Appendix

A0 Mathematical preliminaries

We will make use of functional analysis results on the theory of Hilbert space. We refer to [Lang, 2012] for a comprehensive introduction to the topic. We precise here that, even when not explicitly stated, all Hilbert spaces considered in the present work are real, and all linear operator are bounded.
We will make use of the spectral theory for compact self-adjoint operators. We refer again to [Lang, 2012] for a detailed discussion.

We will now introduce some concepts from the theory of kernels and RKHSs.
Consider a compact $K \subset \mathbb{R}^d$. A function $Q : K \to \mathbb{R}$ is said to be symmetric if for all $x, x' \in K$ we have $Q(x, x') = Q(x', x)$. Let us restate the definition of kernel.

Definition 2 (Kernel)**.** *A kernel Q on K is a symmetric continuous function $K^2 \to \mathbb{R}$ such that, for all $n \in \mathbb{N}$, for any finite subset $\{x_1 \dots x_n\} \subset K$, the matrix $\{Q(x_i, x_j)\}_{i,j}$ is non-negative definite.*

We state here a characterisation of kernels, which is an extension of Lemma 2. Despite being a classical result (see the discussion about Mercer kernels in [Paulsen and Raghupathi, 2016]), we will give a proof, for the sake of completeness.

Lemma A1. *[Extension of Lemma 2] Let $Q : K^2 \to \mathbb{R}$ be a continuous symmetric function. Then, given any finite Borel measure μ on K, we can define the integral operator $T_\mu(Q)$ on $L^2(K, \mu)$, via*

$$T_\mu(Q)\,\varphi(x) = \int_K T(x, x')\,\varphi(x')\,\mathrm{d}\mu(x')\,,$$

for any $\varphi \in L^2(K, \mu)$. The operator $T_\mu(Q)$ is a bounded compact self-adjoint definite operator.
Moreover, Q is a kernel if and only if $T_\mu(Q)$ is non-negative definite for all finite Borel measures μ on K.

Proof. Let $Q : K^2 \to \mathbb{R}$ be a continuous symmetric function. Then $T_\mu(Q)$ is a well defined bounded compact self-adjoint operator [Lang, 2012].
Let us assume that Q is a kernel. By Mercer's theorem [Paulsen and Raghupathi, 2016], we can find continuous functions $\{Y_k\}_{k\in\mathbb{N}}$ such that for all $x, x' \in K$

$$Q(x, x') = \sum_{k=0}^{\infty} Y_k(x) Y_k(x')$$

and the convergence is uniform on K^2.
The continuity of the Y_k's implies that they can be seen as elements of $L^2(K, \mu)$. Moreover, the uniform convergence, along with the fact that $\mu(K) < \infty$, implies the convergence of the sum wrt the $L^2(K, \mu)$ operator norm. In particular $T_\mu(K)$ is a limit of non-negative definite operators and hence non-negative definite.
Now, assume that, for all finite Borel μ, $T_\mu(Q)$ is non-negative definite. Chosen a finite set $\{x_1 \dots x_n\} \subset K$, in particular we have that $\mu = \sum_{i=1}^n \delta_{x_i}$ is a finite Borel measure (where δ_x is the Dirac measure on $x \in K$). Hence $T_\mu(Q)$ is the matrix $\{Q(x_i, x_j)\}_{i,j}$. We conclude that Q is a kernel. □

We will now give a definition of the Reproducing Kernel Hilbert Space associated to a kernel. We refer to [Paulsen and Raghupathi, 2016] for a general and comprehensive introduction to the topic.

Definition A1 (RKHS)**.** *Given a kernel Q on K, we can associate to it a real Hilbert space $\mathcal{H}_Q$, with the following properties:*

- *The elements of $\mathcal{H}_Q$ are functions $K \to \mathbb{R}$.*
- *Denoting as $\langle \cdot, \cdot \rangle_Q$ the inner product of $\mathcal{H}_Q$, for each $x \in K$, there exists a element $k_x \in \mathcal{H}_Q$ such that $h(x) = \langle h, k_x \rangle_Q$, for all $h \in \mathcal{H}_Q$.*
- *For all $x, x' \in K$, $\langle k_x, k_{x'} \rangle_Q = Q(x, x')$.*

Such a Hilbert space exists for each kernel Q and it is unique up to isomorphism, [Paulsen and Raghupathi, 2016]. $\mathcal{H}_Q$ is called the Reproducing Kernel Hilbert Space (RKHS) of Q.

In general, it is not easy to give an explicit form for the RKHS associated to a kernel Q. However, we can say that it contains the linear span of $\{x \mapsto Q(x, x')\}_{x' \in K}$. Actually, this linear span is a dense subset of $\mathcal{H}_Q$, wrt the norm of $\mathcal{H}_Q$ [Paulsen and Raghupathi, 2016].

A kernel on K is said to be universal if its RKHS is dense in the space of continuous functions $C(K)$, wrt the uniform norm.

Definition 4 (Universal Kernel)**.** *Let Q be a kernel on K, and $\mathcal{H}_Q(K)$ its RKHS. We say that Q is universal on K if for any $\varepsilon > 0$ and any continuous function g on K, there exists $h \in \mathcal{H}_Q(K)$ such that $\|h - g\|_\infty < \varepsilon$.*

We can now state a characterization of universal kernels, from [Sriperumbudur et al., 2011].

Lemma A2. *Let $Q : K^2 \to \mathbb{R}$ be a kernel, where $K \subset \mathbb{R}^d$ is compact. Q is a universal kernel if and only if $T_\mu(Q)$ is strictly positive definite for all finite Borel measures μ on K, i.e., $\langle T_\mu(Q)\,\varphi, \varphi\rangle > 0$ for all non-zero $\varphi \in L^2(K, \mu)$.*

As a final note, hereafter we often omit the explicit reference to the measure μ, that is we will speak of the operator $T(Q)$ on $L^2(K)$. Unless otherwise stated, this notation implies the choice of an arbitrary finite Borel measure μ on the compact K.

A1 Residual Neural Networks and Gaussian processes

Consider a standard ResNet architecture with $L+1$ layers, labelled with $l \in [0 : L]$, of dimensions $\{N_l\}_{l \in [0:L]}$.

$$\begin{aligned} &y_0(x) = W_0\, x + B_0\,; \\ &y_l(x) = y_{l-1}(x) + \mathcal{F}((W_l, B_l), y_{l-1}) \quad \text{for } l \in [1 : L]\,, \end{aligned} \tag{1}$$

where $x \in \mathbb{R}^d$ is an input, $y^l(x)$ is the vector of pre-activations, W^l and B^l are respectively the weights and bias of the l^{th} layer, and $\mathcal{F}$ is a mapping that defines the nature of the layer. In general, the mapping $\mathcal{F}$ consists of successive applications of simple activation functions. In this work, for the sake of simplicity, we consider Fully Connected blocks with ReLU activation function $\phi : x \mapsto \max(0, x)$

$$\mathcal{F}((W, B), x) = W\phi(x) + B\,.$$

Hereafter, N_l denotes the number of neurons in the l^{th} layer, ϕ the activation function and $[m : n] := \{m, m+1, ..., n\}$ for $m \leq n$. The components of weights and bias are respectively initialized with $W_l^{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_w^2/N_{l-1})$, and $B_l^i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_b^2)$ where $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution of mean μ and variance σ^2.

In [Yang and Schoenholz, 2017], authors showed that wide deep ResNets might suffer from gradient exploding during backpropagation.

Recent results by [Hayou et al., 2021] suggest that scaling the residual blocks with $L^{-1/2}$ might have some beneficial properties on model pruning at initialization. This is a result of the stabilization effect of scaling on the gradient.

More generally, we introduce the residual architecture:

$$\begin{aligned} &y_0(x) = W_0\, x + B_0\,; \\ &y_l(x) = y_{l-1}(x) + \lambda_{l,L}\, \mathcal{F}((W_l, B_l), y_{l-1})\,, \quad l \in [1 : L]\,, \end{aligned} \tag{2}$$

where $(\lambda_{k,L})_{k \in [1:L]}$ is a sequence of scaling factors. We assume hereafter that there exists $\lambda_{\max} \in (0, \infty)$ such that for all $L \geq 1$ and $k \in [1 : L]$, we have that $\lambda_{k,L} \in (0, \lambda_{\max}]$.

A1.1 Recurrence for the covariance kernel

Recall that in the limit of infinite width, each layer of a ResNet can be seen a centred Gaussian Process. For the layer l we define the covariance kernel Q_l as $Q_l(x, x') = \mathbb{E}[y_l^1(x) y_l^1(x')]$ for $x, x' \in \mathbb{R}^d$.

By a standard approach, introduced by [Schoenholz et al., 2017] for feedforward neural networks, and easily generalizable for ResNets [Yang, 2019b, Hayou et al., 2019b], it is possible to evaluate the covariance kernels layer by layer, recursively. More precisely, consider a ResNet of form (2). Assume that y_{l-1}^i is a Gaussian process for all i. Let $x, x' \in \mathbb{R}^d$. We have that

$$\begin{aligned} Q_l(x,x') &= \mathbb{E}[y_l^1(x)y_l^1(x')] \\ &= \mathbb{E}[y_{l-1}^1(x)y_{l-1}^1(x')] + \sum_{j=1}^{N_{l-1}} \mathbb{E}[(W_l^{1j})^2\phi(y_{l-1}^j(x))\phi(y_{l-1}^j(x'))] + \mathbb{E}[(B_l^1)^2] + \mathbb{E}[B_l^1(y_{l-1}^1(x)+y_{l-1}^1(x'))] \\ &+ \mathbb{E}\left[\sum_{j=1}^{N_{l-1}} W_l^{1j}(y_{l-1}^1(x)\phi(y_{l-1}^1(x')) + y_{l-1}^1(x')\phi(y_{l-1}^1(x)))\right]. \end{aligned}$$

Some terms vanish because $\mathbb{E}[W_l^{1j}] = \mathbb{E}[B_l^j] = 0$. Let $Z_j = \frac{\sqrt{N_{l-1}}}{\sigma_w} W_l^{1j}$. The second term can be written as

$$\mathbb{E}\left[\frac{\sigma_w^2}{N_{l-1}}\sum_j (Z_j)^2\phi(y_{l-1}^j(x))\phi(y_{l-1}^j(x'))\right] \to \sigma_w^2\,\mathbb{E}[\phi(y_{l-1}^1(x))\phi(y_{l-1}^1(x'))]\,,$$

where we have used the Central Limit Theorem. Therefore, we have

$$Q_l(x,x') = Q_{l-1}(x,x') + \lambda_{l,L}^2\Psi_{l-1}(x,x')\,, \tag{A1}$$

where $\Psi_{l-1}(x,x') = \sigma_b^2 + \sigma_w^2\mathbb{E}[\phi(y_{l-1}^1(x))\phi(y_{l-1}^1(x'))]$.

For the ReLU activation function $\phi(x) = \max(0,x)$, the recurrence relation can be written more explicitly, since we can give a simple expression for the expectation $\mathbb{E}[\phi(y_{l-1}^1(x))\phi(y_{l-1}^1(x'))]$, [Daniely et al., 2016]. Let C_l be the correlation kernel, defined as

$$C_l(x,x') = \frac{Q_l(x,x')}{\sqrt{Q_l(x,x)Q_l(x,x')}}$$

and let $f : [-1,1] \to \mathbb{R}$ be given by

$$f : \gamma \mapsto \frac{1}{\pi}(\sqrt{1-\gamma^2} - \gamma\arccos\gamma)\,. \tag{4}$$

Then we have $\mathbb{E}[\phi(y_{l-1}^1(x))\phi(y_{l-1}^1(x'))] = \frac{1}{2}\left(1 + \frac{f(C_{l-1})}{C_{l-1}}\right)Q_{l-1}$ and so we find the recurrence relation (5)

$$\begin{aligned} &Q_l = Q_{l-1} + \lambda_{l,L}^2\left[{\sigma_b}^2 + \tfrac{{\sigma_w}^2}{2}\left(1 + \tfrac{f(C_{l-1})}{C_{l-1}}\right)Q_{l-1}\right]; \\ &Q_0(x,x') = \sigma_b^2 + \sigma_w^2\,\tfrac{x\cdot x'}{d}\,. \end{aligned} \tag{5}$$

For the remainder of this appendix, we define the function

$$\hat{f}(\gamma) = \gamma + f(\gamma) = \frac{1}{\pi}\left(\gamma\arcsin(\gamma) + \sqrt{1-\gamma^2}\right) + \frac{1}{2}\gamma\,. \tag{A2}$$

For all l, the diagonal terms of Q_l have closed-form expressions. We show this in the next lemma.

Lemma A3 (Diagonal elements of the covariance)**.** *Consider a ResNet of the form (2) and let $x \in \mathbb{R}^d$. We have that for all $l \in [1:L]$,*

$$Q_l(x,x) = -\frac{2\sigma_b^2}{\sigma_w^2} + \prod_{k=1}^{l}\left(1 + \frac{\sigma_w^2\lambda_{k,L}^2}{2}\right)\left(Q_0(x,x) + \frac{2\sigma_b^2}{\sigma_w^2}\right).$$

Proof. We know that

$$Q_l(x,x) = Q_{l-1}(x,x) + \lambda_{l,L}^2\left(\sigma_b^2 + \frac{\sigma_w^2}{2}\hat{f}(1)\right),$$

where $\hat{f}$ is given by (A2). It is straightforward that $\hat{f}(1) = 1$. This yields

$$Q_l(x,x) + \frac{2\sigma_b^2}{\sigma_w^2} = \left(1 + \lambda_{l,L}^2 \frac{\sigma_w^2}{2}\right)\left(Q_{l-1}(x,x) + \frac{2\sigma_b^2}{\sigma_w^2}\right).$$

we conclude by telescopic product. □

As a corollary of the previous result, it is easy to show that for a Standard ResNet the diagonal terms explode with depth, which is Lemma 1 in the main paper.

Lemma 1 (Exploding kernel with standard ResNet). *Consider a ResNet of type (1). Then, for all $x \in \mathbb{R}^d$,*

$$Q_L(x,x) \geq \left(1 + \tfrac{\sigma_w^2}{2}\right)^L \left(\sigma_b^2\left(1 + \tfrac{2}{\sigma_w^2}\right) + \tfrac{\sigma_w^2}{d}\|x\|^2\right).$$

Proof. The statement trivially follows from Lemma A3, using that $Q_0(x,x) = \sigma_b^2 + \frac{\sigma_w^2}{d}\|x\|^2$ and the fact that for a Standard ResNet (1), all the coefficients $\lambda_{l,L}$'s are equal to 1. □

In the case of a ResNet with no bias, the correlation kernel follows a simple recursive formula described in the next lemma.

Lemma A4 (Correlation formula with zero bias). *For a ResNet of the form (2) with $\sigma_b = 0$, we have that for all $x, x' \in \mathbb{R}^d$ and $l \leq L$:*

$$C_l(x,x') = \frac{1}{1+\alpha_{l,L}} C_{l-1}(x,x') + \frac{\alpha_{l,L}}{1+\alpha_{l,L}} \hat{f}(C_{l-1}(x,x')),$$

where $\alpha_{l,L} = \frac{\lambda_{l,L}^2 \sigma_w^2}{2}$.

Proof. This is direct result of the covariance recursion formula (5). □

A1.2 Proof of Proposition 1

We use the following result from [Yang, 2020] in order to derive closed form expressions for the second moment of the gradients.

Lemma A5 (Corollary of Theorem D.1. in [Yang, 2020]). *Consider a ResNet of the form (2) with weights W. In the limit of infinite width, we can assume that W^T used in back-propagation is independent from W used for forward propagation, for the calculation of Gradient Covariance and NTK.*

Next we re-state and prove Proposition 1.

Proposition 1 (Stable Gradient). *Consider a ResNet of type (2), and let $\mathcal{L}_y(x) := \ell(y_L^1(x), y)$ for some $(x,y) \in \mathbb{R}^d \times \mathbb{R}$, where $\ell : (z,y) \mapsto \ell(z,y)$ is a loss function satisfying $\sup_{K_1 \times K_2} \left|\frac{\partial \ell(z,y)}{\partial z}\right| < \infty$, for all compacts $K_1, K_2 \subset \mathbb{R}$. Then, in the limit of infinite width, for any compacts $K \subset \mathbb{R}^d$, $K' \subset \mathbb{R}$, there exists a constant $C > 0$ such that for all $(x,y) \in K \times K'$*

$$\sup_{l \in [0:L]} \mathbb{E}\left[\left|\frac{\partial \mathcal{L}_y(x)}{\partial W_l^{11}}\right|^2\right] \leq C \exp\left(\tfrac{\sigma_w^2}{2}\textstyle\sum_{l=1}^L \lambda_{l,L}^2\right).$$

Moreover, if there exists $\lambda_{\min} > 0$ such that for all $L \geq 1$ and $l \in [1:L]$ we have $\lambda_{l,L} \geq \lambda_{\min}$, then, for all $(x,y) \in (\mathbb{R}^d \setminus \{0\}) \times \mathbb{R}$ such that $\left|\frac{\partial \ell(z,y)}{\partial z}\right| \neq 0$, there exists $\kappa > 0$ such that for all $l \in [1:L]$

$$\mathbb{E}\left[\left|\frac{\partial \mathcal{L}_y(x)}{\partial W_l^{11}}\right|^2\right] \geq \kappa\left(1 + \tfrac{\lambda_{\min}^2 \sigma_w^2}{2}\right)^L.$$

Proof. Let $(x,y) \in \mathbb{R}^d \times \mathbb{R}$ and $\bar{q}^l(x,y) = \mathbb{E}\left[\left|\frac{\partial \mathcal{L}_y(x)}{\partial y_l^1}\right|^2\right]$. Using Lemma A6, we have that

$$\bar{q}^l(x,y) = \left(1 + \frac{\sigma_w^2 \lambda_{l+1,L}^2}{2}\right) \bar{q}^{l+1}(x,y)\,.$$

This yields

$$\bar{q}^l(x,y) = \prod_{k=l+1}^{L} \left(1 + \frac{\sigma_w^2 \lambda_{k,L}^2}{2}\right) \bar{q}^l(x,y)\,.$$

Moreover, using Lemma A5, we have that $\mathbb{E}\left[\left|\frac{\partial \mathcal{L}_y(x)}{\partial W_l^{11}}\right|^2\right] = \lambda_{l,L}^2 \bar{q}^l(x,y)\mathbb{E}[\phi(y_{l-1}^1(x))^2]$. We have $\mathbb{E}[\phi(y_{l-1}^1(x))^2] = \frac{1}{2}Q_{l-1}(x,x)$. From Lemma A3 we know that

$$Q_{l-1}(x,x) = -\frac{2\sigma_b^2}{\sigma_w^2} + \prod_{k=1}^{l-1}\left(1 + \frac{\sigma_w^2\lambda_{k,L}^2}{2}\right)\left(Q_0(x,x) + \frac{2\sigma_b^2}{\sigma_w^2}\right) \leq \prod_{k=1}^{l-1}\left(1 + \frac{\sigma_w^2\lambda_{k,L}^2}{2}\right)\left(Q_0(x,x) + \frac{2\sigma_b^2}{\sigma_w^2}\right)\,,$$

This yields

$$\mathbb{E}\left[\left|\frac{\partial \mathcal{L}_y(x)}{\partial W_{11}^l}\right|^2\right] \leq \frac{2}{\sigma_w^2}\prod_{k=1}^{L}\left(1 + \frac{\sigma_w^2\lambda_{k,L}^2}{2}\right)\left(\frac{1}{2}Q_0(x,x) + \frac{\sigma_b^2}{\sigma_w^2}\right)\bar{q}^l(x,y)\,.$$

It is straightforward that $\prod_{k=1}^{L}\left(1 + \frac{\sigma_w^2\lambda_{k,L}^2}{2}\right) \leq \exp\left(\frac{\sigma_w^2}{2}\sum_{k=1}^{L}\lambda_{k,L}^2\right)$. Let $K \subset \mathbb{R}^d$, $K' \subset \mathbb{R}$ be two compact subsets. Using the condition on the loss function ℓ, we have that

$$\mathbb{E}\left[\left|\frac{\partial \mathcal{L}_y(x)}{\partial W_{11}^l}\right|^2\right] \leq C\exp\left(\frac{\sigma_w^2}{2}\sum_{k=1}^{L}\lambda_{k,L}^2\right),$$

where $C = \frac{2}{\sigma_w^2}\left(\sup_{(x,y)\in K\times K'}\bar{q}^l(x,y)\right)\left(\sup_{x\in K}Q_0(x,x) + \frac{2\sigma_b^2}{\sigma_w^2}\right)$. We conclude by taking the supremum over l and x, y.

Let $(x,y) \in (\mathbb{R}^d \setminus \{0\}) \times \mathbb{R}$ such that $\left|\frac{\partial \ell(z,y)}{\partial z}\right| \neq 0$. We have that

$$\begin{aligned}\mathbb{E}\left[\left|\frac{\partial \mathcal{L}_y(x)}{\partial W_{11}^l}\right|^2\right] &\geq \frac{1}{2}\frac{\lambda_{l,L}^2}{1 + \frac{\sigma_w^2}{2}\lambda_{l,L}^2}\prod_{k=2}^{L}\left(1 + \frac{\sigma_w^2\lambda_{k,L}^2}{2}\right)Q_1(x,x)\bar{q}^l(x,y) \\ &\geq \kappa\left(1 + \frac{\sigma_w^2\lambda_{\min}^2}{2}\right)^L,\end{aligned}$$

where $\kappa = \frac{1}{2}\frac{\lambda_{\min}^2}{\left(1+\frac{\sigma_w^2}{2}\lambda_{\max}^2\right)\left(1+\frac{\sigma_w^2}{2}\lambda_{\min}^2\right)}Q_1(x,x)\,\bar{q}^l(x,y) > 0.$ □

Using Lemma A5, we can derive simple recursive formulas for the second moment of the gradient as well as for the Neural Tangent Kernel (NTK). This was previously done in [Schoenholz et al., 2017] for feedforward neural networks, we prove a similar result for ResNet in the next lemma.

Lemma A6 (Gradient Second moment)**.** *In the limit of infinite width, using the same notation as in proposition 1, we have that*

$$\bar{q}^l(x,y) = \left(1 + \frac{\sigma_w^2\lambda_{l+1,L}^2}{2}\right)\bar{q}^{l+1}(x,y)\,.$$

Proof. It is straighforward that

$$\frac{\partial \mathcal{L}_y(x)}{\partial y_l^i} = \frac{\partial \mathcal{L}_y(x)}{\partial y_{l+1}^i} + \lambda_{l+1,L} \sum_j \frac{\partial \mathcal{L}_y(x)}{\partial y_{l+1}^j} W_{l+1}^{ji} \phi'(y_l^i)\,.$$

Using lemma A5 and the Central Limit Theorem, we have that

$$\bar{q}^l(x,y) = \bar{q}^{l+1}(x,y) + \lambda_{l+1,L}^2 \bar{q}^{l+1}(x,y)\sigma_w^2 \mathbb{E}[\phi'(y_l^i(x))^2]\,.$$

We conclude using $\mathbb{E}[\phi'(y_i^l(x))^2] = \mathbb{P}(\mathcal{N}(0,1) > 0) = \frac{1}{2}$. □

Before moving to the next proofs, recall the definition of Stable ResNet.

Definition 1 (Stable ResNet). *A ResNet of type (2) is called a Stable ResNet if and only if* $\lim_{L\to\infty} \sum_{k=1}^{L} \lambda_{k,L}^2 < \infty$.

A1.3 Some general results: Q_l and C_l are kernels

Fix a compact $K \subset \mathbb{R}^d$. If $\sigma_b = 0$, then assume that $0 \notin K$. We will now show that, for all layers l, the covariance function Q_l is a kernel in the sense of Definition 2.
The symmetric property of Q_l is clear by definition as the covariance of a Gaussian Process. Let us now discuss the regularity of Q_l as a function on K^2.

The next result shows that any function $F(\phi) : \gamma \mapsto \mathbb{E}[\phi(X)\phi(Y), (X,Y) \sim \mathcal{N}(0, \left(\begin{smallmatrix} 1 & \gamma \\ \gamma & 1 \end{smallmatrix}\right))]$ is analytic on the segment $[-1,1]$.

Lemma A7 (O'Donnell (2014)). *Let $F(\phi)(\gamma) = \mathbb{E}[\phi(X)\phi(Y), (X,Y) \sim \mathcal{N}(0, \left(\begin{smallmatrix} 1 & \gamma \\ \gamma & 1 \end{smallmatrix}\right))]$. Then for all $\phi \in L^2(\mathcal{N}(0,1))$, there exists a non negative sequence $\{a_n\}_{n\in\mathbb{N}}$ such that $F(\phi)(\gamma) = \sum_{i\in\mathbb{N}} a_i \gamma^i$ for all $\gamma \in [-1,1]$.*

Leveraging the previous result, the function f defined in (4) is analytic. We clarify this in the next lemma.

Lemma A8 (Analytic property of f). *The function $f : [-1,1] \to \mathbb{R}$, defined in (4), is an analytic function on $(-1,1)$, whose expansion $f(\gamma) = \sum_{n\in\mathbb{N}} \alpha_n \gamma^n$ converges absolutely on $[-1,1]$. Moreover, $\alpha_n > 0$ for all even $n \in \mathbb{N}$, $\alpha_1 = -1/2$ and $\alpha_n = 0$ for all odd $n \geq 3$.*

Proof. With the notations of Lemma A7, when ϕ is the ReLU activation function we have that $F(\phi) = \hat{f}$, defined in (A2). Hence, by Lemma A7, we know that $\hat{f}$ is analytic on $(-1,1)$ and its expansion around 0 converges on $[-1,1]$. In particular this will be true for f as well.
For $\gamma \in [-1,1]$, let us write $\hat{f}(\gamma) = \sum_{n\in\mathbb{N}} a_n \gamma^n$. Recalling the explicit form of $\hat{f}$, that is

$$\hat{f}(\gamma) = \frac{1}{\pi}\gamma \arcsin(\gamma) + \frac{1}{\pi}\sqrt{1-\gamma^2} + \frac{1}{2}\gamma\,,$$

we get $a_0 = \frac{1}{\pi}$. Moreover, we have that for all $\gamma \in (-1,1)$

$$\hat{f}'(\gamma) = \frac{1}{\pi}\arcsin\gamma + \frac{1}{2}\,.$$

This yields $a_1 = \hat{f}'(0) = \frac{1}{2}$. Then, noticing that

$$\hat{f}^{(3)}(\gamma) = \frac{\gamma}{\pi(1-\gamma^2)^{3/2}}$$

is an odd function, we get that for all $i \geq 1, a_{2i+1} = 0$. Now let us prove that for all $k \geq 1$, there exist $b_{k,0}, b_{k,1}, ..., b_{k,k-1} > 0$ such that, for all $\gamma \in (-1,1)$,

$$\hat{f}^{(2k)}(\gamma) = \frac{1}{\pi}\sum_{m=0}^{k-1} b_{k,m}\gamma^{2m}(1-\gamma^2)^{-k-m+1/2}\,.$$

We prove this by induction. For $k = 1$, we have that

$$\hat{f}^{(2)}(\gamma) = \frac{1}{\pi}(1-\gamma^2)^{-1/2}\,,$$

so that our claim holds. Assume now that it is true for some $k \geq 1$, let us prove it for $k+1$. It is easy to see that

$$b_{k+1,m} = \begin{cases} 2(2k-1)b_{k,0} + 2b_{k,1} & \text{if } m = 0; \\ 2(4k^2-1)b_{k,0} + 5(2k+1)b_{k,1} + 12b_{k,2} & \text{if } m = 1; \\ 2(m+1)(2m+1)b_{k,m+1} + (4m+1)(2k+2m-1)b_{k,m} & \\ \quad + (2k+2m-3)(2k+2m-1)b_{k,m-1} & \text{if } m \in \{2,3,...,k-1\}; \\ (4k-3)(4k-1)b_{k,k-1} & \text{if } m = k. \end{cases} \tag{A3}$$

The induction is straightforward. In particular, we have shown that $a_{2i} = \frac{\hat{f}^{(2i)}(0)}{(2i)!} = \frac{b_{i,0}}{(2i)!} > 0$.
The conclusion for the coefficients α's of the expansion of f is then trivial. □

Using Lemma A8, it will not be hard to show that Q_l is continuous. The non-negativity of $T(Q_l)$ can be seen as a consequence of the definition of Q_l as the covariance of a Gaussian Process. However, we will give a direct proof of it, so that we can state here a general result which we will need later on.

Lemma A9. *Let C be a kernel on K, such that $|C(z)| \leq 1$ for all $z \in K$. Consider a non-negative real sequence $\{\alpha_n\}_{n\in\mathbb{N}}$, and assume that*

$$g(\gamma) = \sum_{k=0}^{\infty} \alpha_k\, \gamma^k$$

converges uniformly on $[-1,1]$. Then, for all finite Borel measure μ on K, $T_\mu(g(C))$ is a non-negative definite compact operator, and in particular $g(C)$ is a kernel.

Proof. Fix a finite Borel measure μ on K and notice that $g(C)$ is continuous and symmetric (as uniform limit of continuous and symmetric functions). Moreover, since the Taylor expansion of g around 0 converges uniformly on $[-1,1]$, and since $|C(z)| \leq 1$ for all $z \in K$, we have that $T_\mu(g(C)) = \sum_{k\in\mathbb{N}} \alpha_k\, T_\mu(C^k)$, the sum converging wrt the operator norm on $L^2(K,\mu)$.
As a consequence of the Schur product theorem[14], the product of two kernels is still a kernel.
As a consequence, it is easy to prove by induction that $T_\mu(C^k)$ is non-negative definite for all k. Hence $T_\mu(g(C))$ is the converging limit of a sum of compact non-negative definite operator. We conclude by Lemma A1. □

Lemma A10. *For both Standard and Stable ResNet architectures, for any layer l, the covariance function Q_l and the correlation function C_l are kernels on K, in the sense of Definition 2.*

Proof. It is straightforward to prove that Q_0 is a kernel. Now let us show that if Q_l is a kernel for some l, then C_l is a kernel. Since Q_l is symmetric and so C_l is. Moreover, the diagonal elements of Q_l are continuous by Lemma A3 and do not vanish (since if $\sigma_b = 0$ we are assuming that $0 \notin K$). Hence C_l is continuous. It is then trivial to show that the non-negative definiteness of $T(Q_l)$ implies that $T(C_l)$ is non-negative definite, and so C_l is a kernel if Q_l is.
Now we proceed by induction. Suppose that Q_{l-1} and C_{l-1} are kernels and recall the recursion (5), taking the coefficient λ to be 1 in the case of a Standard ResNet. Notice that it can be rewritten as

$$Q_l = Q_{l-1} + \lambda_l^2 \left({\sigma_b}^2 + \tfrac{{\sigma_w}^2}{2} \hat{f}(C_{l-1})\, R_{l-1} \right)\,,$$

where we have omitted the dependence on L for λ, we have defined $R_{l-1}(x,x') = \sqrt{Q_{l-1}(x,x)Q_{l-1}(x',x')}$ and $\hat{f}$ is defined in (A2). Clearly R_{l-1} is a kernel. By Lemma A8 and Lemma A9 we have that $\hat{f}(C_l)$ is a kernel. Using the property that sums and products of kernels are kernels (the sum is trivial, cf Footnote 14 for the product), we conclude that Q_l, and so C_l, is a kernel on K. □

[14] Given two matrices M_1 and M_2, define they're Schur product as the matrix $M = M_1 \circ M_2$, whose elements are $M^{ij} = {M_1}^{ij} {M_2}^{ij}$. If M_1 and M_2 are non-negative definite, then M is non-negative definite.

A1.4 Proof of Proposition 2

As always, consider an arbitrary compact set $K \in \mathbb{R}^d$. Assume that $0 \notin K$ if $\sigma_b = 0$. Recall from Appendix A0 that with the notation $\mathcal{H}_Q(K)$ we refer to the RKHS generated by a kernel Q on K. We will now prove Proposition 2.

Proposition 2. *$\mathcal{H}_{Q_l}(K) \subseteq \mathcal{H}_{Q_{l+1}}(K)$ for all $l \in [0 : L-1]$.*

Proof. We have already shown that $T(Q_l) - T(Q_{l-1})$ is non-negative definite in the proof of Lemma A10. We conclude by using the RKHS hierarchy result (see for instance [Paulsen and Raghupathi, 2016] or page 354 in [Aronszajn, 1950]). □

A1.5 Proof of Lemma 3

We present here the proof of Lemma 3. We have already recalled the Definition 4 of universal kernel in Appendix A0. For convenience of the reader, we restate here the definition of expressive GP.
Let K be a compact in $\mathbb{R}^d$.

Definition 5 (Expressive GP). *A Gaussian Process on K is said to be expressive on $L^2(K)$ if, denoted by ψ a random realisation, for all $\varphi \in L^2(K)$, for all $\varepsilon > 0$,*

$$\mathbb{P}(\|\psi - \varphi\|_2 \leq \varepsilon) > 0\,.$$

Lemma 3. *A universal kernel Q on K induces an expressive GP on $L^2(K)$.*

Proof. First, notice that if Q is universal then $T(Q)$ is strictly positive definite [Sriperumbudur et al., 2011] and so all its eigenvalues are strictly positive.
Recall the spectral theorem for compact self-adjoint operators: there is a orthonormal basis of $L^2(K)$ made of the eigenfunctions $\{\psi_n\}_{n\in\mathbb{N}}$ of $T(Q)$. Denoting by $\mu_n > 0$ the eigenvalue of $T(Q)$ relatively to ψ_n, since $T(Q)$ is compact we have the equality (Karhunen - Loève decomposition [Grenander, 1950])

$$\psi = \sum_{k=0}^{\infty} Z_k \sqrt{\mu_k}\, \psi_k \sim \mathcal{GP}(0, Q)\,,$$

where $\{Z_k\}_{k\in\mathbb{N}}$ is a family of iid normal random variables, and the series is convergent uniformly on K and in L^2 for the stochastic part [Paulsen and Raghupathi, 2016], that is $\lim_{N\to\infty} \sup_{x\in K} \mathbb{E}[(\psi(x) - \sum_{k=0}^{N} Z_k \sqrt{\mu_k}\, \psi_k(x))^2] = 0$ uniformly for $x \in K$. In particular, we get that $\lim_{N\to\infty} \mathbb{E}[\|\psi - \sum_{k=0}^{N} Z_k \sqrt{\mu_k} \psi_k\|_2^2] = 0$. As consequence, for all $\varphi \in L^2(K)$, we have that $\| \sum_{k=0}^{N} Z_k \sqrt{\mu_k}\, \psi_k - \varphi\|_2^2$ converges in squared mean to $\|\psi - \varphi\|_2^2$, for $N \to \infty$.
Now, let $\varphi = \sum_{k=0}^{N} a_k\, \psi_k$ for some finite N and some real coefficients $\{a_0 \dots a_N\}$. We have (with convergence in squared mean)

$$\|\psi - \varphi\|_2^2 = \sum_{k=0}^{N} (Z_k \sqrt{\mu_k} - a_k)^2 + \sum_{k=N+1}^{\infty} \mu_k\, Z_k^2\,.$$

For $k \in [0 : N]$, we can define the interval $I_k = \left[\frac{a_k}{\sqrt{\mu_k}} - \frac{\varepsilon}{\sqrt{2(N+1)\mu_k}}, \frac{a_k}{\sqrt{\mu_k}} + \frac{\varepsilon}{\sqrt{2(N+1)\mu_k}} \right]$, so that, for all $z \in I_k$ we have $(z\sqrt{\mu_k} - a_k)^2 \leq \frac{\varepsilon^2}{2(N+1)}$. Since all these intervals are non empty, we get

$$\mathbb{P}\left(\sum_{k=0}^{N} (Z_k \sqrt{\mu_k} - a_k)^2 \leq \frac{\varepsilon^2}{2} \right) \geq \prod_{k=0}^{N} \mathbb{P}(Z_k \in I_k) > 0\,.$$

On the other hand, we have that

$$\delta_N = \mathbb{E}\left[\sum_{k=N+1}^{\infty} \mu_k\, Z_k^2 \right] = \sum_{k=N+1}^{\infty} \mu_k\,.$$

By Mercer's theorem [Paulsen and Raghupathi, 2016], $T(Q)$ is trace class and hence $\delta_N \to 0$ for diverging N. By Markov's inequality

$$\mathbb{P}\left(\sum_{k=N+1}^{\infty} \mu_k Z_k^2 \geq \frac{\varepsilon^2}{2}\right) \leq \frac{2\delta_N}{\varepsilon^2}$$

and we can conclude that $\mathbb{P}(\|\psi - \varphi\|_2 \leq \varepsilon) > 0$ for N large enough.

For a general $\varphi = \sum_{k=0}^{\infty} a_k \psi_k$, let $\varphi_N = \sum_{k=0}^{N} a_k \psi_k$. Since $\{\psi_k\}_{k\in\mathbb{N}}$ is a basis of $L^2(K)$, fixed $\varepsilon > 0$, it is always possible to find a N such that $\|\varphi - \varphi_N\|_2 \leq \varepsilon/2$ and $\mathbb{P}(\|\varphi_N - \psi\|_2 \leq \varepsilon/2) > 0$, and so we conclude. □

A1.6 Proof of Proposition 3

In order to prove Proposition 3 we first need a preliminary result, which will be at the core of the proof of Theorem 1 as well.

Proposition A1. *Let $K \subset \mathbb{R}^d$ be compact. Assume $\sigma_b > 0$ and let $\tilde{f} : \gamma \mapsto \frac{\gamma}{2} + f(\gamma)$ be defined on $[-1, 1]$. Then the kernel $\tilde{f}(c_0)$, defined point-wise as $\tilde{f}(c_0)(x, x') = \tilde{f}(c_0(x, x'))$, is universal on K.*

Proof. First notice that $c_0(x, x') = \frac{1+\zeta\, x\cdot x'}{\sqrt{(1+\zeta\,\|x\|^2)(1+\zeta\,\|x'\|^2)}}$, where $\zeta = \sigma_w^2/\sigma_b^2$. For $n \in \mathbb{N}$, define $p_n : (x, x') \mapsto c_0(x, x')^{2n}$, with the convention that $p_0 \equiv 1$. It is easy to verify that c_0 is kernel. As a consequence, p_n is a kernel for all n, since it is a product of kernels.[15] From Lemma A8, we can write

$$\tilde{f}(c_0) = \sum_{n\in\mathbb{N}} \alpha_n\, p_n\,,$$

the sum converging uniformly on K^2, with $\alpha_n > 0$ for all $n \in \mathbb{N}$. By Lemma A9, $\tilde{f}(c_0)$ is a kernel.
Now, for each n, we have

$$p_n(x, x') = \frac{1}{(1+\zeta\,\|x\|^2)^n (1+\zeta\,\|x'\|^2)^n} \sum_{k=0}^{2n} \omega_{k,n}\,(x\cdot x')^k\,,$$

where the coefficients $\omega_{k,n}$'s are all strictly positive, explicitly $\omega_{k,n} = \zeta^k \binom{n}{k}$.
Expanding the inner product $x \cdot x'$, we can express p_n in the form

$$p_n(x, x') = \sum_{J\in\mathcal{J}_n} \beta_{J,n}\, A_{J,n}(x) A_{J,n}(x')\,,$$

where $\mathcal{J}_n = \{(j_1 \dots j_d) \in \mathbb{N}^d : \sum_{i=1}^d j_i \in [0 : 2n]\}$, all the coefficients $\beta_{J,n}$'s are strictly positive and the $A_{J,n}$'s are defined as

$$A_{J,n}(x) = \frac{{x_1}^{j_1} \dots {x_d}^{j_d}}{(1+\zeta\,\|x\|^2)^n}\,.$$

Hence we can write $\tilde{f}(c_0)$ as

$$\tilde{f}(c_0)(x, x') = \sum_{n\in\mathbb{N}} \sum_{J\in\mathcal{J}_n} \alpha_n \beta_{J,n} A_{J,n}(x) A_{J,n}(x')\,. \tag{A4}$$

For any $n, n' \in \mathbb{N}$, $J \in \mathcal{J}_n$, $J' \in \mathcal{J}_{n'}$, it is clear that $A_{J,n} A_{J',n'} = A_{J'',n+n'}$, where J'' is some element in $\mathcal{J}_{n+n'}$. As a consequence, the linear span of the family $\{A_{J,n}\}_{n\in\mathbb{N}, J\in\mathcal{J}_n}$ is an algebra $\mathcal{A}$ (which is actually a subalgebra of $C(K)$ since all the $A_{J,n}$'s are continuous). Moreover $A_{(0\dots0),0} \equiv 1$, so that $\mathcal{A}$ contains a constant, and it is straightforward to check that $\mathcal{A}$ separates points, that is for all distinct $x, x' \in K$ there exists $a \in \mathcal{A}$ such that $a(x) \neq a(x')$. Then, from Stone-Weierstrass theorem [Lang, 2012], $\mathcal{A}$ is dense in $C(K)$ wrt the uniform norm.
For all $n \in \mathbb{N}$, $J \in \mathcal{J}_n$, let $\theta_{n,J} = \sqrt{\alpha_n \beta_{n,J}}$. Define a bijection $\iota : \mathbb{N} \to \{(n, J) : n \in \mathbb{N}, J \in \mathcal{J}_n\}$ and let $\Phi_n = \theta_{\iota(n)} A_{\iota(n)}$. For all $x \in K$, we have that $\Phi(x) = \{\Phi_n(x)\}_{n\in\mathbb{N}} \in \ell^2$, since $p_n(x, x) < \infty$. We conclude that Φ is a feature map for $\tilde{f}(c_0)$, and the density of the linear span of $\{\Phi_n\}_{n\in\mathbb{N}}$ allows to claim that the kernel is universal on K, in the sense of Definition 4 (cf Theorem 7 in [Micchelli et al., 2006]). □

[15] See footnote 14.

Let $K \subset \mathbb{R}^d$ be an arbitrary compact set. We are now ready to prove Proposition 3.

Proposition 3. *If $\sigma_b > 0$, then Q_2 is universal on K. From Proposition 2, Q_L is universal for all $L \geq 2$.*

Proof. Assume $\sigma_b > 0$ and let $K \subset \mathbb{R}^d$ be a compact set. With the notation of Proposition A1, we have that

$$Q_1 = Q_0 + \lambda_{1,L}^2 \left(\sigma_b^2 + \frac{\sigma_w^2}{2} \left(\frac{1}{2} + \frac{\tilde{f}(C_0)}{C_0} \right) Q_0 \right).$$

By proposition A1, we know that the kernel $\tilde{f}(C_0)$ given by $\tilde{f}(C_0)(x, x') = \tilde{f}(C_0(x, x'))$ is universal on K. Let us prove that $\frac{\tilde{f}(C_0)}{C_0} Q_0$ is universal. Let $\varepsilon > 0$ and $\varphi \in C(K)$, the space of continuous functions on K. Define $\frac{\varphi}{Q_0}(x) = \frac{\varphi(x)}{Q_0(x,x)}$. By the universality of $\tilde{f}(C_0)$, there exists $g \in \mathcal{H}_{\tilde{f}(C_0)}(K)$ such that

$$\left\| g - \frac{\varphi}{\sqrt{Q_0}} \right\|_\infty \leq \epsilon \,.$$

with g can be written as a finite linear combination of the functions $\{\hat{f}(C_0)(x, .)\}_{x \in K}$. This yields

$$\left\| g\sqrt{Q_0} - \varphi \right\|_\infty \leq \epsilon\kappa \,,$$

where $g\sqrt{Q_0}(x) = g(x)\sqrt{Q_0(x,x)}$ and $\kappa = \sup_{x \in K} \sqrt{Q_0(x,x)}$. It is straightforward that $g\sqrt{Q_0} \in \mathcal{H}_{\frac{\tilde{f}(C_0)}{C_0} Q_0}(K)$,[16] Therefore, $\frac{\tilde{f}(C_0)}{C_0} Q_0$ is universal. Since Q_0 is non-negative, we have that Q_1 is universal by an RKHS hierarchy argument similar to Proposition 2. Using Proposition 2, we conclude that Q_L is universal on K. □

A1.7 Proof of Proposition 4

Proposition 4. *Assume $\sigma_b = 0$. Then for all $L \geq 2$, Q_L is universal on $\mathbb{S}^{d-1}$ for $d \geq 2$.*

Proof. See the proof of Proposition A7 in Appendix A8. □

A1.8 Proof of Proposition 5

Proposition 5 is a well known classical result (see for instance Appendix H in [Yang and Salman, 2019] and the references therein. For completeness we give a proof in Appendix A8.

Proposition 5 (Spectral decomposition on $\mathbb{S}^{d-1}$)**.** *Let Q be a zonal kernel on $\mathbb{S}^{d-1}$, that is $Q(x, x') = p(x \cdot x')$ for a continuous function $p : [-1, 1] \to \mathbb{R}$. Then, there is a sequence $\{\mu_k \geq 0\}_{k \in \mathbb{N}}$ such that for all $x, x' \in \mathbb{S}^{d-1}$*

$$Q(x, x') = \sum_{k \geq 0} \mu_k \sum_{j=1}^{N(d,k)} Y_{k,j}(x) Y_{k,j}(x') \,,$$

where $\{Y_{k,j}\}_{k \geq 0, j \in [1:N(d,k)]}$ are spherical harmonics of $\mathbb{S}^{d-1}$ and $N(d, k)$ is the number of harmonics of order k. With respect to the standard spherical measure, the spherical harmonics form an orthonormal basis of $L^2(\mathbb{S}^{d-1})$ and $T(Q)$ is diagonal on this basis.

Proof. See the proof of Lemma A22 in Appendix A8. □

[16] This is trivial for a function g that can be written as a finite sum of functions of the form $\alpha_i \tilde{f}(C_0)(x_i, .)$, and this would be enough since these functions are dense in $C(K)$ as shown in the proof of Proposition A1. More generally, given two kernels Q and Q', if $h \in \mathcal{H}_Q$ and $h' \in \mathcal{H}_{Q'}$, then $hh' \in \mathcal{H}_{QQ'}$, cf Theorem 5.16 in [Paulsen and Raghupathi, 2016].

A1.9 Proof of Lemma 4

Lemma 4. *Consider a standard ResNet of type (1) and let $K \subset \mathbb{R}^d \setminus \{0\}$ be a compact set. We have that*

$$\lim_{L\to\infty} \sup_{x,x'\in K} |1 - C_L(x,x')| = 0\,.$$

Moreover, if $\sigma_b = 0$, then,

$$\sup_{x,x'\in K} |1 - C_L(x,x')| = \mathcal{O}(L^{-2})\,.$$

Therefore, $\mathcal{H}_{C_\infty}(K)$ is the space of constant functions.

Proof. This result was proven in [Hayou et al., 2019a] in the case of no bias. It was also proven for a slightly different ResNet architecture in [Yang and Schoenholz, 2017].
Consider a ResNet of type (1) and let $K \subset \mathbb{R}^d \setminus \{0\}$ be a compact set. We have that for all $x, x' \in K$

$$Q_L(x,x') = Q_{L-1}(x,x') + \sigma_b^2 + \frac{\sigma_w^2}{2}\hat{f}(C_{L-1}(x,x'))\sqrt{Q_{L-1}(x,x)Q_{L-1}(x',x')}\,.$$

Since $\hat{f}(x) \geq x$, C_L is non-decreasing wrt L and converges to the unique fixed point of $\hat{f}$ which is 1. This convergence is uniform in x, x', i.e. $\lim_{L\to\infty} \sup_{x,x'\in K} 1 - C_L(x,x') = 0$.
Re-writing the recursion yields

$$C_L(x,x') = \delta_L \frac{1}{1+\alpha} C_{L-1}(x,x') + \zeta_L + \delta_L \frac{\alpha}{1+\alpha}\hat{f}(C_{L-1}(x,x'))\,,$$

where $\alpha = \frac{\sigma_w^2}{2}$, $\delta_l = \left(1 + \frac{\sigma_b^2}{(1+\alpha)Q_{L-1}(x,x)}\right)^{-1/2}\left(1 + \frac{\sigma_b^2}{(1+\alpha)Q_{L-1}(x,x)}\right)^{-1/2}$ and $\zeta_L = \sigma_b^2(Q_L(x,x)Q_L(x',x'))^{-1/2}$.
Using Lemma A3, and the boundedness of C_L, a simple Taylor expansion yields

$$\begin{aligned} C_L(x,x') &= \frac{1}{1+\alpha} C_{L-1}(x,x') + \frac{\alpha}{1+\alpha}\hat{f}(C_{L-1}(x,x')) + g_L(x,x') \\ &= C_{L-1}(x,x') + \frac{\alpha}{1+\alpha} f(C_{L-1}(x,x')) + g_L(x,x')\,, \end{aligned}$$

where the expansion is uniform on $x, x' \in K$, and $f(x) = \hat{f}(x) - x$, and $g_L = \mathcal{O}(e^{-\beta L})$ for some $\beta > 0$.
The previous dynamical system can be decomposed in two parts, a first part without the term $\mathcal{O}(e^{-\beta L})$ which is the homogeneous system, i.e. the system without bias, and the term $\mathcal{O}(e^{-\beta L})$ which is the contribution of the bias in the dynamical system.
Assume $\sigma_b = 0$, then the term g_L vanishes. Moreover, a Taylor expansion of $\hat{f}$ near 1 yields

$$f(x) = s(1-x)^{3/2} + \mathcal{O}((1-x)^{5/2})\,.$$

Therefore, uniformly in $x, x' \in K$, we have that

$$C_L(x,x') = C_{L-1}(x,x') + \frac{s\alpha}{1+\alpha}(1 - C_{L-1}(x,x'))^{3/2} + \mathcal{O}((1 - C_{L-1}(x,x'))^{5/2})\,.$$

Letting $\gamma_L = 1 - C_L$, a simple Taylor expansion leads to

$$\gamma_L^{-1/2} = \gamma_{L-1}^{-1/2} + \frac{s\alpha}{2(1+\alpha)} + \mathcal{O}(\gamma_{L-1})\,.$$

Therefore, $\gamma_L \sim \kappa L^{-2}$ where $\kappa = \frac{4(1+\alpha)^2}{s^2\alpha^2}$. This equivalence is uniform in $x, x' \in K$.

It is likely that the rate $\mathcal{O}(L^{-2})$ holds without assuming $\sigma_b = 0$. However, the analysis in this requires unnecessarily complicated details. □

A2 Stable ResNet with uniform scaling

In this section we detail the proofs for the uniform scaling of a Scaled ResNet, that is $\lambda_{l,L} = 1/\sqrt{L}$.
When not otherwise specified, K is a generic compact of $\mathbb{R}^d$. We assume that $0 \notin K$ if $\sigma_b = 0$.

A2.1 Continuous formulation

We provide the results of existence, uniqueness and regularity of the solution of (8) in Lemma A11. Corollary A1 shows that the differential problem can be restated in the operator space. Eventually we give a proof of Lemma 5, assuring uniform convergence to the continuous limit.

We recall that by continuous formulation we mean a rescaling of the layer index l, which becomes a continuous index t, spanning the interval $[0,1]$, as the depth diverges, that is $L \to \infty$.
More precisely, for all $L \geq 1$ and all $l \in [0:L]$, we can define $t(l,L) = l/L$.
Consider a sequence $\{l_n, L_n\}_{n\in\mathbb{N}}$ (where, for all n, $L_n \geq 1$ and $l \in [0:L_n]$), such that L_n diverges but l_n/L_n converges to a finite $t = \lim_{n\to\infty} t(l_n, L_n)$. We will show in this section (Lemma 5) that the kernels $Q_{l_n|L_n}$ (covariance kernel of the layer l_n in a net with $L_n + 1$ layers) converge uniformly to a kernel, q_t, on K.
Moreover we can define a differential problem for the mapping $t \mapsto q_t$, with $q \in [0,1]$, that is

$$\begin{aligned} \dot{q}_t(x,x') &= {\sigma_b}^2 + \tfrac{{\sigma_w}^2}{2}\left(1 + \tfrac{f(c_t(x,x'))}{c_t(x,x')}\right) q_t(x,x')\,, \\ q_0(x,x') &= \sigma_b^2 + \sigma_w^2\,\tfrac{x\cdot x'}{d}\,, \\ c_t(x,x') &= \tfrac{q_t(x,x')}{\sqrt{q_t(x,x)q_t(x',x')}}\,. \end{aligned} \tag{8}$$

Lemma A11 (Existence and uniqueness)**.** *For any x, x' in K, the solution of (8) is unique and well defined for all $t \in [0,1]$. The maps $(x,x') \mapsto q_t(x,x')$ and $(x,x') \mapsto c_t(x,x')$ are Lipschitz continuous on K^2 and c_t takes values in $[-1,1]$. Moreover, both q_t and c_t are kernels in the sense of Definition 2.*

Proof. First notice that from (8) we can find, with few algebraic manipulations, an explicit recurrence relation for the correlation C_l, defined in (3). For any $x, x' \in K$ we have

$$\begin{aligned} C_{l+1}(x,x') &= A_{l+1}(x,x')\,C_l(x,x') + \tfrac{\sigma_w^2}{2L}\left(1 + \tfrac{\sigma_w^2}{2L}\right)^{-1} A_{l+1}(x,x')\,f(c_l(x,x')) + \tfrac{1}{L}\tfrac{\sigma_b^2}{\sqrt{Q_l(x,x)Q_l(x',x')}}\,; \\ A_l(x,x') &= \sqrt{\left(1 - \tfrac{1}{L}\tfrac{\sigma_b^2}{Q_l(x,x)}\right)\left(1 - \tfrac{1}{L}\tfrac{\sigma_b^2}{Q_l(x',x')}\right)}\,. \end{aligned} \tag{A5}$$

We can find a Cauchy problem for the correlation directly from (8) or by noting that $A_l(x,x') = 1 - \frac{\sigma_b^2}{2L}\left(\frac{1}{Q_l(x,x)} + \frac{1}{Q_l(x',x')}\right) + o(1/L)$, for $L \to \infty$. With both approaches, we have

$$\begin{aligned} \dot{c}_t(x,x') &= {\sigma_b}^2\left(\mathcal{G}_t(x,x') - \mathcal{A}_t(x,x')\,c_t(x,x')\right) + \frac{\sigma_w^2}{2} f(c_t(x,x'))\,, \\ c_0(x,x') &= \frac{\sigma_b^2 + \sigma_w^2\,x\cdot x'}{\sqrt{(\sigma_b^2 + \sigma_w^2\,\|x\|^2)(\sigma_b^2 + \sigma_w^2\,\|x'\|^2)}}\,, \end{aligned} \tag{A6}$$

where f is defined in (4) and

$$\mathcal{A}_t(x,x') = \frac{1}{2}\left(\frac{1}{q_t(x,x)} + \frac{1}{q_t(x',x')}\right)\,; \qquad \mathcal{G}_t(x,x') = \sqrt{\frac{1}{q_t(x,x)\,q_t(x',x')}}\,.$$

Note that for the diagonal terms $q_t(x,x)$, (8) reduces to $\dot{q}_t = {\sigma_b}^2 + \frac{{\sigma_w}^2}{2} q_t$, whose solution is

$$q_t(x,x) = e^{\frac{{\sigma_w}^2}{2}t}\,q_0(x,x) + \frac{2{\sigma_b}^2}{{\sigma_w}^2}\left(e^{\frac{{\sigma_w}^2}{2}t} - 1\right) = e^{\frac{{\sigma_w}^2}{2}t}\,(\sigma_b^2 + \sigma_w^2\,\|x\|^2) + \frac{2{\sigma_b}^2}{{\sigma_w}^2}\left(e^{\frac{{\sigma_w}^2}{2}t} - 1\right)\,.$$

Now, fix $z = (x,x') \in K^2$ and let $\gamma_0 = c_0(z) \in [-1,1]$. Consider $\bar{f} : \mathbb{R} \to \mathbb{R}$, an arbitrary Lipschitz extension of f to the whole $\mathbb{R}$ and define $H : [0,\infty) \times \mathbb{R} \to \mathbb{R}$ as

$$H(t,\gamma) = \sigma_b^2(\mathcal{G}_t(z) - \mathcal{A}_t(z)\,\gamma) + \frac{\sigma_w^2}{2}\,\bar{f}(\gamma)\,.$$

H is Lipschitz continuous in γ and C^∞ in t, so there exists $\tau > 0$ such that the Cauchy problem

$$\dot{\gamma}(t) = H(t, \gamma(t))\,;$$
$$\gamma(0) = \gamma_0$$

has a unique C^1 solution defined for $t \in [0, \tau)$.
Noticing that

$$\mathcal{G}_t(x, x') - \mathcal{A}_t(x, x') = -\frac{1}{2}\left(\frac{1}{q_t(x,x)} - \frac{1}{q_t(x',x')}\right)^2 \leq 0\,,$$

we get that for all t_1 such that $\gamma(t_1) = 1$ we have $\dot{\gamma}(t_1) \leq 0$, since $f(1) = 0$, and for all t_{-1} such that $\gamma(t_{-1}) = -1$ we have $\dot{\gamma}(t_{-1}) = \sigma_b^2(\mathcal{G}_t(x, x') + \mathcal{A}_t(x, x')) + \frac{\sigma_w^2}{2} > 0$. As a consequence $\gamma(t) \in [-1, 1]$ for all $t \in [0, \tau)$ and we can take $\tau = \infty$.
In particular we get that (A6) has a unique solution $t \mapsto c_t(z)$, defined for $t \in [0, 1]$ and bounded in $[-1, 1]$.
As a consequence, (8) has a unique and well defined solution for all $t \geq 0$.

Now notice that $z \mapsto c_0(z)$ is Lipschitz on K^2. let us denote as L_0 a Lipschitz constant for c_0.
Since both $\mathcal{G}_t$ and $\mathcal{A}_t$ are C^1, we can find real constants L_G, L_A and M_A such that for all z, z' elements of K^2

$$|\mathcal{G}_t(z) - \mathcal{G}_t(z')| \leq L_G \|z - z'\|\,;$$
$$|\mathcal{A}_t(z) - \mathcal{A}_t(z')| \leq L_A \|z - z'\|\,;$$
$$|\mathcal{A}_t(z)| \leq M_A\,.$$

Let L_f be a Lipschitz constant for f. Using the fact that $|c_t| \leq 1$, we can write

$$|\dot{c}_t(z) - \dot{c}_t(z')| \leq L_1 \|z - z'\| + L_2 |c_t(z) - c_t(z')|\,,$$

where $L_1 = \sigma_b^2(L_G + L_A)$ and $L_2 = \sigma_b^2 M_A + \frac{\sigma_w^2}{2} L_f$.
Now fix z and z' and consider $\Delta(t) = c_t(z) - c_t(z')$. We have

$$|\dot{\Delta}(t)| \leq L_1 \|z - z'\| + L_2 |\Delta(t)|\,;$$
$$|\Delta(0)| \leq L_0 \|z - z'\|\,.$$

So $|\Delta(t)| \leq \left(\frac{L_1}{L_2}\left(e^{L_2 t} - 1\right) + L_0\, e^{L_2 t}\right)\|z - z'\|$, meaning that c_t (and so q_t) is Lipschitz on L^2.

Since the mapping $(x, x') \mapsto q_t(x, x')$ is continuous, it defines a compact integral operator $T(q_t)$ on $L^2(K)$ [Lang, 2012]. Since q_t is real and symmetric under the swap of x and x', the operator is self-adjoint. The same holds true for c_t.

The fact that $T(q_t)$ is a non-negative operator can be seen as a corollary of Lemma 5. Indeed all $T(Q_{l|L})$ is a non-negative definite operator, since it is induced by a kernel. Hence, for each $t \in [0, 1]$ it is enough to find a sequence $\{l_n, L_n\}_{n\in\mathbb{N}}$ (where $L_n \geq 1$ is an integer and $l_n \in [0 : L_n]$) such that $L_n \to \infty$ and $l_n/L_n \to t$. By Lemma 5, $T(Q_{l_n|L_n}) \to T(q_t)$ in the L^∞ norm, and hence in L^2, as we are on a compact set. By Lemma A10, for all $n \in \mathbb{N}$ we have that $T(Q_{l_n|L_n})$ is non-negative definite. Since the subspace of non-negative definite operators in L^2 is closed wrt the L^2 operator norm, we conclude.
Once we have established that $T(q_t)$ is non-negative definite, it follows immediately that $T(c_t)$ is non-negative as well. Since these results hold for any arbitrary finite Borel measure μ on K, we can thus conclude by Lemma A1 that both q_t and c_t are kernels, in the sense of Definition 2. □

Corollary A1. *The maps $t \mapsto T(q_t)$ and $t \mapsto T(c_t)$, defined on $[0, 1]$, are continuous and twice differentiable with respect to the operator norm in $L^2(K)$. Moreover, $\frac{\mathrm{d}}{\mathrm{d}t}T(q_t) = T(\dot{q}_t)$, $\frac{\mathrm{d}}{\mathrm{d}t}T(c_t) = T(\dot{c}_t)$, $\frac{\mathrm{d}^2}{\mathrm{d}t^2}T(q_t) = T(\ddot{q}_t)$ and $\frac{\mathrm{d}^2}{\mathrm{d}t^2}T(c_t) = T(\ddot{c}_t)$.*

Proof. Consider the map $(t,z) \mapsto q_t(z)$, defined on $[0,1] \times K^2$, which is continuous wrt z and C^2 wrt t, as it can be easily checked. Since K^2 and $[0,1]$ are compact sets, it follows that for any t

$$\lim_{s\to t} \sup_{z\in I^2} \left| \frac{q_s(z) - q_t(z)}{s-t} - \dot{q}_t(z) \right| = \sup_{z\in I^2} \lim_{s\to t} \left| \frac{q_s(z) - q_t(z)}{s-t} - \dot{q}_t(z) \right| = 0\,.$$

Hence $\lim_{s\to t} \frac{q_s - q_t}{t-s} = \dot{q}_t$ uniformly on K^2, and hence $\lim_{s\to t} \frac{T(q_s)-T(q_t)}{t-s} = T(\dot{q}_t)$ in the $L^2(K,\mu)$ norm for operators, since K is compact.
The proof for the second derivative works in the same way, using the fact that $(t,z) \mapsto q_t(z)$ is continuous in z and C^1 in t.
As a consequence of the above results, $t \mapsto T(q_t)$ is continuous and twice differentiable, with $\frac{\mathrm{d}}{\mathrm{d}t} T(q_t) = T(\dot{q}_t)$ and $\frac{\mathrm{d}^2}{\mathrm{d}t^2} T(q_t) = T(\ddot{q}_t)$.
The proof for $T(c_t)$ is analogous. □

Lemma 5 (Convergence to the continuous limit)**.** *Let $Q_{l|L}$ be the covariance kernel of the layer l in a net of $L+1$ layers $[0:L]$, and q_t be the solution of (8), then*

$$\lim_{L\to\infty} \sup_{l\in[0:L]} \sup_{(x,x')\in K^2} |Q_{l|L}(x,x') - q_{t=l/L}(x,x')| = 0\,.$$

Proof. We will show that the relation holds for c_t, and hence for q_t.
Let H, defined on $[0,1] \times K^2$, be such that $\dot{c}_t(z) = H(z,t,c_t(z))$. Explicitly, with the same notations as in (A6), we have

$$H(z,t,\gamma) = \sigma_b^2(\mathcal{G}_t(z) - \mathcal{A}_t(z)\,\gamma) + \frac{\sigma_w^2}{2} f(\gamma)\,.$$

Define

$$\tau(h) = \sup_{t,z} \left| \frac{c_{t+h}(z) - c_t(z)}{h} - H(z,t,c_t(z)) \right|.$$

Since t and z takes values on compact sets, by uniform continuity, fixed h we can write, for $h \to 0$

$$\sup_t \sup_{s\in[t,t+h]} |H(z,s,c_s(z)) - H(z,t,c_t(z))| = o(h)\,.$$

Hence, since τ can be rewritten as $\tau(h) = \frac{1}{h} \sup_{t,z} \left| \int_t^{t+h} (H(z,s,c_s(z)) - H(z,t,c_t(z))\,\mathrm{d}s \right|$, it is clear that $\tau(h) \to 0$ for $h \to 0$.
Now, for any integer $L \geq 1$, let $\tilde{H}_L : K^2 \times [0:L-1] \times [-1,1]$ be given by

$$\tilde{H}_L(z,l,\gamma) = (A_{l+1|L}(x,x') - 1)\,L\,\gamma + \frac{\sigma_w^2}{2}\left(1 + \frac{\sigma_w^2}{2L}\right)^{-1} A_{l+1|L}(x,x')\,f(c_l(x,x')) + \frac{\sigma_b^2}{\sqrt{Q_{l|L}(x,x)Q_{l|L}(x',x')}}\,,$$

where,

$$A_{l|L}(x,x') = \sqrt{\left(1 - \tfrac{1}{L}\frac{\sigma_b^2}{Q_{l|L}(x,x)}\right)\left(1 - \tfrac{1}{L}\frac{\sigma_b^2}{Q_{l|L}(x',x')}\right)}\,.$$

It is clear from (A5) that $\tilde{H}_L$ has been defined so that $C_{l+1|L}(z) - C_{l|L}(z) = \frac{1}{L}\tilde{H}_L(z,l,\gamma)$, for all $L \in [0:L-1]$ and all $z \in K^2$. Using the explicit form of the diagonal terms of Q and q, it can be easily shown that, for $L \to \infty$,

$$\sup_{(x,x')\in K^2} \sup_{l\in[0:L-1]} A_{l+1|L}(x,x') = 1 + \tfrac{\sigma_b^2}{L}\mathcal{A}_{t=l/L}(x,x') + O(1/L^2)\,;$$

$$\sup_{(x,x')\in K^2} \sup_{l\in[0:L]} \frac{\sigma_b^2}{\sqrt{Q_{l|L}(x,x)Q_{l|L}(x',x')}} = \mathcal{G}_{t=l/L}(x,x') + O(1/L^2)\,,$$

where $\mathcal{A}_t$ and $\mathcal{G}_t$ are defined as in (A6). As a consequence, we can find a constant $M_1 > 0$ and an integer $L_\star > 0$ such that, for all $\gamma \in [-1,1]$, for all $z \in K^2$, for all $L \geq L^\star$

$$|\tilde{H}_L(z,l,\gamma) - H(z,l/L,\gamma)| \leq \frac{M_1}{L}\,. \tag{A7}$$

Moreover, there exists a constant $M_2 > 0$ such that for all $z \in K^2$, all $t \in [0,1]$ and all pairs $(\gamma, \gamma') \in [-1,1]^2$

$$|H(z,t,\gamma) - H(z,t,\gamma')| \leq M_2 \|\gamma - \gamma'\| \,. \tag{A8}$$

Thanks to the two above uniform inequalities, we will now show that, for $L \geq L_\star$,

$$\sup_{l \in [0:L]} \sup_{z \in K^2} |C_{l|L}(x,x') - c_{t(l,L)}(x,x')| \leq \tilde{\tau}(1/L) \frac{e^{M_2} - 1}{M_2} \,, \tag{A9}$$

where $\tilde{\tau} : h \mapsto \tau(h) + M_1 h$.
To do so, fix $L \geq L_\star$ and define $\Delta_l = \sup_{z \in K^2} |C_{l|L}(x,x') - c_{t(l,L)}(x,x')|$. Using the definition of τ, (A7) and (A8) we get

$$|\Delta_{l+1}| \leq \left(1 + \tfrac{M_2}{L}\right) |\Delta_l| + \tfrac{1}{L} \tau(1/L) + \tfrac{M_1}{L} = \left(1 + \tfrac{M_2}{L}\right) |\Delta_l| + \tfrac{1}{L} \tilde{\tau}(1/L) \,.$$

At this point, using the fact that $\Delta_0 = 0$, it is easy to show by induction that

$$\Delta_l \leq \tilde{\tau}(1/L) \, \frac{\left(1 + \frac{M_2}{L}\right)^l - 1}{M_2} \,,$$

and so (A9) follows.
Finally, the uniform convergence of C to c implies the one of Q to q and so we conclude. □

A2.2 Universality of the covariance kernel

We will now prove the results of universality of Theorem 1 and Proposition 6.

Proof of Theorem 1

The idea is to prove that for any finite Borel measure μ on K, the operator $T_\mu(q_t)$ is strictly positive definite if $t > 0$, and then use the characterization of universal kernels given in Lemma A2.
To prove the strict positive definiteness, we will proceed in two steps. First we show in Proposition A2 that for all non-zero $\varphi \in L^2(K,\mu)$, $\langle T_\mu(q_t)\, \varphi, \varphi \rangle > 0$ for t small enough. Then we use Proposition A3, which shows that $\frac{\mathrm{d}}{\mathrm{d}t} T_\mu(q_t)$ is non-negative definite.

Proposition A2. *Fix any finite Borel measure μ on K, and assume that $\sigma_b > 0$. Given any non-zero $\varphi \in L^2(K,\mu)$, there exists a $t_\varphi \in (0,1]$ such that $\langle T_\mu(q_t)\, \varphi, \varphi \rangle > 0$, for all $t \in (0, t_\varphi)$.*

Proof. From Corollary A1, we can expand $T_\mu(q_t)$ around $t = 0$ as

$$T_\mu(q_t) = T_\mu(q_0) + t\, T_\mu(\dot{q}_0) + o(t) = t\, T_\mu\left(\sigma_b^2 + \frac{\sigma_w^2}{2} q_0\right) + T_\mu((c_0 + t f(c_0)) R_0) + o(t) \,,$$

the $o(t)$ being wrt the operator norm, where we have defined the kernel R_0 via $R_0(x,x') = \frac{\sigma_w^2}{2} \sqrt{(1 + \zeta \|x\|^2)(1 + \zeta \|x'\|^2)}$.
Since $T_\mu(q_0)$ is non-negative, for any $\varphi \in L^2(I)$, we have

$$\langle T_\mu(q_t)\, \varphi, \varphi \rangle \geq \langle T_\mu((c_0 + t f(c_0)) R_0)\, \varphi, \varphi \rangle + o(t) = \left(1 - \frac{t}{2}\right) \langle T_\mu(c_0)\, \psi, \psi \rangle + t \, \langle T_\mu(f(c_0)))\, \psi, \psi \rangle + o(t) \,,$$

where $\psi(x) = \sigma_w \sqrt{(1 + \zeta \|x\|^2)/2}\, \varphi(x)$. We conclude by the strict positivity of $\tilde{f}(c_0)$ on $L^2(K,\mu)$, thanks to Proposition A1 and Lemma A2. □

Proposition A3. *For any finite Borel measure μ on K, for any $t \in [0,1]$, the operator $T_\mu(\dot{q}_t)$ on $L^2(K,\mu)$ is non-negative definite. In particular, for all $\varphi \in L^2(K,\mu)$ we have*

$$\tfrac{\mathrm{d}}{\mathrm{d}t} \langle T_\mu(q_t)\, \varphi, \varphi \rangle \geq 0 \,.$$

Proof. Fix μ and $\varphi \in L^2(K,\mu)$. From (8) we can write

$$T_\mu(\dot{q}_t) = T_\mu\left(\sigma_b^2 + \frac{\sigma_w^2}{2}\, q_t + \frac{\sigma_w^2}{2}\,\frac{f(c_t)}{c_t} q_t\right).$$

By Lemma A11, $T_\mu(q_t)$ is non-negative definite, so we can write

$$\begin{aligned}\langle T_\mu(\dot{q}_t)\,\varphi,\varphi\rangle &= \sigma_b^2|\langle 1,\varphi\rangle|^2 + \frac{\sigma_w^2}{2}\left\langle T_\mu\left(\frac{c_t+f(c_t)}{c_t}\, q_t\right)\varphi,\varphi\right\rangle \\ &\geq \frac{\sigma_w^2}{2}\left\langle T_\mu\left(\tilde{f}(c_t)\,\frac{q_t}{c_t}\right)\varphi,\varphi\right\rangle \\ &= \frac{\sigma_b^2}{2}\langle T_\mu(\tilde{f}(c_t))\,\psi,\psi\rangle\,,\end{aligned}$$

where $\tilde{f}:\gamma\mapsto \frac{\gamma}{2}+f(\gamma)$, for $\gamma\in[-1,1]$, and $\psi(x)=\sqrt{q_t(x,x)}\,\varphi(x)$. By Lemma A8, the Taylor expansion of $\tilde{f}$ around 0 converges uniformly on $[-1,1]$, and all its coefficients are non-negative. We conclude by Lemma A9 that $T_\mu(\dot{q}_t)$ is non-negative definite.
Finally, to prove the inequality, it is enough to recall that $\frac{\mathrm{d}}{\mathrm{d}t}T_\mu(q_t) = T_\mu(\dot{q}_t)$ by Corollary A1, the derivative $\frac{\mathrm{d}}{\mathrm{d}t}$ being wrt the operator norm on $L^2(K,\mu)$. □

Theorem 1 (Universality of q_t)**.** *Let $K\subset\mathbb{R}^d$ be compact and assume $\sigma_b>0$. For any $t\in(0,1]$, the solution q_t of (8) is a universal kernel on K.*

Proof. By Lemma A2, it suffices to show that for any finite Borel measure μ on K, $T_\mu(q_t)$ is strictly positive definite for all $t\in(0,1]$. Fix any nonzero $\varphi\in L^2(K,\mu)$, define the map F on $[0,1]$ by $F(t)=\langle T_\mu(q_t)\,\varphi,\varphi\rangle$. For any fixed $t\in(0,1]$, by Proposition A2 we can find $s\in(0,t)$ such that $F(s)>0$. Since F is non decreasing by Proposition A3, we get that $F_t>0$. Hence $T_\mu(q_t)$ is strictly positive definite. □

Proof of Proposition 6

The proof of Proposition 6 is quite similar to the one of Theorem 1.
Using Lemma A15 instead of Lemma A2, we will not need to consider a generic finite Borel measure μ on $\mathbb{S}^{d-1}$, but it will be enough to show that $T_\nu(q_t)$ is a stricly positive operator on $L^2(\mathbb{S}^{d-1},\nu)$, where ν is the standard unifrom spherical measure on $\mathbb{S}^{d-1}$.
Since $\sigma_b=0$, we will not be able to use Proposition A1. We will hence state some preliminary results.

Lemma A12. *Let $\{A_n\}_{n\in\mathbb{N}}$ be a family of compact non-negative operators on a separable Hilbert space $\mathcal{H}$. Let R_n be the range of A_n and assume that $V=\mathrm{Span}(\bigcup_{n\in\mathbb{N}}R_n)$ is dense in $\mathcal{H}$. Let $\{\alpha_n\}_{n\in\mathbb{N}}$ be a strictly positive sequence such that the sum*

$$A=\sum_{n\in\mathbb{N}}\alpha_n\,A_n$$

converges in the operator norm. Then A is a compact strictly positive definite operator.

Proof. A is the convergent limit of a sum of compact self-adjoint operators and hence it is compact and self-adjoint. Now, fix an arbitrary nonzero $h\in\mathcal{H}$. To show that A is strictly positive it is enough to prove that $\langle A\,h,h\rangle>0$. Denote by V_N the linear span of $\bigcup_{n\in[0:N]}R_n$. Since $V_N\subseteq V_{N+1}$ for all N, and $\bigcup_{N\in\mathbb{N}}V_N=V$ is dense in H, there exists a sequence $\{h_N\}_{N\in\mathbb{N}}$ converging to h and such that $h_N\in V_N$ for all N.
Now let us show that there must exist $n^\star\in\mathbb{N}$ such that $A_{n^\star}\,h\neq 0$. Since $\lim_{N\to\infty}\langle h,h_N\rangle=\langle h,h\rangle>0$, there must be a $N^\star$ such that $\langle h,h_{N^\star}\rangle>0$ and so there exists $n^\star\in[0:N^\star]$ and $h_{n^\star}\in V_{n^\star}$ such that $\langle h,h_{n^\star}\rangle\neq 0$. In particular, h is not orthogonal to $R_{n^\star}$ and can not lie in the nullspace of $A_{n^\star}$, using the fact that $A_{n^\star}$ is compact and self-adjoint and so its range and its nullspace are orthogonal [Lang, 2012].
Using the spectral decomposition of non-negative compact operators, it is straightforward that $A_{n^\star}\,h\neq 0$ implies that $\langle A_{n^\star}\,h,h\rangle>0$. Now, since A_n is non-negative and $\alpha_n>0$ for all n, we have

$$\langle A\,h,h\rangle=\sum_{n\in\mathbb{N}}\alpha_n\langle A_n\,h,h\rangle\geq\alpha_{n^\star}\langle A_{n^\star}\,h,h\rangle>0\,,$$

and so we conclude. □

Lemma A13. *For all $n \in \mathbb{N}$, consider the kernel p_n on $\mathbb{S}^{d-1}$, defined by $p_n(x,x') = (x\cdot x')^n$, and let $T_\nu(p_n)$ be the induced integral operator on $L^2(\mathbb{S}^{d-1},\nu)$. Denoting as R_n the range of $T_\nu(p_n)$, the subspace $V = \mathrm{Span}\left(\bigcup_{n\in\mathbb{N}} R_n\right)$ is dense in $L^2(\mathbb{S}^{d-1},\nu)$.*
Moreover, letting $V' = \mathrm{Span}\left(\bigcup_{n\in\mathbb{N}} R_{2n}\right)$ and $V'' = \mathrm{Span}\left(\bigcup_{n\in\mathbb{N}} R_{2n+1}\right)$, we have $L^2(\mathbb{S}^{d-1},\nu) = \overline{V'} \oplus \overline{V''}$, the overline denoting the closure in $L^2(\mathbb{S}^{d-1},\nu)$.

Proof. To prove that V is dense, first notice that for each spherical harmonic Y, we can find an operator in the form $T_\nu(P(x\cdot x'))$, for a polynomial P, which has Y in its range. Since the range of such an operator is trivially contained in V, it follows that V contains all the spherical harmonics, and so it is dense in $L^2(\mathbb{S}^{d-1},\nu)$.
Now, note that for any even n and odd n' we have

$$\int_{\mathbb{S}^{d-1}} (x\cdot z)^n (z\cdot x')^{n'} \mathrm{d}\nu(z) = 0\,,$$

by an elementary symmetry argument, since it is the integral on the sphere of a homogeneous polynomial of odd degree $n+n'$ in the components z_i's of z.
It follows that V' and V'' are orthogonal. Since their union V is dense, we conclude that $L^2(\mathbb{S}^{d-1},\nu) = \overline{V'}\oplus\overline{V''}$. □

Corollary A2. *With the notations of Lemma A13, assume that a sequence $\{\alpha_{n\in\mathbb{N}}\}$ is such that $A = \sum_{n\in\mathbb{N}} \alpha_n T_\nu(p_n)$ converges wrt the operator norm on $L^2(\mathbb{S}^{d-1},\nu)$. Then $A = A' + A''$, where $A' : \overline{V'} \to \overline{V'}$ and $A'' : \overline{V''} \to \overline{V''}$. Such a decomposition is unique and*

$$A' = \sum_{n\in\mathbb{N}} \alpha_{2n}\, T_\nu(p_{2n})\,; \qquad\qquad A'' = \sum_{n\in\mathbb{N}} \alpha_{2n+1}\, T_\nu(p_{2n+1})\,,$$

both sums converging wrt the operator norm.

Proof. It is clear that $A = A' + A''$, when both A' and A'' are defind on the whole $L^2(\mathbb{S}^{d-1},\nu)$.
Consider any $\varphi \in L^2(\mathbb{S}^{d-1},\nu)$. We have $A'\varphi \in \overline{V'}$, since $T_\nu(p_{2n})\varphi \in \overline{V'}$ for all n. Analogously, we can show that $A''\varphi \in \overline{V''}$. To conclude that we can consider the restrictions of A' and A'' to $\overline{V'}$ and $\overline{V''}$ respectively, it is enough to recall that for compact self adjoint operators the nullspace is the orthogonal of the closure of the range [Lang, 2012], so that the nullspace of A' contains $\overline{V''}$ and the nullspace of A'' contains $\overline{V'}$. □

Lemma A14. *The function $f : [-1,1] \to \mathbb{R}$, defined in (4), is an analytic function on $(-1,1)$, whose expansion $f(\gamma) = \sum_{n\in\mathbb{N}} \alpha_n \gamma^n$ converges absolutely on $[-1,1]$. Moreover, $\alpha_n > 0$ for all even $n \in \mathbb{N}$, $\alpha_1 = -1/2$ and $\alpha_n = 0$ for all odd $n \geq 3$.*
Let $g : [-1,1] \to \mathbb{R}$ be defined as $g(\gamma) = f(\gamma)f'(\gamma)$. g is analytic on $(-1,1)$ and its expansion $g(\gamma) = \sum_{n\in\mathbb{N}} \beta_n \gamma^n$ converges absolutely on $[-1,1]$. Moreover, for all odd $n \in \mathbb{N}$ the coefficient β_n is strictly positive.

Proof. The claims for f have been already proven in Lemma A8. As for g, the analyticity of f implies the one of f', and it is easy to check the convergence on $[-1,1]$. Moreover, all the odd Taylor coefficients of f' are striclty positive, as the even coefficients of f are. It follows that $\beta_n > 0$ for all odd n. □

Proposition A4. *Given any non-zero $\varphi \in L^2(\mathbb{S}^{d-1},\nu)$, there exists a $t_\varphi \in (0,1]$ such that $\langle T_\nu(q_t)\,\varphi, \varphi\rangle > 0$, for all $t \in (0, t_\varphi)$.*

Proof. The case $\sigma_b > 0$ has been already established in Proposition A2, hence suppose that $\sigma_b = 0$.
First recall (A6)

$$\dot{c}_t = \frac{\sigma_w^2}{2} f(c_t)\,. \tag{A10}$$

Deriving once more we have

$$\ddot{c}_t = g(c_t)\,, \tag{A11}$$

where $g = ff'$ as in Lemma A14.
Define the kernels p_n's, and the subspaces V' and V'' of $L^2(\mathbb{S}^{d-1}, \nu)$, as in Lemma A13. By (A10) and (A11) we can write

$$c_t = c_0 + t\,\dot{c}_0 + \tfrac{t^2}{2}\,\ddot{c}_0 + o(t^2) = c_0 + t\,f(c_0) + \tfrac{t^2}{2}\,g(c_0) + o(t^2)\,.$$

Since $\sigma_b = 0$, we have that $c_0(x, x') = x \cdot x'$, so that $c_0 = p_1$.
From Lemma A14, $T_\nu(\dot{c}_0) = \sum_{n\in\mathbb{N}} \alpha_n\, T_\nu(p_n)$ and $T_\nu(\ddot{c}_0) = \sum_{n\in\mathbb{N}} \beta_n\, T_\nu(p_n)$, both sums converging in the operator norm. Moreover, $\alpha_n > 0$ for all even n and $\alpha_n = 0$ for all odd $n \geq 3$, whilst $\beta_n > 0$ for all odd n.
In particular, by Corollary A2 and Lemma A12, we deduce that the restriction of $T_\nu(\dot{c}_0)|_{\overline{V'}} : \overline{V'} \to \overline{V'}$ is well defined and strictly positive, and the same holds true for the restriction $T_\nu(\ddot{c}_0)|_{\overline{V''}} : \overline{V''} \to \overline{V''}$.
Now fix a non-zero $\varphi \in L^2(\mathbb{S}^{d-1}, \nu)$. By Lemma A13, we can write $\varphi = \varphi' + \varphi''$, with $\varphi' \in \overline{V'}$, $\varphi'' \in \overline{V''}$ uniquely determined.
First, suppose that $\varphi' \neq 0$. Using Corollary A1 and recalling that $c_0 = p_1$, we get

$$\langle T_\nu(c_t)\,\varphi, \varphi\rangle = t\langle T_\nu(\dot{c}_0)|_{\overline{V'}}\,\varphi', \varphi'\rangle + \langle (1 + t\,\alpha_1) T_\nu(p_1)\,\varphi'', \varphi''\rangle + o(t) > 0\,,$$

for t small enough.
On the other hand, for $\varphi' = 0$, we have $\varphi = \varphi''$ and so

$$\langle T_\nu(c_t)\,\varphi, \varphi\rangle = \langle (1 + t\,\alpha_1) T_\nu(p_1)\,\varphi'', \varphi''\rangle + \tfrac{t^2}{2}\,\langle T_\nu(\ddot{c}_0)|_{\overline{V''}}\,\varphi'', \varphi''\rangle + o(t^2) > 0$$

for t small enough.
So there is a t_φ such that, for $t \in (0, t_\varphi)$, $\langle T_\nu(c_t)\,\varphi, \varphi\rangle > 0$. It follows immediately that the same property is true for $T_\nu(q_t)$. □

Lemma A15. *Let Q be a kernel on $\mathbb{S}^{d-1}$. Then Q is universal on $\mathbb{S}^{d-1}$ if and only if $T_\nu(Q)$ is strictly positive definite on $L^2(\mathbb{S}^{d-1}, \nu)$.*

Proof. If Q is universal, $T_\nu(Q)$ is strictly positive definite by Lemma A2. On the other hand, if $T_\nu(Q)$ is strictly positive definite, by Proposition 5 its range contains all the spherical harmonics. Since the RKHS generated by Q contains the range of $T_\nu(Q)$ (Proposition 11.17 in [Paulsen and Raghupathi, 2016]), it contains the linear span of the spherical harmonics, which is dense in $C(\mathbb{S}^{d-1})$ [Kounchev, 2001]. Hence Q is universal. □

Proposition 6 (Universality on $\mathbb{S}^{d-1}$)**.** *For any $t \in (0, 1]$, the covariance kernel q_t, solution of (8) with $\sigma_b = 0$, is universal on $\mathbb{S}^{d-1}$, with $d \geq 2$.*

Proof. Proceeding as in the proof of Theorem 1, using Proposition A3 and Proposition A4 we can show that $T_\nu(q_t)$ is strictly posititive definite on $L^2(\mathbb{S}^{d-1}, \nu)$ for all $t \in (0, 1]$. We conclude by Lemma A15 that q_t is universal on $\mathbb{S}^{d-1}$. □

A3 Stable ResNet with decreasing scaling

A3.1 Proof of Proposition 7

Proposition 7 (Uniform Convergence of the Kernel)**.** *Consider a Stable ResNet with a decreasing scaling, i.e. the sequence $\{\lambda_l\}_{l\geq 1}$ is such that $\sum_l \lambda_l^2 < \infty$. Then for all $(\sigma_b, \sigma_w) \in \mathbb{R}^+ \times (\mathbb{R}^+)^*$, there exists a kernel Q_∞ on $\mathbb{R}^d$ such that for any compact set $K \subset \mathbb{R}^d$,*

$$\sup_{x,x'\in K} |Q_L(x, x') - Q_\infty(x, x')| = \Theta\big(\textstyle\sum_{k\geq L} \lambda_k^2\big)\,.$$

Proof. Let $x, x' \in \mathbb{R}^d$. The kernel Q_l is given recursively by the formula

$$Q_l(x, x') = Q_{l-1}(x, x') + \lambda_l^2\sigma_b^2 + \frac{\sigma_w^2\lambda_l^2}{2}\,\hat{f}(C_{l-1}(x, x'))\sqrt{Q_{l-1}(x, x)}\sqrt{Q_{l-1}(x', x')}\,,$$

where $\hat{f}(t) = 2\mathbb{E}[\phi'(Z_1)\phi'(tZ_1 + \sqrt{1-t^2}Z_2)] = t + f(t)$ and Z_1, Z_2 are iid standard Gaussian variables. In particular, we have

$$Q_l(x,x) = \lambda_l^2\sigma_b^2 + (1 + \frac{\sigma_w^2\lambda_l^2}{2})Q_{l-1}(x,x)\,.$$

which brings

$$Q_l(x,x) + \frac{2\sigma_b^2}{\sigma_w^2} = (1 + \frac{\sigma_w^2\lambda_l^2}{2})(Q_{l-1}(x,x) + \frac{2\sigma_b^2}{\sigma_w^2})\,,$$

Therefore, we can assume without loss of generality that $\sigma_b = 0$. This yields

$$C_l(x,x') = \frac{1}{1 + \frac{\lambda_l\sigma_w^2}{2}}C_{l-1}(x,x') + \frac{\frac{\sigma_w^2\lambda_l^2}{2}}{1 + \frac{\lambda_l\sigma_w^2}{2}}\hat{f}(C^{l-1}(x,x'))\,.$$

Letting $\alpha_l = \frac{\sigma_w^2\lambda_l^2}{2}$ and $C_l := C_l(x,x')$, we have that

$$C_l = \frac{1}{1+\alpha_l}C_{l-1} + \frac{\alpha_l}{1+\alpha_l}\hat{f}(C_{l-1})\,.$$

Since $\hat{f}$ is non decreasing, C^l is non-decreasing and has a limit $C_\infty(x,x') \leq 1$.

Now let us prove that the convergence of C_l to C_∞ happens uniformly with a rate $\sum_{k\geq l}\lambda_l^2$. Using the recursive formula of C_l, and knowing that we have that

$$C_\infty - C_l = \frac{1}{1+\alpha_l}(C_\infty - C_{l-1}) + \frac{\alpha_l}{1+\alpha_l}(C_\infty - f(C_{l-1}))\,.$$

Letting $\delta_l = C_\infty - C_l$, it is easy to see that, uniformly in $x, x' \in \mathbb{R}^d$, we have that

$$\delta_l = \delta_{l-1} + \alpha_l + o(\alpha_l)\,.$$

Therefore, using the fact that $C_l \leq C_\infty$, we have

$$\sup_{(x,x')\in\mathbb{R}^d} |C_l(x,x') - C_\infty(x,x')| = \mathcal{O}\left(\sum_{k\geq l}\alpha_k\right).$$

Moreover, we know that

$$Q_l(x,x) = Q_0(x,x)\prod_{k=1}^{l}(1+\alpha_k)\,,$$

so that for any compact set $K \subset \mathbb{R}^d$

$$\sup_{x\in K}|Q_l(x,x) - Q_\infty(x,x)| \sim \sum_{k\geq l}\alpha_k\,.$$

Moreover, since $C_\infty(x,x') \geq C_l(x,x')$ and $Q_\infty(x,x) \geq Q_l(x,x)$ for all $x \in \mathbb{R}^d$, we can use the fact that

$$\begin{aligned}Q_\infty(x,x') - Q_l(x,x') &= \sqrt{Q_\infty(x,x)Q_\infty(x',x')}(C_\infty(x,x') - C_l(x,x')) \\ &+ C_l(x,x')(\sqrt{Q_\infty(x,x)Q_\infty(x',x')} - \sqrt{Q_l(x,x)Q_l(x',x')})\end{aligned}$$

and hence conclude. □

A3.2 Proof of Corollary 1

Corollary 1. *The following statements hold*
- *Let K be a compact set of $\mathbb{R}^d$ and assume $\sigma_b > 0$. Then, Q_∞ is universal on K.*
- *Assume $\sigma_b = 0$. Then Q_∞ is universal on $\mathbb{S}^{d-1}$.*

Proof. Corollary 1 is a direct result of Propositions 3, 4 and 2. Indeed, for any compact $K \subset \mathbb{R}^d$, $\mathcal{H}_{Q_L}(K) \subset \mathcal{H}_{Q_\infty}(K)$ for all $L \geq 0$. Therefore, the universality of Q_L for some finite L is sufficient to conclude that Q_∞ is universal. □

A4 Neural Tangent Kernel

Throughout this section, we will consider ResNets with NTK parameterization [Jacot et al., 2018]. This simply means that all the components of the biases and the weights will be initialized as iid standard normal random variables. In order to compensate this change of parameterization, the propagation through the network needs to be slightly modified. Hence (2) will be replaced by

$$\begin{aligned} y_0(x) &= \tfrac{\sigma_w}{\sqrt{d}} W_0\, x + \sigma_b\, B_0\,; \\ y_l(x) &= y_{l-1}(x) + \lambda_{l,L}\, \tfrac{\sigma_w}{\sqrt{N_{l-1}}} W_l + \sigma_b\, B_l\,. \end{aligned} \tag{A12}$$

However, it is strightforward to verify that the recurrence (5) for the covariance kernels keeps unchanged. Clearly, the dynamics of a standard ResNet with NTK parameterization can be recovered from (A12) by setting $\lambda_{l+1,L} = 1$ for all l, L.

The Neural Tangent Kernel, introduced by [Jacot et al., 2018], is defined as

$$\tilde{\Theta}_L^{ij}(x, x') = \nabla_{\text{par}}\, y_L^i(x) \cdot \nabla_{\text{par}}\, y_L^j(x')\,,$$

where ∇_{par} denotes the gradient wrt the parameters of the network.
The NTK of a Stable ResNet can be evaluated recursively. We will now prove the recurrence formula (9). The following result was proven in Lemma 3 in [Hayou et al., 2019b] for the case of a standard ResNet without bias. We extend it to ResNet with bias.

Lemma A16 (Recurrence relatino for the NTK)**.** *For a Stable ResNet, the NTK can be evaluated recursively, layer by layer, as*

$$\Theta_0 = Q_0\,; \qquad \Theta_{l+1} = \Theta_l + \lambda_{l+1,L}^2\,(\Psi_l + \Psi_l'\,\Theta_l)\,, \tag{9}$$

where $\Psi_l(x, x') = \sigma_b^2 + \sigma_w^2 \mathbb{E}[\phi(y_1^l(x))\phi(y_1^l(x'))]$ *and* $\Psi_l'(x, x') = \sigma_w^2 \mathbb{E}[\phi'(y_1^l(x))\phi'(y_1^l(x'))]$.

Proof. The first result is the same as in the FFNN case [Jacot et al., 2018], since we assume there is no residual connections between the first layer and the input. Let $x, x' \in \mathbb{R}^d$. We have

$$\Theta_0(x, x') = \sum_{j=0}^{d} \frac{\partial y_0^1(x)}{\partial w_0^{1j}} \frac{\partial y_0^1(x)}{\partial w_0^{1j}} + \frac{\partial y_0^1(x)}{\partial b_0^1} \frac{\partial y_0^1(x)}{\partial b_0^1} = \frac{\sigma_w^2}{d} x \cdot x' + \sigma_b^2\,.$$

We prove the second result by induction. The proof is similar to the one of ResNet in [Hayou et al., 2019b]. Let $\theta_k = (W_k, B_k)$. For $l \geq 1$ and $i \in [1 : N_{l+1}]$

$$\partial_{\theta_{0:l}} y_{l+1}^i(x) = \partial_{\theta_{0:l}} y_l^i(x) + \lambda_{l+1,L} \frac{\sigma_w}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{l+1}^{ij} \phi'(y_l^j(x)) \partial_{\theta_{1:l}} y_l^j(x)\,.$$

Therefore, we obtain

$$\begin{aligned} (\partial_{\theta_{0:l}} y_{l+1}^i(x))(\partial_{\theta_{0:l}} y_{l+1}^i(x'))^t &= (\partial_{\theta_{0:l}} y_l^i(x))(\partial_{\theta_{0:l}} y_l^i(x'))^t \\ &\quad + \lambda_{l+1,L}^2 \frac{\sigma_w^2}{N_l} \sum_{j,j'}^{N_l} W_{l+1}^{ij} W_{l+1}^{ij'} \phi'(y_l^j(x))\phi'(y_l^{j'}(x')) \partial_{\theta_{0:l}} y_l^j(x) (\partial_{\theta_{0:l}} y_l^{j'}(x'))^t + I\,, \end{aligned}$$

where

$$I = \lambda_{l+1,L} \frac{\sigma_w}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{l+1}^{ij} (\phi'(y_l^j(x)) \partial_{\theta_{0:l}} y_l^i(x) (\partial_{\theta_{0:l}} y_l^j(x'))^t + \phi'(y_l^j(x')) \partial_{\theta_{0:l}} y_l^j(x) (\partial_{\theta_{0:l}} y_l^i(x'))^t)\,.$$

We prove the result by induction. Assume the result is true for layers $1, 2, ..., l$ and let us prove it for $l+1$. Using the induction hypothesis, as $N_1, N_2, ..., N_{l-1} \to \infty$ recursively, we have that

$$(\partial_{\theta_{0:l}} y_{l+1}^i(x))(\partial_{\theta_{0:l}} y_{l+1}^i(x'))^t + \lambda_{l+1,L}^2 \frac{\sigma_w^2}{N_l} \sum_{j,j'}^{N_l} W_{l+1}^{ij} W_{l+1}^{ij'} \phi'(y_l^j(x))\phi'(y_l^{j'}(x')) \partial_{\theta_{0:l}} y_l^j(x) (\partial_{\theta_{0:l}} y_l^{j'}(x'))^t + I$$

$$\to \Theta_l(x, x') + \lambda_{l+1,L}^2 \frac{\sigma_w^2}{N_l} \sum_{j}^{N_l} (W_{l+1}^{ij})^2 \phi'(y_l^j(x))\phi'(y_l^j(x')) \Theta_l(x, x') + I'\,,$$

where $I' = \frac{\sigma_w^2}{N_l} W_{l+1}^{ii} (\phi'(y_l^i(x)) + \phi'(y_l^i(x'))) \, \Theta_l(x, x')$.

As $N_l \to \infty$, we have that $I' \to 0$. Using the law of large numbers, as $N_l \to \infty$

$$\frac{\sigma_w^2}{N_l} \sum_j^{N_l} (W_{l+1}^{ij})^2 \phi'(y_l^j(x)) \phi'(y_l^j(x')) \Theta_l(x, x') \to \Psi_l' \Theta_l(x, x').$$

Moreover, we have that

$$\begin{aligned}
&(\partial_{W_{l+1}} y_{l+1}^i(x))(\partial_{W_{l+1}} y_{l+1}^i(x'))^t + (\partial_{B_{l+1}} y_{l+1}^i(x))(\partial_{B_{l+1}} y_{l+1}^i(x'))^t \\
&= \frac{\sigma_w^2}{N_l} \sum_j \phi(y_l^j(x)) \phi(y_l^j(x')) + \sigma_b^2 \underset{N_l \to \infty}{\to} \sigma_w^2 \mathbb{E}[\phi(y_l^1(x)) \phi(y_l^1(x'))] + \sigma_b^2 = \Psi_l \,,
\end{aligned}$$

and so we conclude. □

As a corollary of the above result, using the results in [Daniely et al., 2016] for the ReLU activation function, we can express the recursion more explicitly. We have

$$\Psi_l = \sigma_b^2 + \tfrac{\sigma_w^2}{2} \left(1 + \tfrac{f(C_l)}{C_l}\right) Q_l \,; \qquad \Psi_l' = \tfrac{\sigma_w^2}{2} (1 + f'(C_l)) \,,$$

where f is defined in (4) and $f' : \gamma \mapsto -\frac{1}{\pi} \arccos \gamma$ is the first derivative of f. So we can write

$$\Theta_{l+1} = \Theta_l + \lambda_{l+1,L}^2 \left(\sigma_b^2 + \tfrac{\sigma_w^2}{2} \left(1 + \tfrac{f(C_l)}{C_l}\right) Q_l + \tfrac{\sigma_w^2}{2} (1 + f'(C_l)) \, \Theta_l\right). \tag{A13}$$

We can now easily check that the NTK is a kernel in the sense of Definition 2.

Lemma A17 (Θ_l is a kernel)**.** *For all layer l, Θ_L is a kernel in the sense of definition (2).*

Proof. It's clear that $\Theta_0 = Q_0$ is a kernel. Now fix any layer l. We have already proved in Lemma A10 that $\left(1 + \frac{f(C_l)}{C_l}\right) Q_l$ is a kernel. With a similar argument, noting that $1 + f'$ can be expressed as a power series with only non negative coefficients on $[-1, 1]$, we conclude by Lemma A9 that $1 + f'(C_l)$ is a kernel. Using the usual argument that sums and product of kernels are kernels, we conclude by induction that Θ_l is a kernel. □

As a final remark, note that from (A1), we have that $\lambda_{l,L}^2 \Psi_l = Q_{l+1} - Q_l$. Hence we can rewrite (A13) as

$$\Theta_{l+1} - \Theta_l = Q_{l+1} - Q_l + \lambda_{l+1,L}^2 \tfrac{\sigma_w^2}{2} (1 + f'(C_l)) \, \Theta_l \,. \tag{A14}$$

Since $1 + f'$ is non negative on $[-1, 1]$, it is easy to show by induction that $\Theta_l \geq Q_l$, point-wise, for all l. This is done explicitly in the next Lemma, which is a Corollary of Lemma 1 and show the divergence of the NTK for a Standard ResNet.

Lemma A18 (Exploding NTK)**.** *Consider a ResNet of form (1). For all $x \in \mathbb{R}^d$,*

$$\Theta_L(x, x) \geq \left(1 + \tfrac{\sigma_w^2}{2}\right)^L \left(Q_0(x, x) + \tfrac{2\sigma_b^2}{\sigma_w^2}\right). \tag{A15}$$

Proof. By Lemma 1, it suffices to show that $\Theta_L(x, x) \geq Q_L(x, x)$.
Recall (A13), noticing that $1 + f' \geq 0$ on $[-1, 1]$, $\left(1 + \frac{f(C_l)}{C_l}\right) Q_l \geq 0$ and that $\Theta_0 = Q_0 \geq 0$, by an easy induction we have that $\Theta_l \geq 0$ for all l. As a consequence, from (A14), we get that $\Theta_{l+1} - \Theta_l \geq Q_{l+1} - Q_l$. Hence, again with a straightforward induction we have that $\Theta_l \geq Q_l$ for all l and the the whole K^2. In particular $\Theta_L(x, x) \geq Q_L(x, x)$ for all $x \in K$. □

Lemma A19 (Normalized NTK recursion)**.** *Consider a ResNet of type (1) without bias, and let $\alpha = \frac{\sigma_w^2}{2}$. The NTK recursion formula can be written in terms of normalized NTK $\kappa^l(x, x') = \Theta_l(x, x')/(1 + \alpha)^{l-1}$*

$$\kappa_l(x, x') = \left(\frac{1 + \alpha \hat{f}'(C_{l-1}(x, x'))}{1 + \alpha}\right) \kappa_{l-1}(x, x') + \alpha \hat{f}(C_{l-1}(x, x')) \sqrt{Q_0(x, x) Q_0(x', x')} \,,$$

where $\hat{f}$ is given by (A2), $\hat{f}(t) = \frac{1}{\pi}(t \arcsin t + \sqrt{1 - t^2}) + \frac{1}{2} t$.

Proof. Let $x, x' \in \mathbb{R}^d$. For a ResNet of type (1), we have that

$$\Theta_l = \Theta_{l-1} + \left(\Psi_{l-1} + \Psi'_{l-1}\,\Theta_{l-1}\right)\,,$$

where $\Psi_{l-1} = \alpha Q_{l-1}(x, x')$ and $\Psi'_{l-1} = \alpha \hat{f}'(C_{l-1})$. Using the recursive formula for the diagonal elements, we have that $\Psi_{l-1} = \alpha(1+\alpha)^{l-1}\hat{f}(C_{l-1}(x,x'))\sqrt{Q_0(x,x)Q_0(x',x')}$. We conclude by dividing both sides by $(1+\alpha)^{l-1}$. □

A4.1 Proof of Proposition 8

Proposition 8. *Fix a compact $K \subset \mathbb{R}^d$ ($0 \notin K$ if $\sigma_b = 0$) and consider a Stable ResNet with decreasing scaling. Then Θ_L converges uniformly over K^2 to a kernel Θ_∞. Moreover Θ_∞ is universal on K if $\sigma_b > 0$. If $K = \mathbb{S}^{d-1}$, then the universality holds for $\sigma_b = 0$.*

Proof. Let $K \subset \mathbb{R}^d$ ($0 \notin K$ if $\sigma_b = 0$) be a compact. From (A13), with a decreasing scaling, we have that

$$\begin{aligned}\Theta_l &= \Theta_{l-1} + \lambda_l^2\left(\Psi_{l-1} + \Psi'_l\,\Theta_{l-1}\right)\\ &= \left(1 + \lambda_l^2\frac{\sigma_w^2}{2}f'(C_{l-1})\right)\Theta_{l-1} + \lambda_l^2\Psi_{l-1}\,.\end{aligned}$$

Therefore, the NTK can be expressed exclusively in terms of the covariance kernels $(Q_k)_{k\in[0:l-1]}$, more precisely we have that

$$\Theta_l = \prod_{k=1}^{l}\left(1 + \lambda_k^2\frac{\sigma_w^2}{2}f'(C_{k-1})\right)Q_0 + \sum_{k=1}^{l}\lambda_l^2\prod_{j=k}^{l}\left(1 + \lambda_j^2\frac{\sigma_w^2}{2}f'(C_{j-1})\right)\Psi_{k-1}\,.$$

It is straightforward that Θ_l converges pointwise to a limiting kernel Θ_∞. Let us prove that the convergence is uniform over K. By observing that $|f'| \le 1$, we have that for all $x, x' \in K$

$$\begin{aligned}|\Theta_\infty(x,x') - \Theta_l(x,x')| &\le \prod_{k=1}^{l}\left(1 + \lambda_k^2\frac{\sigma_w^2}{2}\right)\left|\prod_{k=l+1}^{\infty}\left(1 + \lambda_k^2\frac{\sigma_w^2}{2}\right) - 1\right|Q_0(x,x')\\ &\quad + \sum_{k=l+1}^{\infty}\lambda_k^2\prod_{j=k}^{l}\left(1 + \lambda_j^2\frac{\sigma_w^2}{2}\right)\Psi_{k-1}(x,x')\\ &\le \kappa\sum_{k=l+1}^{\infty}\lambda_k^2\,.\end{aligned}$$

where κ is a constant that depends on the compact K. This proves the uniform convergence with a rate of $\mathcal{O}\left(\sum_{k=l+1}^{\infty}\lambda_k^2\right)$. As a consequence, being a uniform limit of kernels, Θ_∞ is a kernel.

Proceeding as in the proof of Lemma A17, it's easy to prove by induction that for all l, $\Theta_l - Q_l$ is a kernel. In particular,

$$T(\Theta_l) \succeq T(Q_l)\,,$$

where $\succeq$ is in the operator sense, that is $T(\Theta_l) - T(Q_l)$ is non-negative definite. This yields

$$T(\Theta_\infty) \succeq T(Q_\infty)\,.$$

Therefore Θ_∞ inherits the universality of Q_∞ naturally by the RKHS hierarchy [Paulsen and Raghupathi, 2016]. We conclude that Θ_∞ is universal (for both cases). □

For the rest of this section, let $K \subset \mathbb{R}^d$ by a compact set. If $\sigma_b = 0$, assume that $0 \notin K$.

With the uniform scaling, for arbitrary $x, x' \in K$, the continuous version of (9) reads

$$\begin{aligned}&\dot{\theta}_t(x,x') = \dot{q}_t(x,x') + \tfrac{\sigma_w^2}{2}(1 + f'(c_t(x,x')))\,\theta_t(x,x')\,;\\ &\theta_0 = q_0\,,\end{aligned} \tag{A16}$$

where $f' : \gamma \mapsto -\frac{1}{\pi}\arccos\gamma$ is the first derivative of f, defined in (4).

Lemma A20. *For any x, x' in K, the solution $t \mapsto \Theta_t$ of (A16) is unique and well defined for all $t \in [0,1]$. Moreover, the map $(x,x') \mapsto \Theta_t(x,x')$ is a kernel in the sense of Definition 2 for all $t \in [0,1]$.*
We have the $L^2(K)$ convergence of the discrete model to the continuous one:

$$\lim_{L\to\infty} \sup_{l\in[0:L]} \left\| T(\Theta_{l|L}) - T(\theta_{t=l/L}) \right\|_2 = 0\,.$$

Proof. The existence and the uniqueness are clear, since it is a homogeneous first order Cauchy problem, with continuous coefficients. We can write explicitly the solution as

$$\theta_t = e^{G_t} \left(q_0 + \int_0^t \dot{q}_s \, e^{-G_s} \, \mathrm{d}s \right) , \tag{A17}$$

where $G_t(z) = \frac{\sigma_w^2}{2} \int_0^t (1 + f'(c_s(z)))\,\mathrm{d}s$ for $z \in K^2$. It becomes then clear that $z \mapsto \Theta_t(z)$ is a continuous and symmetric function on K^2.
it is easy to check that the uniform convergence of C and Q to c and q implies that for all $z \in K$, $\lim_{L\to\infty} \sup_{l\in[0:L]} |\Theta_{l|L}(z) - \theta_{l/L}(z)| = 0$. As consequence, by dominated convergence,

$$\lim_{L\to\infty} \sup_{l\in[0:L]} \left\| T(\Theta_{l|L}) - T(\theta_{t=l/L}) \right\|_2 = 0\,.$$

Hence, $T(\theta_t)$ is the limit of a sequence of non-negative definite operators and hence it is non-negative definite, so that θ_t is a kernel on K for all $t \in [0,1]$. □

Proposition 9. *Let $K \subset \mathbb{R}^d$ and fix $t \in (0,1]$. If $\sigma_b > 0$, then θ_t is universal on K. The same holds true if $\sigma_b = 0$ and $K = \mathbb{S}^{d-1}$.*

Proof. Fix $t \in (0,1]$. The solution of (A16) can be written as $\theta_t = q_t + r_t$, where

$$r_t = \tfrac{\sigma_w^2}{2} \int_0^t (1 + f'(c_s))\, \theta_s \, \mathrm{d}s\,.$$

Now, let us show that r_t. First, by Lemma A14 it is easy to check that $1 + f'$ is analytic on $(-1,1)$ and its Taylor expansion around 0 converges on $[-1,1]$. Moreover all the Taylor coefficients are non negative. Hence, Lemma A9 shows that $(1 + f'(c_s))$ is a kernel for all $s \in [0,s)$. It follows that $(1 + f'(c_s))\,\theta_s$ is a kernel.[17] Now, $(1 + f'(c_s)\,\theta_s$ is continuous and symmetric on Z^2, and it is easy to check from (A17) that it is uniformly bounded for $s \in [0,t)$. It follows that r_t is continuous and symmetric. Now, fix an arbitrary finite Borel measure μ on K. We have to show that $T_\mu(r_t)$ is non-negative definite, so that we can conclude by Lemma A1. Fixed $\varphi \in L^2(K,\mu)$, by simple standard arguments we have

$$\langle T_\mu(r_t)\,\varphi, \varphi\rangle = \int_0^t \langle T_\mu((1 + f'(c_s))\,\theta_s)\,\varphi, \varphi\rangle \,\mathrm{d}s \geq 0$$

and so r_t is a kernel.
Now, given two kernels Q and R, it is a classical result that $Q + R$ is a kernel and its RKHS contains the RKHS of Q and R, [Paulsen and Raghupathi, 2016]. We conclude that the RKHS of θ_t contains the RKHS of q_t. Since q_t is universal, θ_t is universal. □

A5 A PAC-Bayes Generalization result

In this section, we study the PAC-Bayes upper bound of a GP with kernel Q_L. We consider a dataset S with N iid training examples $\{(x_i, y_i) \in X \times Y, i \in [1:N]\}$, and a hypothesis space $\mathcal{H}$ from which we want to learn an optimal hypothesis according to some bounded loss function $\ell : Y \times Y \to [0,1]$. The empirical loss of a hypothesis $h \in \mathcal{H}$ is given by

$$r_S(h) = \frac{1}{N} \sum_{i=1}^N \ell(h(x_i), y_i)\,.$$

[17] See footnote 14.

Assuming that the samples are distributed as $(x, y) \sim \nu$ where ν is a probability distribution on $X \times Y$, we define the generalization (true) loss by

$$r(h) = \mathbb{E}_\nu[\ell(f(x), y)]\,.$$

For some randomized learning algorithm $\mathcal{A}$, the empirical and generalization loss are given by

$$r_S(\mathcal{A}) = \mathcal{E}_{h\sim\mathcal{A}}[r_s(h)]\,; \qquad r(\mathcal{A}) = \mathcal{E}_{h\sim\mathcal{A}}[r(h)]\,.$$

The PAC-Bayes theorem gives a probabilistic upper bound on the generalization loss $r(\mathcal{A})$ of a randomized learning algorithm $\mathcal{A}$ in terms of the empirical loss $r_S(\mathcal{A})$. Fix a prior distribution $\mathcal{P}$ on the hypothesis set $\mathcal{H}$. The Kullback-Leibler divergence between $\mathcal{A}$ and $\mathcal{P}$ is defined as $KL(\mathcal{A}\|\mathcal{P}) = \int \log \frac{\mathcal{A}(h)}{P(h)} \mathcal{A}(h)dh \in [0, \infty]$. The Bernoulli KL-divergence is given by $kl(a||p) = a \log \frac{a}{p} + (1-a) \log \frac{1-a}{1-p}$ for $a, p \in [0, 1]$. We define the inverse Bernoulli KL-divergence kl^{-1} by

$$kl^{-1}(a, \varepsilon) = \sup\{p \in [0, 1] : kl(a, p) \leq \varepsilon\}\,.$$

Theorem 2 (PAC-Bayesian theorem)**.** *For any loss function ℓ that is $[0, 1]$ valued, for any distribution ν, for any $N \in \mathbb{N}$, for any prior P, and any $\delta \in (0, 1]$, with probability at least $1 - \delta$ over the sample S, we have*

$$\forall \mathcal{A}, \quad r(\mathcal{A}) \leq kl^{-1}\left(r_S(\mathcal{A}), \frac{KL(\mathcal{A}\|P) + \log \frac{2\sqrt{N}}{\delta}}{N}\right).$$

The PAC-Bayesian theorem gives can also be stated as

$$kl(r_S(\mathcal{A}), r(\mathcal{A})) \leq \frac{KL(\mathcal{A}\|P) + \log \frac{2\sqrt{N}}{\delta}}{N}\,.$$

The KL-divergence term $KL(\mathcal{A}\|P)$ plays a major role as it controls the generalization gap, i.e. the difference (in terms of Bernoulli KL-divergence) between the empirical loss and the generalization loss. In our setting, we consider an ordinary GP regression with prior $P(f) = \mathcal{GP}(f|0, Q(x, x'))$. Under the standard assumption that the outputs $y_N = (y_i)_{i\in[1:N]}$ are noisy versions of $f_N = (f(x_i))_{i\in[1:N]}$ with $y_N|f_N \sim \mathcal{N}(y_N|f_N, \sigma^2 I)$, the Bayesian posterior $\mathcal{A}$ is also a GP and is given by

$$\mathcal{A}(f) = \mathcal{GP}(f|Q_N(x)(Q_{NN} + \sigma^2 I)^{-1} y_N, Q(x, x') - Q_N(x)(Q_{NN} + \sigma^2 I)^{-1})Q_N(x')^T)\,, \tag{A18}$$

where $Q_N(x) = (Q(x, x_i))_{i\in[1:N]}$ and $Q_{NN} = (Q(x_i, x_j))_{1\leq i,j\leq N}$. In this setting, we have the following result

Proposition 10 (Stability of PAC-Bayes bound)**.** *Let Q_L be the kernel of a ResNet. Let P_L be a GP with kernel Q_L and $\mathcal{A}_L$ be the corresponding Bayesian posterior for some fixed noise level $\sigma > 0$. Then, in a fixed setting (fixed sample size N), the following results hold:*

1. *With a standard ResNet, we have*
$$KL(\mathcal{A}_L\|P_L) \gtrsim L\,.$$
2. *With a Stable ResNet, we have*
$$KL(\mathcal{A}_L\|P_L) = \mathcal{O}_L(1)\,.$$

Proof. The proof relies on the simple observation that $P_L(f|f_N) = \mathcal{A}_L(f|f_N)$. This yields

$$\begin{aligned} KL(\mathcal{A}_L\|P_L) &= KL(\mathcal{A}_L(f_N)\mathcal{A}_L(f|f_N)\|P_L(f_N)P_L(f|f_N)) \\ &= KL(\mathcal{A}_L(f_N)\|P_L(f_N)) \\ &= \frac{1}{2}\log(\det(Q_{L,NN} + \sigma^2 I)) - \frac{N}{2}\log(\sigma^2) - \frac{1}{2}\operatorname{Tr}(Q_{L,NN}(Q_{L,NN} + \sigma^2 I)^{-1}) \\ &+ \frac{1}{2} y_N^T (Q_{L,NN} + \sigma^2 I)^{-1} Q_{L,NN} (Q_{L,NN} + \sigma^2 I)^{-1} y_N\,, \end{aligned} \tag{A19}$$

where $Q_{L,NN} = (Q_L(x_i, x_j))_{1\leq i,j\leq N}$.

Since $Q_{L,NN}$ is symmetric and strictly positive definite, it is straightforward that the largest eigenvalue of $Q_{L,NN}(Q_{L,NN}+\sigma^2 I)^{-1}$ is smaller than 1. This yields

$$\mathrm{Tr}(Q_{L,NN}(Q_{L,NN}+\sigma^2 I)^{-1}) \leq N$$

and

$$y_N^T(Q_{L,NN}+\sigma^2 I)^{-1}Q_{L,NN}(Q_{L,NN}+\sigma^2 I)^{-1}y_N \leq \sigma^{-2}\|y_N\|_2\,.$$

Both quantities are bounded independently from L and the scaling factors $(\lambda_{k,L})_{k\in[2:L]}$.
Now let us analyse the first term $\frac{1}{2}\log(\det(Q_{L,NN}+\sigma^2 I))$. Let $\mu_{L,0}\geq\mu_{L,1}\geq\cdots\geq\mu_{L,N}$ be the eigenvalues of $Q_{L,NN}$. For a simplification purpose, we assume the inputs belong to the unit sphere $\mathbb{S}^{d-1}$. The proof extends to any compact set.
Let us study the behaviour of the first term for both cases.

Case 1. Assume we have a standard ResNet architecture. On the unit sphere $\mathbb{S}^{d-1}$, we have that $Q_L(x,x')\geq q_L C_L(x,x')$, where $q_L=(1+\frac{\sigma_2}{2})^L\delta$ with $\delta=(\sigma_b^2+\frac{\sigma_w^2}{d})/(1+\frac{\sigma_w^2}{2})$. Using Lemma 4, we know that $\lim_{L\to\infty}\hat{\mu}_{L,0}=\hat{\mu}_{\infty,0}\in(0,\infty)$ and for all $k\geq 1$, $\lim_{L\to\infty}\hat{\mu}_{L,k}=0$. This yields

$$\begin{aligned}\log(\det(Q_{L,NN}+\sigma^2 I)) &\geq \sum_{k=1}^{N}\log(q_L\hat{\mu}_{L,k}+\sigma^2)\\ &\geq \log(q_L\hat{\mu}_{L,0}+\sigma^2)+(N-1)\log(\sigma^2)\\ &\gtrsim L\log(1+\frac{\sigma_w^2}{2})\,,\end{aligned}$$

where the last inequality holds for sufficiently large L.

Case 2. In the case of Stable ResNet, we know that as $L\to\infty$, the kernel Q_L converges to a strictly positive definite kernel Q_∞, therefore the first term $\log(\det(Q_{L,NN}+\sigma^2 I))$ remains bounded as $L\to\infty$, which concludes the proof. □

A6 NNGP correlation kernel without bias as a modified NNGP kernel

Unscaled ResNets suffer from the exploding variance problem, which needs to be avoided in order to isolate the disadvantages of inexpressivity in their NNGP kernel. In order to do so, we use the NNGP correlation kernel C instead of NNGP covariance kernel Q, noting that Lemma A4 provides a simple recursion formula for C if $\sigma_b=0$, at depth $l\leq L$:

$$C_l(x,x')=\frac{1}{1+\alpha_{l.L}}C_{l-1}(x,x')+\frac{\alpha_{l,L}}{1+\alpha_{l,L}}\hat{f}(C_{l-1}(x,x'))\,, \tag{A20}$$

where $\alpha_{l,L}=\frac{\lambda_{l,L}^2\sigma_w^2}{2}$ and $\hat{f}$ defined in (A2). In order to combine this with open-source packages [Novak et al., 2020, Bradbury et al., 2018] designed for NNGP calculation, we note that (A20) can be viewed as the NNGP kernel of the following modified ResNet layer, using the same notation as in (2):

$$y_l(x)=\sqrt{1-\hat{\alpha}_{l,L}}y_{l-1}(x)+\sqrt{\hat{\alpha}_{l,L}}\,\mathcal{F}((W_l,B_l),y_{l-1})\,,\quad l\in[1:L]\,, \tag{A21}$$

with $\hat{\alpha}_{l,L}=\frac{\alpha_{l,L}}{1+\alpha_{l,L}}$

A7 Experimental details and additional results

A7.1 NNGP results

For our Vanilla ResNet NNGP results, we preprocess all training, validation and test data by first centering the training set and then normalizing all images to lie on the pixel dimension sphere. For our Wide ResNet NNGP results we normalise all data so that the training set is centered and has channel-wise unit variance. We use Kaiming [He et al., 2015] initialisation throughout, with $\sigma_w^2=2$ and $\sigma_b^2=0$. Vanilla ResNets have the same

structure as type (2) in Table 2 and we use the same WRN kernel architecture as [Lee et al., 2019] in Table 1 but omit the final average pooling step, which is known to improve kernel performance but dramatically increase computational costs [Novak et al., 2019, Lee et al., 2020]. Throughout this work, where there are residual blocks with multiple layers, we calculate our scaling factors for uniform and decreasing scaled Stable ResNets by the number of residual connections. For example, a WRN-202 has only 99 residual connections, so we set $\lambda_{l,L}^{-1} = \sqrt{99}$ for the uniform scaling factors. We tune the noise variance σ^2, which is akin to the regularisation parameter in kernel ridge regression. To do so, we compute validation accuracy on a validation set of size 5000, selecting the best $\sigma^2 = \lambda \times \text{Trace}(Q_{NN})/N$ from a logarithmic scale of $\lambda = [0.001, 0.01, 0.1]$, where N is the training set size and Q_{NN} is the $N \times N$ training set Gram matrix for NNGP Q.

A7.2 Trained ResNet results

For all our trained ResNet experiments we use a similar setup to the open-source code for [Wang et al., 2020] in PyTorch [Paszke et al., 2019]. We repeat each experiment 3 times and report the best test accuracy and error intervals. All ResNets are initialised with Kaiming initialisation [He et al., 2015] and like [Wang et al., 2020] we adopt ResNets architectures where we double the number of filters in each convolutional layer. For experiments with BatchNorm, on CIFAR-10/100 we use batch size 64 across all depths and on TinyImageNet we used batch size 128 for depths 32 & 50, and batch size 100 for depth 104 in order to allow the model to fit onto a single 11GB VRAM GPU. We use SGD with momentum parameter 0.9 and weight decay parameter 10^{-4} throughout.

We also present results for ResNets trained without BatchNorm [Ioffe and Szegedy, 2015]. BatchNorm is a normalization layer commonly used with modern ResNets that is known to improve performance and allows deeper ResNets to be trained, though the precise reasons for this are not well understood. Several recent works [De and Smith, 2020, Zhang et al., 2019] have studied the possibility of removing the need for BatchNorm layers, by introducing trainable uniform scalings to the residual connection to stabilise variance at initialisation & gradients, demonstrating promising results. Note, our work additionally introduces decreasing scaling and also uses the infinite-width NNGP/NTK connection to assess the theoretical advantages of scaled Stable ResNets in the limit of infinite depth.

Moreover, our focus is not towards the possibility of removing BatchNorm and we show in Table 3 that our scalings can improve BatchNorm ResNets. However, we also present results without BatchNorm in Table 4, where again we see that our scaled stable ResNets improve performance compared to their unscaled counterparts: for example both Decreasing and Uniform scaling outperform the unscaled ResNet by over 3% test accuracy on CIFAR-100 with ResNet-104.

For ResNets trained without BatchNorm, for a fair comparison we tuned the initial learning rate on a small logarithmic scale, using batch size 128.

Table 4: Test accurracies (%) of trained deep ResNets **without BatchNorm** of various scalings and depths on CIFAR-10 (C-10), CIFAR-100 (C-100).

Dataset	Depth	Scaled (D)	Scaled (U)	Unscaled
C-10	32	$92.64_{\pm 0.19}$	$92.78_{\pm 0.18}$	$92.11_{\pm 0.17}$
	50	$92.33_{\pm 0.05}$	$\mathbf{92.72}_{\pm 0.12}$	$92.10_{\pm 0.17}$
	104	$92.81_{\pm 0.09}$	$\mathbf{93.28}_{\pm 0.17}$	$92.70_{\pm 0.08}$
C-100	ResNet32	$\mathbf{67.73}_{\pm 0.42}$	$67.06_{\pm 0.38}$	$65.37_{\pm 0.32}$
	ResNet50	$\mathbf{69.38}_{\pm 0.20}$	$68.76_{\pm 0.18}$	$66.02_{\pm 0.41}$
	ResNet104	$70.60_{\pm 0.52}$	$\mathbf{70.95}_{\pm 0.13}$	$67.41_{\pm 0.41}$

A8 Some results on the Sphere $\mathbb{S}^{d-1}$

On the sphere $\mathbb{S}^{d-1}$, the kernel Q_2 is analytic as a result of lemma A8. Moreover, the coefficient of the analytic decomposition are all positive.

Lemma A21 (Analytic decomposition of 2 layer ReLU ResNet)**.** *For all* $(x, x') \in \mathbb{S}^{d-1}$, $Q_2(x, x') = g(x \cdot x')$ *where* $g(z) = \sum_{i \geq 0} a_i z^i$ *and* $a_i > 0$ *for all* $i \geq 0$.

Proof. Let $x, x' \in \mathbb{S}^{d-1}$. We have

$$Q_0(x, x') = \sigma_b^2 + \frac{\sigma_w^2}{d} x \cdot x' \,.$$

As a result, for all x, x', $Q_0(x, x) = Q_0(x', x') = \sigma_b^2 + \frac{\sigma_w^2}{d}$. The diagonal term of the kernel is the same for all $x \in \mathbb{S}^{d-1}$. We note $\beta_l = Q_l(x, x)$ and $z = x \cdot x'$. Using this observation, we have that

$$Q_1(x, x') = Q_0(x, x') + \lambda_1^2(\sigma_b^2 + \frac{\sigma_2}{2} \hat{f}(C_0(x, x'))\beta_0) \,.$$

It can be easily deduced from lemma A8 that there exist $\{b_i\}_{i \geq 0}$ such that

$$C_1(x, x') = b_0 + b_1 z + \sum_{i \geq 0} b_{2i} z^{2i} \,,$$

where $b_0, b_1, b_{2i} > 0$.
Following the same approach, we have that

$$Q_2(x, x') = Q_1(x, x') + \lambda_2(\sigma_b^2 + \frac{\sigma_w^2}{2} f(C_1(x, x'))\beta_2)$$

and

$$\hat{f}(C_1(x, x')) = a_0 + a_1 C_1(x, x') + \sum_{i \geq 1} a_{2i} (C_1(x, x'))^{2i} \,.$$

Having the terms of orders 0 and 1 in $C_1(x, x')$ ensures having a positive coefficient for all terms z^i for $i \geq 1$, which concludes the proof. □

The previous result can be easily extended to general $L \geq 2$. We have that

$$Q_L(x, x') = g_L(x \cdot x') \,,$$

where $g_L : [-1, 1] \to \mathbb{R}$ is a continuous function. Kernels that can be written in this form are known as the dot-product kernels (or zonal kernels on the unit sphere). In our setting, we have a stronger property; we prove in the next result we show a that the kernel Q_L is analytic on the sphere $\mathbb{S}^{d-1}$ in the sense that the function g_L is analytic on $[-1, 1]$.

Proposition A5 (Q_L is analytic)**.** *Let $L \geq 2$, there exists $(\alpha_{L,i})_{i \geq 0}$ such that for all $x, x' \in \mathbb{S}^{d-1}$*

$$Q_L(x, x') = \sum_{i \geq 0} \alpha_{L,i} (x \cdot x')^i \,.$$

Moreover, $(\alpha_{L+1,i})_{i \geq 0}$ can be expressed in terms of $(\alpha_{L,i})_{i \geq 0}$

$$\alpha_{L+1,i} = \alpha_{L,i} + \lambda_{L+1,L+1} \times \gamma_{L,i} \,, \tag{A22}$$

with

$$\gamma_{L,i} = \begin{cases} \sigma_b^2 + \beta_L \frac{\sigma_w^2}{2} \sum_{m \geq 0} \frac{a_m}{\beta_L^m} \alpha_{L,0}^m & \text{if } i = 0 \,; \\ \beta_L \frac{\sigma_w^2}{2} \sum_{m \geq 0} \frac{a_m}{\beta_L^m} \sum_{k_1 + \ldots + k_m = i} \prod_{j=1}^{m} \alpha_{L,k_j} & \text{if } i \geq 1 \,. \end{cases}$$

where $\beta_L = Q_L(x, x) = Q_L(x', x') = \sum_{i \geq 0} \alpha_{L,i}$ and $(a_m)_{m \geq 0}$ is such that $a_0, a_1 > 0$ and $a_{2i} > 0$ and $a_{2i+1} = 0$ for all $i \geq 1$.
As a result, for all $L \geq 2, i \geq 0, \alpha_{L,i} > 0$.

Proof. The result is true for $L = 2$ by lemma A21. Let us prove the result for all $L \geq 3$ by induction.
Let $L \geq 3$, $x, x' \in \mathbb{S}^{d-1}$, $z = x \cdot x'$ and $\beta_l = Q_l(x, x) = Q_l(x', x')$. Assume the result is true for L and let us prove it for $L + 1$. We have that

$$Q_{L+1}(x, y) = Q_L(x, y) + \lambda_{L+1,L+1}^2(\sigma_b^2 + \frac{\sigma_w^2}{2} f(C_l(x, y))\beta_l) \,.$$

Knowing that $C_l(x,y) = \frac{1}{\beta_l} Q_l(x,y)$, we have that

$$\begin{aligned} f(C_l(x,y)) &= \sum_{m\geq 0} \frac{a_m}{\beta_l^m} C_l(x,y)^m \\ &= \sum_{m\geq 0} \frac{a_m}{\beta_l^m} (\sum_{i\geq 0} \alpha_{l,i} z^i)^m \\ &= \sum_{m\geq 0} \frac{a_m}{\beta_l^m} \sum_{i\geq 0} \sum_{k_1+\ldots+k_m=i} \prod_{j=1}^m \alpha_{l,k_j} z^i \\ &= \sum_{i\geq 0} \Big[\sum_{m\geq 0} \frac{a_m}{\beta_l^m} \sum_{k_1+\ldots+k_m=i} \prod_{j=1}^m \alpha_{l,k_j} \Big] z^i \,, \end{aligned}$$

which gives the recursive formulas for the coefficients of the analytic decomposition. Observe that the coefficients are non-decreasing wrt L. Using lemma A21 we conclude that $\alpha_{L,i} > 0$. □

For depth $L \geq 2$, proposition A5 shows that all coefficient $(\alpha_{L,i})_{i\geq 0}$ are (strictly) positive. It turns out that this is a sufficient condition for the kernel Q_L to be strictly positive definite. We state this in the next proposition. The result can be seen as a consequence of Lemma A12 and Lemma A13. However we will give here a more direct proof.

Proposition A6 (Q_L is strictly p.d. for $L \geq 2$)**.** *Let Q be an analytic kernel on the unit sphere $\mathbb{S}^{d-1}$, i.e. there exist a sequence of real numbers $(\alpha_i)_{i\geq 0}$ such that for all $x, x' \in \mathbb{S}^{d-1}$*

$$Q(x,x') = \sum_{i\geq 0} \alpha_i (x\cdot x')^i$$

Assume $\alpha_i > 0$ for all $i \in \mathbb{N}$. Then, Q is strictly positive definite.
As a result, for all $L \geq 2$, $T_\nu(Q_L)$ is strictly positive definite, i.e. for any non-zero function $\varphi \in L^2(\mathbb{S}^{d-1} l, \nu)$

$$\langle T_\nu(Q_L)\varphi, \varphi\rangle > 0$$

ν is the standard uniform measure on the sphere $\mathbb{S}^{d-1}$.

Proof. Let Q be an analytic kernel on the unit sphere $\mathbb{S}^{d-1}$, that is there exists a sequence of real numbers $(\alpha_i)_{i\geq 0}$ such that for all $x, x' \in \mathbb{S}^{d-1}$

$$Q(x,x') = \sum_{i\geq 0} \alpha_i (x\cdot x')^i \,,$$

and assume $\alpha_i > 0$ for all $i \in \mathbb{N}$. The map $(x,x') \mapsto x\cdot x'$ is trivially a kernel in the sense of Definition 2. For all $i \geq 0$, $(x,x') \mapsto (x\cdot x')^i$ is a kernel as well.[18] It follows that $T_\nu(Q)$ is non-negative definite, as a converging sum of non-negative operators. Let us prove that it is strictly positive definite.
Let $\varphi \in L_2(\mathbb{S}^{d-1}, \nu)$ such that $\langle T_\nu(Q)\,\varphi, \varphi\rangle = 0$. Since $\alpha_i > 0$ for all i, we have that for all $i \geq 0$

$$\int\int (x\cdot x')^i \varphi(x)\varphi(x')\,\mathrm{d}\nu(x)\mathrm{d}\nu(x') = 0\,,$$

recalling that ν is the uniform measure on the sphere $\mathbb{S}^{d-1}$. This yields

$$\int\int P(x\cdot x')\varphi(x)\varphi(x')\,\mathrm{d}\nu(x)\mathrm{d}\nu(x') = 0 \tag{A23}$$

for any polynomial function P.
Since φ is a function on the sphere $\mathbb{S}^{d-1}$, it can be decomposed in the Spherical Harmonics orthonormal basis $(Y_{k,j})_{k,j}$ (see e.g. [MacRobert, 1967]) as

$$\forall x \in \mathbb{S}^{d-1}, \quad \varphi(x) = \sum_{k\geq 0} \sum_{j=1}^{N(d,k)} b_{k,j} Y_{k,j}(x)$$

[18] See footnote 14.

where $b_{k,j} = \int_{\mathbb{S}^{d-1}} \varphi(w) Y_{k,j}(w)\, \mathrm{d}\nu(w)$.
In particular, equation (A23) is true for the Associated Legendre Polynomials P_k. Knowing that $N(d,k)P_k(x\cdot x') = \sum_{j=1}^{N(d,k)} Y_{k,j}(x)Y_{k,j}(x')$, (A23) yields

$$\int\int \sum_{j=1}^{N(d,k)} Y_{k,j}(x)Y_{k,j}(x')\varphi(x)\varphi(x')\, \mathrm{d}\nu(x)\mathrm{d}\nu(x') = 0$$

for all $k \geq 0$. Therefore,

$$\sum_{j=1}^{N(d,k)} b_{k,j}^2 = 0$$

for all $k \geq 0$. We conclude that $\varphi = 0$. □

By Mercer's theorem [Paulsen and Raghupathi, 2016], the kernel Q_L can be decomposed in an orthonormal basis of $L^2(\mathbb{S}^{d-1})$. It turns out that this orthonormal basis is the so-called Spherical Harmonics of $\mathbb{S}^{d-1}$. This is a corollary of the next lemma, which is a classical result [Yang and Salman, 2019].

Lemma A22 (Spectral decomposition on $\mathbb{S}^{d-1}$)**.** *Let Q be a zonal kernel on $\mathbb{S}^{d-1}$, that is $Q(x,x') = p(x\cdot x')$ for a continuous function $p : [-1,1] \to \mathbb{R}$. Then, there is a sequence $\{\mu_k \geq 0\}_{k\in\mathbb{N}}$ such that for all $x, x' \in \mathbb{S}^{d-1}$*

$$Q(x,x') = \sum_{k\geq 0} \mu_k \sum_{j=1}^{N(d,k)} Y_{k,j}(x)Y_{k,j}(x')\,,$$

where $\{Y_{k,j}\}_{k\geq 0, j\in[1:N(d,k)]}$ are spherical harmonics of $\mathbb{S}^{d-1}$ and $N(d,k)$ is the number of harmonics of order k. With respect to the standard spherical measure ν on $\mathbb{S}^{d-1}$, the spherical harmonics form an orthonormal basis of $L^2(\mathbb{S}^{d-1},\nu)$ and $T_\nu(Q)$ is diagonal on this basis.

Proof. We start by giving a brief review of the theory of Spherical Harmonics ([MacRobert, 1967]). For some $k \geq 1$, let $(Y_{k,j})_{1\leq j\leq N(d,k)}$ be the set of Spherical Harmonics of degree k. We have $N(d,k) = \frac{2k+d-2}{k}\binom{k+d-3}{d-2}$.

The set of functions $(Y_{k,j})_{k\geq 1, j\in[1:N(d,k)]}$ form an orthonormal basis of $L^2(\mathbb{S}^{d-1},\nu)$, where ν is the uniform measure on $\mathbb{S}^{d-1}$.

For some function p, the Hecke-Funk formula reads

$$\int_{\mathbb{S}^{d-1}} p(\langle x,w\rangle) Y_{k,j}(w)\, \mathrm{d}\nu(w) = \frac{\Omega_{d-1}}{\Omega_d} Y_{k,j}(x) \int_{-1}^{1} p(t) P_k^d(t)(1-t^2)^{(d-3)/2} dt$$

where Ω_d is the volume of the unit sphere $\mathbb{S}^{d-1}$, and P_k^d is the multi-dimensional Legendre polynomials given explicitly by Rodrigues' formula

$$P_k^d(t) = \left(-\frac{1}{2}\right)^k \frac{\Gamma(\frac{d-1}{2})}{\Gamma(k+\frac{d-1}{2})} (1-t^2)^{\frac{3-d}{2}} \frac{\mathrm{d}^k}{\mathrm{d}t^k} (1-t^2)^{k+\frac{d-3}{2}}\,.$$

$(P_k^d)_{k\geq 0}$ form an orthogonal basis of $L^2([-1,1],(1-t^2)^{\frac{d-3}{2}}dt)$, i.e.

$$\langle P_k^d, P_{k'}^d \rangle_{L^2([-1,1],(1-t^2)^{\frac{d-3}{2}}dt)} = \delta_{k,k'}\,,$$

where δ_{ij} is the Kronecker symbol.

Using the Heck-Funk formula, we prove that Q can be decomposed on the Spherical Harmonics basis. Indeed, for any $x, x' \in \mathbb{S}^{d-1}$, the decomposition on the spherical harmonics basis yields

$$Q(x,x') = \sum_{k\geq 0} \sum_{j=1}^{N(d,k)} \left[\int_{\mathbb{S}^{d-1}} p(\langle w,x'\rangle) Y_{k,j}(w)\mathrm{d}\nu(w)\right] Y_{k,j}(x)\,.$$

Using the Hecke-Funk formula yields

$$Q(x,x') = \sum_{k\geq 0}\sum_{j=1}^{N(d,k)} \left[\frac{\Omega_{d-1}}{\Omega_d}\int_{-1}^{1} p(t)P_k^d(t)(1-t^2)^{(d-3)/2}dt\right] Y_{k,j}(x)Y_{k,j}(x')\,.$$

We conclude that

$$Q(x,x') = \sum_{k\geq 0}\mu_k \sum_{j=1}^{N(d,k)} Y_{k,j}(x)Y_{k,j}(x')\,.$$

where $\mu_k = \frac{\Omega_{d-1}}{\Omega_d}\int_{-1}^{1} p(t)P_k^d(t)(1-t^2)^{(d-3)/2}dt$. We also have that $\mu_k \geq 0$ since Q is non-negative by definition. The last statement, follows from the spectral theory of compact self-adjoint operators and the orthonormality of the spherical harmonics (see the appendix of [Yang and Salman, 2019] for details). □

Corollary A3 (Spectral decomposition of Q_L)**.** *For $L \geq 1$, there exist $(\mu_{L,k})_{k\geq 0}$ such that $\mu_{L,k} > 0$ for all $k \geq 0$, and for all $x, x' \in \mathbb{S}^{d-1}$ we have*

$$Q_L(x,x') = \sum_{k\geq 0}\mu_{L,k} \sum_{j=1}^{N(d,k)} Y_{k,j}(x)Y_{k,j}(x')\,,$$

where $(Y_{k,j})_{k\geq 0, j\in[1:N(d,k)]}$ are spherical harmonics of $\mathbb{S}^{d-1}$ and $N(d,k)$ is the number of harmonics of order k.

Corollary A3 shows that for any depth L, the Spherical Harmonics are the eigenfunctions of the kernel Q_L. The fact that $\mu_{L,k} > 0$ is a direct result of Proposition A6. Leveraging this result, we can prove a stronger result, which is the universality of the kernel Q_L.

Proposition A7 (Universality on $\mathbb{S}^{d-1}$)**.** *For all $L \geq 2$, Q_L is universal on $\mathbb{S}^{d-1}$ for $d \geq 2$.*

Proof. The result is a consequence of Lemma A15 and Proposition A6. An alternative proof is the following. It is a classical result that the set Spherical Harmonics form an orthonormal basis on $L^2(\mathbb{S}^{d-1}, \nu)$. Leveraging the result from corollary A3, it is straightforward that any continuous function in $L^2(\mathbb{S}^{d-1}, \nu)$ can be approximated by a function of the form $\sum_i Q_L(x_i, .)$ which belongs to the RKHS of Q_L. Therefore, Q_L is universal on $\mathbb{S}^{d-1}$. Note that we have not made the assumption that $\sigma_b > 0$. □

References

G. Yang and S. Schoenholz. Mean field residual networks: On the edge of chaos. In *Advances in Neural Information Processing Systems*, pages 7103–7114, 2017.

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*, 2019a.

R.M. Neal. *Bayesian Learning for Neural Networks*, volume 118. Springer Science & Business Media, 1995.

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.

S.S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein. Deep information propagation. In *International Conference on Learning Representations*, 2017.

J. Lee, L. Xiao, S. Schoenholz, Y. Bahri, R. Novak, 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.

A.G. Matthews, J. Hron, M. Rowland, R.E. Turner, and Z. Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018.

J. Lee, Y. Bahri, R. Novak, S.S. Schoenholz, J. Pennington, and J. Sohl-Dickstein. Deep neural networks as Gaussian processes. In *International Conference on Learning Representations*, 2018.

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

S. Hayou, A. Doucet, and J. Rousseau. Mean-field behaviour of neural tangent kernel for deep neural networks. *arXiv preprint arXiv:1905.13654*, 2019b.

S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International Conference on Machine Learning*, 2015.

S. Hayou, J.F. Ton, A. Doucet, and Y.W. Teh. Robust pruning at initialization. In *International Conference on Learning Representations*, 2021.

G. Yang. Tensor programs i: Wide feedforward or recurrent neural networks of any architecture are Gaussian processes. *arXiv preprint arXiv:1910.12478*, 2019b.

A. Daniely, R. Frostig, and Y. Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. *Advances in Neural Information Processing Systems 29*, 2016.

V.I. Paulsen and M. Raghupathi. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Cambridge University Press, 2016.

R. M. Dudley. *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2 edition, 2002.

I. Steinwart. Convergence types and rates in generic Karhunen-Loeve expansions with applications to sample path properties. *Potential Analysis*, 51(3):361–395, 2019.

S. Lang. *Real and Functional Analysis*. Graduate Texts in Mathematics. Springer, New York, 3rd edition, 2012.

U. Grenander. Stochastic processes and statistical inference. *Arkiv Matematik*, 1(3):195–277, 10 1950.

G. Yang and H. Salman. A fine-grained spectral perspective on neural networks. *arXiv preprint 1907.10599*, 2019.

B. Sriperumbudur, K. Fukumizu, and G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(70):2389–2410, 2011.

M. Kanagawa, P. Hennig, D. Sejdinovic, and B.K. Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.

L. Chizat and F. Bach. A note on lazy training in supervised differentiable programming. In *Advances in Neural Information Processing Systems*, 2019.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018.

M. Seeger. PAC-Bayesian generalisation error bounds for Gaussian process classification. *Journal of Machine Learning Research*, page 233–269, 02 2002.

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.

V. Shankar, A. Fang, W. Guo, S. Fridovich-Keil, L. Schmidt, J. Ragan-Kelley, and B. Recht. Neural kernels without tangents. *International Conference on Machine Learning*, 2020.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference 2016*, 2016.

B. He, B. Lakshminarayanan, and Y. W. Teh. Bayesian deep ensembles via the neural tangent kernel. *Advances in Neural Information Processing Systems*, 2020.

R. Novak, L. Xiao, J. Hron, J. Lee, A. Alemi, J. Sohl-Dickstein, and S. Schoenholz. Neural tangents: Fast and easy infinite neural networks in Python. In *International Conference on Learning Representations*, 2020. URL https://github.com/google/neural-tangents.

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

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

G. Yang. Tensor programs iii: Neural matrix laws. *arXiv preprint arXiv:2009.10685*, 2020.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, pages 337–404, 1950.

C. Micchelli, Y. Xu, and H. Zhang. Universal kernels. *Journal of Machine Learning Research*, pages 2651–2667, 2006.

O. Kounchev. *Multivariate Polysplines: Applications to Numerical and Wavelet Analysis*. Elsevier Science, 2001.

J. Bradbury, R. Frostig, P. Hawkins, M. Johnson, C. Leary, D. Maclaurin, and S. Wanderman-Milne. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015.

R. Novak, L. Xiao, J. Lee, Y. Bahri, G. Yang, J. Hron, D. A Abolafia, J. Pennington, and J. Sohl-Dickstein. Bayesian deep convolutional networks with many channels are Gaussian processes. In *International Conference on Learning Representations*, 2019.

C. Wang, G. Zhang, and R. Grosse. Picking winning tickets before training by preserving gradient flow. In *International Conference on Learning Representations*, 2020.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8026–8037, 2019.

S De and SL Smith. Batch normalization biases residual blocks towards the identity function in deep networks. *Advances in Neural Information Processing Systems*, 2020.

H. Zhang, Y. N Dauphin, and T. Ma. Fixup initialization: Residual learning without normalization. In *International Conference on Learning Representations*, 2019.

T.M. MacRobert. *Spherical Harmonics: An Elementary Treatise on Harmonic Functions with Applications*. Pergamon Press, 1967.