A General Notation

In this section, we formally define our general notation used in the paper. We use the notations $\mathbf{0}_n$ and $\boldsymbol{I}_n$ to denote the zero vector and the square identity matrix of dimension n, respectively. For a matrix M (a vector v), $M^\top$ ($v^\top$) denotes its transpose. In addition, det and log det are used to denote determinant of M and its logarithm, respectively. The notation $\|v\|_{l^2}$ is used to denote the l^2 norm of a vector v. For $v \in \mathbb{R}^d$ and a symmetric positive definite matrix $\Sigma \in \mathbb{R}^{d\times d}$, $\mathcal{N}(v, \Sigma)$ denotes a normal distribution with mean v and covariance Σ. The Kronecker delta is denoted by $\delta_{i,j}$. The notation $\mathbb{S}^{d-1}$ denotes the d dimensional hypersphere in $\mathbb{R}^d$. For example, $\mathbb{S}^2 \subset \mathbb{R}^3$ is the usual sphere. The notations o and $\mathcal{O}$ denote the standard mathematical orders, while $\tilde{\mathcal{O}}$ is used to denote $\mathcal{O}$ up to logarithmic factors. For two sequences $a_n, b_n : \mathbb{N} \to \mathbb{R}$, we use the notation $a_n \sim b_n$, when $a_n = \mathcal{O}(b_n)$ and $b_n = \mathcal{O}(a_n)$. For $s \in \mathbb{N}$, we define $(2s-1)!! = \prod_{i=1}^{s}(2i-1)$. For example, $3!! = 3$ and $5!! = 15$. For a normed space $\mathcal{H}$, we use $\|.\|_{\mathcal{H}}$ to denote the norm associated with $\mathcal{H}$. For two normed spaces $\mathcal{H}_1, \mathcal{H}_2$, we write $\mathcal{H}_1 \subset \mathcal{H}_2$, if the following two conditions are satisfied. First, $\mathcal{H}_1 \subset \mathcal{H}_2$ as sets. Second, there exist a constant $c_1 > 0$, such that $\|f\|_{\mathcal{H}_2} \le c_1 \|f\|_{\mathcal{H}_1}$, for all $f \in \mathcal{H}_1$. We also write $\mathcal{H}_1 \equiv \mathcal{H}_2$, when both $\mathcal{H}_1 \subset \mathcal{H}_2$ and $\mathcal{H}_2 \subset \mathcal{H}_1$. The derivative of $a : \mathbb{R} \to \mathbb{R}$ is denoted by a'. We define $0^0 = 0$, so that $a_0(x) = (\max(0, x))^0$ corresponds to the step function: $a_0(x) = 0$, when $x \le 0$, and $a_0(x) = 1$, when $x > 0$.

B Mercer's Theorem

In this section, we give an overview of Mercer's theorem, as well as the the reproducing kernel Hilbert spaces (RKHSs) associated with the kernels. Mercer's theorem (Mercer 1909) provides a spectral decomposition of the kernel in terms of an infinite dimensional feature map (see, e.g. (Steinwart and Christmann 2008), Theorem 4.49).

Theorem 5 (Mercer's Theorem) *Let $\mathcal{X}$ be a compact domain. Let k be a continuous square integrable kernel with respect to a finite Borel measure μ. Define a positive definite operator T_k*

$$(T_k f)(.) = \int_{\mathcal{X}} k(., x) f(x) d\mu.$$

Then, there exists a sequence of eigenvalue-eigenfunction pairs $\{(\lambda_i, \phi_i)\}_{i=1}^{\infty}$ such that $\lambda_i \in \mathbb{R}^+$, and $T_k\phi_i = \lambda_i\phi_i$, for $i \ge 1$. Moreover, the kernel function can be represented as

$$k(x, x') = \sum_{i=1}^{\infty} \lambda_i \phi_i(x)\phi_i(x'),$$

where the convergence of the series holds uniformly on $\mathcal{X} \times \mathcal{X}$.

The λ_i and ϕ_i defined in Mercer's theorem are referred to as Mercer eigenvalues and Mercer eigenfunctions, respectively.

Let $\mathcal{H}_k$ denote the RKHS corresponding to k, defined as a Hilbert space equipped with an inner product $\langle ., .\rangle_{\mathcal{H}_k}$ satisfying the following: $k(., x) \in \mathcal{H}_k, \forall x \in \mathcal{X}$, and $\langle f, k(., x)\rangle_{\mathcal{H}_k} = f(x), \forall x \in \mathcal{X}, \forall f \in \mathcal{H}_k$ (reproducing property). As a consequence of Mercer's theorem, $\mathcal{H}_k$ can be represented in terms of $\{(\lambda_i, \phi_i)\}_{i=1}^{\infty}$, that is often referred to as Mercer's representation theorem (see, e.g., (Steinwart and Christmann 2008), Theorem 4.51).

Theorem 6 (Mercer's Representation Theorem) *Let $\{(\lambda_i, \phi_i)\}_{i=1}^{\infty}$ be the Mercer eigenvalue-eigenfunction pairs. Then, the RKHS corresponding to k is given by*

$$\mathcal{H}_k = \left\{ f(\cdot) = \sum_{i=1}^{\infty} w_i \lambda_i^{\frac{1}{2}} \phi_i(\cdot) : w_i \in \mathbb{R}, ||f||^2_{\mathcal{H}_k} := \sum_{i=1}^{\infty} w_i^2 < \infty \right\}.$$

Mercer's representation theorem indicates that $\{\lambda_i^{\frac{1}{2}}\phi_i\}_{i=1}^{\infty}$ form an orthonormal basis for $\mathcal{H}_k$: $\langle \lambda_i^{\frac{1}{2}}\phi_i, \lambda_{i'}^{\frac{1}{2}}\phi_{i'}\rangle_{\mathcal{H}_k} = \delta_{i,i'}$. It also provides a constructive definition for the RKHS as the span of this orthonormal basis, and a constructive definition for the $\|f\|_{\mathcal{H}_k}$ as the l^2 norm of the weights $[w_i]_{i=1}^{\infty}$ vector.

C Proof of Lemma 1

Lemma 1 offers a recursive relation over s for the RF kernel. We prove the lemma by taking the derivative of $\kappa_s(.)$, and applying the Stein's lemma (given at the end of this section).

Let $x = [0, 0, \ldots, 0, 1]^\top$ and $x' = [0, 0, \ldots, \sqrt{1-u^2}, u]^\top$, so that $x^\top x' = u$, and $x, x' \in \mathbb{S}^{d-1}$. We have

$$\begin{aligned}
\frac{\partial}{\partial u}\mathbb{E}_{w\sim\mathcal{N}(\mathbf{0}_d,\mathbf{I}_d)}\left[a_s(w^\top x)a_s(w^\top x')\right] &= \mathbb{E}_{w\sim\mathcal{N}(\mathbf{0}_d,\mathbf{I}_d)}\left[a_s(w^\top x)sa_{s-1}(w^\top x')(w_d - \frac{uw_{d-1}}{\sqrt{1-u^2}})\right] \\
&= s\mathbb{E}_{w\sim\mathcal{N}(\mathbf{0}_d,\mathbf{I}_d)}\Bigg[sa_{s-1}(w^\top x)a_{s-1}(w^\top x') \\
+(s-1)ua_s(w^\top x)a_{s-2}(w^\top x')\Bigg] &- s\mathbb{E}_{w\sim\mathcal{N}(\mathbf{0}_d,\mathbf{I}_d)}\left[\frac{(s-1)u}{\sqrt{1-u^2}}\sqrt{1-u^2}a_s(w^\top x)a_{s-2}(w^\top x')\right] \\
&= s^2\mathbb{E}_{w\sim\mathcal{N}(\mathbf{0}_d,\mathbf{I}_d)}\left[a_{s-1}(w^\top x)a_{s-1}(w^\top x')\right].
\end{aligned}$$

The first equation is obtained by taking derivative of the term inside the expectation. The second equation is obtained by applying Stein's lemma to $\mathbb{E}[a_s(w^\top x)sa_{s-1}(w^\top x')w_d]$, where w_d is the normally distributed random variable, also to $\mathbb{E}[a_s(w^\top x)sa_{s-1}(w^\top x')\frac{uw_{d-1}}{\sqrt{1-u^2}}]$, where w_{d-1} is the normally distributed random variable.

Taking into account the constant normalization of the RF kernel $c^2 = \frac{2}{(2s-1)!!}$, we have

$$\begin{aligned}\kappa_s'(u) &= \frac{2}{(2s-1)!!}\frac{\partial}{\partial u}\mathbb{E}_{w\sim\mathcal{N}(\mathbf{0}_d,\mathbf{I}_d)}\left[a_s(w^\top x)a_s(w^\top x')\right]\\ &= \frac{2s^2}{(2s-1)!!}\mathbb{E}_{w\sim\mathcal{N}(\mathbf{0}_d,\mathbf{I}_d)}\left[a_{s-1}(w^\top x)a_{s-1}(w^\top x')\right]\\ &= \frac{s^2}{2s-1}\kappa_{s-1},\end{aligned}$$

which proves the lemma.

Lemma 2 (Stein's Lemma) *Suppose $X \sim \mathcal{N}(0,1)$ is a normally distributed random variable. Consider a function $g : \mathbb{R} \to \mathbb{R}$ such that both $\mathbb{E}[Xg(X)]$ and $\mathbb{E}[g'(X)]$ exist. We then have*

$$\mathbb{E}[Xg(X)] = \mathbb{E}[g'(X)].$$

The proof of Stein's lemma follows from an integration by parts (see, e.g., Stein 1986; Nourdin and Peccati 2009; Chen, Goldstein, and Shao 2011).

D Proof of Theorem 1

This theorem follows from Theorem 1 of (Bietti and Bach 2021), which proved that the eigendecay of a rotationally invariant kernel $\kappa : [-1,1] \to \infty$ can be determined based on its asymptotic expansions around the endpoints ± 1. Recall that a rotationally invariant kernel k can be represented as $\kappa(x^\top x') = k(x,x')$. Their result is formally given in the following lemma.

Lemma 3 (Theorem 1 in (Bietti and Bach 2021)) *Assume $\kappa : [-1,1] \to \mathbb{R}$ is C^∞ and has the following asymptotic expansions around ± 1*

$$\begin{aligned}\kappa(1-t) &= p_{+1}(t) + c_{+1}t^\theta + o(t^\theta),\\ \kappa(-1+t) &= p_{-1}(t) + c_{-1}t^\theta + o(t^\theta),\end{aligned}$$

for $t > 0$, where $p_{\pm 1}$ are polynomials, and $\theta > 0$ is not an integer. Also, assume that the derivatives of κ admit similar expansions obtained by differentiating the above ones. Then, there exists $C_{d,\theta}$ such that, when i is even, if $c_{+1} \neq -c_{-1}$: $\tilde{\lambda}_i \sim (c_{+1} + c_{-1})C_{d,\theta}i^{-d-2\theta+1}$, and, when i is odd, if $c_{+1} \neq c_{-1}$: $\tilde{\lambda}_i \sim (c_{+1} - c_{-1})C_{d,\theta}l^{-d-2\theta+1}$. In the case $|c_{+1}| = |c_{-1}|$, then we have $\tilde{\lambda}_i = o(l^{-d-2\theta+1})$ for one of the two parities. If κ is infinitely differentiable on $[-1,1]$ so that no such θ exists, the $\tilde{\lambda}_i$ decays faster than any polynomial.

Building on this result, the eigendecays can be obtained by deriving the endpoint expansions for the RF and NT kernels. That is derived in (Bietti and Bach 2021) for the ReLU activation function. For completeness, here, we derive the endpoint expansions for RF and NT kernels associated with $s-1$ times differentiable activation functions $a_s(.)$.

Recall that for a 2 layer network, the corresponding RF and NT kernels are given by

$$\begin{aligned}\kappa_{\mathrm{NT},s}(x^\top x') &= c^2(x^\top x')\mathbb{E}_{w\sim\mathcal{N}(\mathbf{0}_d,\mathbf{I}_d)}[a_s'(w^\top x)a_s'(w^\top x')] + \kappa_s(x^\top x'),\\ \kappa_s(x^\top x') &= c^2\mathbb{E}_{w\sim\mathcal{N}(\mathbf{0}_d,\mathbf{I}_d)}[a_s(w^\top x)a_s(w^\top x')].\end{aligned}$$

For the special cases of $s=0$ and $s=1$, the following closed form expressions can be derived by taking the expectations.

$$\begin{aligned}\kappa_0(u) &:= \mathbb{E}_{w\sim\mathcal{N}(\mathbf{0}_d,\mathbf{I}_d)}[a_0(w^\top x)a_0(w^\top x')] = \frac{1}{\pi}(\pi - \arccos(u)),\\ \kappa_1(u) &= \frac{1}{\pi}\left(u(\pi - \arccos(u)) + \sqrt{1-u^2}\right),\end{aligned}$$

where $u = x^\top x'$. At the endpoints ± 1, $\kappa_0(u), \kappa_1(u)$ have the following asymptotic expansions.

$$\begin{aligned}\kappa_0(1-t) &= 1 - \frac{\sqrt{2}}{\pi}t^{\frac{1}{2}} + o(t^{\frac{1}{2}}),\\ \kappa_0(-1+t) &= \frac{\sqrt{2}}{\pi}t^{\frac{1}{2}} + o(t^{\frac{1}{2}}),\\ \kappa_1(1-t) &= 1 - t + \frac{2\sqrt{2}}{3\pi}t^{\frac{3}{2}} + o(t^{\frac{3}{2}}),\\ \kappa_1(-1+t) &= \frac{2\sqrt{2}}{3\pi}t^{\frac{3}{2}} + o(t^{\frac{3}{2}}).\end{aligned} \tag{9}$$

These endpoint expansions where used in (Bietti and Bach 2021) to obtain the eigendecay of the RF and NT kernels with ReLU activation functions.

We first extend the results on the endpoint expansions and the eigendecay to a 2 layer neural network with $s > 1$, in Subsection D.1. Then, we extend the derivation to $l > 2$ layer neural networks, in Subsection D.2. In Subsection D.3, we show how the eigendecays in terms of $\tilde{\lambda}_i$ can be translated to the ones in terms of λ_i.

D.1 Endpoint Expansions for 2 Layer Networks with $s > 1$

We first note that the normalization constant $c^2 = \frac{2}{(2s-1)!!}$, suggested in Section 2, ensures $\kappa_s(1) = 1$, for all $s \geq 1$. The reason is that, by the well known values for the even moments of the normal distribution, we have

$$\mathbb{E}_{w\sim\mathcal{N}(\mathbf{0}_d,\mathbf{I}_d)}[a_s(w^\top x)a_s(w^\top x')] = \frac{(2s-1)!!}{2},$$

when $x^\top x' = 1$. Also notice that $\kappa_s(-1) = 0$, for all $s \geq 1$. The reason is that when $x^\top x' = -1$, at least one of $w^\top x$ and $w^\top x'$ is non-positive.

We note that the normalization constant is considered only for the convenience of some calculations, and does not affect the exponent in the eigendecays. Specifically, scaling the kernel with a constant factor, scales the corresponding Mercer eigenvalues with the same constant. The constant factor scaling does not affect the Mercer eigenfunctions.

From Lemma 1, recall

$$\kappa_s'(.) = \frac{s^2}{2s-1}\kappa_{s-1}(.).$$

This is a key component in deriving the endpoint expansions for $s > 1$. In particular, this allows us to recursively obtain the endpoint expansions of $\kappa_s(.)$ from those of $\kappa_{s-1}(.)$, by integration. That results in the following expansions for κ_s.

$$\begin{aligned}\kappa_s(1-t) &= p_{+1,s}(t) + c_{+1,s}t^{\frac{2s+1}{2}} + o(t^{\frac{2s+1}{2}}),\\ \kappa_s(-1+t) &= p_{-1,s}(t) + c_{-1,s}t^{\frac{2s+1}{2}} + o(t^{\frac{2s+1}{2}}),\end{aligned} \tag{10}$$

Here, $p_{+1,s}(t) = -\frac{s^2}{2s-1}\int p_{+1,s-1}(t)dt$, subject to $p_{+1,s}(0) = 1$ (that follows form $\kappa_s(1) = 1$). Thus, $p_{+1,s}(t)$ can be obtained recursively, starting from $p_{+1,1}(t) = 1 - t$ (see (9)). For example,

$$\begin{aligned}p_{+1,2}(t) &= -\frac{4}{3}(-\frac{3}{4} + t - \frac{t^2}{2}),\\ p_{+1,3}(t) &= \frac{36}{15}(\frac{15}{36} - \frac{3}{4}t + \frac{1}{2}t^2 - \frac{1}{6}t^3),\end{aligned}$$

and so on.

Similarly, $p_{-1,s}(t)$ can be expressed in closed form using $p_{-1,s}(t) = \frac{s^2}{2s-1}\int p_{-1,s-1}(t)dt$ subject to $p_{-1,s}(0) = 0$ (that follows from $\kappa_s(-1) = 0$), and starting from $p_{-1,1}(t) = 0, \forall t$ (see (9)). That leads to $p_{-1,s}(t) = 0, \forall s > 1, t$.

The exact expressions of $p_{\pm1,s}(t)$ however do not affect the eigendecay. Instead, the constants $c_{\pm1,s}$ are important for the eigendecay, based on Lemma 3. The constants can also be obtained recursively. Starting from $c_{+1,1} = \frac{2\sqrt{2}}{3\pi}$ (see (9)), $c_{+1,s} = \frac{-2s^2c_{+1,s-1}}{(2s-1)(2s+1)}$. Also starting from $c_{-1,1} = \frac{2\sqrt{2}}{3\pi}$ (see (9)), $c_{-1,s} = \frac{2s^2c_{+1,s-1}}{(2s-1)(2s+1)}$. This recursive relation leads to

$$c_{-1,s} = \frac{2^s\sqrt{2}}{\pi}\prod_{r=1}^{s}\frac{r^2}{(4r^2-1)}, \tag{11}$$

and $c_{+1,s} = (-1)^{s-1}c_{-1,s}$.

In the case of RF kernel, applying Lemma 3, we have $\tilde{\lambda}_i \sim i^{-d-2s}$, for i having the opposite parity of s, and $\tilde{\lambda}_i = o(i^{-d-2s})$ for i having the same parity of s.

For the NT kernel, using 2, we have

$$\begin{aligned}\kappa_{\text{NT},s}(1-t) &= c^2(1-t)s^2\kappa_{s-1}(1-t) + \kappa_s(1-t)\\ \kappa_{\text{NT},s}(-1+t) &= c^2(-1+t)s^2\kappa_{s-1}(-1+t) + \kappa_s(-1+t)\end{aligned} \tag{12}$$

Using the expansion of the RF kernel, we get

$$\begin{aligned}\kappa_{\text{NT},s}(1-t) &= p''_{+1,s}(t) + c''_{+1,s}t^{\frac{2s-1}{2}} + o(t^{\frac{2s-1}{2}}),\\ \kappa_{\text{NT},s}(-1+t) &= p''_{-1,s}(t) + c''_{-1,s}t^{\frac{2s-1}{2}} + o(t^{\frac{2s-1}{2}}),\end{aligned} \tag{13}$$

where

$$\begin{aligned} p''_{+1,s}(t) &= c^2(1-t)s^2 p_{+1,s-1}(t) + p_s(t), \\ p''_{-1,s}(t) &= c^2(-1+t)s^2 p_{-1,s}(t) + p_{-1,s}(-1+t), \end{aligned}$$

and

$$\begin{aligned} c''_{+1,s} &= c_{+1,s-1}, \\ c''_{-1,s} &= c_{-1,s-1}. \end{aligned}$$

Thus, in the case of NT kernel, applying Lemma 3, we have $\tilde{\lambda}_i \sim i^{-d-2s+2}$, for i having the same parity of s, and $\tilde{\lambda}_i = o(i^{-d-2s+2})$ for i having the opposite parity of s.

D.2 Endpoint Expansions for $l > 2$ Layer Networks

We here extend the endpoint expansions and the eigendecays to $l > 2$ later networks. Recall $\kappa_s^l(u) = \kappa_s(\kappa_s^{l-1}(u))$. Using this recursive relation over l, we prove the following expressions

$$\begin{aligned} \kappa_s^l(1-t) &= p^l_{+1,s}(t) + c^l_{+1,s} t^{\frac{2s+1}{2}} + o(t^{\frac{2s+1}{2}}) \\ \kappa_s^l(-1+t) &= p^l_{-1,s}(t) + c^l_{-1,s} t^{\frac{2s+1}{2}} + o(t^{\frac{2s+1}{2}}) \end{aligned}$$

where $p^l_{\pm 1,s}$ are polynomials and $c^l_{\pm 1,s}$ are constants.

The notations $p_{+1,s}$ and $c_{+1,s}$ in Subsection D.1 correspond to $p^2_{+1,s}$ $c^2_{+1,s}$, where we drop the superscript specifying the number l of layers, for 2 layer networks.

For the endpoint expansion at 1, we have

$$\begin{aligned} \kappa_s^l(1-t) &= \kappa_s(\kappa_s^{l-1}(1-t)) \\ &= \kappa_s(1 - 1 + p^{l-1}_{+1,s}(t) + c^{l-1}_{+1,s} t^{\frac{2s+1}{2}} + o(t^{\frac{2s+1}{2}})) \\ &= p_{+1,s}(1 - p^{l-1}_{+1,s}(t) - c^{l-1}_{+1,s} t^{\frac{2s+1}{2}} - o(t^{\frac{2s+1}{2}})) \\ &\quad + c_{+1,s}(1 - p^{l-1}_{+1,s}(t) - c^{l-1}_{+1,s} t^{\frac{2s+1}{2}} - o(t^{\frac{2s+1}{2}}))^{\frac{2s+1}{2}} \\ &\quad + o\left((1 - p^{l-1}_{+1,s}(t) - c^{l-1}_{+1,s} t^{\frac{2s+1}{2}} - o(t^{\frac{2s+1}{2}}))^{\frac{2s+1}{2}}\right) \end{aligned}$$

Thus, we have

$$c^l_{+1,s} = (-q^{l-1,1}_{+1,s})^{\frac{2s+1}{2}} c_{+1,s} - q^{2,1}_{+1,s} c^{l-1}_{+1,s}, \tag{14}$$

where $q^{l,i}_{+1,s}$ is the coefficient of t^i in $p^l_{+1,s}(t)$. For these coefficients, we have

$$q^{l,1}_{+1,s} = -q^{2,1}_{+1,s} q^{l-1,1}_{+1,s}. \tag{15}$$

Starting from $q^{2,1}_{+1,s} = -\frac{s^2}{2s-1}$ (which can be seen from the recursive expression of $p_{+1,s}$ given in Section D.1), we get $q^{l,1}_{+1,s} = -\frac{s^{2(l-1)}}{(2s-1)^{l-1}}$. We thus have

$$c^l_{+1,s} = \left(\frac{s^{(2s+1)}}{(2s-1)^{\frac{(2s+1)}{2}}} \right)^{l-2} c_{+1,s} + \frac{s^2}{2s-1} c^{l-1}_{+1,s}. \tag{16}$$

That implies

$$c^l_{+1,s} \sim \left(\frac{s^{(2s+1)}}{(2s-1)^{\frac{(2s+1)}{2}}} \right)^{l-2} c_{+1,s}, \tag{17}$$

when $s > 1$.

The characterization of $c^l_{+1,s}$ shows that our results do not apply to deep neural networks when $s > 1$, in the sense that the constants grow exponentially in l (as stated in Remark 2).

For the endpoint expansion at -1, we have

$$\begin{aligned} \kappa_s^l(-1+t) &= \kappa_s(\kappa_s^{l-1}(-1+t)) \\ &= \kappa_s(p^{l-1}_{-1,s}(t) + c^{l-1}_{-1,s} t^{\frac{2s+1}{2}} + o(t^{\frac{2s+1}{2}})) \\ &= \kappa_s(q^{l-1,0}_{-1,s}) + \kappa'_s(q^{l-1,0}_{-1,s})(p^{l-1}_{-1,s}(t) - q^{l-1,0}_{-1,s} + c^{l-1}_{-1,s} t^{\frac{2s+1}{2}}) + o(t^{\frac{2s+1}{2}}). \end{aligned}$$

where $q^{l,i}_{-1,s}$ is the coefficient of t^i in $p^l_{-1,s}(t)$. From the expression above we can see that

$$q^{l,0}_{-1,s} = \kappa_s(q^{l-1,0}_{-1,s}) \tag{18}$$

Thus, $0 \leq q^{l,0}_{-1,s} \leq 1$. In addition, the expression above implies

$$c^l_{-1,s} = c^{l-1}_{-1,s}\kappa'_s(q^{l-1,0}_{-1,s}). \tag{19}$$

From Lemma 1, $\kappa'_s(u) \leq \frac{s^2}{2s-1}$, for all u. Therefore,

$$\begin{aligned} c^l_{-1,s} &\leq \left(\frac{s^2}{2s-1}\right)^{l-2} c_{-1,s} \\ &= o(c^l_{+1,s}), \end{aligned}$$

when, $s > 1$.

Comparing $c^l_{\pm 1,a}$, we can see that for $l > 2$, we have $|c^l_{+1,s}| \neq |c^l_{-1,s}|$. Thus, for the RF kernel with $l > 2$, applying Lemma 3, we have $\tilde{\lambda}_i \sim i^{-d-2s}$.

For the NT kernel, recall

$$\kappa^l_{\mathrm{NT},s}(u) = c^2\kappa^{l-1}_{\mathrm{NT},s}(u)\kappa'_s(\kappa^{l-1}_s(u)) + \kappa^l_s(u).$$

The second term is exactly the same as the RF kernel. Recall the expression of κ'_s based on κ_{s-1} from Lemma 1. To find the endpoint expansions of the first term, we prove the following expressions for $\kappa_{s-1}(\kappa^{l-1}_s(u))$.

$$\begin{aligned} \kappa_{s-1}(\kappa^{l-1}_s(1-t)) &= p'^l_{+1,s}(t) + c'^l_{+1,s}t^{\frac{2s-1}{2}} + o(t^{\frac{2s-1}{2}}) \\ \kappa_{s-1}(\kappa^{l-1}_s(-1+t)) &= p'^l_{-1,s}(t) + c'^l_{-1,s}t^{\frac{2s+1}{2}} + o(t^{\frac{2s+1}{2}}) \end{aligned}$$

We use the notation used for the RF kernel to write the following expansion around 1

$$\begin{aligned} \kappa_{s-1}(\kappa^{l-1}_s(1-t)) &= \kappa_{s-1}(1-1+p^{l-1}_{+1,s}(t) + c^{l-1}_{+1,s}t^{\frac{2s+1}{2}} + o(t^{\frac{2s+1}{2}})) \\ &= p_{+1,s-1}(1-p^{l-1}_{+1,s}(t) - c^{l-1}_{+1,s}t^{\frac{2s+1}{2}} - o(t^{\frac{2s+1}{2}})) \\ &\quad + c_{+1,s-1}(1-p^{l-1}_{+1,s}(t) - c^{l-1}_{+1,s}t^{\frac{2s+1}{2}} - o(t^{\frac{2s+1}{2}}))^{\frac{2s-1}{2}} \\ &\quad + o\left((1-p^{l-1}_{+1,s}(t) - c^{l-1}_{+1,s}t^{\frac{2s+1}{2}} - o(t^{\frac{2s+1}{2}}))^{\frac{2s-1}{2}}\right). \end{aligned}$$

Thus, we have

$$c'^l_{+1,s} = (-q^{l-1,1}_{+1,s})^{\frac{2s-1}{2}}c_{+1,s-1} - q^{2,1}_{+1,s-1}c^{l-1}_{+1,s} \tag{20}$$

Recall, form the analysis of the RF kernel that $q^{l,1}_{+1,s} = -\frac{s^2(l-1)}{(2s-1)^{l-1}}$. We thus have

$$c'^l_{+1,s} = \left(\frac{s^{(2s-1)}}{(2s-1)^{\frac{(2s-1)}{2}}}\right)^{l-2} c_{+1,s-1} + \frac{(s-1)^2}{2s-3}c^{l-1}_{+1,s}. \tag{21}$$

That implies

$$c'^l_{+1,s} \sim \left(\frac{s^{(2s-1)}}{(2s-1)^{\frac{(2s-1)}{2}}}\right)^{l-2} c_{+1,s-1}. \tag{22}$$

For the endpoint expansion at -1, we have

$$\begin{aligned} \kappa_{s-1}(\kappa^{l-1}_s(-1+t)) &= \kappa_{s-1}(p^{l-1}_{-1,s}(t) + c^{l-1}_{-1,s}t^{\frac{2s+1}{2}} + o(t^{\frac{2s+1}{2}})) \\ &= \kappa_{s-1}(q^{l-1,0}_{-1,s}) + \kappa'_{s-1}(q^{l-1,0}_{-1,s})(p^{l-1}_{-1,s}(t) - q^{l-1,0}_{-1,s} + c^{l-1}_{-1,s}t^{\frac{2s+1}{2}}) + o(t^{\frac{2s+1}{2}}). \end{aligned}$$

We thus have

$$c'^l_{-1,s} = c^{l-1}_{-1,s-1}\kappa'_{s-1}(q^{l-1,0}_{-1,s}). \tag{23}$$

Since $\kappa'_{s-1} \leq \frac{(s-1)^2}{2s-3}$, we have

$$c^{\prime l}_{-1,s} \leq \left(\frac{s^2}{2s-3}\right)^{l-2} c_{-1,s-1}.$$

Now we use the end point expansions of κ^l_s and $\kappa_{s-1}(\kappa^{l-1}_s)$ to obtain the following endpoint expansions for $\kappa^l_{\mathrm{NT},s}$.

$$\begin{aligned} \kappa^l_{\mathrm{NT},s}(1-t) &= p^{\prime\prime l}_{+1,s}(t) + c^{\prime\prime l}_{+1,s} t^{\frac{2s-1}{2}} + o(t^{\frac{2s-1}{2}}), \\ \kappa^l_{\mathrm{NT},s}(-1+t) &= p^{\prime\prime l}_{-1,s}(t) + c^{\prime\prime l}_{-1,s} t^{\frac{2s-1}{2}} + o(t^{\frac{2s-1}{2}}). \end{aligned}$$

We have

$$c^{\prime\prime l}_{+1,s} = \frac{c^2 s^2}{2s-1} q^{\prime l,0}_{+1,s} c^{\prime\prime l-1}_{+1,s} + \frac{c^2 s^2}{2s-1} q^{\prime\prime l-1,0}_{+1,s} c^{\prime l}_{+1,s},$$

and,

$$q^{\prime\prime l,0}_{+1,s} = \frac{c^2 s^2}{2s-1} q^{\prime\prime l-1,0}_{+1,s} q^{\prime l,0}_{+1,s} + q^{l,0}_{+1,s}. \tag{24}$$

Since $\kappa_s(1) = 1$, by induction we have $\kappa^l_s(1) = 1, \forall l \geq 2$. In addition, $\kappa_{s-1}(\kappa^{l-1}_s(1)) = \kappa_{s-1}(1) = 1, \forall l \geq 3$. Therefore $q^{\prime l,0}_{+1,s}, q^{l,0}_{+1,s} = 1$. Thus,

$$q^{\prime\prime l,0}_{+1,s} = \frac{c^2 s^2}{2s-1} q^{\prime\prime l-1,0}_{+1,s} + 1. \tag{25}$$

Starting from $q^{\prime\prime 2,0}_{+1,s} = \frac{s^2}{2s-1} + 1$, we can derive the following expression for $q^{\prime\prime l,0}_{+1,s}$, when $s > 1$,

$$\begin{aligned} q^{\prime\prime l,0}_{+1,s} = \frac{1}{c^2}\left(\frac{c^2 s^2}{2s-1}\right)^{l-1} &+ \frac{\left(\frac{c^2 s^2}{2s-1}\right)^{l-1} - 1}{\left(\frac{c^2 s^2}{2s-1}\right) - 1} \\ &\sim \frac{1}{c^2}\left(\frac{c^2 s^2}{2s-1}\right)^{l-1}. \end{aligned}$$

When $s = 1$, $q^{\prime\prime l,0}_{+1,s} = l$, that is consistent with the results in (Bietti and Bach 2021).
For $c^{\prime\prime l}_{+1,s}$, we thus have

$$c^{\prime\prime l}_{+1,s} = \frac{c^2 s^2}{2s-1} c^{\prime\prime l-1}_{+1,s} + \frac{c^2 s^2}{2s-1} q^{\prime\prime l-1,0}_{+1,s} c^{\prime l}_{+1,s}.$$

Replacing the expressions for $q^{\prime\prime l-1,0}_{+1,s}$ and $c^{\prime l}_{+1,s}$ derived above, we obtain

$$c^{\prime\prime l}_{+1,s} \sim \left(\frac{1}{c^2}\right)^{(\frac{2s-1}{2})(l-2)+1} \left(\frac{c^2 s^2}{2s-1}\right)^{(\frac{2s+1}{2})(l-2)+1} c_{+1,s-1}. \tag{26}$$

For the constant in the expansion around -1, we have

$$c^{\prime\prime l}_{-1,s} = \frac{c^2 s^2}{2s-1} q^{\prime l,0}_{-1,s} c^{\prime\prime l-1}_{-1,s}, \tag{27}$$

where $q^{\prime l,0}_{-1,s} = \kappa_{s-1}(q^{l-1,0}_{-1,s})$ is bounded between 0 and 1. Thus, starting from $c^{\prime\prime 2}_{-1,s} = c_{-1,s-1}$, we obtain

$$c^{\prime\prime l}_{-1,s} \leq \left(\frac{c^2 s^2}{2s-1}\right)^{l-2} c_{-1,s-1} \tag{28}$$

Comparing $c^{\prime\prime l}_{\pm 1,a}$, we can see that for $l > 2$, we have $|c^{\prime\prime l}_{+1,s}| \neq |c^{\prime\prime l}_{-1,s}|$. Thus, for the NT kernel with $l > 2$, applying Lemma 3, we have $\tilde{\lambda}_i \sim i^{-d-2s+2}$.

D.3 The Eigendecays in Terms of λ_i

Recall the multiplicity $N_{d,i} = \frac{2i+d-2}{i}\binom{i+d-3}{d-2}$ of $\tilde{\lambda}_i$. Here we take into account this multiplicity to give the eigendecay expressions in terms of λ_i.

Using $(\frac{n}{k})^k \leq \binom{n}{k} \leq (\frac{ne}{k})^k$, for all $k, n \in \mathbb{N}$, we have, for all $i > 1$ and $d > 2$,

$$2(i-1)^{d-2}(\frac{1}{d-2}+\frac{1}{i-1})^{d-2} \leq N_{d,i} \leq \frac{(d+2)e^{d-2}}{2}(i-1)^{d-2}(\frac{1}{d-2}+\frac{1}{i-1})^{d-2},$$

where we used $2 < \frac{2i+d-2}{i} < \frac{d+2}{2}$. We thus have $N_{d,i} \sim (i-1)^{d-2}$, for the scaling of $N_{d,i}$ with i, with constants given above.

Recall we define $\lambda_i = \tilde{\lambda}_{i'}$, for i and i' which satisfy $\sum_{i''=1}^{i'-1} N_{d,i''} < i \leq \sum_{i''=1}^{i'} N_{d,i''}$. Using $N_{d,i} \sim (i-1)^{d-2}$, we have $\sum_{i''=1}^{i'} N_{d,i''} \sim i'^{d-1}$. Thus $\lambda_i \sim \tilde{\lambda}_{i'} = \tilde{\lambda}_{i^{\frac{1}{d-1}}}$. Replacing i with $i^{\frac{1}{d-1}}$, in the expressions derived for $\tilde{\lambda}_i$ in this section, we obtain the eigendecay expressions for λ_i reported in Table 2.

E Proof of Theorem 2

The Matérn kernel can also be decomposed in the basis of spherical harmonics. In particular, (Borovitskiy et al. 2020) proved the following expression for the Matérn kernel with smoothness parameter ν on the hypersphere $\mathbb{S}^{d-1}$ (see, also Dutordoir, Durrande, and Hensman 2020, Appendix B)

$$k_\nu(x,x') = \sum_{i=1}^{\infty}\sum_{j=1}^{N_{d,i}}(\frac{2\nu}{\kappa^2}+i(i+d-2))^{-(\nu+\frac{d-1}{2})}\tilde{\phi}_{i,j}(x)\tilde{\phi}_{i,j}(x'),$$

where $\tilde{\phi}_{i,j}$ are the spherical harmonics.

Theorem 6 implies that the RKHS of Metérn kernel is also constructed as a span of spherical harmonics. Thus, in order to show the equivalence of the RKHSs of the neural kernels with various activation functions and a Matérn kernel with the corresponding smoothness, we show that the ratio between their norms is bounded by absolute constants.

Let $f \in \mathcal{H}_{k_{\mathrm{NT},s}}^{l}$, with $l \geq 2$. As a result of Mercer's representation theorem, we have

$$\begin{aligned} f(\cdot) &= \sum_{i=1}^{\infty}\sum_{j=1}^{N_{d,i}} w_{i,j}\tilde{\lambda}_i^{\frac{1}{2}}\tilde{\phi}_{i,j}(\cdot) \\ &= \sum_{i=1}^{\infty}\sum_{j=1}^{N_{d,i}} w_{i,j}\frac{\tilde{\lambda}_i^{\frac{1}{2}}}{(\frac{2\nu}{\kappa^2}+i(i+d-2))^{-\frac{1}{2}(\nu+\frac{d-1}{2})}}(\frac{2\nu}{\kappa^2}+i(i+d-2))^{-\frac{1}{2}(\nu+\frac{d-1}{2})}\tilde{\phi}_{i,j}(\cdot). \end{aligned}$$

Note that

$$\frac{\tilde{\lambda}_i}{(\frac{2\nu}{\kappa^2}+i(i+d-2))^{-(\nu+\frac{d-1}{2})}} = \mathcal{O}(\frac{\tilde{\lambda}_i}{i^{-2\nu-d+1}}).$$

For the NT kernel, from Theorem 1, $\tilde{\lambda}_i = \mathcal{O}(i^{-d-2s+2})$. Thus, when $\nu = s - \frac{1}{2}$,

$$\frac{\tilde{\lambda}_i}{(\frac{2\nu}{\kappa^2}+i(i+d-2))^{-(\nu+\frac{d-1}{2})}} = \mathcal{O}(1).$$

So, with $\nu = s - \frac{1}{2}$ we have

$$\begin{aligned} \|f\|^2_{\mathcal{H}_{k_\nu}} &= \sum_{i=1}^{\infty}\sum_{j=1}^{N_{d,i}} w_{i,j}^2\frac{\tilde{\lambda}_i}{(\frac{2\nu}{\kappa^2}+i(i+d-2))^{-(\nu+\frac{d-1}{2})}} \\ &= \mathcal{O}(\sum_{i=1}^{\infty}\sum_{j=1}^{N_{d,i}} w_{i,j}^2) \\ &= \mathcal{O}(\|f\|^2_{\mathcal{H}_{\kappa_{\mathrm{NT},s}^{l}}}). \end{aligned}$$

That proves $\mathcal{H}_{k_{\mathrm{NT},s}^{l}} \subset \mathcal{H}_{k_\nu}$.

A similar proof shows that when $\nu = s + \frac{1}{2}$, $\mathcal{H}^l_{k_s} \subset \mathcal{H}_{k_\nu}$.

Now, let $f \in \mathcal{H}_{k_\nu}$. As a result of Mercer's representation theorem, we have

$$\begin{aligned} f(\cdot) &= \sum_{i=1}^{\infty}\sum_{j=1}^{N_{d,i}} w_{i,j}(\frac{2\nu}{\kappa^2} + i(i+d-2))^{-\frac{1}{2}(\nu+\frac{d-1}{2})}\tilde{\phi}_{i,j}(\cdot) \\ &= \sum_{i=1}^{\infty}\sum_{j=1}^{N_{d,i}} w_{i,j}\frac{(\frac{2\nu}{\kappa^2} + i(i+d-2))^{-\frac{1}{2}(\nu+\frac{d-1}{2})}}{\tilde{\lambda}_i^{\frac{1}{2}}}\tilde{\lambda}_i^{\frac{1}{2}}\tilde{\phi}_{i,j}(\cdot). \end{aligned}$$

For the NT kernel with $l > 2$, from Theorem 1, $\tilde{\lambda}_l \sim l^{-d-2s+2}$. Thus, when $\nu = s - \frac{1}{2}$,

$$\frac{(\frac{2\nu}{\kappa^2} + l(l+d-2))^{-(\nu+\frac{d-1}{2})}}{\tilde{\lambda}_l} = \mathcal{O}(1).$$

So, with $\nu = s - \frac{1}{2}$ we have

$$\begin{aligned} \|f\|^2_{\mathcal{H}_{\kappa^l_{\mathrm{NT},s}}} &= \sum_{i=1}^{\infty}\sum_{j=1}^{N_{d,i}} w^2_{i,j}\frac{(\frac{2\nu}{\kappa^2} + i(i+d-2))^{-(\nu+\frac{d-1}{2})}}{\tilde{\lambda}_i} \\ &= \mathcal{O}(\sum_{i=1}^{\infty}\sum_{j=1}^{N_{d,i}} w^2_{i,j}) \\ &= \mathcal{O}(\|f\|^2_{\mathcal{H}_{k_\nu}}). \end{aligned}$$

That proves, for $l > 2$, when $\nu = s - \frac{1}{2}$, $\mathcal{H}_{k_\nu} \subset \mathcal{H}^l_{k_{\mathrm{NT},s}}$. A similar proof shows that for $l > 2$, when $\nu = s + \frac{1}{2}$, $\mathcal{H}_{k_\nu} \subset \mathcal{H}^l_{k_s}$, which completes the proof of Theorem 2.

F Proof of Theorem 3

Recall the definition of the information gain for a kernel κ,

$$\mathcal{I}(Y_n; F) = \frac{1}{2}\log\det\left(\boldsymbol{I}_n + \frac{1}{\lambda^2}\mathbf{K}_n\right).$$

To bound the information gain, we define two new kernels resulting from the partitioning of the eigenvalues of the neural kernels to the first M eigenvalues and the remainder. In particular, we define

$$\begin{aligned} \tilde{\kappa}(x^\top x) &= \sum_{i=1}^{M}\sum_{j=1}^{N_{d,i}} \tilde{\lambda}_i\tilde{\phi}_{i,j}(x)\tilde{\phi}_{i,j}(x'), \\ \tilde{\tilde{\kappa}}(x^\top x) &= \sum_{i=M+1}^{\infty}\sum_{j=1}^{N_{d,i}} \tilde{\lambda}_i\tilde{\phi}_{i,j}(x)\tilde{\phi}_{i,j}(x'). \end{aligned} \tag{29}$$

We present the proof for the NT kernel. A similar proof applies to the RF kernel.

The truncated kennel at M eigenvalues, $\tilde{\kappa}(x^\top x)$, corresponds to the projection of the RKHS of κ onto a finite dimensional space with dimension $\sum_{i=1}^{M}\sum_{j=1}^{N_{d,i}}$.

We use the notations $\tilde{\mathbf{K}}_n$ and $\tilde{\tilde{\mathbf{K}}}_n$ to denote the kernel matrices corresponding to $\tilde{\kappa}$ and $\tilde{\tilde{\kappa}}$, respectively.

As proposed in (Vakili, Khezeli, and Picheny 2021), we write

$$\begin{aligned} \log\det\left(\mathbf{I}_n + \frac{1}{\lambda^2}\mathbf{K}_n\right) &= \log\det\left(\mathbf{I}_n + \frac{1}{\lambda^2}(\tilde{\mathbf{K}}_n + \tilde{\tilde{\mathbf{K}}}_n)\right) \\ = \log\det\left(\mathbf{I}_n + \frac{1}{\lambda^2}\tilde{\mathbf{K}}_n\right) &+ \log\det\left(\mathbf{I}_n + \frac{1}{\lambda^2}(\mathbf{I}_n + \frac{1}{\lambda^2}\tilde{\mathbf{K}}_n)^{-1}\tilde{\tilde{\mathbf{K}}}_n\right). \end{aligned} \tag{30}$$

The analysis in (Vakili, Khezeli, and Picheny 2021) crucially relies on an assumption that the eigenfunctions are uniformly bounded. In the case of spherical harmonics however (see, e.g., Stein and Weiss 2016, Corollary 2.9)

$$\sup_{\substack{i=1,2,\ldots,\\ j=1,2,\ldots,N_{d,i},\\ x\in\mathcal{X}}} \tilde{\phi}_{i,j}(x)=\infty. \tag{31}$$

Thus, their analysis does not apply to the case of neural kernels on the hypersphere.

To bound the two terms on the right hand side of (30), we first prove in Lemma 4 that $\tilde{\tilde{\kappa}}(x,x')$ approaches 0 as M grows (uniformly in x and x'), with a certain rate.

Lemma 4 *For the NT kernel, we have $\tilde{\tilde{\kappa}}(x^\top x')=\mathcal{O}(M^{-2s+1})$, for all $x,x'\in\mathcal{X}$.*

Proof of Lemma 4: Recall that the eigenspace corresponding to $\tilde{\lambda}_i$ has dimension $N_{d,i}$ and consists of spherical harmonics of degree i. Legendre addition theorem (Maleček and Nádeník 2001) states

$$\sum_{j=1}^{N_{d,i}}\phi_{i,j}(x)\phi_{i,j}(x')=c_{i,d}C_i^{(d-2)/2}(\cos(\mathtt{d}(x,x'))), \tag{32}$$

where $\mathtt{d}$ is the geodesic distance on the hypersphere, $C_i^{(d-2)/2}$ are Gegenbauer polynomials (those are the same as Legendre polynomials up to a scaling factor), and the constant $c_{i,d}$ is

$$c_{i,d}=\frac{N_{d,i}\Gamma((d-2)/2)}{2\pi^{(d-2)/2}C_k^{(d-2)/2}(1)}. \tag{33}$$

We thus have

$$\begin{aligned}\tilde{\tilde{\kappa}}(x^\top x') &= \sum_{i=M+1}^{\infty}\tilde{\lambda}_i\sum_{j=1}^{N_{d,i}}\phi_{i,j}(x)\phi_{i,j}(x')\\ &= \sum_{i=M+1}^{\infty}\tilde{\lambda}_i c_{i,d}C_i^{(d-2)/2}(\cos(\mathtt{d}(x,x')))\\ &\le \sum_{i=M+1}^{\infty}\tilde{\lambda}_i N_{d,i}\frac{\Gamma((d-2)/2)}{2\pi^{((d-2)/2)}}\\ &\sim \sum_{i=M+1}^{\infty} i^{-d-2s+2}i^{d-2}\\ &\sim M^{-2s+1}.\end{aligned}$$

Here, the inequality follows from $C_i^{(d-2)/2}(\cos(\mathtt{d}(x,x')))\le C_i^{(d-2)/2}(1)$, because the Gegenbauer polynomials attain their maximum at the endpoint 1. For the fourth line we used $N_{d,i}\sim i^{d-2}$. The implied constants include the implied constants in $N_{d,i}\sim i^{d-2}$ given in Section D.3, and $\frac{\Gamma((d-2)/2)}{2\pi^{((d-2)/2)}}$.

□

We also introduce a notation $N_M=\sum_{i=1}^M N_{d,i}$ for the number of eigenvalues corresponding to the spherical harmonics of degree up to M, taking into account their multiplicities, that satisfies $N_M\sim M^{d-1}$ (see Section D.3).

We are now ready to bound the two terms on the right hand side of (30). Let us define $\mathbf{\Phi}_{n,N_M}=[\boldsymbol{\phi}_{N_M}(x_1),\boldsymbol{\phi}_{N_M}(x_2),\ldots,\boldsymbol{\phi}_{N_M}(x_n)]^\top$, an $n\times N_M$ matrix which stacks the feature vectors $\boldsymbol{\phi}_{N_M}(x_i)=[\phi_j(x_i)]_{j=1}^{N_M}$, $i=1,\ldots,n$, at the observation points, as its rows. Notice that

$$\tilde{\mathbf{K}}_n=\mathbf{\Phi}_{n,N_M}\Lambda_{N_M}\mathbf{\Phi}_{n,N_M}^\top,$$

where Λ_{N_M} is the diagonal matrix of the eigenvalues defined as $[\Lambda_{N_M}]_{i,j}=\lambda_i\delta_{i,j}$.

Now, consider the Gram matrix

$$\boldsymbol{G}=\Lambda_{N_M}^{\frac{1}{2}}\mathbf{\Phi}_{n,N_M}^\top\mathbf{\Phi}_{n,N_M}\Lambda_{N_M}^{\frac{1}{2}}.$$

As it was shown in (Vakili, Khezeli, and Picheny 2021), by matrix determinant lemma, we have

$$\begin{aligned}\log\det(\mathbf{I}_n+\frac{1}{\lambda^2}\tilde{\mathbf{K}}_n) &= \log\det(\mathbf{I}_{N_M}+\frac{1}{\lambda^2}\boldsymbol{G})\\ &\le N_M\log\left(\frac{1}{N_M}\operatorname{tr}(\mathbf{I}_{N_M}+\frac{1}{\lambda^2}\boldsymbol{G})\right)\\ &= N_M\log(1+\frac{n}{\lambda^2 N_M}).\end{aligned}$$

To upper bound the second term on the right hand side of (30), we use Lemma 4. In particular, since

$$\mathrm{tr}\left((\mathbf{I}_n + \frac{1}{\lambda^2}\tilde{\mathbf{K}}_n)^{-1}\tilde{\tilde{\mathbf{K}}}_n\right) \leq \mathrm{tr}(\tilde{\tilde{\mathbf{K}}}_n),$$

and $[\tilde{\tilde{\mathbf{K}}}_n]_{i,i} = \mathcal{O}(M^{-2s+1})$, we have

$$\mathrm{tr}\left(\mathbf{I}_n + \frac{1}{\lambda^2}(\mathbf{I}_n + \frac{1}{\lambda^2}\tilde{\mathbf{K}}_n)^{-1}\tilde{\tilde{\mathbf{K}}}_n\right) = \mathcal{O}\left(n(1+\frac{1}{\lambda^2}M^{-2s+1})\right).$$

Therefore

$$\begin{aligned}\log\det\left(\mathbf{I}_n + \frac{1}{\lambda^2}(\mathbf{I}_n + \frac{1}{\lambda^2}\tilde{\mathbf{K}}_n)^{-1}\tilde{\tilde{\mathbf{K}}}_n\right) &\leq n\log\left(\mathcal{O}(1+\frac{1}{\lambda^2}M^{-2s+1})\right)\\ &\leq \frac{nM^{-2s+1}}{\lambda^2} + \mathcal{O}(1),\end{aligned}$$

where for the last line we used $\log(1+z) \leq z$ which holds for all $z \in \mathbb{R}$.

We thus have $\gamma_k(n) = \mathcal{O}(M^{d-1}\log(n) + nM^{-2s+1})$. Choosing $M \sim n^{\frac{1}{d+2s-2}}(\log(n))^{\frac{-1}{d+2s-2}}$, we obtain

$$\gamma_{\kappa^l_{\mathrm{NT},s}}(n) = \mathcal{O}\left(n^{\frac{d-1}{d+2s-2}}(\log(n))^{\frac{2s-1}{d+2s-2}}\right). \tag{34}$$

For example, with ReLU activation functions, we have

$$\gamma_{\kappa^l_{\mathrm{NT},1}}(n) = \mathcal{O}\left(n^{\frac{d-1}{d}}(\log(n))^{\frac{1}{d}}\right).$$

G Proof of Theorem 4

As stated in the paper, the proof follows the same steps as in the proof of Theorem 3 of (Vakili et al. 2021). Their theorem holds for general kernels, provided the bound on MIG. We have adopted their theorem for the special case of neural kernels, inserting our novel bounds on MIG of the neural kernels given in Theorem 3. For completeness, we include a detailed proof here.

The proof consists of tow components. First, we bound the uncertainty estimate $\sigma_n(x)$ in terms of $\gamma_k(n)$. Our novel bounds on $\gamma_k(n)$ of neural kernels given in Theorem 3 allow us to derive explicit bounds on $\sigma_n(x)$. Then, we bound the error $|f(x) - \hat{f}_n(x)|$ in terms of $\sigma_n(x)$, using implicit error bounds.

Recall the way that the dataset $\tilde{\mathcal{D}}_n$ is collected: $x_i = \arg\max_{x\in\mathcal{X}}\sigma_{i-1}(x)$. This ensures that, $\forall x \in \mathcal{X}$, $\sigma_{i-1}(x) \leq \sigma_{i-1}(x_i)$. Due to positive definiteness of the kernel matrix, conditioning on a larger dataset reduces the uncertainty. Thus $\forall x \in \mathcal{X}$ and $\forall i \leq n$, $\sigma_n(x) \leq \sigma_i(x)$. Therefore $\sigma^2_n(x)$ is upper bounded by the average of $\sigma^2_{i-1}(x_i)$ over i: $\forall x \in \mathcal{X}$

$$\sigma^2_n(x) \leq \frac{1}{n}\sum_{i=1}^n \sigma^2_{i-1}(x_i). \tag{35}$$

It is shown that (e.g., see, Srinivas et al. 2010, Lemmas 5.3, 5.4)

$$\sum_{i=1}^n \sigma^2_{i-1}(x_i) \leq \frac{2\mathcal{I}(Y_n;F)}{\log(1+\frac{1}{\lambda^2})}. \tag{36}$$

Thus, combining (35) and (36), and by definition of $\gamma_k(n)$, we have, $\forall x \in \mathcal{X}$

$$\sigma_n(x) \leq \sqrt{\frac{2\gamma_k(n)}{n\log(1+\frac{1}{\lambda^2})}}. \tag{37}$$

Now, inserting our bounds on $\gamma_{\kappa^l_{\mathrm{NT},s}}(n)$ and $\gamma_{\kappa^l_s}(n)$ of the neural kernels from Theorem 3, we obtain, $\forall x \in \mathcal{X}$

$$\begin{aligned}\sigma_n(x) &\leq n^{\frac{-2s+1}{2d+4s-4}}(\log(n))^{\frac{2s-1}{2d+4s-4}}\sqrt{\frac{2}{\log(1+\frac{1}{\lambda^2})}}, \quad \text{in the case of NT kernel,}\\ \sigma_n(x) &\leq n^{\frac{-2s-1}{2d+4s}}(\log(n))^{\frac{2s+1}{2d+4s}}\sqrt{\frac{2}{\log(1+\frac{1}{\lambda^2})}}, \quad \text{in the case of RF kernel.}\end{aligned} \tag{38}$$

The second component of the proof is summarized in the following lemma.

Lemma 5 *Under Assumption 1, we have, with probability at least* $1-\delta$*,* $\forall x \in \mathcal{X}$

$$|f(x) - \hat{f}_n(x)| \leq \left(B + C(\log(\frac{n^{d-1}}{\delta}))^{\frac{1}{2}}\right)\sigma_n(x) + \frac{2}{\sqrt{n}}, \tag{39}$$

where C *is an absolute constant, and* B *is the upper bound on the RKHS norm of* f *given in Assumption 1.*

Lemma 5 follows from equation 7, and a probability union bound over a discretization of the domain. A detailed proof of this lemma is given at the end of this section.

Inserting the bounds on $\sigma_n(x)$ from (38) in Lemma 5, we have, with probability at least $1-\delta$, uniformly in x,

$$\begin{aligned}
|f(x) - \hat{f}_n(x)| &= \mathcal{O}\left(n^{\frac{-2s+1}{2d+4s-4}}(\log(n))^{\frac{2s-1}{2d+4s-4}}(\log(\frac{n^{d-1}}{\delta}))^{\frac{1}{2}}\right), \quad \text{in the case of NT kernel,}\\
|f(x) - \hat{f}_n(x)| &= \mathcal{O}\left(n^{\frac{-2s-1}{2d+4s}}(\log(n))^{\frac{2s+1}{2d+4s}}(\log(\frac{n^{d-1}}{\delta}))^{\frac{1}{2}}\right), \quad \text{in the case of RF kernel.}
\end{aligned}$$

That completes the proof.

Proof of Lemma 5:

Recall the implicit error bound given in equation 7: For a fixed $x \in \mathcal{X}$, under Assumption 1, we have, with probability at least $1-\delta$ (Vakili et al. 2021),

$$|f(x) - \hat{f}_n(x)| \leq \beta(\delta)\sigma_n(x), \tag{40}$$

where $\beta(\delta) = B + \frac{R}{\lambda}\sqrt{2\log(\frac{2}{\delta})}$. Lemma 5 extends this inequality to a uniform bound in x, using a probability union bound over a discretization of the domain.

For $f \in \mathcal{H}_k$, with $\|f\|_{\mathcal{H}_k} \leq B$, and for $n \in \mathbb{N}$, there exists a fine discretization $\mathbb{X}_n$ of $\mathcal{X}$ with size $|\mathbb{X}_n|$ such that $f(x) - f([x]_n) \leq \frac{1}{\sqrt{n}}$, where $[x]_n = \arg\min_{x' \in \mathbb{X}_n} \|x' - x\|_{l^2}$ is the closest point in $\mathbb{X}_n$ to x, and $|\mathbb{X}_n| \leq cB^{d-1}n^{(d-1)/2}$, where c is an absolute constant independent of n and B (Chowdhury and Gopalan 2017; Vakili et al. 2021).

Under Assumption 1, it can be shown that: with probability at least $1-\delta/2$ (Vakili et al. 2021, Lemma 4)

$$\|\hat{f}_n\|_{\mathcal{H}_k} \leq \underbrace{B + \frac{R\sqrt{n}}{\lambda}\sqrt{2\log(\frac{4n}{\delta})}}_{U(\delta)}. \tag{41}$$

Note that $\hat{f}_n$ is a random function, where the randomness comes from the randomness in observation noise.

We thus conclude that there is a discretization $\tilde{\mathbb{X}}_n$ with size $|\tilde{\mathbb{X}}_n| \leq c(U(\delta))^{d-1}n^{(d-1)/2}$ such that $|f(x) - f([x]_n)| \leq \frac{1}{\sqrt{n}}$, and with probability at least $1-\delta/2$,

$$|\hat{f}_n(x) - \hat{f}_n([x]_n)| \leq \frac{1}{\sqrt{n}}. \tag{42}$$

Let $\delta' = \frac{\delta}{2c(U(\delta))^{d-1}n^{(d-1)/2}}$. A probability union bound over the discretization $\tilde{\mathbb{X}}_n$ implies that: With probability at least $1-\delta/2$, uniformly in $x \in \tilde{\mathbb{X}}_n$,

$$|f(x) - \hat{f}_n(x)| \leq \beta(\delta')\sigma_n(x). \tag{43}$$

Accounting for the discretization error from (42), and a probability union bound over (41) and (43), we have, with probability at least $1-\delta$, uniformly in $x \in \mathcal{X}$,

$$\begin{aligned}
|f(x) - \hat{f}_n(x)| &\leq |f(x) - f([x]_n)| + |f([x]_n) - \hat{f}_n([x]_n)| + |\hat{f}_n([x]_n) - \hat{f}_n(x)|\\
&\leq |f([x_n]) - \hat{f}_n([x_n])| + \frac{2}{\sqrt{n}}\\
&\leq \beta(\delta')\sigma_n(x) + \frac{2}{\sqrt{n}}.
\end{aligned}$$

The first line holds by triangle inequality, the second line comes from the discretization error, and the third line holds by (43).

Inserting the value of $U(\delta) = B + \frac{R\sqrt{n}}{\lambda}\sqrt{2\log(\frac{2n}{\delta})}$ in δ', and the value of $\delta' = \frac{\delta}{2c(U(\delta))^{d-1}n^{(d-1)/2}}$ in $\beta(\cdot) = B + \frac{R}{\lambda}\sqrt{2\log(\frac{2}{\cdot})}$, we arrive at the lemma.

H Details on the Experiments

In this section, we provide further details on the experiments shown in the main text, Section 6. The code will be made available upon the acceptance of the paper.

We consider NT kernels $\kappa_{\mathrm{NT},s}(.)$, with $s = 1, 2, 3$, which correspond to wide fully connected 2 layer neural networks with activation functions $a_s(.)$. In the first step, we create a synthetic function f belonging to the RKHS of a NT kernel κ. For this purpose, we randomly generate $n_0 = 100$ points on the hypersphere $\mathbb{S}^{d-1}$. Let $\hat{X}_{n_0} = [\hat{x}_i]_{i=1}^{n_0}$ denote the vector of these points. We also randomly sample $\hat{Y}_{n_0} = [\hat{y}_i]_{i=1}^{n_0}$ from a multivariate Gaussian distribution $\mathcal{N}(\mathbf{0}_{n_0}, \mathbf{K}_{n_0})$, where $[\hat{\mathbf{K}}_{n_0}]_{i,j} = \kappa(\hat{x}_i^\top \hat{x}_j)$. We define a function $g(.) = \hat{\mathbf{k}}_{n_0}^\top(.)(\hat{\mathbf{K}}_{n_0} + \delta^2\mathbf{I}_{n_0})^{-1}\hat{Y}_{n_0}$, where $\delta^2 = 0.01$ and $[\hat{\mathbf{K}}_{n_0}(x)]_i = \kappa(x^\top \hat{x}_i)$. We then normalize g with its range to obtain $f(.) = \frac{g(.)}{\max_{x\in\mathcal{X}} g(x) - \min_{x\in\mathcal{X}} g(x)}$. For a fixed $\hat{X}_{n_0}$ and $\hat{Y}_{n_0}$, g is a linear combination of partial applications of the kernel. Thus g is in the RKHS of κ, and its RKHS norm can be bounded as follows.

$$\begin{aligned}
\|g\|^2_{\mathcal{H}_\kappa} &= \left\langle \hat{\mathbf{k}}_{n_0}^\top(.)(\hat{\mathbf{K}}_{n_0} + \delta^2\mathbf{I}_{n_0})^{-1}\hat{Y}_{n_0}, \hat{\mathbf{k}}_{n_0}^\top(.)(\hat{\mathbf{K}}_{n_0} + \delta^2\mathbf{I}_{n_0})^{-1}\hat{Y}_{n_0} \right\rangle_{\mathcal{H}_\kappa}\\
&= \hat{Y}_{n_0}^\top(\hat{\mathbf{K}}_{n_0} + \delta^2\mathbf{I}_{n_0})^{-1}\hat{\mathbf{K}}_{n_0}(\hat{\mathbf{K}}_{n_0} + \delta^2\mathbf{I}_{n_0})^{-1}\hat{Y}_{n_0}\\
&= \hat{Y}_{n_0}^\top(\hat{\mathbf{K}}_{n_0} + \delta^2\mathbf{I}_{n_0})^{-1}\hat{Y}_{n_0} - \hat{Y}_{n_0}^\top(\hat{\mathbf{K}}_{n_0} + \delta^2\mathbf{I}_{n_0})^{-2}\hat{Y}_{n_0}\\
&\leq \hat{Y}_{n_0}^\top(\hat{\mathbf{K}}_{n_0} + \delta^2\mathbf{I}_{n_0})^{-1}\hat{Y}_{n_0}\\
&\leq \frac{\|\hat{Y}_{n_0}\|^2_{l^2}}{\delta^2}.
\end{aligned}$$

The second line follows from the reproducing property. We thus can see that for each fixed $\hat{X}_{n_0}$ and $\hat{Y}_{n_0}$, g (and consequently f) belong to the RKHS of κ.

The values of $\max_{x\in\mathcal{X}} g(x)$ and $\min_{x\in\mathcal{X}} g(x)$ are numerically approximated by sampling $10,000$ points on the hypersphere, and choosing the maximum and minimum over the sample.

We then generate the training datasets $\mathcal{D}_n$ of the sizes $n = 2^i$, with $i = 1, 2, \dots, 13$, by sampling n points $X_n = [x_i]_{i=1}^n$ on the hypersphere, uniformly at random. The values $Y_n = [y_i]_{i=1}^n$ are generated according to f. We then train the neural network model to obtain $\hat{f}_n(.)$. The error $\max_{x\in\mathcal{X}} |f(x) - \hat{f}_n(x)|$ is then numerically approximated by sampling $10,000$ random points on the hypersphere and choosing the maximum of the sample.

We have considered 9 different cases for the pairs of the kernel and the input domain. In particular, the experiments are run for each $\kappa_{\mathrm{NT},s}$, $s = 1, 2, 3$ on all $\mathbb{S}^{d-1}$, $d = 2, 3, 4$. In addition, each one of these 9 experiments is repeated 20 times (180 experiments in total).

In Figure 1 we plot $\max_{x\in\mathcal{X}} |f(x) - \hat{f}_n(x)|$ versus n, averaged over 20 repetitions, for each one of the 9 experiments. Note that for $\max_{x\in\mathcal{X}} |f(x) - \hat{f}_n(x)| \sim n^\alpha$, we have $\log(\max_{x\in\mathcal{X}} |f(x) - \hat{f}_n(x)|) = \alpha \log(n) + \text{constant}$. Thus, in our log scale plots, the slope of the line represents the exponent of the error rate. As predicted analytically, we see all the exponents are negative (the error converges to 0). In addition, the absolute value of the exponent is larger, when s is larger or d is smaller. The bars in the plot on the left in Figure 1, show the standard deviation of the exponents.

For training of the model, we have used *neural-tangents* library (Novak et al. 2019) that is based on *JAX* (Bradbury et al. 2018). The library is primarily suitable for $\kappa_{\mathrm{NT},1}(.)$ corresponding to the ReLU activation function. We thus made an amendment by directly feeding the expressions of the RF kernels, κ_s, $s = 2, 3$, to the *stax.Elementwise* layer provided in the library. Below we give these expressions

$$\begin{aligned}
\kappa_2(u) &= \frac{1}{3\pi}\left[3\sin(\theta)\cos(\theta) + (\pi - \theta)(1 + 2\cos^2(\theta))\right],\\
\kappa_3(u) &= \frac{1}{15\pi}\left[15\sin(\theta) - 11\sin^3(\theta) + (\pi - \theta)(9\cos(\theta) + 6\cos^3(\theta))\right],
\end{aligned}$$

where $\theta = \arccos(u)$.

We derived these expressions using Lemma 1 in a recursive way starting from $\kappa_1(.)$. Also, see (Cho and Saul 2009), which provides a similar expression for $\kappa_2(.)$ and a general method to obtain $\kappa_s(.)$ for other values of s. We note that we only need to supply $\kappa_s(.)$ to the *neural-tangents* library. The NT kernel $\kappa_{\mathrm{NT},s}(.)$ will then be automatically created.

Our experiments run on a single GPU machine (*GeForce RTX 2080 Ti*) with 11 GB of VRAM memory. Each one of the 180 experiments described above takes approximately 4 minutes to run.