A Formulas for NTK

We begin by providing the recursive definition of NTK for fully connected (FC) networks with bias initialized at zero. The formulation includes a parameter \( \beta \) that when set to zero the recursive formula coincides with the formula given in \([1]\) for bias-free networks.

The network model. We consider a \( L \)-hidden-layer fully-connected neural network (in total \( L+1 \) layers) with bias. Let \( x \in \mathbb{R}^d \) (and denote \( d_0 = d \)), we assume each layer \( l \in [L] \) of hidden units includes \( d_l \) units. The network model is expressed as

\[
\begin{align*}
g^{(0)}(x) &= x \\
f^{(l)}(x) &= W^{(l)} g^{(l-1)}(x) + \beta b^{(l)} \in \mathbb{R}^{d_l}, \quad l = 1, \ldots, L \\
g^{(l)}(x) &= \sqrt{c_\sigma/d_l} \sigma(f^{(l)}(x)) \in \mathbb{R}^{d_l}, \quad l = 1, \ldots, L \\
f(\theta, x) &= f^{(L+1)}(x) = W^{(L+1)} \cdot g^{(L)}(x) + \beta b^{(L+1)}
\end{align*}
\]

The network parameters \( \theta \) include \( W^{(L+1)}, W^{(L)}, \ldots, W^{(1)} \), where \( W^{(l)} \in \mathbb{R}^{d_l \times d_{l-1}}, b^{(l)} \in \mathbb{R}^{d_l \times 1} \), \( W^{(L+1)} \in \mathbb{R}^{1 \times d_L}, b^{(L+1)} \in \mathbb{R} \), \( \sigma \) is the activation function and \( c_\sigma = 1/\left(\mathbb{E}_{z \sim \mathcal{N}(0,1)}[\sigma(z)^2]\right) \). The network parameters are initialized with \( \mathcal{N}(0, I) \), except for the biases \( \{b^{(1)}, \ldots, b^{(L)}, b^{(L+1)}\} \), which are initialized with zero.

The recursive formula for NTK. The recursive formula in \([9]\) assumes the bias is initialized with a normal distribution. Here we assume the bias is initialized at zero, yielding a slightly different formulation, which can be readily derived from \([9]\)’s formulation.

Given \( x, z \in \mathbb{R}^d \), we denote the NTK for this fully connected network with bias by \( \kappa^{FC_{(L+1)}}(x, z) := \Theta^{(L)}(x, z) \). The kernel \( \Theta^{(L)}(x, z) \) is defined using the following recursive definition. Let \( h \in [L] \) then

\[
\Theta^{(h)}(x, z) = \Theta^{(h-1)}(x, z) \Sigma^{(h)}(x, z) + \Sigma^{(h)}(x, z) + \beta^2,
\]

where

\[
\Sigma^{(0)}(x, z) = x^T z \\
\Theta^{(0)}(x, z) = \Sigma^{(0)}(x, z) + \beta^2.
\]

and we define

$$\Sigma^{(h)}(x, z) = c_\sigma E_{(u,v) \sim \mathcal{N}(0,\Lambda^{(h-1)})} (\sigma(u)\sigma(v))$$

$$\hat{\Sigma}^{(h)}(x, z) = c_\sigma E_{(u,v) \sim \mathcal{N}(0,\Lambda^{(h-1)})} (\hat{\sigma}(u)\hat{\sigma}(v))$$

$$\Lambda^{(h-1)} = \begin{pmatrix} \Sigma^{(h-1)}(x, x) & \Sigma^{(h-1)}(x, z) \\ \Sigma^{(h-1)}(z, x) & \Sigma^{(h-1)}(z, z) \end{pmatrix}.$$ 

Now, let

$$\lambda^{(h-1)}(x, z) = \frac{\Sigma^{(h-1)}(x, z)}{\sqrt{\Sigma^{(h-1)}(x, x)\Sigma^{(h-1)}(z, z)}}. \quad (2)$$

By definition $|\lambda^{(h-1)}| \leq 1$, and for ReLU activation we have $c_\sigma = 2$ and

$$\Sigma^{(h)}(x, z) = c_\sigma \frac{\lambda^{(h-1)}(\pi - \arccos(\lambda^{(h-1)})) + \sqrt{1 - (\lambda^{(h-1)})^2}}{2\pi} \sqrt{\Sigma^{(h-1)}(x, x)\Sigma^{(h-1)}(z, z)}.$$ 

$$\hat{\Sigma}^{(h)}(x, z) = c_\sigma \frac{\pi - \arccos(\lambda^{(h-1)})}{2\pi}. \quad (4)$$

The parameter $\beta$ allows us to consider a fully-connected network either with ($\beta > 0$) or without bias ($\beta = 0$). When $\beta = 0$, the recursive formulation is the same as existing derivations, e.g., [9]. Finally, the normalized NTK of a FC network with $L + 1$ layers, without bias, is given by

$$\frac{1}{L+1} k_{\text{FC}}^{L+1}(x, x_j).$$

**NTK for a two-layer FC network on $\mathbb{S}^{d-1}$.** Using the recursive formulation above, for points on the hypersphere $\mathbb{S}^{d-1}$ NTK for a two-layer FC network with bias initialized at 0, is as follows. Let $u = x^T z$, with $x, z \in \mathbb{S}^{d-1}$. Then,

$$k_{\text{FC}^\beta(2)}(x, z) = \Theta^{(1)}(x, z)$$

$$= \Theta^{(0)}(x, z) \Sigma^{(1)}(x, z) + \Sigma^{(1)}(x, z) + \beta^2$$

$$= (u + \beta^2) \frac{\pi - \arccos(u)}{\pi} + u(\pi - \arccos(u)) + \sqrt{1 - u^2} + \beta^2.$$ 

Rearranging, we get

$$k_{\text{FC}^\beta(2)}(x, z) = k_{\text{FC}^\beta(2)}(u) = \frac{1}{\pi} \left( (2u + \beta^2)(\pi - \arccos(u)) + \sqrt{1 - u^2} \right) + \beta^2. \quad (5)$$

**B NTK on $\mathbb{S}^{d-1}$**

This section provides a characterization of NTK on the hypersphere $\mathbb{S}^{d-1}$ under the uniform measure. The recursive formulas of the kernels are given in Appendix [A].

**Lemma 1.** Let $k_{\text{FC}^\beta(L)}(x, z), x, z \in \mathbb{S}^{d-1}$, denote the NTK kernels for FC networks with $L \geq 2$ layers, possibly with bias initialized with zero. This kernel is zonal, i.e., $k_{\text{FC}^\beta(L)}(x, z) = k_{\text{FC}^\beta(L)}(x^T z)$.

**Proof.** See Appendix [D].

To prove the next theorem, we recall several results on the the arithmetics of RKHS, following [8][15].

**B.1 RKHS for sums and products of kernels.**

Let $k_1, k_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be kernels with RKHS $\mathcal{H}_{k_1}$ and $\mathcal{H}_{k_2}$, respectively. Then,

1. **Aronszajn’s kernel sum theorem.** The RKHS for $k = k_1 + k_2$ is given by $\mathcal{H}_{k_1 + k_2} = \{f_1 + f_2 | f_1 \in \mathcal{H}_{k_1}, f_2 \in \mathcal{H}_{k_2}\}$
2. This yields the kernel sum inclusion. \( \mathcal{H}_{k_1}, \mathcal{H}_{k_2} \subseteq \mathcal{H}_{k_1+k_2} \)
3. Norm addition inequality. \( \|f_1 + f_2\|_{\mathcal{H}_{k_1+k_2}} \leq \|f_1\|_{\mathcal{H}_{k_1}} + \|f_2\|_{\mathcal{H}_{k_2}} \)
4. Norm product inequality. \( \|f_1 \cdot f_2\|_{\mathcal{H}_{k_1+k_2}} \leq \|f_1\|_{\mathcal{H}_{k_1}} \cdot \|f_2\|_{\mathcal{H}_{k_2}} \)
5. Aronszajn’s inclusion theorem. \( \mathcal{H}_{k_1} \subseteq \mathcal{H}_{k_2} \) if and only if \( \exists s > 0, \) such that \( k_1 \ll s^2 k_2, \) where the latter notation means that \( s^2 k_2 - k_1 \) is a positive definite kernel over \( \mathcal{X} \).

B.2 The decay rate of the eigenvalues of NTK

Theorem 1. Let \( x, z \in S^{d-1} \). With bias initialized at zero and \( \beta > 0 \):

1. \( k^{\text{FC}_\beta}(L) \) can be decomposed according to

\[
k_{\text{FC}_\beta}(L)(x, z) = \sum_{k=0}^{\infty} \sum_{j=1}^{N(d,k)} \lambda_k Y_{k,j}(x)Y_{k,j}(z),
\]

with \( \lambda_k > 0 \) for all \( k \geq 0 \) and into \( Y_{k,j} \) are the spherical harmonics of \( S^{d-1} \), and

2. \( \exists k_0 \) and constants \( C_1, C_2, C_3 > 0 \) that depend on the dimension \( d \) such that \( \forall k > k_0 \)

(a) \( C_1 k^{-\beta} \leq \lambda_k \leq C_2 k^{-\beta} \) if \( L = 2 \), and

(b) \( C_3 k^{-\beta} \leq \lambda_k \) if \( L \geq 3 \).

We split the theorem into the next two lemmas. The first lemma handles NTK of two-layer FC networks with bias, and the second lemma handles NTK for deep networks.

Lemma 2. Let \( x, z \in S^{d-1} \) and \( k^{\text{FC}_\beta}(2)(x^T z) \) as defined in (6) with \( \beta > 0 \). Then, \( k^{\text{FC}_\beta}(2) \) decomposes according to (6) where \( \lambda_k > 0 \) for all \( k \geq 0 \) and \( \exists k_0 \) such that \( \forall k \geq k_0 \)

\[
C_1 k^{-d} \leq \lambda_k \leq C_2 k^{-d},
\]

where \( C_1, C_2 > 0 \) are constants that depend on the dimension \( d \).

Proof. To prove the lemma we leverage the results of (6). First, under the assumption of the uniform measure on \( S^{d-1} \), we can apply Mercer decomposition to \( k^{\text{FC}_\beta}(2)(x^T z) \), where the eigenfunctions are the spherical harmonics. This is due to the observation that \( k^{\text{FC}_\beta}(2)(x^T z) \) is positive and zonal in \( S^{d-1} \). It is zonal by Lemma [1] and positive, since \( k^{\text{FC}_\beta}(2) \) can be decomposed as

\[
k^{\text{FC}_\beta}(2)(u) = \frac{1}{\pi} \left( (2u + \beta^2)(\pi - \arccos(u)) + \sqrt{1 - u^2} \right) + \beta^2
\]

\[
= \frac{1}{\pi} \left( 2u(\pi - \arccos(u)) + \sqrt{1 - u^2} \right) + \frac{1}{\pi} \beta^2 \left( \pi - \arccos(u) \right) + \beta^2
\]

\[
:= \kappa(x^T z) + \beta^2 \kappa_0(x^T z) + \beta^2,
\]

where \( \kappa(x^T z) \) is the NTK for a bias-free, two-layer network introduced in (5) and \( \kappa_0(x^T z) \) is known to be the zero-order arc-cosine kernel (6). By kernel arithmetic, this yields another kernel and this means that \( k^{\text{FC}_\beta}(2) \) is a positive kernel.

Furthermore, according to Proposition 5 in (5)

\[
\kappa(x^T z) = \sum_{k=0}^{\infty} \mu_k \sum_{j=1}^{N(d,k)} Y_{k,j}(x)Y_{k,j}(z),
\]

where \( Y_{k,j}, j = 1, \ldots, N(d,k) \) are spherical harmonics of degree \( k \), and the eigenvalues \( \mu_k \) satisfy \( \mu_0, \mu_1 > 0, \mu_k = 0 \) if \( k = 2j + 1 \) with \( j \geq 1 \) and otherwise, \( \mu_k > 0 \) and \( \mu_k \sim C(d)k^{-d} \) as \( k \to \infty \), with \( C(d) \) a constant depending only on \( d \). Next, following Lemma 17 in (5) the eigenvalues of \( \kappa_0(x^T z) \), denoted \( \eta_k \) satisfy \( \eta_0 > 0, \eta_k > 0 \) if \( k = 2j + 1 \), with \( j \geq 1 \) and behave asymptotically as \( C_0(d)k^{-d} \). Consequently, \( k^{\text{FC}_\beta}(2) = \kappa + \beta^2 \kappa_0 + \beta^2 \), and since both \( \kappa \) and \( \kappa_0 \) have the spherical
harmonics as their eigenfunctions, their eigenvalues are given by \( \lambda_k = \mu_k + \beta^2 \eta_k > 0 \) for \( k > 0 \) and \( \lambda_0 = \mu_0 + \beta^2 \eta_0 + \beta^2 > 0 \), and asymptotically \( \lambda_k \sim \mathcal{C}(d) k^{-d} \), where \( \mathcal{C}(d) = C(d) + \beta^2 C_0(d) \).

To conclude, this implies that \( \exists k_0, C_1(d) > 0 \) and \( C_2(d) > 0 \), such that for all \( k \geq k_0 \) it holds that
\[
C_1 k^{-d} \leq \lambda_k \leq C_2 k^{-d}
\]
and also, unless \( \beta = 0 \), for all \( k \geq 0 \)
\[
\lambda_k > 0.
\]

Next, we prove the second part of Theorem 1, which relates to deep FC networks with bias, \( \mathcal{K}^{FC_\beta(L)} \), i.e. we prove the following lemma.

**Lemma 3.** Let \( x, z \in \mathbb{S}^{d-1} \) and \( \mathcal{K}^{FC_\beta(L)}(xz) \) as defined in Appendix A. Then

1. \( \mathcal{K}^{FC_\beta(L)} \) decomposes according to (6) with \( \lambda_k > 0 \) for all \( k \geq 0 \)
2. \( \exists k_0 \) such that \( \forall k > k_0 \) it holds that \( C_3 k^{-d} \leq \lambda_k \) in which \( C_3 > 0 \) depends on the dimension \( d \)
3. \( \mathcal{H}^{FC_\beta(L-1)} \subseteq \mathcal{H}^{FC_\beta(L)} \)

**Proof.** Following Lemma 1, it holds that \( \mathcal{K}^{FC_\beta(L)} \) is zonal, and therefore can be decomposed according to (6). In order to prove the lemma we look at the recursive formulation of the NTK kernel, i.e.,
\[
\mathcal{K}^{FC_\beta(L+1)} = \mathcal{K}^{FC_\beta(L)} \Sigma^{(l)} + \Sigma^{(l)} + \beta^2.
\]

Now, following Lemma 17 in [5] all of the eigenvalues of \( \Sigma^{(l)} \) are positive, including \( \lambda_0 > 0 \). This implies that the constant function \( g(x) \equiv 1 \in \mathcal{H}^{\Sigma^{(l)}} \).

Now, we use the norm multiplicity inequality in Sec. B.1 and show that \( \mathcal{H}^{\mathcal{K}^{FC_\beta(L)}} \subseteq \mathcal{H}^{\mathcal{K}^{FC_\beta(L)} \Sigma^{(l)}} \). Let \( f \in \mathcal{H}^{\mathcal{K}^{FC_\beta(L)}} \), i.e., \( \|f\|_{\mathcal{H}^{\mathcal{K}^{FC_\beta(L)}}} < \infty \). We showed that \( 1 \in \mathcal{H}^{\Sigma^{(l)}} \). Therefore, \( \|f \cdot 1\|_{\mathcal{H}^{\mathcal{K}^{FC_\beta(L)} \Sigma^{(l)}}} \leq \|f\|_{\mathcal{H}^{\mathcal{K}^{FC_\beta(L)} \Sigma^{(l)}}} \|1\|_{\mathcal{H}^{\Sigma^{(l)}}} < \infty \), implying that \( f \in \mathcal{H}^{\mathcal{K}^{FC_\beta(L)} \Sigma^{(l)}} \).

Finally, according to the kernel sum inclusion in Sec. B.1 relying on the recursive formulation (7) we have \( \mathcal{H}^{\mathcal{K}^{FC_\beta(L)}} \subseteq \mathcal{H}^{\mathcal{K}^{FC_\beta(L)} \Sigma^{(l)}} \subseteq \mathcal{H}^{\mathcal{K}^{FC_\beta(L+1)}} \). Therefore,
\[
\mathcal{H}^{\mathcal{K}^{FC_\beta(2)}} \subseteq \ldots \subseteq \mathcal{H}^{\mathcal{K}^{FC_\beta(L-1)}} \subseteq \mathcal{H}^{\mathcal{K}^{FC_\beta(L)}}.
\]

This completes the proof, by using Aronszan’s inclusion theorem as follows. Since \( \mathcal{H}^{\mathcal{K}^{FC_\beta(2)}} \subseteq \mathcal{H}^{\mathcal{K}^{FC_\beta(L)}} \), then by Aronszain’s inclusion theorem \( \exists s > 0 \) such that \( \mathcal{K}^{FC_\beta(2)} \ll s^2 \mathcal{K}^{FC_\beta(L)} \). Since the kernels are zonal on the sphere (with uniform distribution of the data) their corresponding RKHS share the same eigenfunctions, namely the spherical harmonics.

Therefore, for all \( k \geq 0 \) it holds
\[
s^2 \lambda_k^{FC_\beta(L)} \geq \lambda_k^{FC_\beta(2)} > 0
\]
and for \( k \to \infty \) it holds that
\[
s^2 \lambda_k^{FC_\beta(L)} \geq \lambda_k^{FC_\beta(2)} \geq \frac{C_1}{k^d}
\]
completing the proof. \( \square \)
C Laplace Kernel in $\mathbb{S}^{d-1}$

The Laplace kernel $k(x, y) = e^{-c\|x-y\|}$ restricted to the sphere $\mathbb{S}^{d-1}$ is defined as

$$K(x, y) = k(x^T y) = e^{-c\sqrt{1-x^T y}}$$

(9)

where $c > 0$ is a tuning parameter. We next prove an asymptotic bound on its eigenvalues.

**Theorem 2.** Let $x, y \in \mathbb{S}^{d-1}$ and $k(x^T y) = e^{-c\sqrt{1-x^T y}}$ be the Laplace kernel, restricted to $\mathbb{S}^{d-1}$. Then $k$ can be decomposed as in (6) with the eigenvalues $\lambda_k$ satisfying $\lambda_k > 0$ for all $k \geq 0$ and $\exists k_0$ such that $\forall k > k_0$ it holds that:

$$B_1 k^{-d} \leq \lambda_k \leq B_2 k^{-d}$$

where $B_1, B_2 > 0$ are constants that depend on the dimension $d$ and the parameter $c$.

Our proof relies on several supporting lemmas.

**Lemma 4.** ([17] Thm. 1.14 page 6) For all $\alpha > 0$ it holds that

$$\int_{\mathbb{R}^d} e^{-\alpha \|x\|} e^{-2\pi i x \cdot t} dx = c_d \frac{\alpha}{(\alpha^2 + \|t\|^2)^{(d+1)/2}},$$

(10)

where $c_d = \Gamma\left(\frac{d+1}{2}\right)/(\pi^{(d+1)/2})$.

**Lemma 5.** Let $f(x) = e^{-c\|x\|}$ with $x \in \mathbb{R}^d$. Then, its Fourier transform $\Phi(w)$ with $w \in \mathbb{R}^d$ is

$$\Phi(w) = \Phi(||w||) = C(1 + \|w\|^2/c^2)^{-(d+1)/2}$$

for some constant $C > 0$.

**Proof.** To calculate the Fourier transform we need to calculate the following integral

$$\Phi(w) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-c\|x\|} e^{-i x \cdot w} dx.$$

According to the Lemma 4, plugging $\alpha = \frac{c}{2}$ and $t = \frac{w}{2c}$ into (10) yields

$$\Phi(w) = c_d \frac{c}{(c^2 + \|w\|^2)^{(d+1)/2}} = \frac{c_d}{c^{(d+1)}} \frac{1}{\left(1 + \|w\|^2/c^2\right)^{(d+1)/2}} = C \left(1 + \frac{\|w\|^2}{c^2}\right)^{-(d+1)/2}$$

with $C = \frac{c_d}{c^{(d+1)}} > 0$. 

**Lemma 6.** ([17] Thm. 4.1) Let $f(x)$ be defined as $f(||x||)$ for all $x \in \mathbb{R}^d$, and let $\Phi(w) = \Phi(||w||)$ denote its Fourier Transform in $\mathbb{R}^d$. Then, its corresponding kernel on $\mathbb{S}^{d-1}$ is defined as the restriction $k(x^T y) = f(||x-y||)$ with $x, y \in \mathbb{S}^{d-1}$. By Mercer’s Theorem the spherical harmonic expansion of $k(x^T y)$ is of the form

$$k(x^T y) = \sum_{k=0}^{\infty} \lambda_k \sum_{j=1}^{N(d, k)} Y_{k, j}(x)Y_{k, j}(y).$$

Then, the eigenvalues in the spherical harmonic expansion $\lambda_k$ are related to the Fourier coefficients of $f$, $\Phi(t)$, as follows

$$\lambda_k = \int_0^\infty t \Phi(t) J_{k + \frac{d-1}{2}}(t) dt,$$

(11)

where $J_v(t)$ is the usual Bessel function of the first kind of order $v$.

Having, these supporting Lemmas, we can now prove **Theorem 2**

**Proof.** First, $k(\cdot, \cdot)$ is a positive zonal kernel and hence can be written as

$$k(x^T y) = \sum_{k=0}^{\infty} \lambda_k \sum_{j=1}^{N(d, k)} Y_{k, j}(x)Y_{k, j}(y).$$
Next, to derive the bounds we plug the Fourier coefficients, \( \Phi(\omega) \), computed in Lemma 5 into the expression for the harmonic coefficients, \( \lambda_k \), obtaining

\[
\lambda_k = C \int_0^\infty \frac{t}{(1 + \frac{t^2}{2})^{d/2}} J_{k+d/2}^2(t) dt.
\]

Applying a change of variables \( t = cx \) we get

\[
\lambda_k = c^2 C \int_0^\infty \frac{x}{(1 + x^2)^{d/2}} J_{k+d/2}^2(cx) dx.
\]

We next bound this integral from both above and below. To get an upper bound we observe that for \( x \in [0, \infty) \) \( x^2 < 1 + x^2 \), implying that \( x(1 + x^2)^{-(d+1)/2} \leq c \), and consequently

\[
\lambda_k < c^2 C \int_0^\infty x^{-d} J_{k+d/2}^2(cx) dx := c^2 CA(k, d, c).
\]

The above integral \( A(k, d, c) \) was computed in [13] (Sec. 13.41 page 402 with \( a := c, \lambda := d \), and \( \mu = \nu := k + (d-2)/2 \)) which gives

\[
A(k, d, c) = \int_0^\infty x^{-d} J_{k+d/2}^2(cx) dx = \frac{\left(\frac{c}{2}\right)^{d-1} \Gamma(d(1 - \frac{1}{2}))}{2\Gamma^2\left(\frac{d+1}{2}\right) \Gamma(k + d - \frac{1}{2})}.
\]

Using Stirling’s formula \( \Gamma(x) = \sqrt{2\pi x^{x-1/2}} e^{-x} (1 + O(x^{-1})) \) as \( x \to \infty \). Consequently, for sufficiently large \( k \gg d \)

\[
\lambda_k < c^2 C A(k, d, c) = c^2 C \frac{\left(\frac{c}{2}\right)^{d-1} \Gamma(d(1 - \frac{1}{2}))}{2\Gamma^2\left(\frac{d+1}{2}\right) \Gamma(k + d - \frac{1}{2})} \\
\sim c^2 C \frac{\left(\frac{c}{2}\right)^{d-1} \Gamma(d)}{2\Gamma^2\left(\frac{d+1}{2}\right)} \cdot \frac{(k - \frac{1}{2})^{k-1} e^{-k + \frac{1}{2}}}{(k + d - \frac{1}{2})^{k-1} e^{-k + \frac{1}{2} + \frac{1}{2}} (1 + O(k^{-1}))} = B_2 k^{-d},
\]

where \( B_2 \) depends on \( c, C \) and the dimension \( d \).

We use again the relation (12) to derive a lower bound for \( \lambda_k \). First, note that since \( t, 1 + t^2, J_{k}^2(t) \) are all non-negative for \( t \in [0, \infty) \) and therefore

\[
\lambda_k \geq c^2 C \int_1^\infty \frac{x}{(1 + x^2)^{d/2}} J_{k+d/2}^2(cx) dx \geq c^2 C \int_1^\infty \frac{1}{2^{d/2}} x^d J_{k+d/2}^2(cx) dx
\]

\[
= \frac{C^2}{2^{d/2}} \left( \int_0^\infty x^{-d} J_{k+d/2}^2(cx) dx - \int_0^1 x^{-d} J_{k+d/2}^2(cx) dx \right)
\]

\[
= \frac{C^2}{2^{d/2}} \int_0^\infty x^{-d} J_{k+d/2}^2(cx) dx \left( 1 - \int_0^1 x^{-d} J_{k+d/2}^2(cx) dx \right)
\]

\[
= \frac{C^2}{2^{d/2}} A(k, d, c) \left( \frac{1}{A(k, d, c)} - \frac{B(k, d, c)}{A(k, d, c)} \right),
\]

where \( B(k, d, c) := \int_0^1 x^{-d} J_{k+d/2}^2(cx) dx \). The first integral, \( A(k, d, c) \), was shown in (14) to converge asymptotically to \( B_2 k^{-d} \). To bound the second integral, \( B(k, d, c) \), we use an inequality from [13] (Section 3.31, page 49), which states that for \( v, t \in \mathbb{R}, v > -\frac{1}{2} \),

\[
|J_v(t)| \leq \frac{2^{-v} t^v}{\Gamma(v+1)}.
\]

This gives an upper bound for \( B(k, d, c) \)

\[
B(k, d, c) = \int_0^1 x^{-d} J_{k+d/2}^2(cx) dx \leq \int_0^1 x^{-d} \frac{2^{-2(k+d/2)} (cx)^{2(k+d/2)}}{\Gamma^2(k + \frac{d}{2})} dx \leq \frac{\left(\frac{c}{2}\right)^{2(k+d/2)} \Gamma^2(k + \frac{d}{2})}{\Gamma\left(\frac{d}{2}\right)}.
\]
Applying Stirling’s formula we obtain \( B(k, d, c) \leq O \left( \frac{c^2}{2^{k+d}} \right) \), which implies that as \( k \) grows, \( \frac{B(k, d, c)}{A(k, d, c)} \to 0 \). Therefore, asymptotically for large \( k \)

\[
\lambda_k \geq \frac{C_k^2}{2^{k+d}} A(k, d, c) \left( 1 - \frac{B(k, d, c)}{A(k, d, c)} \right) \geq \frac{C_k^2}{2^{k+d}} A(k, d, c),
\]

from which we conclude that \( \lambda_k > B_1 k^{-d} \), where the constant \( B_1 \) depends on \( c, C, \) and \( d \). We have therefore shown that there exists \( k_0 \) such that \( \forall k > k_0 \)

\[
B_1 k^{-d} \leq \lambda_k \leq B_2 k^{-d}.
\]

Finally, to show that \( \lambda_k > 0 \) for all \( k \geq 0 \) we use again (11) in Lemma 6 which states that

\[
\lambda_k = \int_0^\infty t \Phi(t) J_k^2 \left( \frac{x}{t} \right) dt.
\]

Note that in the interval \((0, \infty)\) it holds that \( t > 0 \) and \( \Phi(t) > 0 \) due to Lemma 3. Therefore \( \lambda_k = 0 \) implies that \( J_k^2 \left( \frac{x}{t} \right) \) is identically 0 on \((0, \infty)\), contradicting the properties of the Bessel function of the first kind. Hence, \( \lambda_k > 0 \) for all \( k \).

\[\square\]

### C.1 Proof of main theorem

**Theorem 3.** Let \( H_{\text{Lap}} \) denote the RKHS for the Laplace kernel restricted to \( S^{d-1} \), and let \( H_{\text{FC}(2)} \) denote the NTK corresponding to a FC network with \( L \) layers with bias, restricted to \( S^{d-1} \), then \( H_{\text{Lap}} \subseteq H_{\text{FC}(2)} \).

**Proof.** Let \( \lambda_k^{\text{Lap}}, \) \( \lambda_k^{\text{FC}(2)} \), and \( \lambda_k^{\text{FC}(1)} \) denote the eigenvalues of the three kernel, \( k_{\text{Lap}}, k_{\text{FC}(2)}, \) and \( k_{\text{FC}(1)} \) in their Mercer’s decomposition, i.e.,

\[
k(x^T z) = \sum_{k=0}^{\infty} \lambda_k \sum_{j=1}^{N(d, k)} Y_{k,j}(x) Y_{k,j}(z).
\]

Denote by \( k_0 \) the smallest \( k \) for which Theorems 1 and 2 hold simultaneously. We first show that \( H_{\text{Lap}} \subseteq H_{\text{FC}(2)} \). Let \( f(x) \in H_{\text{Lap}} \), and let \( f(x) = \sum_{k=0}^{\infty} \sum_{j=0}^{N(d, k)} \alpha_{k,j} Y_{k,j}(x) \) denote its spherical harmonic decomposition. Then \( \|f\|_{H_{\text{Lap}}} < \infty \) implies, due to Theorem 2, that

\[
\sum_{k=0}^{\infty} \sum_{j=0}^{N(d, k)} \frac{1}{B_2} k^d \alpha_{k,j}^2 \leq \sum_{k=0}^{\infty} \sum_{j=0}^{N(d, k)} \frac{\alpha_{k,j}^2}{\lambda_k^{\text{Lap}}} < \infty.
\]

Combining this with Theorem 1 and recalling that \( \lambda_k^{\text{FC}(2)} > 0 \) for all \( k \geq 0 \), we have

\[
\sum_{k=k_0}^{\infty} \sum_{j=0}^{N(d, k)} \frac{\alpha_{k,j}^2}{\lambda_k^{\text{FC}(2)}} \leq \sum_{k=k_0}^{\infty} \sum_{j=0}^{N(d, k)} \frac{k^d \alpha_{k,j}^2}{C_1} = \frac{B_2}{C_1} \sum_{k=k_0}^{\infty} \sum_{j=0}^{N(d, k)} \frac{1}{k^d \alpha_{k,j}^2} < \infty,
\]

implying that \( \|f\|_{H_{\text{FC}(2)}}^2 < \infty \), and so \( H_{\text{Lap}} \subseteq H_{\text{FC}(2)} \). Similar arguments can be used to show that \( H_{\text{FC}(1)} \subseteq H_{\text{Lap}} \), proving that \( H_{\text{FC}(1)} = H_{\text{Lap}} \). Finally, following the inclusion relation, the theorem is proved. \[\square\]

### D NTK in \( \mathbb{R}^d \)

In this section we denote \( r_x = \|x\|, r_z = \|z\| \) and by \( \tilde{x} = x/r_x, \tilde{z} = z/r_z \). We first prove Theorem 4 and as a consequence Lemma 7 is proved.

**Theorem 4.** Let \( k_{\text{FC}(L)}(x, 0), k_{\text{FC}(L)}(x, z) \), \( x, z \in \mathbb{R}^d \), denote the NTK kernel with \( L \) layers without bias and with bias initialized to zero, respectively. It holds that (1) Bias-free \( k_{\text{FC}(L)} \) is homogeneous of order 1. (2) Let \( k_{\text{Bias}(L)} = k_{\text{FC}(L)} - k_{\text{FC}(L)} \). Then, \( k_{\text{Bias}(L)} \) is homogeneous of order 0.
Lemma 7. Let \( k_{\text{FC}\beta(L)}(\mathbf{x}, \mathbf{z}) \), \( \mathbf{x}, \mathbf{z} \in \mathbb{S}^{d-1} \), denote the NTK kernels for FC networks with \( L \geq 2 \) layers, possibly with bias initialized with zero. This kernel is zonal, i.e., \( k_{\text{FC}\beta(L)}(\mathbf{x}, \mathbf{z}) = k_{\text{FC}\beta(L)}(\mathbf{x}^T \mathbf{z}) \).

To that end, we first prove the following supporting Lemma.

Lemma 8. For \( \mathbf{x}, \mathbf{z} \in \mathbb{R}^d \) it holds that
\[
\Theta^{(L)}(\mathbf{x}, \mathbf{z}) = r_x r_z \Theta^{(0)}(\hat{\mathbf{x}}, \hat{\mathbf{z}}) = r_x r_z \Theta^{(L)}(\hat{\mathbf{x}}^T \hat{\mathbf{z}}),
\]
where \( \Theta^{(L)} = k_{\text{FC}_0(L+1)} \), as defined in Appendix A.

Proof. We prove this by induction over the recursive definition of \( k_{\text{FC}_0(L+1)} = \Theta^{(L)}(\mathbf{x}, \mathbf{z}) \). Let \( \mathbf{x}, \mathbf{z} \in \mathbb{R}^d \), then by definition
\[
\Theta^{(0)}(\mathbf{x}, \mathbf{z}) = \mathbf{x}^T \mathbf{z} = r_x r_z \Theta^{(0)}(\hat{\mathbf{x}}, \hat{\mathbf{z}}) = r_x r_z \Theta^{(0)}(\hat{\mathbf{x}}^T \hat{\mathbf{z}})
\]
and
\[
\Sigma^{(0)}(\mathbf{x}, \mathbf{z}) = \mathbf{x}^T \mathbf{z} = r_x r_z \Sigma^{(0)}(\hat{\mathbf{x}}, \hat{\mathbf{z}}) = r_x r_z \Sigma^{(0)}(\hat{\mathbf{x}}^T \hat{\mathbf{z}})
\]
Assuming the induction hypothesis holds for \( l \), i.e.,
\[
\Theta^{(l)}(\mathbf{x}, \mathbf{z}) = r_x r_z \Theta^{(l)}(\hat{\mathbf{x}}, \hat{\mathbf{z}}) = r_x r_z \Theta^{(l)}(\hat{\mathbf{x}}^T \hat{\mathbf{z}})
\]
and
\[
\Sigma^{(l)}(\mathbf{x}, \mathbf{z}) = r_x r_z \Sigma^{(l)}(\hat{\mathbf{x}}, \hat{\mathbf{z}}) = r_x r_z \Sigma^{(l)}(\hat{\mathbf{x}}^T \hat{\mathbf{z}})
\]
we prove that those equalities are also true for \( l + 1 \).

By the definition of \( \lambda^{(l)} \) and the induction hypothesis for \( \Sigma^{(l)} \) we have that
\[
\lambda^{(l)}(\mathbf{x}, \mathbf{z}) = \frac{\Sigma^{(l)}(\mathbf{x}, \mathbf{z})}{\sqrt{\Sigma^{(l)}(\mathbf{x}, \mathbf{x}) \Sigma^{(l)}(\mathbf{z}, \mathbf{z})}} = \frac{\Sigma^{(l)}(\hat{\mathbf{x}}, \hat{\mathbf{z}})}{\sqrt{\Sigma^{(l)}(\hat{\mathbf{x}}, \hat{\mathbf{x}}) \Sigma^{(l)}(\hat{\mathbf{z}}, \hat{\mathbf{z}})}} = \lambda^{(l)}(\hat{\mathbf{x}}, \hat{\mathbf{z}}) = \lambda^{(l)}(\hat{\mathbf{x}}^T \hat{\mathbf{z}})
\]
Plugging this result in the definitions of \( \Sigma \) and \( \hat{\Sigma} \), using the induction hypothesis we obtain
\[
\Sigma^{(l+1)}(\mathbf{x}, \mathbf{z}) = r_x r_z \Sigma^{(l+1)}(\hat{\mathbf{x}}, \hat{\mathbf{z}}) = r_x r_z \Sigma^{(l+1)}(\hat{\mathbf{x}}^T \hat{\mathbf{z}})
\]
Finally, using the recursion formula \( r = 0 \) and the induction hypothesis for \( \Theta^{(l)} \), we obtain
\[
\Theta^{(l+1)}(\mathbf{x}, \mathbf{z}) = r_x r_z \Theta^{(l+1)}(\hat{\mathbf{x}}, \hat{\mathbf{z}}) = r_x r_z \Theta^{(l+1)}(\hat{\mathbf{x}}^T \hat{\mathbf{z}})
\]
\boxdot

A corollary of this Lemma is that \( k_{\text{FC}_0(L)} \) is homogeneous of order 1 in \( \mathbb{R}^d \), proving the first part of Theorem 4. Also, it is homogeneous of order 0 in \( \mathbb{S}^{d-1} \), proving Lemma 7 for \( \beta = 0 \).

We next turn to proving the second part of Theorem 4, i.e., that \( k_{\text{Bias}(L)} = k_{\text{FC}\beta(L)} - k_{\text{FC}_0(L)} \) is homogeneous of order 0 in \( \mathbb{R}^d \). By rewriting the recursive definition of \( k_{\text{FC}\beta(L)} \), shown in Appendix A, we can express \( k_{\text{Bias}(1)} \) in the following recursive manner \( k_{\text{Bias}(1)} = \beta^2 \), and \( k_{\text{Bias}(L+1)} = k_{\text{Bias}(L)} + \beta^2 \). Therefore, \( k_{\text{Bias}(L)} \) is homogeneous of order zero, since it depends only on \( \hat{\Sigma} \), which is by itself homogeneous of order zero \( \langle 15 \rangle \). This concludes Theorem 4.

Finally, Lemma 7 is proved, since \( k_{\text{FC}\beta(L)} = k_{\text{FC}_0(L)} + k_{\text{Bias}(L)} \), and when restricted to \( \mathbb{S}^{d-1} \) both components are homogeneous of order 0.

Theorem 5. Let \( p(r) \) be a decaying density on \([0, \infty)\) such that \( 0 < \int_0^\infty p(r)r^2dr < \infty \) and \( \mathbf{x}, \mathbf{z} \in \mathbb{R}^d \).

1. Let \( k_0(\mathbf{x}, \mathbf{z}) \) be homogeneous of order 1 such that \( k_0(\mathbf{x}, \mathbf{z}) = r_x r_z \hat{k}_0(\hat{\mathbf{x}}^T \hat{\mathbf{z}}) \). Then its eigenfunctions with respect to \( p(r_x) \) are given by \( \hat{\Psi}_{k,j} = ar_x \hat{Y}_{k,j}(\hat{\mathbf{x}}) \), where \( \hat{Y}_{k,j} \) are the spherical harmonics in \( \mathbb{S}^{d-1} \) and \( a \in \mathbb{R} \).
2. Let $k(x, z) = k_0(x, z) + k_1(x, z)$ so that $k_0$ as in 1 and $k_1$ is homogeneous of order 0. Then the eigenfunctions of $k$ are of the form $\Psi_{k,j} = (ar_x + b)Y_{k,j}(\hat{x})$.

Proof. 1. Since $\hat{k}_0$ is zonal, its Mercer’s representation reads

$$
\hat{k}_0(\hat{x}, \hat{z}) = \sum_{k=0}^{\infty} \lambda_k \sum_{j=1}^{N(d,k)} Y_{k,j}(\hat{x})Y_{k,j}(\hat{z}),
$$

where the spherical harmonics $Y_{k,j}$ are the eigenfunctions of $\hat{k}_0$. Consequently, as noted also in [5],

$$
k_0(x, z) = a^2 \sum_{k=0}^{\infty} \lambda_k \sum_{j=1}^{N(d,k)} r_x Y_{k,j}(\hat{x})r_z Y_{k,j}(\hat{z}).
$$

The orthogonality of the eigenfunctions $\Psi_{k,j}(x) = ar_x Y_{k,j}(\hat{x})$ is verified as follows. Let $\bar{p}(x)$ denote a probability density on $\mathbb{R}^d$ such that $\bar{p}(x) = p(r_x)/A(r_x)$, where $A(r_x)$ denotes the surface area of a sphere of radius $r_x$ in $\mathbb{R}^d$. Then,

$$
\int_{\mathbb{R}^d} \Psi_{k,j}(x)\Psi_{k',j'}(x)\bar{p}(x)dx = a^2 \int_0^{\infty} r_x^{d+1} p(r_x)/A(r_x) dr_x \int_{S^{d-1}} Y_{k,j}(\hat{x})Y_{k',j'}(\hat{x})d\hat{x} = \delta_{k,k'}\delta_{j,j'},
$$

where the rightmost equality is due to the orthogonality of the spherical harmonics and by setting

$$
a^2 = \left( \int_0^{\infty} r_x^{d+1} p(r_x)/A(r_x) dr_x \right)^{-1}.
$$

Clearly this integral is positive, and the conditions of the theorem guarantee that it is finite.

2. By the conditions of the theorem we can write

$$
k(x, z) = r_x r_z \hat{k}_0(\hat{x}^T \hat{z}) + \hat{k}_1(\hat{x}^T \hat{z}),
$$

where $\hat{x}, \hat{z} \in S^{d-1}$. On the hypersphere the spherical harmonics are the eigenfunctions of $k_0$ and $k_1$. Denote their eigenvalues respectively by $\lambda_k$ and $\mu_k$, so that

$$
\int_{S^{d-1}} k_0(\hat{x}^T \hat{z})\bar{Y}_k(\hat{z})d\hat{z} = \lambda_k \bar{Y}_k(\hat{x}) \quad (16)
$$

$$
\int_{S^{d-1}} k_1(\hat{x}^T \hat{z})\bar{Y}_k(\hat{z})d\hat{z} = \mu_k \bar{Y}_k(\hat{x}) \quad (17)
$$

where $\bar{Y}_k(\hat{x})$ denote the zonal spherical harmonics. We next show that the space spanned by the functions $r_x \bar{Y}_k(x)$ and $\bar{Y}_k(x)$ is fixed under the following integral transform

$$
\int_{\mathbb{R}^d} k(x, z)(\alpha r_z + \beta)\bar{Y}_k(\hat{z})\bar{p}(\hat{z})d\hat{z} = (\alpha r_x + b)\bar{Y}_k(\hat{x}), \quad (18)
$$

$\alpha, \beta, a, b \in \mathbb{R}$ are constants. The left hand side can be written as the application of an integral operator $T(x, z)$ to a function $\Phi^{k}_{\alpha, \beta}(z) = (\alpha r_z + \beta)\bar{Y}_k(\hat{z})$. Expressing this operator application in spherical coordinates yields

$$
T(x, z)\Phi^{k}_{\alpha, \beta}(z) = \int_0^{\infty} \frac{p(r_z)r_z^{d-1}}{A(r_z)} dr_z \int_{\hat{z} \in S^{d-1}} (r_x r_z k_0(\hat{x}^T \hat{z}) + k_1(\hat{x}^T \hat{z})) (\alpha r_z + \beta)\bar{Y}_k(\hat{z})d\hat{z}.
$$

We use (16) and (17) to substitute for the inner integral, obtaining

$$
T(x, z)\Phi^{k}_{\alpha, \beta}(z) = \int_0^{\infty} \frac{p(r_z)r_z^{d-1}}{A(r_z)} (\lambda_k r_x + \mu_k) (\alpha r_z + \beta)\bar{Y}_k(\hat{x}) dr_z.
$$

Together with (18), this can be written as

$$
T(x, z)\Phi^{k}_{\alpha, \beta}(z) = \Phi^k_{\alpha, \beta}(x),
$$

where

$$
\Phi^k_{\alpha, \beta}(x) = (\alpha r_x + b)\bar{Y}_k(\hat{x}).
$$
where
\[
\begin{pmatrix}
a \\
b
\end{pmatrix} =
\begin{pmatrix}
\lambda_k & 0 \\
0 & \mu_k
\end{pmatrix}
\begin{pmatrix}
M_2 & M_1 \\
M_1 & M_0
\end{pmatrix}
\begin{pmatrix}
\alpha \\
\beta
\end{pmatrix}
\]
where \( M_q = \int_0^\infty \frac{r^{q-d+\frac{1}{2}}}{A(r)} dr \), \( 0 \leq q \leq 2 \). By the conditions of the theorem these moments are finite. This proves that the space spanned by \( \{ r_x \hat{Y}(\hat{x}), \bar{Y}(\hat{x}) \} \) is fixed under \( T(x, z) \), and therefore the eigenfunctions of \( k^{FC_\beta}(L) \) take the form \( (\bar{a}r_x + \bar{b}) \bar{Y}(\hat{x}) \) for some constants \( \bar{a}, \bar{b} \).

The implication of Theorem 5 is that the eigenvectors of \( k^{FC_0}(L) \) are the spherical harmonic functions, scaled by the norm of their arguments. With bias, \( k^{FC_\beta}(L) \) has up to \( 2N(d, k) \) eigenfunctions for every frequency \( k \), of the general form \( (ar_x + b)Y_{k,j}(\hat{x}) \) where \( a, b \) are constants that differ from one eigenfunction to the next.

E Experimental Details

E.1 The UCI dataSet

In this section, we provide experimental details for the UCI dataset. We use precisely the same pre-processed datasets, and follow the same performance comparison protocol as in [2].

NTK Specifications We reproduced the results of [2] using the publicly available code\footnote{https://github.com/LeoYu/neural-tangent-kernel-UCI} and followed the same protocol as in [2]. The total number of kernels evaluated in [2] are 15 and the SVM cost value parameter \( C \) is tuned from \( 10^{-2} \) to \( 10^4 \) by powers of 10. Hence, the total number of hyper-parameter combinations searched using cross-validation is \( 10^5 \) \((15 \times 7)\).

Exponential Kernels Specifications For the Laplace and Gaussian kernels, we searched for 10 kernel width values \((1/c)\) from \( 2^{-2} \times \nu \) to \( \nu \) in the log space with base 2, where \( \nu \) is chosen heuristically as the median of pairwise \( l_2 \) distances between data points (known as the median trick [7]). So, the total number of kernel evaluations is 10. For \( \gamma \)-exponential, we searched through 5 equally spaced values of \( \gamma \) from 0.5 to 2. Since we wanted to keep the number of the kernel evaluations the same as for NTK in [2], we searched through only three kernel bandwidth values \((1/c)\) which are \( 1, \nu \) and \#features (default value in the sklearn package\footnote{https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.rbf_kernel.html}). So, the total number of kernel evaluations is 15 \((5 \times 3)\).

For a fair comparison with [2], we swept the same range of SVM cost value parameter \( C \) as in [2], i.e., from \( 10^{-2} \) to \( 10^4 \) by powers of 10. Hence, the total number of hyper-parameter search using cross-validation is \( 70 \) \((10 \times 7)\) for Laplace and \( 105 \) \((15 \times 7)\) for \( \gamma \)-exponential which is the same as for NTK in [2].

E.2 Large scale datasets

We used the experimental setup mentioned in [14] and the publicly available code\footnote{https://github.com/LCSL/FALKON_paper} solves kernel ridge regression (KRR [16]) using the FALKON algorithm, which solves the following linear system
\[
(K_{nn} + \lambda n I) \alpha = \hat{y},
\]
where \( K \) is an \( n \times n \) kernel matrix defined by \((K)_{ij} = K(x_i, x_j), \hat{y} = (y_1, \ldots, y_n)^T \), and \( \lambda \) is the regularization parameter. Refer to [14] for more details.

In Table\footnote{https://github.com/LeoYu/neural-tangent-kernel-UCI}, we provide the hyper parameters chosen with cross validation.
We set $\beta$ where $C$-Exp.

$C$-Exp Gaussian

$C$-Exp Laplace

$C$-Exp Laplace

$C$-Exp Gaussian

E.3 C-Exp: Convolutional Exponential Kernels

Let $x = (x_1, ..., x_d)^T$ and $z = (z_1, ..., z_d)^T$ denote two vectorized images. Let $P$ denote a window function (we used $3 \times 3$ windows). Our hierarchical exponential kernels are defined by $\Theta(x, z)$ as follows:

$$\Theta_{ij}^{[0]}(x, z) = x_iz_j$$

$$s_{ij}^{[h]}(x, z) = \sum_{m \in P} \Theta_{ij}^{[h]}(x_{i+m}, z_{j+m}) + \beta^2$$

$$\Theta_{ij}^{[h+1]}(x, z) = K(s_{ij}^{[h]}(x, z), s_{ii}^{[h]}(x, x), s_{jj}^{[h]}(z, z))$$

$$\bar{\Theta}(x, z) = \sum_i \Theta^{[h]}_{ii}(x, z)$$

where $\beta \geq 0$ denotes the bias and the last step is analogous to a fully connected layer in networks, and we set

$$K(s_{ij}, s_{ii}, s_{jj}) = \sqrt{s_{ii}s_{jj}} k\left(\frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}\right)$$

where $k$ can be any kernel defined on the sphere. In the experiments we applied this scheme to the three exponential kernels, Laplace, Gaussian and $\gamma$-exponential.

Technical details We used the following four kernels:

CNTK \[1\] $L = 6, \beta = 3$.

C-Exp Laplace. $L = 3, \beta = 3, k(x^Tz) = a + be^{-c\sqrt{2-2x^Tz}}$ with $a = -11.491, b = 12.606, c = 0.048$.

C-Exp $\gamma$-exponential. $L = 8, \beta = 3, k(x^Tz) = a + be^{-c(2-2x^Tz)^{\gamma/2}}$ with $a = -0.276, b = 1.236, c = 0.424, \gamma = 1.888$.

C-Exp Gaussian. $L = 12, \beta = 3, k(x^Tz) = a + be^{-c(2-2x^Tz)}$ with $a = -0.22, b = 1.166, c = 0.435$.

We set $\beta$ in these experiments with cross validation in $\{1, ..., 10\}$. For each kernel $k$ above, the parameters $a, b, c$ and $\gamma$ were chosen using non-linear least squares optimization with the objective $\sum_{u \in U} (k(u) - k^{FC,(b)}(u))^2$, where $k^{FC,(b)}$ is the NTK for a two-layer network defined in \[5\] with bias $\beta = 1$, and the set $U$ included (inner products between) pairs of normalized $3 \times 3 \times 3$ patches drawn uniformly from the CIFAR images. The number of layers $L$ is chosen by cross validation.

For the training phase we used 1-hot vectors from which we subtracted 0.1, as in \[12\]. For the classification phase, as in \[10\], we normalized the kernel matrices such that all the diagonal elements are ones. To avoid ill conditioned kernel matrices we applied ridge regression with a regularization factor of $\lambda = 5 \cdot 10^{-5}$. Finally, to reduce overall running times, we parallelized the kernel computations on NVIDIA Tesla V100 GPUs.

References


