(1-z) & \text{if } z\in[c,1]\end{cases},$$

where $z=c$ is the intersection of the secant line $1-(\check{\phi}'(1)-\hat{\epsilon})(1-z)$ and $\check{\phi}$. We visualize $\tilde{g}_{\hat{\epsilon}}(z)$ in Fig. 3c. By Lemma 4, we know that for $z\in(c,1)$:

$$\lim_{L\to\infty}\frac{\tilde{g}_{\hat{\epsilon}}^{(L)}(z)}{a_1^L}=\frac{R_{\hat{\epsilon}}(z)}{(1-z)^{-\log_{\check{\phi}'(1)-\hat{\epsilon}}\phi'(0)}},$$

where $R_{\hat{\epsilon}}(z)$ is non-negative for $z\in[0,1]$, bounded from above, and bounded away from 0 around $z=1$. Since $\hat{\epsilon}$ is arbitrary, we conclude that for some ϵ'', for $z\in(\epsilon'',1)$:

$$\lim_{L\to\infty}\frac{f_\epsilon^{(L)}(z)}{a_1^L}\leq\lim_{L\to\infty}\frac{\check{\phi}^{(L)}(z)}{a_1^L}\leq\frac{R_1(z)}{(1-z)^{-\log_{\check{\phi}'(1)}\check{\phi}'(0)}},\tag{15}$$

where $R_1(z)$ is non-negative for $z\in[0,1]$, bounded from above, and bounded away from 0 around $z=1$. By substituting back the above inequalities into Eq. (13), we conclude that

$$\lim_{L\to\infty}\frac{K^{(L)}(z)}{a_1^L(L+1)}=\lim_{\epsilon\to0}\lim_{L\to\infty}\frac{K^{(L)}(z)-k_\epsilon^{(L)}(z)}{a_1^L(L+1)}+\lim_{\epsilon\to0}\lim_{L\to\infty}\frac{k_\epsilon^{(L)}(z)}{a_1^L(L+1)}\leq\frac{R_1(z)}{(1-z)^{-\log_{\check{\phi}'(1)}\check{\phi}'(0)}}.\tag{16}$$

To conclude the proof, we need to establish a similar lower bound on the above limit. We will construct the lower bound by first establishing the order of the singularity of the iteration of $\check{\phi}$ and then showing that this order is a lower bound on the order of the singularity for the iterated, normalized NTK. Note that we have already established an upper bound on the order of the singularity of the iteration of $\check{\phi}$ in Eq. (15). Now, we alternatively lower bound $\check{\phi}$ via the following function:

$$\tilde{f}(z)=\begin{cases}a_1 z & \text{if } x\in[0,c)\\ 1-\check{\phi}'(1)(1-z) & \text{if } x\in[c,1]\end{cases},$$

where $z=c$ corresponds to the intersection of the tangent lines of $\check{\phi}$ at $z=0$ and $z=1$. We visualize $\tilde{f}(z)$ in Fig. 3d. By Lemma 4, we have that for $z\in(c,1)$:

$$\lim_{L\to\infty}\frac{\check{\phi}^{(L)}(z)}{a_1^L}\geq\lim_{L\to\infty}\frac{\tilde{f}^{(L)}(z)}{a_1^L}=\frac{Q(z)}{(1-z)^{-\log_{\check{\phi}'(1)}\check{\phi}'(0)}},$$

where $Q(z)$ is non-negative for $z \in [0,1]$, bounded from above, and bounded away from 0 around $z = 1$. Hence, we conclude that:

$$\lim_{L\to\infty} \frac{\check{\phi}^{(L)}(z)}{a_1^L} = \frac{R_2(z)}{(1-z)^{-\log_{\check{\phi}'(1)} \check{\phi}'(0)}},$$

where $R_2(z)$ is non-negative for $z \in [0,1]$, bounded from above, and bounded away from 0 around $z = 1$. Lastly, we utilize Eq. (8) to show that:

$$\lim_{L\to\infty} \frac{\check{\phi}^{(L)}(z)}{a_1^L} \leq \lim_{L\to\infty} \frac{K^{(L)}(z)}{a_1^L(L+1)}.$$

In particular, Eq. (8) states that

$$K^{(L)}(z) = \sum_{i=0}^{L} \check{\phi}^{(i)}(z) \prod_{k=i}^{L-1} \check{\phi}' \left(\check{\phi}^{(k)}(z)\right).$$

We next write $\check{\phi}^{(i)}(z)$ as a product and substitute the computed product back into Eq. (8). Namely, using the power series representation for $\check{\phi}$ and unrolling the iteration, we obtain:

$$\begin{aligned}
\check{\phi}^{(i)}(z) &= \sum_{j=1}^{\infty} a_j \left(\check{\phi}^{(i-1)}(z)\right)^j \\
&= a_1 \check{\phi}^{(i-1)}(z) \left(1 + \sum_{j=2}^{\infty} \frac{a_j}{a_1} \left(\check{\phi}^{(i-1)}(z)\right)^j\right) \\
&= a_1^i z \prod_{k=0}^{i-1} \left(1 + \sum_{j=2}^{\infty} \frac{a_j}{a_1} \left(\check{\phi}^{(k)}(z)\right)^j\right).
\end{aligned}$$

We similarly use the power series for $\check{\phi}'(z)$ to conclude that

$$\begin{aligned}
\check{\phi}' \left(\check{\phi}^{(k)}(z)\right) &= \sum_{j=1}^{\infty} j a_j \left(\check{\phi}^{(k)}(z)\right)^{j-1} \\
&\geq \sum_{j=1}^{\infty} a_j \left(\check{\phi}^{(k)}(z)\right)^{j-1} \\
&\geq a_1 \left(1 + \sum_{j=2}^{\infty} \frac{a_j}{a_1} \left(\check{\phi}^{(k)}(z)\right)^{j-1}\right) \\
&\geq a_1 \left(1 + \sum_{j=2}^{\infty} \frac{a_j}{a_1} \left(\check{\phi}^{(k)}(z)\right)^{j}\right) \quad \left(\text{as } \check{\phi}^{(k)}(z) \leq 1.\right).
\end{aligned}$$

Therefore, we can simplify Eq. (8) as follows:

$$\begin{aligned}
K^{(L)}(z) &= \sum_{i=0}^{L} \check{\phi}^{(i)}(z) \prod_{k=i}^{L-1} \check{\phi}' \left(\check{\phi}^{(k)}(z)\right) \\
&\geq \sum_{i=0}^{L} a_1^i z \prod_{k=0}^{i-1} \left(1 + \sum_{j=2}^{\infty} \frac{a_j}{a_1} \left(\check{\phi}^{(k)}(z)\right)^j\right) \prod_{k'=i}^{L-1} a_1 \left(1 + \sum_{j=2}^{\infty} \frac{a_j}{a_1} \left(\check{\phi}^{(k')}(z)\right)^j\right) \\
&= \sum_{i=0}^{L} a_1^L z \prod_{k=0}^{L-1} \left(1 + \sum_{j=2}^{\infty} \frac{a_j}{a_1} \left(\check{\phi}^{(k)}(z)\right)^j\right) \\
&= (L+1)\check{\phi}^{(L)}(z),
\end{aligned}$$

and conclude that

$$\frac{K^{(L)}(z)}{a_1^L(L+1)} \geq \frac{\check{\phi}^{(L)}(z)}{a_1^L}.$$

As a consequence,

$$\lim_{L\to\infty} \frac{K^{(L)}(z)}{a_1^L(L+1)} \geq \lim_{L\to\infty} \frac{\check{\phi}^{(L)}(z)}{a_1^L} = \frac{R_2(z)}{(1-z)^{-\log_{\check{\phi}'(1)} \check{\phi}'(0)}}. \tag{17}$$

Lastly, we combine Eq. (16) and (17) to conclude that there exists some ϵ such that for $z \in (\epsilon, 1)$:

$$\frac{R_2(z)}{(1-z)^{-\log_{\check{\phi}'(1)} \check{\phi}'(0)}} \leq \lim_{L\to\infty} \frac{K^{(L)}(z)}{a_1^L(L+1)} \leq \frac{R_1(z)}{(1-z)^{-\log_{\check{\phi}'(1)} \check{\phi}'(0)}} \ .$$

Thus, we conclude that

$$\lim_{L\to\infty} \frac{K^{(L)}(z)}{a_1^L(L+1)} = \frac{R(z)}{(1-z)^{-\log_{\check{\phi}'(1)} \check{\phi}'(0)}},$$

where $R(z)$ is non-negative for $z \in [0,1]$, bounded from above, and bounded away from 0 around $z = 1$. This concludes the proof of Theorem 1. □

To prove Theorem 2, we will use the result of Theorem 1 and that of [9], which analyzes the optimality of singular kernel smoothers. To connect infinitely wide and deep networks with kernel smoothers, we next prove that the infinite depth limit of the NTK corresponds to a kernel smoother under the conditions of Theorem 1.

Lemma 5. *Let* $\psi(z) = \lim_{L\to\infty} \frac{K^{(L)}(z)}{a_1^L(L+1)}$. *Then under the setting of Theorem 2,*

$$m_n(x) = \text{sign}\left(\sum_{i=1}^n y^{(i)} \psi(x^T x^{(i)})\right),$$

assuming $\left|\sum_{i=1}^n y^{(i)} \psi(x^T x^{(i)})\right| > 0$ *for almost all* $x \in \mathcal{S}_+^{d-1}$.

Proof. Let $m_n^{(L)}(x)$ be defined as follows:

$$m_n^{(L)}(x) = \text{sign}\left(y\Big(K_n^{(L)}\Big)^{-1} K^{(L)}(X,x)\right).$$

We first note that multiplying the argument to the sign function by a positive constant does not affect the value. Hence, we have:

$$\lim_{L\to\infty} m_n^{(L)}(x) = \lim_{L\to\infty} \text{sign}\left(y\Big(K_n^{(L)}\Big)^{-1} \frac{K^{(L)}(X,x)}{a_1^L(L+1)}\right).$$

Now we compare the argument of the sign function above to the corresponding kernel smoother. Namely, we have:

$$\left|y\Big(K_n^{(L)}\Big)^{-1} \frac{K^{(L)}(X,x)}{a_1^L(L+1)} - y\frac{K^{(L)}(X,x)}{a_1^L(L+1)}\right| \leq \|y\|_2 \left\|\Big(K_n^{(L)}\Big)^{-1} - I\right\|_2 \left\|\frac{K^{(L)}(X,x)}{a_1^L(L+1)}\right\|_2,$$

where the inequality follows from the Cauchy-Schwarz inequality and $\|Av\|_2 \leq \|A\|_2\|v\|_2$ for $A \in \mathbb{R}^{n\times n}, v \in \mathbb{R}^n$. Now since 0 is an attractor for $\check{\phi}$, then for any $h > 0$, there exists L_1 such that for $L > L_1$, the

spectrum of $\hat{K}^{(L)}$ is contained in $[1-hn^2, 1+hn^2]$ by Weyl's inequalities. Hence, the spectrum of $\left(K_n^{(L)}\right)^{-1}$ is contained in $\left[\frac{1}{1+hn^2}, \frac{1}{1-hn^2}\right]$. Thus, we conclude that

$$\left\|\left(K_n^{(L)}\right)^{-1} - I\right\|_2 \leq \left(\frac{1}{1-hn^2} - 1\right).$$

Hence by selecting h appropriately small, we conclude that for any ϵ_1, there exists L_1 such that for $L > L_1$, $\left\|\left(K_n^{(L)}\right)^{-1} - I\right\|_2 < \epsilon_1$. Next, since $\lim_{L\to\infty} \frac{K^{(L)}(x^{(i)},x)}{a^L(L+1)} = \psi(x^T x^{(i)})$, for any ϵ_2, we can select L_2 such that for $L > L_2$,

$$\left| y\frac{K^{(L)}(X,x)}{a^L(L+1)} - \sum_{i=1}^n y^{(i)}\psi(x^T x^{(i)}) \right| < \epsilon_2.$$

Next under the assumption in the lemma, we may thus select ϵ_1, ϵ_2 small enough such that the argument of $m_n^{(L)}(x)$ is not exactly 0 for $L > \max(L_1, L_2)$. Thus we can interchange the limit and the sign function. As a consequence, for any $x \neq x^{(i)}$ for $i \in \{1, 2, \ldots, n\}$ satisfying $\sum_{i=1}^n y^{(i)}\psi(x^T x^{(i)}) \neq 0$, we obtain that

$$\begin{aligned}
\lim_{L\to\infty} m_n^{(L)}(x) &= \lim_{L\to\infty} \operatorname{sign}\left(y\left(K_n^{(L)}\right)^{-1}\frac{K^{(L)}(X,x)}{a^L(L+1)}\right) \\
&= \operatorname{sign}\left(\lim_{L\to\infty} y\left(K_n^{(L)}\right)^{-1}\frac{K^{(L)}(X,x)}{a^L(L+1)}\right) \\
&= \operatorname{sign}\left(\sum_{i=1}^n y^{(i)}\psi(x^T x^{(i)})\right).
\end{aligned}$$

Lastly, if $x = x^{(i)}$ for some $i \in \{1, 2, \ldots, n\}$, then since $\psi(z)$ has a singularity at $z = 1$,

$$\lim_{L\to\infty} m_n^{(L)}(x) = \operatorname{sign}\left(\lim_{z\to 1}\sum_{i=1}^n \frac{1}{\psi(z)} y^{(i)}\psi(x^T x^{(i)})\right) = \operatorname{sign}(y^{(i)}),$$

which completes the proof. □

We lastly utilize Theorem 1, Lemma 5, and the result of [9] to prove Theorem 2 (expressed below in terms of dual activations), which identifies infinitely wide and deep classifiers that achieve optimality.

Theorem. *Let m_n denote the classifier in Eq. (3) corresponding to training an infinitely wide and deep network with activation function ϕ on n training points. Let m denote the Bayes optimal classifier, i.e. $m(x) = \underset{\tilde{y}\in\{-1,1\}}{\arg\max}\, \mathbb{P}(y = \tilde{y}|x)$. If the dual activation, $\check{\phi}$ satisfies:*

1) $\check{\phi}(0) = 0$, $\check{\phi}(1) = 1$,

2) $0 < \check{\phi}'(0) < 1$ and $\check{\phi}'(1) < \infty$,

3) $-\frac{\log(\check{\phi}'(0))}{\log(\check{\phi}'(1))} = \frac{d}{2}$,

then m_n satisfies $\lim_{n\to\infty} \mathbb{P}_X(|m_n(x) - m(x)| > \epsilon) = 0$ for almost all $x \in \mathcal{S}_+^d$ and for any $\epsilon > 0$.

Proof. Thus far, we proved that under the conditions of Theorem 1, the classifier m_n corresponds to taking the sign of a kernel smoother using a singular kernel with singularity of order $-\frac{\log(\check{\phi}'(0))}{\log(\check{\phi}'(1))}$. For data with density in $\mathbb{R}^d$, kernel smoothers with singular kernels of the form $K_h(x,\tilde{x}) = \frac{1}{\|x-\tilde{x}\|^d}$ (i.e., the Hilbert estimate) converge to the Bayes optimal classifier in probability for almost all samples as $n \to \infty$ [9]. We note that multiplying $K_h(x,\tilde{x})$ by a non-negative function that is bounded away from 0 around 1 and bounded from above such that the kernel is still monotonically increasing also yields optimality in the same

sense (see Supplementary Information C). Returning to our setting, for any $x, \tilde{x} \in \mathcal{S}_+^d$, we can re-write the kernel $K_h(x, \tilde{x})$ as

$$K_h(x, \tilde{x}) = \frac{1}{\|x - \tilde{x}\|^d} = \frac{1}{2^{\frac{d}{2}}(1 - x^T\tilde{x})^{\frac{d}{2}}}.$$

The constant $\frac{1}{2^{\frac{d}{2}}}$ again does not affect the sign function. Lastly, assumption 3 selects the order of the singularity such that the limiting kernel from Theorem 1 can be written up to constant factors as a Hilbert estimate, which concludes the proof of Theorem 2. □

C Extension of Hilbert estimate optimality from [9]

We utilize the following extension of the result from [9] to prove Theorem 2. In this section, we follow the notation from [9] in our statements and proofs.

Corollary. *For $x \in \mathcal{S}_+^d$, let $m(x)$ denote the Bayes optimal regressor. For $x, \tilde{x} \in \mathcal{S}_+^d$, let $K(x^T\tilde{x}) = \frac{R(x^T\tilde{x})}{2^{\frac{d}{2}}(1-x^T\tilde{x})^{\frac{d}{2}}}$, where $R(z) \geq 0$ for $z \in [0, 1]$ is bounded from above, bounded away from 0 around $z = 1$, and $K(\cdot)$ is monotonically increasing in $[0, 1]$. Given a dataset $\{X_i, Y_i\}_{i=1}^n \subset \mathcal{S}_+^d \times \mathbb{R}$, let*

$$m_n(x) = \frac{\sum_{i=1}^n Y_i K(x^T X_i)}{\sum_{i=1}^n K(x^T X_i)}.$$

Let X have any density f on $\mathcal{S}_+^d$ and let Y be bounded. Then, at almost all x with $f(x) > 0$, $m_n(x) \to m(x)$ in probability as $n \to \infty$.

Proof. The proof closely follows that of the theorem in [9] with the differences that (1) we map from densities on $\mathcal{S}_+^d$ to densities on $\mathbb{R}^d$, and (2) we simply verify that the function $R(z)$ does not change the asymptotic analyses of the original proof. We begin by noting that the kernel K involves chordal distances on the sphere, i.e.,

$$K(x^T\tilde{x}) = \frac{R\left(x^T\tilde{x}\right)}{\|x - \tilde{x}\|^d}.$$

We first define the random variable $W := \|P(x) - P(X)\|^d V_d$, where V_d is the volume of the unit sphere in d dimensions and $P : \mathcal{S}^d \to \mathbb{R}^d$ is the stereographic projection such that $\mathcal{S}_+^d$ maps to a bounded region. We let f_P denote the density of the points $P(x)$ for $x \in \mathcal{S}_+^d$. We note that Euclidean distances after stereographic projection can be related to chordal distances, $\|x - X\|$, via the following formula (up to isometries of the sphere):

$$\|x - X\|^2 = \frac{\|P(x) - P(X)\|^2}{(1 + \|P(x)\|^2)(1 + \|P(X)\|^2)}.$$

Since we select the projection such that $\|P(x)\| < \infty$ for $x \in \mathcal{S}_+^d$, we have that $(1 + \|P(x)\|^2)(1 + \|P(X)\|^2)$ is bounded and nonzero, i.e., it is again a factor that simply scales the kernel function. We thus define

$$Q(x, \tilde{x}) = R(x^T\tilde{x})(1 + \|P(x)\|^2)^{\frac{d}{2}}(1 + \|P(\tilde{x})\|^2)^{\frac{d}{2}},$$

which is bounded away from zero for some $\epsilon > 0$ and $x, \tilde{x}$ such that $x^T\tilde{x} > 1 - \epsilon$. Letting $W_i := V_d\|P(x) - P(X_i)\|^d$, the regressor $m_n(x)$ is given by

$$m_n(x) = \frac{\sum_{i=1}^n Y_i \frac{Q(x, X_i)}{W_i}}{\sum_{i=1}^n \frac{Q(x, X_i)}{W_i}}.$$

Hence, we can utilize the proof strategy of [9] for points $P(x)$ in $\mathbb{R}^d$. Namely as in [9], we analyze the term:

$$|m_n(x) - m(x)| \leq \left| \frac{\sum_{i=1}^n (Y_i - m(X_i)) \frac{Q(x,X_i)}{W_i}}{\sum_{i=1}^n \frac{Q(x,X_i)}{W_i}} \right| + \frac{\sum_{i=1}^n |m(X_i) - m(X))| \frac{Q(x,X_i)}{W_i}}{\sum_{i=1}^n \frac{Q(x,X_i)}{W_i}} := I + II.$$

To simplify notation, we let $Q_i = Q(x, X_i)$ and we let $\frac{Q_{(i)}}{W_{(i)}}$ denote the i^{th} order statistic ordered such that $W_{(1)} \leq W_{(2)} \ldots \leq W_{(n)}$. Now, the proof strategy of [9] is to show that the terms I and II respectively converge to 0 in probability for almost all x as $n \to \infty$. To prove that I converges in this manner, following the proof of [9], we have that:

$$\mathbb{E}[I^2 | \{X_i\}_{i=1}^n] \leq C_1 \frac{\frac{1}{W_{(1)}}}{\sum_{j=1}^k \frac{Q_{(j)}}{W_{(j)}}} \leq C_2 \frac{\frac{1}{W_{(1)}}}{\sum_{j=1}^k \frac{1}{W_{(j)}}},$$

where k such that $W_{(k)} > V_d \delta^d$ for small δ, and $C_1, C_2 > 0$ are constants since $\{Q_{(j)}\}_{j=1}^k$ are non-negative and bounded away from 0. Hence, the convergence of I follows directly from the proof of [9]. To establish the convergence of II, we follow the proof of [9] and first establish that

$$A_n := \frac{\sum_{i \leq \theta n} \frac{Q_{(i)}}{W_{(i)}}}{\sum_{i=1}^n \frac{Q_{(i)}}{W_{(i)}}} \to 1$$

in probability as $n \to \infty$, for all θ fixed in $(0, 1)$. Let χ denote the indicator function, and following the notation of [9], let $U_{(i)}$ denote uniform order statistics. The work of [9] establishes that for any fixed $\epsilon \in (0, 1)$ there exists δ such that for all $W_{(i)} \leq V_d \delta^d$:

$$(1 - \epsilon) f_P(P(x)) W_{(i)} \leq U_{(i)} \leq (1 + \epsilon) f_P(P(x)) W_{(i)}.$$

Hence, we consider the event $B = [W_{\lfloor \theta n \rfloor} \leq V_d \delta^d]$, and then as in the proof of [9], we obtain

$$A_n \chi_B \geq 1 - \frac{2\epsilon}{1+\epsilon} - \frac{\frac{nC_3}{W_{\lfloor \theta n \rfloor}}}{f_P(P(x)) \sum_{i \leq \theta n} \frac{Q_{(i)}}{U_{(i)}}} \geq 1 - \frac{2\epsilon}{1+\epsilon} - C_4 \frac{\frac{n}{W_{\lfloor \theta n \rfloor}}}{f_P(P(x)) \sum_{i \leq \theta n} \frac{1}{U_{(i)}}},$$

where $C_3, C_4 > 0$ are constants since $Q_{(i)}$ is bounded and positive for $i \leq \lfloor \theta n \rfloor$. The convergence of A_n then follows by continuing the proof from [9]. Next, again following the proof of [9], for any $\epsilon > 0$, we also select δ such that:

$$\sup_{r \leq \delta} \frac{\int_{S_{P(x),r}} |m(y) - m(x)| f_P(y) dy}{\int_{S_{P(x),r}} f_P(y) dy} \leq \epsilon,$$

where $S_{P(x),r}$ denotes the closed ball in $\mathbb{R}^d$ of radius r centered at $P(x)$. Then as in [9], select $A = \{y : m(y) - m(x) > \epsilon\}$ and select $\theta \in (0, 1)$ small enough such that $\mathbb{P}(\|P(X_{(\lfloor \theta n \rfloor)}) - P(x)\| > \delta) \to 0$ as $n \to \infty$. Then, we have:

$$\begin{aligned}
II &= \frac{\sum_{i=1}^n |m(X_i) - m(x))| \frac{Q_i}{W_i}}{\sum_{i=1}^n \frac{Q_i}{W_i}} \\
&\leq 2 \frac{\sum_{i > \theta n} \frac{Q_i}{W_i}}{\sum_{i=1}^n \frac{Q_i}{W_i}} + 2\chi_{\|P(X_{\lfloor \theta n \rfloor}) - P(x)\| > \delta} + \epsilon + \frac{\sum_{i: P(X_i) \in S_{P(x),\delta} \cap A} \frac{Q_i}{W_i}}{\sum_{i=1}^n \frac{Q_i}{W_i}} \\
&:= V_1 + V_2 + V_3 + V_4.
\end{aligned}$$

Now as in [9], we have that $V_1 \to 0$ in probability, as we showed $A_n \to 1$ in probability above. Then, $V_2 \to 0$ in probability and V_3 can be made as small as possible by the choice of ϵ. Lastly, $V_4 \to 0$ since, following the proof of [9]:

$$\frac{\sum_{i: P(X_i) \in S_{P(x),\delta} \cap A} \frac{Q_i}{W_i}}{\sum_{i=1}^n \frac{Q_i}{W_i}} \leq 2\epsilon + C_5 \frac{\frac{1}{W_{(1)}}}{\sum_{i=1}^n \frac{Q_{(i)}}{W_{(i)}}},$$

where $C_5 > 0$ is a constant. The above term goes to 0 in probability by the analysis of part I and the arbitrary choice of ϵ. This concludes the proof of this extension of the result of [9]. □

D Proof of Corollary 1

For ease of reading, we repeat Corollary 1 below.

Corollary. *Let m_n denote the classifier in Eq. (3) corresponding to training an infinitely wide and deep network with activation function*

$$\phi(x) = \begin{cases} \frac{1}{12\sqrt{70}} h_7(x) + \frac{1}{\sqrt{2}} x & \text{if } d = 1 \\ \frac{1}{2^{d/4}} \left(\frac{x^3 - 3x}{\sqrt{6}} \right) + \sqrt{1 - \frac{2}{2^{d/2}}} \left(\frac{x^2 - 1}{\sqrt{2}} \right) + \frac{1}{2^{d/4}} x & \text{if } d \geq 2 \end{cases},$$

where $h_7(x)$ is the 7^{th} probabilist's Hermite polynomial.[10] *Then the classifier m_n is Bayes optimal.*

Proof. We need only check that $\check{\phi}(z)$ satisfies the conditions of Theorem 2. We first consider the case $d \geq 2$. In particular, since $\frac{x^2-1}{\sqrt{2}}$ is the 2nd normalized probabilist's Hermite polynomial and $\frac{x^3-3x}{\sqrt{6}}$ is the third normalized probabilist's Hermite polynomial, we have by [8, Lemma 11] that

$$\check{\phi}(z) = \frac{1}{2^{\frac{d}{2}}} z^3 + \left(1 - \frac{2}{2^{\frac{d}{2}}} \right) z^2 + \frac{1}{2^{\frac{d}{2}}} z.$$

We thus have by direct computation that

$$\check{\phi}'(1) = \frac{3}{2^{\frac{d}{2}}} + \left(2 - \frac{4}{2^{\frac{d}{2}}} \right) + \frac{1}{2^{\frac{d}{2}}} = 2 \;\; ; \;\; \check{\phi}'(0) = \frac{1}{2^{\frac{d}{2}}},$$

and so, the result follows from Theorem 2 since

$$-\log_{\check{\phi}'(1)} \check{\phi}'(0) = \log_2 2^{\frac{d}{2}} = \frac{d}{2}.$$

Now for the case of $d = 1$, we have again by [8, Lemma 11] that

$$\check{\phi}(z) = \frac{z^7}{2} + \frac{z}{2}.$$

By direct computation,

$$\check{\phi}'(1) = \frac{7}{2} + \frac{1}{2} = 4 \;\text{ and }\; \check{\phi}'(0) = \frac{1}{2}.$$

Hence, the result follows from Theorem 2 since

$$-\log_{\check{\phi}'(1)} \check{\phi}'(0) = \frac{1}{2} = \frac{d}{2}.$$

□

E Proof of Theorem 3

We repeat Theorem 3 below in terms of dual activations.

Theorem. *Let m_n denote the classifier in Eq. (3) corresponding to training an infinitely wide and deep network with activation function $\phi(\cdot)$ on n training examples. If the dual activation, $\check{\phi}$, satisfies:*

[10] For $d = 1$, this activation function can be written in closed form as $\frac{x^7 - 21x^5 + 105x^3 + (12\sqrt{35} - 105)x}{12\sqrt{70}}$.

1) $\check{\phi}(0) = 0$, $\check{\phi}(1) = 1$,

2) $\check{\phi}'(0) = 0$, $\check{\phi}'(1) < \infty$,

then $m_n(x)$ *is the 1-NN classifier for* $x \in \mathcal{S}_+^d$.

Proof. Let $m_n^{(L)}(x)$ be defined as follows:

$$m_n^{(L)}(x) = \text{sign}\left(y\left(K_n^{(L)}\right)^{-1} K^{(L)}(X, x)\right).$$

By the proof of Lemma 5, we analogously have that the Gram matrix converges to the identity matrix as depth approaches infinity, i.e. $\lim_{L\to\infty} K_n^{(L)} = I$. For $x, \tilde{x} \in \mathcal{S}_+^d$, let $z = x^T\tilde{x}$ and consider the radial kernel $K^{(L)}(z) = K^{(L)}(x, \tilde{x})$. Let $\check{\phi}(z) = \sum_{i=2}^{\infty} a_i z^i$ for $a_i \geq 0$, as given by Eq. (7). Without loss of generality, we assume $a_2 > 0$. The proof will follow by using induction to establish:

$$\check{\phi}^{(L)}(z) = z^{2^L} h_L(z) \quad \text{and} \quad K^{(L)}(z) = z^{2^L} g_L(z), \tag{18}$$

where h_L, g_L are positive, increasing functions on $(0, 1]$. The base case follows for $L = 0$ since $\check{\phi}^{(0)}(z) = K^{(0)}(z) = z$. Hence, we assume the statement is true for $L = T - 1$ and prove the statement for $L = T$. We have

$$\check{\phi}^{(T)}(z) = \check{\phi}\left(\check{\phi}^{(T-1)}(z)\right) = \sum_{i=2}^{\infty} a_i \left(\check{\phi}^{(T-1)}(z)\right)^i = \left(\check{\phi}^{(T-1)}(z)\right)^2 \left[\sum_{i=2}^{\infty} a_i \left(\check{\phi}^{(T-1)}(z)\right)^{i-2}\right],$$

and hence using the inductive hypothesis, we can conclude that

$$\check{\phi}^{(T)}(z) = z^{2^T} h_{T-1}(z)^2 \left[\sum_{i=2}^{\infty} a_i \left(\check{\phi}^{(T-1)}(z)\right)^{i-2}\right] = z^{2^T} h_T(z),$$

where h_T is positive and increasing since h_{T-1} and the term in brackets are positive and increasing. We proceed similarly for $K^{(T)}$. Namely, we have:

$$\begin{aligned}
K^{(T)}(z) &= K^{(T-1)}(z)\check{\phi}'(\check{\phi}^{(T-1)}(z)) + \check{\phi}^{(T)}(z) \\
&= z^{2^{T-1}} g_{T-1}(z) \left[\sum_{i=2} i a_i \left(\check{\phi}^{(T-1)}(z)\right)^{i-1}\right] + z^{2^T} h_T(z) \\
&= z^{2^{T-1}} \check{\phi}^{(T-1)}(z) g_{T-1}(z) \left[\sum_{i=2} i a_i \left(\check{\phi}^{(T-1)}(z)\right)^{i-2}\right] + z^{2^T} h_T(z) \\
&= z^{2^T} \left(h_{T-1}(z) g_{T-1}(z) \left[\sum_{i=2} i a_i \left(\check{\phi}^{(T-1)}(z)\right)^{i-2}\right] + h_T(z)\right) \\
&= z^{2^T} g_T(z),
\end{aligned}$$

where $g_T(z)$ is positive and increasing since h_T, h_{T-1}, g_{T-1} and the term in brackets are positive and increasing, which completes the induction argument.

Now let $z_i = x^T x^{(i)}$ for $i \in \{1, 2, \ldots, n\}$. Without loss of generality assume that $z_1 > z_j$ for all $j \neq 1$. To show that $\lim_{L\to\infty} m_n^{(L)}(x)$ is equivalent to the 1-NN classifier, we need only show that $\lim_{L\to\infty} m_n^{(L)}(x) = y^{(1)}$. By Eq. (18) for $j \neq 1$, we have that

$$\begin{aligned}
\lim_{L\to\infty} \frac{K^{(L)}(z_j)}{K^{(L)}(z_1)} &= \lim_{L\to\infty} \frac{z_j^{2^L} g_L(z_j)}{z_1^{2^L} g_L(z_1)} \\
&\leq \lim_{L\to\infty} \frac{z_j^{2^L}}{z_1^{2^L}} \quad \text{(since } z_j < z_1 \text{ and } g_L \text{ are positive and increasing)} \\
&= 0.
\end{aligned}$$

As a consequence, since $K^{(L)}(z_1) > 0$, we obtain that

$$\lim_{L\to\infty} m_n^{(L)}(x) = \lim_{L\to\infty} \operatorname{sign}\left(y\Big(K_n^{(L)}\Big)^{-1} \frac{K^{(L)}(X,x)}{K^{(L)}(x^{(1)},x)}\right) = y^{(1)},$$

which establishes that $\lim_{L\to\infty} m_n^{(L)}(x)$ converges to the 1-NN classifier, thereby completing the proof. □

F Proof of Proposition 1

We repeat Proposition 1 below for ease of reading.

Proposition. *Let m_n denote the classifier in Eq. (3) corresponding to training an infinitely wide and deep network with activation function $\phi(\cdot)$ on n training examples. For $x, \tilde{x} \in \mathcal{S}_+^d$ with $x \neq \tilde{x}$, if the NTK $K^{(L)}$ satisfies*

$$\lim_{L\to\infty} \frac{K^{(L)}(x,\tilde{x})}{C(L)} > C_1 \quad \textit{and} \quad \lim_{L\to\infty} \frac{K^{(L)}(x,\tilde{x})}{C(L)} \neq \lim_{L\to\infty} \frac{K^{(L)}(x,x)}{C(L)} \tag{19}$$

with $C_1 > 0$ and $0 < C(L) < \infty$ for any L, then m_n implements the majority vote classifier, i.e.,

$$m_n(x) = \operatorname{sign}\Big(\sum_{i=1}^n y^{(i)}\Big)\ .$$

Proof. Let $C_2 = \lim_{L\to\infty} \frac{K^{(L)}(x,x)}{C(L)}$. We consider two cases: (1) when $C_2 = \infty$, and (2) when $C_2 < \infty$. When $C_2 = \infty$, we have:

$$\begin{aligned}
\lim_{L\to\infty} m_n^{(L)}(x) &= \lim_{L\to\infty} \operatorname{sign}\Big(y(K_n^{(L)})^{-1}K^{(L)}(X,x)\Big) \\
&= \lim_{L\to\infty} \operatorname{sign}\left(y\left(\frac{K_n^{(L)}}{K^{(L)}(x,x)}\right)^{-1} \frac{K^{(L)}(X,x)}{C(L)}\right) \\
&= \operatorname{sign}\left(\sum_{i=1}^n y^{(i)} C_1\right) \\
&= \operatorname{sign}\left(\sum_{i=1}^n y^{(i)}\right),
\end{aligned}$$

which corresponds to the majority vote classifier. When $C_2 < \infty$, we use the Sherman-Morrison formula to compute the inverse of the Gram matrix $\lim_{L\to\infty}(K_n^{(L)})^{-1}$. In particular, since the inverse is a continuous map on invertible matrices,

$$\lim_{L\to\infty} (K_n^{(L)})^{-1} = \frac{1}{(C_2 - C_1)} I - \frac{C_1}{(C_2 - C_1)(C_2 - C_1 + C_1 n)} J,$$

where I is the identity matrix and J is the all-ones matrix. Hence, we have that for $x \neq x^{(i)}$ for $i \in \{1, 2, \ldots, n\}$:

$$\begin{aligned}
\lim_{L\to\infty} y(K_n^{(L)})^{-1} \frac{K^{(L)}(X,x)}{C(L)} &= y\left(\frac{1}{(C_2 - C_1)} I - \frac{C_1}{(C_2 - C_1)(C_2 - C_1 + C_1 n)} J\right) C_1 \mathbf{1} \\
&= \frac{C_1}{C_2 - C_1 + C_1 n} \sum_{i=1}^n y^{(i)},
\end{aligned}$$

where $\mathbf{1} \in \mathbb{R}^n$ is the all-ones vector. Assuming that $\sum_{i=1}^n y^{(i)} \neq 0$, we can swap the limit and sign function to conclude that:

$$\begin{aligned}\lim_{L\to\infty} m_n^{(L)}(x) &= \text{sign}\left(\lim_{L\to\infty} y(K_n^{(L)})^{-1}K^{(L)}(X,x)\right)\\ &= \text{sign}\left(\frac{C_1}{C_2 - C_1 + C_1 n}\sum_{i=1}^n y^{(i)}\right)\\ &= \text{sign}\left(\sum_{i=1}^n y^{(i)}\right),\end{aligned}$$

which completes the proof. □

G Proofs for when Infinitely Wide and Deep Networks are Majority Vote Classifiers

The following lemma implies that any activation function satisfying $\check{\phi}(0) > 0$ and $\check{\phi}'(1) > 1$ yields a NTK satisfying Eq. (19) and thus, the infinite depth classifier is the majority vote classifier by Proposition 1.

Lemma 6. *Let m_n denote the classifier in Eq. (3) corresponding to training an infinitely wide and deep network with activation function $\phi(\cdot)$ on n training examples. If $\check{\phi}$ satisfies:*

1) $\check{\phi}(0) > 0$, $\check{\phi}(1) = 1$,

2) $1 < \check{\phi}'(1) < \infty$,

then m_n is the majority vote classifier.

Proof. We show that the limiting kernel satisfies the properties of Proposition 1 with $C_2 = \infty$. Note that we must have $\check{\phi}'(0) < 1$ by Lemma 2. Now, since $\check{\phi}(0) < 1$ and $\check{\phi}(1) = 1$, by the intermediate value theorem, there exists some $c \in (0,1)$ such that $\check{\phi}(c) = c$.

We claim that $\check{\phi}'(c) < 1$. Suppose for the sake of contradiction that $\check{\phi}'(c) \geq 1$. Then, since $\check{\phi}(z)$ can be written as a convergent power series with non-negative coefficients, we have that $\check{\phi}(z) \geq z$ for $z \in (c,1]$. Hence either $\check{\phi}(z) = z$ on some subset of $(c,1]$ or $\check{\phi}(z) > z$ for $z \in (c,1]$. In the former case, analytic continuation implies that $\check{\phi}(z) = z$ on $[0,1]$, and in the latter case, $\check{\phi}(1) > 1$. Thus, in either case we reach a contradiction and thus we can conclude that $\check{\phi}'(c) < 1$. Therefore,it follows that c is the unique fixed point attractor of $\check{\phi}(z)$.

Lastly, since $c \in (0,1)$, we can conclude that the infinite depth NTK solves the equilibrium equation corresponding to the recursive formula for the NTK in Eq. (5). Namely, for any $z \in (0,1)$ and $K^*(z) := \lim_{L\to\infty} K^{(L)}(z)$:

$$K^*(z) = K^*(z)\check{\phi}'(c) + c \implies K^*(z) = \frac{c}{1 - \check{\phi}'(c)}.$$

Hence, for any $z \in (0,1)$, it holds that $\lim_{L\to\infty} K^{(L)}(z) = \frac{c}{1-\check{\phi}'(c)}$. Lastly, letting $a = \check{\phi}'(1)$, for $z = 1$, we have that

$$K^{(L)}(1) = \frac{a^L - 1}{a - 1},$$

and so $\lim_{L\to\infty} K^{(L)}(1) = \infty$. Thus, $\lim_{L\to\infty} K^{(L)}(x,\tilde{x})$ satisfies the conditions of Proposition 1, which concludes the proof of the lemma. □

We next show that if $\check{\phi}$ falls under case 3 with $\check{\phi}'(1) < 1$, then under ridge regularization, the corresponding infinitely wide and deep classifier also implements majority vote classification.

Lemma 7. *Let $m_{n,\lambda}^{(L)}$ denote the ridge-regularized kernel machine with regularization term λ and with the NTK of a fully connected network with L hidden layers and activation function ϕ on n training points. If $\check{\phi}$ satisfies:*

1) $\check{\phi}(0) > 0$, $\check{\phi}(1) = 1$,

2) $\check{\phi}'(1) < 1$,

then $\lim_{\lambda \to 0^+} \lim_{L \to \infty} m_{n,\lambda}^{(L)}(x)$ *is the majority vote classifier.*

Proof. The proof follows that of Proposition 1. Since $\check{\phi}'(1) < 1$, $z = 1$ is the unique fixed point attractor of $\check{\phi}$. Then as in the proof of Lemma 6, for all $x, \tilde{x} \in \mathcal{S}_+^d$, it holds that

$$\lim_{L \to \infty} K^{(L)}(x, \tilde{x}) = \frac{1}{1 - \check{\phi}'(1)}.$$

Letting $c = \frac{1}{1-\check{\phi}'(1)}$, we obtain that

$$\begin{aligned}
\lim_{L \to \infty} m_{n,\lambda}^{(L)}(x) &= \operatorname{sign}\left(y(K_n^{(L)} + \lambda I)^{-1} K^{(L)}(X, x)\right) \\
&= \operatorname{sign}\left(y \left[\lim_{L \to \infty} (K_n^{(L)} + \lambda I)^{-1}\right] \left[\lim_{L \to \infty} K^{(L)}(X, x)\right]\right) \\
&= \operatorname{sign}\left(y \left[\frac{1}{\lambda} I - \frac{c}{\lambda(\lambda + cn)} J\right] c\mathbf{1}\right) \\
&= \operatorname{sign}\left(\frac{c}{\lambda + cn} \sum_{i=1}^{n} y^{(i)}\right),
\end{aligned}$$

where $J \in \mathbb{R}^{n \times n}$ is the all-ones matrix, $\mathbf{1} \in \mathbb{R}^n$ is the all-ones vector, and the third equality follows from the Sherman-Morrison formula. Hence, Proposition 1 implies that $\lim_{\lambda \to 0^+} \lim_{L \to \infty} m_{n,\lambda}^{(L)}(x)$ is again the majority vote classifier, thereby completing the proof. □