Appendix

0 Setup and notations

0.1 Neural Tangent Kernel

Consider a neural network model consisting of L layers $(y^l)_{1\leq l\leq L}$, with $y^l : \mathbb{R}^{n_{l-1}} \to \mathbb{R}^{n_l}$, $n_0 = d$ and let $\theta = (\theta^l)_{1\leq l\leq L}$ be the flattened vector of weights and bias indexed by the layer's index and p be the dimension of θ. Recall that θ^l has dimension $n_l + 1$. The output f of the neural network is given by some transformation $s : \mathbb{R}^{n_L} \to \mathbb{R}^o$ of the last layer $y^L(x)$; o being the dimension of the output (e.g. number of classes for a classification problem). For any input $x \in \mathbb{R}^d$, we thus have $f(x, \theta) = s(y^L(x)) \in \mathbb{R}^o$. As we train the model, θ changes with time t and we denote by θ_t the value of θ at time t and $f_t(x) = f(x, \theta_t) = (f_j(x, \theta_t), j \leq o)$. Let $D = (x_i, z_i)_{1\leq i\leq N}$ be the data set and let $\mathcal{X} = (x_i)_{1\leq i\leq N}$, $\mathcal{Z} = (z_j)_{1\leq j\leq N}$ be the matrices of input and output respectively, with dimension $d \times N$ and $o \times N$. For any function $g : \mathbb{R}^{d\times o} \to \mathbb{R}^k$, $k \geq 1$, we denote by $g(\mathcal{X}, \mathcal{Z})$ the matrix $(g(x_i, z_i))_{1\leq i\leq N}$ of dimension $k \times N$.

Jacot et al. (2018) studied the behaviour of the output of the neural network as a function of the training time t when the network is trained using a gradient descent algorithm. Lee et al. (2019) built on this result to linearize the training dynamics. We recall hereafter some of these results.

For a given θ, the empirical loss is given by $\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^N \ell(f(x_i, \theta), z_i)$. The full batch GD algorithm is given by

$$\hat{\theta}_{t+1} = \hat{\theta}_t - \eta\nabla_\theta\mathcal{L}(\hat{\theta}_t), \tag{1}$$

where $\eta > 0$ is the learning rate.
Let $T > 0$ be the training time and $N_s = T/\eta$ be the number of steps of the discrete GD (1). The continuous time system equivalent to (1) with step $\Delta t = \eta$ is given by

$$d\theta_t = -\nabla_\theta\mathcal{L}(\theta_t)dt. \tag{2}$$

This differs from the result by Lee et al. (2019) since we use a discretization step of $\Delta t = \eta$. It is well known that this discretization scheme leads to an error of order $\mathcal{O}(\eta)$ (see Appendix). Equation (2) can be re-written as

$$d\theta_t = -\frac{1}{N}\nabla_\theta f(\mathcal{X}, \theta_t)^T\nabla_{z'}\ell(f(\mathcal{X}, \theta_t), \mathcal{Z})dt.$$

where $\nabla_\theta f(\mathcal{X}, \theta_t)$ is a matrix of dimension $oN \times p$ and $\nabla_{z'}\ell(f(\mathcal{X}, \theta_t), \mathcal{Z})$ is the flattened vector of dimension oN constructed from the concatenation of the vectors $\nabla_{z'}\ell(z', z_i)_{|z'=f(x_i,\theta_t)}, i \leq N$. As a result, the output function $f_t(x) = f(x, \theta_t) \in \mathbb{R}^o$ satisfies the following ODE

$$df_t(x) = -\frac{1}{N}\nabla_\theta f(x, \theta_t)\nabla_\theta f(\mathcal{X}, \theta_t)^T\nabla_{z'}\ell(f_t(\mathcal{X}), \mathcal{Z})dt. \tag{3}$$

The Neural Tangent Kernel (NTK) K^L_θ is defined as the $o \times o$ dimensional kernel satisfying: for all $x, x' \in \mathbb{R}^d$,

$$\begin{aligned} K^L_{\theta_t}(x, x') &= \nabla_\theta f(x, \theta_t)\nabla_\theta f(x', \theta_t)^T \in \mathbb{R}^{o\times o} \\ &= \sum_{l=1}^L \nabla_{\theta^l} f(x, \theta_t)\nabla_{\theta^l} f(x', \theta_t)^T. \end{aligned} \tag{4}$$

We also define $K^L_{\theta_t}(\mathcal{X}, \mathcal{X})$ as the $oN \times oN$ matrix defined blockwise by

$$K^L_{\theta_t}(\mathcal{X}, \mathcal{X}) = \begin{pmatrix} K^L_{\theta_t}(x_1, x_1) & \cdots & K^L_{\theta_t}(x_1, x_N) \\ K^L_{\theta_t}(x_2, x_1) & \cdots & K^L_{\theta_t}(x_2, x_N) \\ \vdots & \ddots & \vdots \\ K^L_{\theta_t}(x_N, x_1) & \cdots & K^L_{\theta_t}(x_N, x_N) \end{pmatrix}.$$

By applying (3) to the vector $\mathcal{X}$, one obtains

$$df_t(\mathcal{X}) = -\frac{1}{N}K^L_{\theta_t}(\mathcal{X}, \mathcal{X})\nabla_{z'}\ell(f_t(\mathcal{X}), \mathcal{Z})dt, \tag{5}$$

meaning that for all $j \leq N$

$$df_t(x_j) = -\frac{1}{N}K^L_{\theta_t}(x_j, \mathcal{X})\nabla_{z'}\ell(f_t(\mathcal{X}), \mathcal{Z})dt.$$

Infinite width dynamics. In the case of an FFNN, Jacot et al. (2018) proved that, with GD, the kernel $K^L_{\theta_t}$ converges to a kernel K^L which depends only on L (number of layers) for all $t < T$ when $n_1, n_2, ..., n_L \to \infty$, where T is an upper bound on the training time, under the technical assumption $\int_0^T ||\nabla_z \ell(f_t(\mathcal{X}, \mathcal{Z}))||_2 dt < \infty$ a.s. with respect to the initialization weights. The infinite width limit of the training dynamics is given by

$$df_t(\mathcal{X}) = -\frac{1}{N} K^L(\mathcal{X}, \mathcal{X}) \nabla_{z'} \ell(f_t(\mathcal{X}), \mathcal{Z}) dt, \tag{6}$$

We note hereafter $\hat{K}^L = K^L(\mathcal{X}, \mathcal{X})$. As an example, with the quadratic loss $\ell(z', z) = \frac{1}{2}||z' - z||^2$, (6) is equivalent to

$$df_t(\mathcal{X}) = -\frac{1}{N} \hat{K}^L (f_t(\mathcal{X}) - \mathcal{Z}) dt, \tag{7}$$

which is a simple linear model that has a closed-form solution given by

$$f_t(\mathcal{X}) = e^{-\frac{1}{N}\hat{K}^L t} f_0(\mathcal{X}) + (I - e^{-\frac{1}{N}\hat{K}^L t})\mathcal{Z}. \tag{8}$$

For general input $x \in \mathbb{R}^d$, we have

$$f_t(x) = f_0(x) + \gamma(x, \mathcal{X})(I - e^{-\frac{1}{N}\hat{K}^L t})(\mathcal{Z} - f_0(\mathcal{X})). \tag{9}$$

where $\gamma(x) = K^L(x, \mathcal{X}) K^L(\mathcal{X}, \mathcal{X})^{-1}$.

0.2 The Architecture

Let ϕ be the ReLU activation function. We consider the following architecture

FeedForward Fully-Connected Neural Network (FFNN). Consider an FFNN of depth L, widths $(n_l)_{1 \le l \le L}$, weights w^l and bias b^l. For some input $x \in \mathbb{R}^d$, the forward propagation using the NTK parameterization is given by

$$\begin{aligned} y^1_i(x) &= \frac{\sigma_w}{\sqrt{d}} \sum_{j=1}^{d} w^1_{ij} x_j + \sigma_b b^1_i \\ y^l_i(x) &= \frac{\sigma_w}{\sqrt{n_{l-1}}} \sum_{j=1}^{n_{l-1}} w^l_{ij} \phi(y^{l-1}_j(x)) + \sigma_b b^l_i, \quad l \ge 2. \end{aligned} \tag{10}$$

1 Proof techniques

The techniques used in the proofs range from simple algebraic manipulation to tricky inequalities.

Lemmas 1, 2. The proofs of these lemmas are simple and follow the same inductive argument as in the proof of the original NTK result in Jacot et al. (2018). Note that these results can also be obtained by simple application of the Master Theorem in Yang (2020) using the framework of Tensor Programs.

Proposition 1, Theorems 1, 2. The proof of these results follow two steps; Firstly, estimating the asymptotic behaviour of the NTK in the limit of large depth; secondly, controling these behaviour using upper/lower bounds. We analyse the asymptotic behaviour of the NTK of FFNN using existing results on signal propagation in deep FFNN.
It is relatively easy to control the dynamics of the NTK in the Ordered/Chaotic phase, however, the dynamics become a bit complicated on the Edge of Chaos and technical lemmas which we call Appendix Lemmas are introduced for this purpose.

Proposition 2. The spectral decomposition of zonal kernels on the sphere is a classical result in spectral theory which was recently applied to Neural Tangent Kernel Geifman et al. (2020); Cao et al. (2020); **?**. In order to prove the convergence of the eigenvalues, we use Dominated Convergence Theorem, leveraging the asymptotic results in Proposition 1 and Theorems 1, 2.

2 The infinite width limit

2.1 Forward propagation

FeedForward Neural Network. For some input $x \in \mathbb{R}^d$, the propagation of this input within the network is given by

$$y_i^1(x) = \frac{\sigma_w}{\sqrt{d}} \sum_{j=1}^{d} w_{ij}^1 x_j + \sigma_b b_i^1$$

$$y_i^l(x) = \frac{\sigma_w}{\sqrt{n_{l-1}}} \sum_{j=1}^{n_{l-1}} w_{ij}^l \phi(y_j^{l-1}(x)) + \sigma_b b_i^l, \quad l \geq 2$$

Where $\phi : \mathbb{R} \to \mathbb{R}$ is the activation function. When we take the limit $n_{l-1} \to \infty$ recursively over l, this implies, using Central Limit Theorem, that $y_i^l(x)$ is a Gaussian variable for any input x. This gives an error of order $\mathcal{O}(1/\sqrt{n_{l-1}})$ (standard Monte Carlo error). More generally, an approximation of the random process $y_i^l(.)$ by a Gaussian process was first proposed by Neal (1995) in the single layer case and has been extended to the multiple layer case by Lee et al. (2018) and Matthews et al. (2018). The limiting Gaussian process kernels follow a recursive formula given by, for any inputs $x, x' \in \mathbb{R}^d$

$$\begin{aligned}\kappa^l(x, x') &= \mathbb{E}[y_i^l(x) y_i^l(x')] \\ &= \sigma_b^2 + \sigma_w^2 \mathbb{E}[\phi(y_i^{l-1}(x))\phi(y_i^{l-1}(x'))] \\ &= \sigma_b^2 + \sigma_w^2 \Psi_\phi(\kappa^{l-1}(x, x), \kappa^{l-1}(x, x'), \kappa^{l-1}(x', x')),\end{aligned}$$

where Ψ_ϕ is a function that only depends on ϕ. This provides a simple recursive formula for the computation of the kernel κ^l; see, e.g., Lee et al. (2018) for more details.

Residual Neural Networks. The infinite width limit approximation for ResNet yields similar results with an additional residual terms. It is straighforward to see that, in the case of a ResNet with FFNN-type layers, we have that

$$\kappa^l(x, x') = \kappa^{l-1}(x, x') + \sigma_b^2 + \sigma_w^2 F_\phi(\kappa^{l-1}(x, x), \kappa^{l-1}(x, x'), \kappa^{l-1}(x', x')),$$

2.2 Gradient Independence

In the mean-field literature of DNNs, an omnipresent approximation in prior literature is that of the gradient independence which is similar in nature to the practice of feedback alignment (Lillicrap et al., 2016). This approximation states that, for wide neural networks, the weights used for forward propagation are independent from those used for back-propagation. When used for the computation of Neural Tangent Kernel, this approximation was proven to give the exact computation for standard architectures such as FFNN, CNN and ResNets Yang (2020) (Theorem D.1).

This result has been extensively used in the literature as an approximation before being proved to yields exact computation for the NTK, and theoretical results derived under this approximation were verified empirically; see references below.

Gradient Covariance back-propagation. Analytical formulas for gradient covariance back-propagation were derived using this result, in (Hayou et al., 2019; Schoenholz et al., 2017; Yang and Schoenholz, 2017b; Lee et al., 2018; Poole et al., 2016; Xiao et al., 2018; Yang, 2019). Empirical results showed an excellent match for FFNN in Schoenholz et al. (2017), for Resnets in Yang (2019) and for CNN in Xiao et al. (2018).

Neural Tangent Kernel. The Gradient Independence approximation was implicitly used in Jacot et al. (2018) to derive the infinite width Neural Tangent Kernel (See Jacot et al. (2018), Appendix A.1). Authors have found that this infinite width NTK computed with the Gradient Independence approximation yields excellent match with empirical (exact) NTK.

We use this result in our proofs and we refer to it simply by the Gradient Independence.

3 Warmup: Results from the theory of signal propagation in DNNs

3.1 Notation

For FFNN layers, let $q^l(x) := q^l(x, x)$ be the variance of $y_1^l(x)$ (the choice of the index 1 is not important since, in the infinite width limit, the random variables $(y_i^l(x))_{i \in [1:N_l]}$ are iid). Let $q^l(x, x')$, resp. $c_1^l(x, x')$ be

the covariance, resp. the correlation between $y_1^l(x)$ and $y_1^l(x')$. For Gradient back-propagation, let $\tilde{q}^l(x,x')$ be the Gradient covariance defined by $\tilde{q}^l(x,x') = \mathbb{E}\left[\frac{\partial \mathcal{L}}{\partial y_1^l}(x)\frac{\partial \mathcal{L}}{\partial y_1^l}(x')\right]$ where $\mathcal{L}$ is some loss function. Similarly, let $\tilde{q}^l(x)$ be the Gradient variance at point x. We also define $\dot{q}^l(x,x') = \sigma_w^2\mathbb{E}[\phi'(y_1^{l-1}(x))\phi'(y_1^{l-1}(x'))]$.

3.1.1 Covariance propagation

Covariance propagation for FFNN. In Section 2.1, we derived the covariance kernel propagation in an FFNN. For two inputs $x, x' \in \mathbb{R}^d$, we have

$$q^l(x,x') = \sigma_b^2 + \sigma_w^2\mathbb{E}[\phi(y_i^{l-1}(x))\phi(y_i^{l-1}(x'))] \tag{11}$$

this can be written as

$$q^l(x,x') = \sigma_b^2 + \sigma_w^2\mathbb{E}\left[\phi\left(\sqrt{q^l(x)}Z_1\right)\phi\left(\sqrt{q^l(x')}(c^{l-1}Z_1 + \sqrt{1-(c^{l-1})^2}Z_2\right)\right], \quad Z_1, Z_2 \overset{iid}{\sim} \mathcal{N}(0,1),$$

with $c^{l-1} := c^{l-1}(x,x')$.
With ReLU, and since ReLU is positively homogeneous (i.e. $\phi(\lambda x) = \lambda\phi(x)$ for $\lambda \geq 0$), we have that

$$q^l(x,x') = \sigma_b^2 + \frac{\sigma_w^2}{2}\sqrt{q^l(x)}\sqrt{q^l(x')}f(c^{l-1})$$

where f is the ReLU correlation function given by Hayou et al. (2019)

$$f(c) = \frac{1}{\pi}(c\arcsin c + \sqrt{1-c^2}) + \frac{1}{2}c.$$

Covariance propagation for ResNet with ReLU. In the case of ResNet, only an added residual term shows up in the recursive formula. For a ResNet with FFNN layers, the recursion reads

$$q^l(x,x') = q^{l-1}(x,x') + \sigma_b^2 + \frac{\sigma_w^2}{2}\sqrt{q^l(x)}\sqrt{q^l(x')}f(c^{l-1}) \tag{12}$$

3.1.2 Gradient Covariance back-propagation

Gradient back-propagation for FFNN. The gradient back-propagation is given by

$$\frac{\partial \mathcal{L}}{\partial y_i^l} = \phi'(y_i^l)\sum_{j=1}^{N_{l+1}}\frac{\partial \mathcal{L}}{\partial y_j^{l+1}}W_{ji}^{l+1}.$$

where $\mathcal{L}$ is some loss function. Using the Gradient Independence 2.2, we have as in Schoenholz et al. (2017)

$$\tilde{q}^l(x) = \tilde{q}^{l+1}(x)\frac{N_{l+1}}{N_l}\chi(q^l(x)).$$

where $\chi(q^l(x)) = \sigma_w^2\mathbb{E}[\phi(\sqrt{q^l(x)}Z)^2]$.

3.1.3 Edge of Chaos (EOC)

Let $x \in \mathbb{R}^d$ be an input. The convergence of $q^l(x)$ as l increases has been studied by Schoenholz et al. (2017) and Hayou et al. (2019). In particular, under weak regularity conditions, it is proven that $q^l(x)$ converges to a point $q(\sigma_b, \sigma_w) > 0$ independent of x as $l \to \infty$. The asymptotic behaviour of the correlations $c^l(x,x')$ between $y^l(x)$ and $y^l(x')$ for any two inputs x and x' is also driven by (σ_b, σ_w): the dynamics of c^l is controlled by a function f i.e. $c^{l+1} = f(c^l)$ called the correlation function. The authors define the EOC as the set of parameters (σ_b, σ_w) such that $\sigma_w^2\mathbb{E}[\phi'(\sqrt{q(\sigma_b,\sigma_w)}Z)^2] = 1$ where $Z \sim \mathcal{N}(0,1)$. Similarly the Ordered, resp. Chaotic, phase is defined by $\sigma_w^2\mathbb{E}[\phi'(\sqrt{q(\sigma_b,\sigma_w)}Z)^2] < 1$, resp. $\sigma_w^2\mathbb{E}[\phi'(\sqrt{q(\sigma_b,\sigma_w)}Z)^2] > 1$. On the Ordered phase, the gradient will vanish as it backpropagates through the network, and the correlation $c^l(x,x')$ converges exponentially to 1. Hence the output function becomes constant (hence the name 'Ordered phase'). On the Chaotic phase, the gradient explodes and the correlation converges exponentially to some limiting value $c < 1$ which results in the output function being discontinuous everywhere (hence the 'Chaotic' phase name). On the EOC, the second moment of the gradient remains constant throughout the backpropagation and the correlation converges to 1 at a sub-exponential rate, which allows deeper information propagation. Hereafter, f **will always refer to the correlation function**.

We initialize the model with $w_{ij}^l, b_i^l \overset{iid}{\sim} \mathcal{N}(0,1)$, where $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution of mean μ and variance σ^2. In the remainder of this appendix, we assume that the following conditions are satisfied

- The input data is a subset of a compact set E of $\mathbb{R}^d$, and no two inputs are co-linear.
- All results are derived in the limit of infinitely wide networks.

3.2 Some results from the information propagation theory

Results for FFNN with Tanh activation.

Fact 1. *For any choice of $\sigma_b, \sigma_w \in \mathbb{R}^+$, there exist $q, \lambda > 0$ such that for all $l \geq 1$, $\sup_{x \in \mathbb{R}^d} |q^l(x,x) - q| \leq e^{-\lambda l}$. (Equation (3) and conclusion right after in Schoenholz et al. (2017)).*

Fact 2. *On the Ordered phase, there exists $\gamma > 0$ such that $\sup_{x,x' \in \mathbb{R}^d} |c^l(x,x') - 1| \leq e^{-\gamma l}$. (Equation (8) in Schoenholz et al. (2017))*

Fact 3. *Let $(\sigma_b, \sigma_w) \in$ EOC. Using the same notation as in fact 4, we have that $\sup_{(x,x') \in B_\epsilon} |1 - c^l(x,x')| = \mathcal{O}(l^{-1})$. (Proposition 3 in Hayou et al. (2019)).*

Fact 4. *Let $B_\epsilon = \{(x,x') \in \mathbb{R}^d : c^1(x,x') < 1 - \epsilon\}$. On the chaotic phase, there exist $c < 1$ such that for all $\epsilon \in (0,1)$, there exists $\gamma > 0$ such that $\sup_{(x,x') \in B_\epsilon} |c^l(x,x') - c| \leq e^{-\gamma l}$. (Equations (8) and (9) in Schoenholz et al. (2017))*

Fact 5 (Correlation function)**.** *The correlation function f is defined by $f(x) = \frac{\sigma_b^2 + \sigma_w^2 \mathbb{E}[\phi(\sqrt{q}Z_1)\phi(\sqrt{q}(xZ_1 + \sqrt{1-x^2}Z_2))]}{q}$ where q is given in Fact 1 and Z_1, Z_2 are iid standard Gaussian variables.*

Fact 6. *f has a derivative of any order $j \geq 1$ given by*

$$f^{(j)}(x) = \sigma_w^2 q^{j-1} \mathbb{E}[\phi^{(j)}(Z_1)\phi^{(j)}(xZ_1 + \sqrt{1-x^2}Z_2)], \quad \forall x \in [-1,1]$$

As a result, we have that $f^{(j)}(1) = \sigma_w^2 q^{j-1} \mathbb{E}[\phi^{(j)}(Z_1)^2] > 0$ for all $j \geq 1$.

The proof of the previous fact is straightforward following the same integration by parts technique as in the proof of Lemma 1 in Hayou et al. (2019). The result follows by induction.

Fact 7. *Let $(\sigma_b, \sigma_w) \in$ EOC. We have that $f'(1) = 1$ (by definition of EOC). As a result, the Taylor expansion of f near 1 is given by*

$$f(c) = c + \alpha(1-c)^2 - \zeta(1-c)^3 + O((1-c)^4).$$

where $\alpha, \zeta > 0$.

Proof. The proof is straightforward using fact 6, and integral-derivative interchanging. □

Results for FFNN with ReLU activation.

Fact 8. *The ordered phase for ReLU is given by Ord $= \{(\sigma_b, \sigma_w) \in (\mathbb{R}^+)^2 : \sigma_w < \sqrt{2}\}$. Moreover, for any $(\sigma_b, \sigma_w) \in$ Ord, there exist λ such that for all $l \geq 1$, $\sup_{x \in \mathbb{R}^d} |q^l(x,x) - q| \leq e^{-\lambda l}$, where $q = \frac{\sigma_b^2}{1 - \sigma_w^2/2}$.*

The proof is straightforward using equation (11).

Fact 9. *For any (σ_b, σ_w) in the Ordered phase, there exist λ such that for all $l \geq 1$, $\sup_{(x,x') \in \mathbb{R}^d} |c^l(x,x') - 1| \leq e^{-\lambda l}$.*

The proof of this claim follows from standard Banach Fixed point theorem in the same fashion as for Tanh in Schoenholz et al. (2017).

Fact 10. *The Chaotic phase for ReLU is given by Ch $= \{(\sigma_b, \sigma_w) \in (\mathbb{R}^+)^2 : \sigma_w > \sqrt{2}\}$. Moreover, for any $(\sigma_b, \sigma_w) \in$ Ch, for all $l \geq 1$, $x \in \mathbb{R}^d$, $q^l(x,x) \gtrsim (\sigma_w^2/2)^l$.*

The variance explodes exponentially on the Chaotic phase, which means the output of the Neural Network can grow arbitrarily in this setting. Hereafter, when no activation function is mentioned, and when we choose "(σ_b, σ_w) on the Ordered/Chaotic phase", it should be interpreted as "(σ_b, σ_w) on the Ordered phase" for ReLU and "(σ_b, σ_w) on the Ordered/Chaotic phase" for Tanh.

Fact 11. *For ReLU FFNN on the EOC, we have that $q^l(x,x) = \frac{\sigma_w^2}{d}||x||^2$ for all $l \geq 1$.*

The proof is straightforward using equation 11 and that $(\sigma_b, \sigma_w) = (0, \sqrt{2})$ on the EOC.

Fact 12. *The EOC of ReLU is given by the singleton* $\{(\sigma_b, \sigma_w) = (0, \sqrt{2})\}$. *In this case, the correlation function of an FFNN with ReLU is given by*

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

(Proof of Proposition 1 in Hayou et al. (2019)).

Fact 13. *Let* $(\sigma_b, \sigma_w) \in EOC$. *Using the same notation as in fact 4, we have that*

$$\sup_{(x,x') \in B_\epsilon} |1 - c^l(x, x')| = \mathcal{O}(l^{-2})$$

(Follows straightforwardly from Proposition 1 in Hayou et al. (2019)).

Fact 14. *We have that*

$$f(c) = c + s(1 - c)^{3/2} + b(1 - c)^{5/2} + O((1 - c)^{7/2} \tag{13}$$

with $s = \frac{2\sqrt{2}}{3\pi}$ *and* $b = \frac{\sqrt{2}}{30\pi}$.

This result was proven in Hayou et al. (2019) (in the proof of Proposition 1) for order $3/2$, the only difference is that here we push the expansion to order $5/2$.

General results on the correlation function.

Fact 15. *Let* f *be either the correlation function of Tanh or ReLU. We have that*

- $f(1) = 1$ *(Lemma 2 in Hayou et al. (2019)).*
- *On the ordered phase* $0 < f'(1) < 1$ *(By definition).*
- *On the Chaotic phase* $f'(1) > 1$ *(By definition).*
- *On the EOC,* $f'(1) = 1$ *(By definition).*
- *On the Ordered phase and the EOC,* 1 *is the unique fixed point of* f *(Hayou et al. (2019)).*
- *On the Chaotic phase,* f *has two fixed points,* 1 *which is unstable, and* $c < 1$ *which is a stable fixed point Schoenholz et al. (2017).*

Fact 16. *Let* $\epsilon \in (0, 1)$. *On the Ordered/Chaotic phase, with either ReLU or Tanh, there exists* $\alpha \in (0, 1)$,$\gamma > 0$ *such that*

$$\sup_{(x,x') \in B_\epsilon} |f'(c^l(x, x')) - \alpha| \leq e^{-\gamma l}$$

Proof. This result follows from a simple first order expansion inequality. For Tanh on the Ordered phase, we have that

$$\sup_{(x,x') \in B_\epsilon} |f'(c^l(x, x')) - f'(1)| \leq \zeta_l \sup_{(x,x') \in B_\epsilon} |c^l(x, x') - 1|$$

where $\zeta_l = \sup_{t \in (\min_{(x,x') \in B_\epsilon} c^l(x,x'), 1)} |f''(t)| \to |f''(1)|$. We conclude for Ordered phase with Tanh using fact 2. The same argument can be used for Chaotic phase with Tanh using fact 4; in this case, $\alpha = f'(c)$ where c is the unique stable fixed point of the correlation function f.

On the Ordered phase with ReLU, let $\tilde{f}$ be the correlation function. It is easy to see that $\tilde{f}'(c) = \frac{\sigma_w^2}{2} f'(c)$ where f is given in fact 12. $f'(x) = 1 - \frac{\sqrt{2}}{\pi}(1 - x)^{1/2} + \mathcal{O}((1 - x)^{3/2})$. Therefore, there exists $l_0, \zeta > 0$ such that for $l > l_0$,

$$\sup_{(x,x') \in B_\epsilon} |f'(c^l(x, x')) - f'(1)| \leq \zeta \sup_{(x,x') \in B_\epsilon} |c^l(x, x') - 1|^{1/2}$$

We conclude using fact 9. □

Asymptotic behaviour of the correlation in FFNN.

Appendix Lemma 1 (Asymptotic behaviour of c^l for ReLU)**.** *Let* $(\sigma_b, \sigma_w) \in EOC$ *and* $\epsilon \in (0, 1)$. *There exist universal constants* $\kappa, \kappa', \kappa'' > 0$ *(that do not depend on any parameter) such that*

$$\sup_{(x,x') \in B_\epsilon} \left| c^l(x, x') - 1 + \frac{\kappa}{l^2} - \kappa' \frac{\log(l)}{l^3} \right| = \mathcal{O}(l^{-3})$$

and,

$$\sup_{(x,x') \in B_\epsilon} \left| f'(c^l(x, x')) - 1 + \frac{3}{l} - \kappa'' \frac{\log(l)}{l^2} \right| = \mathcal{O}(l^{-2}).$$

Proof. Let $(x, x') \in B_\epsilon$ and $s = \frac{2\sqrt{2}}{3\pi}$. From the preliminary results, we have that $\lim_{l\to\infty} \sup_{x,x'\in\mathbb{R}^d} 1 - c^l(x, x') = 0$ (fact 13). Using fact 14, we have uniformly over B_ϵ,

$$\gamma_{l+1} = \gamma_l - s\gamma_l^{3/2} - b\gamma_l^{5/2} + O(\gamma_l^{7/2})$$

where $s, b > 0$, this yields

$$\gamma_{l+1}^{-1/2} = \gamma_l^{-1/2} + \frac{s}{2} + \frac{3s^2}{8}\gamma_l^{1/2} + (\frac{b}{2} + \frac{5}{16}s^3)\gamma_l + O(\gamma_l^{3/2}).$$

Thus, letting $b' = \frac{b}{2} + \frac{5}{16}s^3$, as l goes to infinity

$$\gamma_{l+1}^{-1/2} - \gamma_l^{-1/2} \sim \frac{s}{2},$$

and by summing and equivalence of positive divergent series

$$\gamma_l^{-1/2} \sim \frac{s}{2}l.$$

Moreover, since $\gamma_{l+1}^{-1/2} = \gamma_l^{-1/2} + \frac{s}{2} + \frac{3s^2}{8}\gamma_l^{1/2} + b'\gamma_l + O(\gamma_l^{3/2})$, using the same argument multiple times and inverting the formula yields

$$c^l(x, x') = 1 - \frac{\kappa}{l^2} + \kappa'\frac{\log(l)}{l^3} + \mathcal{O}(l^{-3})$$

where $\kappa = \frac{9\pi^2}{2}$. Note that, by Appendix Lemma 3 (section **??**), the $\mathcal{O}$ bound can be chosen in a way that it does not depend on (x, x'), it depends only on ϵ; this concludes the proof for the first part of the result.
Using fact 12, we have that

$$\begin{aligned} f'(x) &= \frac{1}{\pi}\arcsin(x) + \frac{1}{2} \\ &= 1 - \frac{\sqrt{2}}{\pi}(1-x)^{1/2} + O((1-x)^{3/2}). \end{aligned}$$

Thus, it follows that

$$f'(c^l(x, x')) = 1 - \frac{3}{l} + \kappa''\frac{\log(l)}{l^2} + \mathcal{O}(l^{-2}).$$

for some universal constant κ'' uniformly over the set B_ϵ, which concludes the proof.

□

We prove a similar result for an FFNN with Tanh activation.

Appendix Lemma 2 (Asymptotic behaviour of c^l for Tanh)**.** *Let $(\sigma_b, \sigma_w) \in EOC$ and $\epsilon \in (0, 1)$. We have*

$$\sup_{(x,x')\in B_\epsilon} \left| c^l(x, x') - 1 + \frac{\kappa}{l} - \kappa(1 - \kappa^2\zeta)\frac{\log(l)}{l^3} \right| = \mathcal{O}(l^{-3})$$

where $\kappa = \frac{2}{f''(1)} > 0$ and $\zeta = \frac{f^3(1)}{6} > 0$. Moreover, we have that

$$\sup_{(x,x')\in B_\epsilon} \left| f'(c^l(x, x')) - 1 + \frac{2}{l} - 2(1 - \kappa^2\zeta)\frac{\log(l)}{l^2} \right| = \mathcal{O}(l^{-2}).$$

Proof. Let $(x, x') \in B_\epsilon$ and $\lambda_l := 1 - c^l(x, x')$. Using a Taylor expansion of f near 1 (fact 7), there exist $\alpha, \zeta > 0$ such that

$$\lambda_{l+1} = \lambda_l - \alpha\lambda_l^2 + \zeta\lambda_l^3 + O(\lambda_l^4)$$

Here also, we use the same technique as in the previous lemma. We have that

$$\begin{aligned} \lambda_{l+1}^{-1} &= \lambda_l^{-1}(1 - \alpha\lambda_l + \zeta\lambda^2 + O(\lambda_l^3))^{-1} = \lambda_l^{-1}(1 + \alpha\lambda_l + (\alpha^2 - \zeta)\lambda_l^2 + O(\lambda_l^3)) \\ &= \lambda_l^{-1} + \alpha + (\alpha^2 - \zeta)\lambda_l + O(\lambda_l^2). \end{aligned}$$

By summing (divergent series), we have that $\lambda_l^{-1} \sim \alpha l$. Therefore,

$$\lambda_{l+1}^{-1} - \lambda_l^{-1} - \alpha = (\alpha^2 - \beta)\alpha^{-1}l^{-1} + o(l^{-1})$$

By summing a second time, we obtain

$$\lambda_l^{-1} = \alpha l + (\alpha - \beta\alpha^{-1})\log(l) + o(\log(l)),$$

Using the same technique once again, we obtain

$$\lambda_l^{-1} = \alpha l + (\alpha - \beta\alpha^{-1})\log(l) + O(1).$$

This yields

$$\lambda_l = \alpha^{-1}l^{-1} - \alpha^{-1}(1-\alpha^{-2}\beta)\frac{\log(l)}{l^2} + O(l^{-2}).$$

In a similar fashion to the previous proof, we can force the upper bound in $\mathcal{O}$ to be independent of x using Appendix Lemma 3. This way, the bound depends only on ϵ. This concludes the first part of the proof.

For the second part, observe that $f'(x) = 1 + (x-1)f''(1) + O((x-1)^2)$, hence

$$f'(c^l(x,x')) = 1 - \frac{2}{l} + 2(1-\alpha^{-2}\zeta)\frac{\log(l)}{l^2} + O(l^{-2})$$

which concludes the proof. □

4 A technical lemma for the derivation of uniform bounds

Results in Theorem 1 and 2 and Proposition 1 involve a supremum over the set B_ϵ. To obtain such results, we need a 'uniform' Taylor analysis of the correlation $c^l(x,x')$ (see the next section) where uniformity is over $(x,x') \in B_\epsilon$. It turns out that such result is trivial when the correlation follows a dynamical system that is controlled by a non-decreasing function. We clarify this in the next lemma.

Appendix Lemma 3 (Uniform Bounds)**.** *Let $A \subset \mathbb{R}$ be a compact set and g a non-decreasing function on A. Define the sequence ζ_l by $\zeta_l = g(\zeta_{l-1})$ and $\zeta_0 \in A$. Assume that there exist α_l, β_l that do not depend on ζ_l, with $\beta_l = o(\alpha_l)$, such that for all $\zeta_0 \in A$,*

$$\zeta_l = \alpha_l + \mathcal{O}_{\zeta_0}(\beta_l)$$

where $\mathcal{O}_{\zeta_0}$ means that the $\mathcal{O}$ bound depends on ζ_0. Then, we have that

$$\sup_{\zeta_0 \in A} |\zeta_l - \alpha_l| = \mathcal{O}(\beta_l)$$

i.e. we can choose the bound $\mathcal{O}$ to be independent of ζ_0.

Proof. Let $\zeta_{0,m} = \min A$ and $\zeta_{0,M} = \max A$. Let $(\zeta_{m,l})$ and $(\zeta_{M,l})$ be the corresponding sequences. Since g is non-decreasing, we have that for all $\zeta_0 \in A$, $\zeta_{m,l} \leq \zeta_l \leq \zeta_{M,l}$. Moreover, by assumption, there exists $M_1, M_2 > 0$ such that

$$|\zeta_{m,l} - \alpha_l| \leq M_1|\beta_l|$$

and

$$|\zeta_{M,l} - \alpha_l| \leq M_2|\beta_l|$$

therefore,

$$|\zeta_l - \alpha_l| \leq \max(|\zeta_{m,l} - \alpha_l|, |\zeta_{M,l} - \alpha_l|) \leq \max(M_1, M_2)|\beta_l|$$

which concludes the proof. □

Note that Appendix Lemma 3 can be easily extended to Taylor expansions with 'o' instead of '$\mathcal{O}$'. We will use this result in the proofs, by refereeing to Appendix Lemma 3.

5 Proofs of Section 3: Large Depth Behaviour of Neural Tangent Kernel

5.1 Proofs of the results of Section 3.1

In this section, we provide proofs for the results of Section 3.1 in the paper.

Recall that Lemma 1 in the paper is a generalization of Theorem 1 in Jacot et al. (2018) and is reminded here. The proof is simple and follows similar induction techniques as in Jacot et al. (2018).

Lemma 1 (Generalization of Th. 1 in Jacot et al. (2018))**.** *Consider an FFNN of the form (3). Then, as $n_1, n_2, ..., n_{L-1} \to \infty$, we have for all $x, x' \in \mathbb{R}^d$, $i, i' \leq n_L$, $K^L_{ii'}(x,x') = \delta_{ii'}K^L(x,x')$, where $K^L(x,x')$ is given by the recursive formula*

$$K^L(x,x') = \dot{q}^L(x,x')K^{L-1}(x,x') + q^L(x,x'),$$

where $q^l(x,x') = \sigma_b^2 + \sigma_w^2\mathbb{E}[\phi(y_1^{l-1}(x))\phi(y_1^{l-1}(x'))]$ and $\dot{q}^l(x,x') = \sigma_w^2\mathbb{E}[\phi'(y_1^{l-1}(x))\phi'(y_1^{l-1}(x'))]$.

Proof. The proof for general σ_w is similar to when $\sigma_w = 1$ (Jacot et al. (2018)) which is a proof by induction.

For $l \geq 2$ and $i \in [1 : n_l]$

$$\partial_{\theta_{1:l}} y_i^{l+1}(x) = \frac{\sigma_w}{\sqrt{n_l}} \sum_{j=1}^{n_l} w_{ij}^{l+1} \phi'(y_j^l(x)) \partial_{\theta_{1:l}} y_j^l(x).$$

Therefore,

$$(\partial_{\theta_{1:l}} y_i^{l+1}(x))(\partial_{\theta_{1:l}} y_i^{l+1}(x'))^t = \frac{\sigma_w^2}{n_l} \sum_{j,j'}^{n_l} w_{ij}^{l+1} w_{ij'}^{l+1} \phi'(y_j^l(x)) \phi'(y_{j'}^l(x')) \partial_{\theta_{1:l}} y_j^l(x) (\partial_{\theta_{1:l}} y_{j'}^l(x'))^t$$

Using the induction hypothesis, namely that as $n_0, n_1, ..., n_{l-1} \to \infty$, for all $j, j' \leq n_l$ and all x, x'

$$\partial_{\theta_{1:l}} y_j^l(x) (\partial_{\theta_{1:l}} y_{j'}^l(x'))^t \to K^l(x, x') \mathbf{1}_{j=j'}$$

we then obtain for all n_l, as $n_0, n_1, ..., n_{l-1} \to \infty$

$$\frac{\sigma_w^2}{n_l} \sum_{j,j'}^{n_l} w_{ij}^{l+1} w_{ij'}^{l+1} \phi'(y_j^l(x)) \phi'(y_{j'}^l(x')) \partial_{\theta_{1:l}} y_j^l(x) (\partial_{\theta_{1:l}} y_{j'}^l(x'))^t \to \quad \frac{\sigma_w^2}{n_l} \sum_{j}^{n_l} (w_{ij}^{l+1})^2 \phi'(y_j^l(x)) \phi'(y_j^l(x')) K^l(x, x')$$

and letting n_l go to infinity, the law of large numbers, implies that

$$\frac{\sigma_w^2}{n_l} \sum_{j}^{n_l} (w_{ij}^{l+1})^2 \phi'(y_j^l(x)) \phi'(y_j^l(x')) K^l(x, x') \to \dot{q}^{l+1}(x, x') K^l(x, x').$$

Moreover, we have that

$$(\partial_{w^{l+1}} y_i^{l+1}(x))(\partial_{w^{l+1}} y_i^{l+1}(x'))^t + (\partial_{b^{l+1}} y_i^{l+1}(x))(\partial_{b^{l+1}} y_i^{l+1}(x'))^t = \frac{\sigma_w^2}{n_l} \sum_j \phi(y_j^l(x)) \phi(y_j^l(x')) + \sigma_b^2$$

$$\underset{n_l \to \infty}{\to} \sigma_w^2 \mathbb{E}[\phi(y_i^l(x)) \phi(y_i^l(x'))] + \sigma_b^2 = q^{l+1}(x, x').$$

which ends the proof.

□

The following proposition establishes that any initialization on the Ordered or Chaotic phase, leads to a trivial limiting NTK as the number of layers L becomes large.

Proposition 1 (Limiting Neural Tangent Kernel with Ordered/Chaotic Initialization)**.** *Let (σ_b, σ_w) be either in the ordered or in the chaotic phase. Then, there exist $\lambda > 0$ such that for all $\epsilon \in (0, 1)$, there exists $\gamma > 0$ such that*

$$\sup_{(x,x') \in B_\epsilon} |K^L(x, x') - \lambda| \leq e^{-\gamma L}.$$

We will use the next lemma in the proof of proposition 1.

Appendix Lemma 4. *Let (a_l) be a sequence of non-negative real numbers such that $\forall l \geq 0, a_{l+1} \leq \alpha a_l + k e^{-\beta l}$, where $\alpha \in (0, 1)$ and $k, \beta > 0$. Then there exists $\gamma > 0$ such that $\forall l \geq 0$, $a_l \leq e^{-\gamma l}$.*

Proof. Using the inequality on a_l, we can easily see that

$$a_l \leq a_0 \alpha^l + k \sum_{j=0}^{l-1} \alpha^j e^{-\beta(l-j)}$$

$$\leq a_0 \alpha^l + k \frac{l}{2} e^{-\beta l/2} + k \frac{l}{2} \alpha^{l/2}$$

where we divided the sum into two parts separated by index $l/2$ and upper-bounded each part. The existence of γ is straightforward. □

Now we prove Proposition 1

Proof. We prove the result for FFNN first. Let x, x' be two inputs. From lemma 1, we have that

$$K^l(x,x') = K^{l-1}(x,x')\dot{q}^l(x,x') + q^l(x,x')$$

where $q^1(x,x') = \sigma_b^2 + \frac{\sigma_w^2}{d}x^T x'$ and $q^l(x,x') = \sigma_b^2 + \sigma_w^2 \mathbb{E}_{f\sim\mathcal{N}(0,q^{l-1})}[\phi(f(x))\phi(f(x'))]$ and $\dot{q}^l(x,x') = \sigma_w^2 \mathbb{E}_{f\sim\mathcal{N}(0,q^{l-1})}[\phi'(f(x))\phi'(f(x'))]$. From facts 1, 2, 4, 9, 16, in the ordered/chaotic phase, there exist $k, \beta, \eta, l_0 > 0$ and $\alpha \in (0,1)$ such that for all $l \geq l_0$ we have

$$\sup_{(x,x')\in B_\epsilon} |q^l(x,x') - k| \leq e^{-\beta l}$$

and

$$\sup_{(x,x')\in B_\epsilon} |\dot{q}^l(x,x') - \alpha| \leq e^{-\eta l}.$$

Therefore, there exists $M > 0$ such that for any $l \geq l_0$ and $x, x' \in \mathbb{R}^d$

$$K^l(x,x') \leq M.$$

Letting $r_l = \sup_{(x,x')\in B_\epsilon} |K^l(x,x') - \frac{k}{1-\alpha}|$, we have

$$r_l \leq \alpha r_{l-1} + Me^{-\eta l} + e^{-\beta l}$$

We conclude using Appendix Lemma 4.

□

Now, we show that the Initialization on the EOC improves the convergence rate of the NTK wrt L. We first prove two preliminary lemmas that will be useful for the proof of the next proposition. Hereafter, the notation $g(x) = \Theta(m(x))$ means there exist two constants $A, B > 0$ such that $Am(x) \leq g(x) \leq Bm(x)$.

Appendix Lemma 5. *Let $A, B, \Lambda \subset \mathbb{R}+$ be three compact sets, and $(a_l), (b_l), (\lambda_l)$ be three sequences of non-negative real numbers such that for all $(a_0, b_0, \lambda_0) \in A \times B \times \Lambda$*

$$a_l = a_{l-1}\lambda_l + b_l, \quad \lambda_l = 1 - \frac{\alpha}{l} + \mathcal{O}(l^{-1-\beta}), \quad b_l = q(b_0) + o(l^{-1}),$$

where $\alpha \in \mathbb{N}^*$ independent of a_0, b_0, λ_0, $q(b_0) \geq 0$ is a limit that depends on b_0, and $\beta \in (0,1)$. Assume the '$\mathcal{O}$' and 'o' depend only on $A, B, \Lambda \subset \mathbb{R}$. Then, we have

$$\sup_{(a_0,b_0,\lambda_0)\in A\times B\times\Lambda} \left| \frac{a_l}{l} - \frac{q}{1+\alpha} \right| = \mathcal{O}(l^{-\beta}).$$

Proof. Let $A, B, \Lambda \subset \mathbb{R}$ be three compact sets and $(a_0, b_0, \lambda_0) \in A \times B \times \Lambda$. It is easy to see that there exists a constant $G > 0$ independent of a_0, b_0, λ_0 such that $|a_l| \leq G \times l + |a_0|$ for all $l \geq 0$. Letting $r_l = \frac{a_l}{l}$, we have that for $l \geq 2$

$$\begin{aligned} r_l &= r_{l-1}(1 - \frac{1}{l})(1 - \frac{\alpha}{l} + \mathcal{O}(l^{-1-\beta})) + \frac{q}{l} + o(l^{-2}) \\ &= r_{l-1}(1 - \frac{1+\alpha}{l}) + \frac{q}{l} + \mathcal{O}(l^{-1-\beta}). \end{aligned}$$

where $\mathcal{O}$ bound depends only on A, B, Λ. Letting $x_l = r_l - \frac{q}{1+\alpha}$, there exists $M > 0$ that depends only on A, B, Λ, and $l_0 > 0$ that depends only on α such that for all $l \geq l_0$

$$x_{l-1}(1 - \frac{1+\alpha}{l}) - \frac{M}{l^{1+\beta}} \leq x_l \leq x_{l-1}(1 - \frac{1+\alpha}{l}) + \frac{M}{l^{1+\beta}}.$$

Let us deal with the right hand inequality first. By induction, we have that

$$x_l \leq x_{l_0-1} \prod_{k=l_0}^{l} (1 - \frac{1+\alpha}{k}) + M \sum_{k=l_0}^{l} \prod_{j=k+1}^{l} (1 - \frac{1+\alpha}{j}) \frac{1}{k^{1+\beta}}.$$

By taking the logarithm of the first term in the right hand side and using the fact that $\sum_{k=l_0}^{l} \frac{1}{k} = \log(l) + O(1)$, we have

$$\prod_{k=l_0}^{l} (1 - \frac{1+\alpha}{k}) = \Theta(l^{-1-\alpha}).$$

where the bound Θ does not depend on l_0. For the second part, observe that

$$\prod_{j=k+1}^{l}(1-\frac{1+\alpha}{j}) = \frac{(l-\alpha-1)!}{l!}\frac{k!}{(k-\alpha-1)!}$$

and

$$\frac{k!}{(k-\alpha-1)!}\frac{1}{k^{1+\beta}} \sim_{k\to\infty} k^{\alpha-\beta}.$$

Since $\alpha \geq 1$ ($\alpha \in \mathbb{N}^*$), then the serie with term $k^{\alpha-\beta}$ is divergent and we have that

$$\begin{aligned}\sum_{k=l_0}^{l}\frac{k!}{(k-\alpha-1)!}\frac{1}{k^2} &\sim \sum_{k=1}^{l}k^{\alpha-\beta}\\ &\sim \int_1^l t^{\alpha-\beta}dt\\ &\sim \frac{1}{\alpha-\beta+1}l^{\alpha-\beta+1}.\end{aligned}$$

Therefore, it follows that

$$\begin{aligned}\sum_{k=l_0}^{l}\prod_{j=k+1}^{l}(1-\frac{1+\alpha}{j})\frac{1}{k^{1+\beta}} &= \frac{(l-\alpha-1)!}{l!}\sum_{k=l_0}^{l}\frac{k!}{(k-\alpha-1)!}\frac{1}{k^{1+\beta}}\\ &\sim \frac{1}{\alpha}l^{-\beta}.\end{aligned}$$

This proves that

$$x_l \leq \frac{M}{\alpha}l^{-\beta} + o(l^{-\beta}).$$

where the 'o' bound depends only on A, B, Λ. Using the same approach for the left-hand inequality, we prove that

$$x_l \geq -\frac{M}{\alpha}l^{-\beta} + o(l^{-\beta}).$$

This concludes the proof.

□

The next lemma is a different version of the previous lemma which will be useful for other applications.

Appendix Lemma 6. *Let $A, B, \Lambda \subset \mathbb{R}+$ be three compact sets, and $(a_l), (b_l), (\lambda_l)$ be three sequences of non-negative real numbers such that for all $(a_0, b_0, \lambda_0) \in A \times B \times \Lambda$*

$$\begin{aligned}a_l &= a_{l-1}\lambda_l + b_l, \quad b_l = q(b_0) + \mathcal{O}(l^{-1}),\\ \lambda_l &= 1 - \frac{\alpha}{l} + \kappa\frac{\log(l)}{l^2} + \mathcal{O}(l^{-2}),\end{aligned}$$

where $\alpha \in \mathbb{N}^*, \kappa \neq 0$ both do not depend on a_0, b_0, Λ_0, $q(b_o) \in \mathbb{R}^+$ is a limit that depends on b_0. Assume the '$\mathcal{O}$' and 'o' depend only on $A, B, \Lambda \subset \mathbb{R}$. Then, we have

$$\sup_{(a_0,b_0,\lambda_0)\in A\times B\times\Lambda}\left|\frac{a_l}{l} - \frac{q}{1+\alpha}\right| = \Theta(\log(l)l^{-1})$$

Proof. Let $A, B, \Lambda \subset \mathbb{R}$ be three compact sets and $(a_0, b_0, \lambda_0) \in A \times B \times \Lambda$. Similar to the proof of Appendix Lemma 5, there exists a constant $G > 0$ independent of a_0, b_0, λ_0 such that $|a_l| \leq G \times l + |a_0|$ for all $l \geq 0$, therefore (a_l/l) is bounded. Let $r_l = \frac{a_l}{l}$. We have

$$\begin{aligned}r_l &= r_{l-1}(1-\frac{1}{l})(1-\frac{\alpha}{l}+\kappa\frac{\log(l)}{l^2}+O(l^{-1-\beta})) + \frac{q}{l} + \mathcal{O}(l^{-2})\\ &= r_{l-1}(1-\frac{1+\alpha}{l}) + r_{l-1}\kappa\frac{\log(l)}{l^2} + \frac{q}{l} + \mathcal{O}(l^{-2}).\end{aligned}$$

Let $x_l = r_l - \frac{q}{1+\alpha}$. It is clear that $\lambda_l = 1 - \alpha/l + \mathcal{O}(l^{-3/2})$. Therefore, using appendix lemma 5 with $\beta = 1/2$, we have $r_l \to \frac{q}{1+\alpha}$ uniformly over a_0, b_0, λ_0. Thus, assuming $\kappa > 0$ (for $\kappa < 0$, the analysis is the same), there exists $\kappa_1, \kappa_2, M, l_0 > 0$ that depend only on A, B, Λ such that for all $l \geq l_0$

$$x_{l-1}(1-\frac{1+\alpha}{l}) + \kappa_1\frac{\log(l)}{l^2} - \frac{M}{l^2} \leq x_l \leq x_{l-1}(1-\frac{1+\alpha}{l}) + \kappa_2\frac{\log(l)}{l^2} + \frac{M}{l^2}.$$

It follows that

$$x_l \leq x_{l_0} \prod_{k=l_0}^{l} (1 - \frac{1+\alpha}{k}) + \sum_{k=l_0}^{l} \prod_{j=k+1}^{l} (1 - \frac{1+\alpha}{j}) \frac{\kappa_2 \log(k) + M}{k^2}$$

and

$$x_l \geq x_{l_0} \prod_{k=l_0}^{l} (1 - \frac{1+\alpha}{k}) + \sum_{k=l_0}^{l} \prod_{j=k+1}^{l} (1 - \frac{1+\alpha}{j}) \frac{\kappa_1 \log(k) - M}{k^2}.$$

Recall that we have

$$\prod_{k=l_0}^{l} (1 - \frac{1+\alpha}{k}) = \Theta(l^{-1-\alpha})$$

and

$$\prod_{j=k+1}^{l} (1 - \frac{1+\alpha}{j}) = \frac{(l-\alpha-1)!}{l!} \frac{k!}{(k-\alpha-1)!}$$

so that

$$\frac{k!}{(k-\alpha-1)!} \frac{\kappa_1 \log(k) - M}{k^2} \sim_{k\to\infty} \log(k) k^{\alpha-1}.$$

Therefore, we obtain

$$\begin{aligned} \sum_{k=l_0}^{l} \frac{k!}{(k-\alpha-1)!} \frac{\kappa_1 \log(k) - M}{k^2} &\sim \sum_{k=1}^{l} \log(k) k^{\alpha-1} \\ &\sim \int_1^l \log(t) t^{\alpha-1} dt \\ &\sim C_1 l^{\alpha} \log(l), \end{aligned}$$

where $C_1 > 0$ is a constant. Similarly, there exists a constant $C_2 > 0$ such that

$$\sum_{k=1}^{l} \frac{k!}{(k-\alpha-1)!} \frac{\kappa_2 \log(k) + M}{k^2} \sim C_2 l^{\alpha} \log(l).$$

Moreover, having that $\frac{(l-\alpha-1)!}{l!} \sim l^{-1-\alpha}$ yields

$$x_l \leq C' l^{-1} \log(l) + o(l^{-1} \log(l))$$

where C' and 'o' depend only on A, B, Λ. Using the same analysis, we get

$$x_l \geq C'' l^{-1} \log(l) + o(l^{-1} \log(l))$$

where C'' and 'o' depend only on A, B, Λ, which concludes the proof.

□

Theorem 1 (Neural Tangent Kernel on the Edge of Chaos)**.** *Let ϕ be ReLU or Tanh, $(\sigma_b, \sigma_w) \in$ EOC and $\tilde{K}^L = K^L/L$. We have that*

$$\sup_{x \in E} |\tilde{K}^L(x,x) - \tilde{K}^{\infty}(x,x)| = \mathcal{O}(L^{-1})$$

Moreover, there exists a constant $\lambda \in (0,1)$ such that for all $\epsilon \in (0,1)$

$$\sup_{(x,x') \in B_\epsilon} \left|\tilde{K}^L(x,x') - \tilde{K}^{\infty}(x,x')\right| = \Theta(\log(L) L^{-1}).$$

where
- *if ϕ is ReLU, then $\tilde{K}^{\infty}(x,x') = \frac{\sigma_w^2 \|x\| \, \|x'\|}{d}(1 - (1-\lambda)\mathbb{1}_{x \neq x'})$.*
- *if ϕ is Tanh, then $\tilde{K}^{\infty}(x,x') = q(1 - (1-\lambda)\mathbb{1}_{x \neq x'})$ where $q > 0$ is a constant.*

Proof. Let $\epsilon \in (0,1)$, $E \subset \mathbb{R}^d$, $(\sigma_b, \sigma_w) \in$ EOC, and $x, x' \in \mathbb{R}^d$. Recall that $c^l(x,x') = \frac{q^l(x,x')}{\sqrt{q^l(x,x) q^l(x',x')}}$. Let $\gamma_l := 1 - c^l(x,x')$ and f be the *correlation function* defined by the recursive equation $c^{l+1} = f(c^l)$ (See appendix 3). By definition, we have that $\dot{q}^l(x,x) = f'(c^{l-1}(x,x'))$. Let us first prove the result for ReLU.

- ϕ =*ReLU*: From fact 11 in the appendix, we know that, when choosing the hyper-parameters (σ_w, σ_b) on the EOC for ReLU, the variance $q^l(x,x)$ is constant w.r.t l and is given by $q^l(x,x) = q^1(x,x) = \frac{\sigma_w^2}{d}||x||^2$. Moreover, from fact 15, we have that $\dot{q}^l(x,x) = 1$. Therefore

$$K^l(x,x) = K^{l-1}(x,x) + \frac{\sigma_w^2}{d}||x||^2 = l\frac{\sigma_w^2}{d}||x||^2 = l\tilde{K}^{\infty}(x,x)$$

which concludes the proof for $K^L(x,x)$. Note that the results is 'exact' for ReLU, which means the upper bound $\mathcal{O}(L^{-1})$ is valid but not optimal in this case. However, we will see that this bound is optimal for Tanh.

From Appendix Lemma 1, we have that

$$\sup_{(x,x')\in B_\epsilon} \left| c^l(x,x') - 1 + \frac{\kappa}{l^2} - \kappa' \frac{\log(l)}{l^3} \right| = \mathcal{O}(l^{-3})$$

and

$$\sup_{(x,x')\in B_\epsilon} \left| f'(c^l(x,x')) - 1 + \frac{3}{l} - \kappa'' \frac{\log(l)}{l^2} \right| = \mathcal{O}(l^{-2}).$$

Using Appendix Lemma 6 with $a_l = K^{l+1}(x,x')$, $b_l = q^{l+1}(x,x')$, $\lambda_l = f'(c^l(x,x'))$, we conclude that

$$\sup_{(x,x')\in B_\epsilon} \left| \frac{K^{l+1}(x,x')}{l} - \frac{1}{4}\frac{\sigma_w^2}{d}\|x\|\|x'\| \right| = \Theta(\log(l)l^{-1})$$

Using the compactness of B_ϵ, we conclude that

$$\sup_{(x,x')\in B_\epsilon} \left| \frac{K^{l}(x,x')}{l} - \frac{1}{4}\frac{\sigma_w^2}{d}\|x\|\|x'\| \right| = \Theta(\log(l)l^{-1})$$

- $\phi = Tanh$: The proof in the case of Tanh is slightly different from that of ReLU. We use different technical lemmas to conclude.

  From Appendix Lemma 2, we have that

$$\sup_{(x,x')\in B_\epsilon} \left| c^l(x,x') - 1 + \frac{\kappa}{l} - \kappa(1-\kappa^2\zeta)\frac{\log(l)}{l^3} \right| = \mathcal{O}(l^{-3})$$

  where $\kappa = \frac{2}{f''(1)} > 0$ and $\zeta = \frac{f^3(1)}{6} > 0$. Moreover, we have that

$$\sup_{(x,x')\in B_\epsilon} \left| f'(c^l(x,x')) - 1 + \frac{2}{l} - 2(1-\kappa^2\zeta)\frac{\log(l)}{l^2} \right| = \mathcal{O}(l^{-2}).$$

  We conclude in the same way as in the case of ReLU using Appendix Lemma 6. The only difference is that, in this case, the limit of the sequence $b_l = q^{l+1}(x,x')$ is the limiting variance q (from facts 3, 1) does not depend on (x,x').

□

5.2 Proofs of the results on ResNets

In this section, we provide proofs for lemma 2 together with Theorem 2 and proposition 2.

Lemma 2 (NTK of a ResNet with Fully Connected layers in the infinite width limit)**.** *Let x, x' be two inputs and $K^{res,1}$ be the exact NTK for the Residual Network with 1 layer. Then, we have*

- *For the first layer (without residual connections), we have for all $x, x' \in \mathbb{R}^d$*

$$K^{res,1}_{ii'}(x,x') = \delta_{ii'}\left(\sigma_b^2 + \frac{\sigma_w^2}{d}x\cdot x'\right),$$

  where $x \cdot x'$ is the inner product in $\mathbb{R}^d$.

- *For $l \geq 2$, as $n_1, n_2, ..., n_{L-1} \to \infty$, we have for all $i, i' \in [1 : n_l]$, $K^{res,l}_{ii'}(x, x') = \delta_{ii'} K^l_{res}(x, x')$, where $K^l_{res}(x, x')$ is given by the recursive formula have for all $x, x' \in \mathbb{R}^d$ and $l \geq 2$, as $n_1, n_2, ..., n_l \to \infty$ recursively, we have*

$$K^l_{res}(x, x') = K^{l-1}_{res}(x, x')(\dot{q}^l(x, x') + 1) + \hat{q}^l(x, x').$$

Proof. The first result is the same as in the FFNN case since we assume there is no residual connections between the first layer and the input. We prove the second result by induction.

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

$$K^1_{res}(x, x') = \sum_j \frac{\partial y^1_1(x)}{\partial w^1_{1j}} \frac{\partial y^1_1(x)}{\partial w^1_{1j}} + \frac{\partial y^1_1(x)}{\partial b^1_1} \frac{\partial y^1_1(x)}{\partial b^1_1} = \frac{\sigma^2_w}{d} x \cdot x' + \sigma^2_b.$$

- The proof is similar to the FeedForward network NTK. For $l \geq 2$ and $i \in [1 : n_l]$

$$\partial_{\theta_{1:l}} y^{l+1}_i(x) = \partial_{\theta_{1:l}} y^l_i(x) + \frac{\sigma_w}{\sqrt{n_l}} \sum_{j=1}^{n_l} w^{l+1}_{ij} \phi'(y^l_j(x)) \partial_{\theta_{1:l}} y^l_j(x).$$

  Therefore, we obtain

$$\begin{aligned}(\partial_{\theta_{1:l}} y^{l+1}_i(x))(\partial_{\theta_{1:l}} y^{l+1}_i(x'))^t &= (\partial_{\theta_{1:l}} y^l_i(x))(\partial_{\theta_{1:l}} y^l_i(x'))^t \\ &+ \frac{\sigma^2_w}{n_l} \sum_{j,j'}^{n_l} w^{l+1}_{ij} w^{l+1}_{ij'} \phi'(y^l_j(x)) \phi'(y^l_{j'}(x')) \partial_{\theta_{1:l}} y^l_j(x) (\partial_{\theta_{1:l}} y^l_{j'}(x'))^t + I\end{aligned}$$

  where

$$I = \frac{\sigma_w}{\sqrt{n_l}} \sum_{j=1}^{n_l} w^{l+1}_{ij} (\phi'(y^l_j(x)) \partial_{\theta_{1:l}} y^l_i(x) (\partial_{\theta_{1:l}} y^l_j(x'))^t + \phi'(y^l_j(x')) \partial_{\theta_{1:l}} y^l_j(x) (\partial_{\theta_{1:l}} y^l_i(x'))^t).$$

  Using the induction hypothesis, as $n_0, n_1, ..., n_{l-1} \to \infty$, we have that

$$(\partial_{\theta_{1:l}} y^{l+1}_i(x))(\partial_{\theta_{1:l}} y^{l+1}_i(x'))^t + \frac{\sigma^2_w}{n_l} \sum_{j,j'}^{n_l} w^{l+1}_{ij} w^{l+1}_{ij'} \phi'(y^l_j(x)) \phi'(y^l_{j'}(x')) \partial_{\theta_{1:l}} y^l_j(x) (\partial_{\theta_{1:l}} y^l_{j'}(x'))^t + I$$

$$\to K^l_{res}(x, x') + \frac{\sigma^2_w}{n_l} \sum_j^{n_l} (w^{l+1}_{ij})^2 \phi'(y^l_j(x)) \phi'(y^l_j(x')) K^l_{res}(x, x') + I',$$

  where $I' = \frac{\sigma^2_w}{n_l} w^{l+1}_{ii} (\phi'(y^l_i(x)) + \phi'(y^l_i(x'))) K^l_{res}(x, x')$.

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

$$\frac{\sigma^2_w}{n_l} \sum_j^{n_l} (w^{l+1}_{ij})^2 \phi'(y^l_j(x)) \phi'(y^l_j(x')) K^l_{res}(x, x') \to \dot{q}^{l+1}(x, x') K^l_{res}(x, x').$$

  Moreover, we have that

$$\begin{aligned}(\partial_{w^{l+1}} y^{l+1}_i(x))(\partial_{w^{l+1}} y^{l+1}_i(x'))^t + (\partial_{b^{l+1}} y^{l+1}_i(x))(\partial_{b^{l+1}} y^{l+1}_i(x'))^t &= \frac{\sigma^2_w}{n_l} \sum_j \phi(y^l_j(x)) \phi(y^l_j(x')) + \sigma^2_b \\ &\underset{n_l \to \infty}{\to} \sigma^2_w \mathbb{E}[\phi(y^l_i(x)) \phi(y^l_i(x'))] + \sigma^2_b = q^{l+1}(x, x').\end{aligned}$$

□

The proof of main theorem on ResNets requires the following lemma on the asymptotic behaviour of c^l for ResNet.

Appendix Lemma 7 (Asymptotic expansion of c^l for ResNet)**.** *Let $\epsilon \in (0, 1)$ and $\sigma_w > 0$. We have for FFNN*

$$\sup_{(x,x') \in B_\epsilon} \left| c^l(x, x') - 1 + \frac{\kappa_{\sigma_w}}{l^2} - \kappa'_{\sigma_w} \frac{\log(l)}{l^3} \right| = \mathcal{O}(l^{-3})$$

where $\kappa_{\sigma_w}, \kappa'_{\sigma_w} > 0$ are two constants that depend on σ_w.
Moreover, we have that

$$\sup_{(x,x')\in B_\epsilon} \left| f'(c^l(x,x')) - 1 + \frac{3(1+\frac{2}{\sigma_w^2})}{l} - \kappa''_{\sigma_W} \frac{\log(l)}{l^2} \right| = \mathcal{O}(l^{-2}).$$

where f is the ReLU correlation function given in fact 12 and $\kappa''_{\sigma_W} > 0$ is a constant that depends on σ_w.

Proof. Let $\epsilon \in (0,1)$.

- Let $x \neq x' \in \mathbb{R}^d$, and $c^l := c^l(x,x')$. It is straightforward that the variance terms follow the recursive form
$$q^l(x,x) = q^{l-1}(x,x) + \sigma_w^2/2 q^{l-1}(x,x) = (1+\sigma_w^2/2)^{l-1} q^1(x,x)$$
Leveraging this observation, we have that
$$c^{l+1} = \frac{1}{1+\alpha} c^l + \frac{\alpha}{1+\alpha} f(c^l),$$
where f is the ReLU correlation function given in fact 12 and $\alpha = \frac{\sigma_w^2}{2}$. Recall that
$$f(c) = \frac{1}{\pi} c \arcsin(c) + \frac{1}{\pi}\sqrt{1-c^2} + \frac{1}{2}c.$$
As in the proof of Appendix Lemma 1, let $\gamma_l = 1 - c^l$, therefore, using Taylor expansion of f near 1 given in fact 14 yields
$$\gamma_{l+1} = \gamma_l - \frac{\alpha s}{1+\alpha}\gamma_l^{3/2} - \frac{\alpha b}{1+\alpha}\gamma_l^{5/2} + O(\gamma_l^{7/5}).$$
This form is exactly the same as in the proof of Appendix Lemma 1 with $s' = \frac{\alpha s}{1+\alpha}$ and $b' = \frac{\alpha b}{1+\alpha}$. Thus, following the same analysis we conclude.
For the second result, observe that the derivation is the same as in Appendix Lemma 1.

□

The next theorem shows that no matter what the choice of $\sigma_w > 0$, the normalized NTK of a ResNet will always have a subexponential convergence rate to a limiting $\bar{K}^\infty_{res}$.

Theorem 2 (NTK for ResNet)**.** *Consider a ResNet satisfying*

$$y^l(x) = y^{l-1}(x) + \mathcal{F}(w^l, y^{l-1}(x)), \quad l \geq 2, \tag{14}$$

where $\mathcal{F}$ is a dense layer with ReLU activation. Let K^L_{res} be the corresponding NTK and $\bar{K}^L_{res} = K^L_{res}/\alpha_L$ (Normalized NTK) with $\alpha_L = L(1+\frac{\sigma_w^2}{2})^{L-1}$. Then, we have

$$\sup_{x\in E} |\bar{K}^L_{res}(x,x) - \bar{K}^\infty_{res}(x,x)| = \Theta(L^{-1})$$

Moreover, there exists a constant $\lambda \in (0,1)$ such that for all $\epsilon \in (0,1)$

$$\sup_{x,x'\in B_\epsilon} \left|\bar{K}^L_{res}(x,x') - \bar{K}^\infty_{res}(x,x')\right| = \Theta(L^{-1}\log(L)),$$

where $\bar{K}^\infty_{res}(x,x') = \frac{\sigma_w^2 \|x\| \, \|x'\|}{d}(1-(1-\lambda)\mathbb{1}_{x\neq x'})$.

Proof. Let $\epsilon \in (0,1)$, $E \subset \mathbb{R}^d$, and $x, x' \in \mathbb{R}^d$. We first prove the result for the diagonal terms $K^L_{res}(x,x)$, we deal afterwards with off-diagonal terms $K^L_{res}(x,x')$.

- Diagonal terms: from fact 12, we have that $\dot{q}^l(x,x) = \frac{\sigma_w^2}{2} f(1) = \frac{\sigma_w^2}{2}$. Moreover, it is easy to see that the variance terms for a ResNet follow the recursive formula $q^l(x,x) = q^{l-1}(x,x) + \sigma_w^2/2 \times q^{l-1}(x,x)$, hence
$$q^l(x,x) = (1+\sigma_w^2/2)^{l-1} \frac{\sigma_w^2}{d}\|x\|^2 \tag{15}$$

Recall that the recursive formula of NTK of a ResNet with fully-connected layers is given by (Appendix Lemma 2)

$$K^l_{res}(x,x') = K^{l-1}_{res}(x,x')(\dot{q}^l(x,x')+1) + q^l(x,x')$$

Hence, for the diagonal terms we obtain

$$K^l_{res}(x,x) = K^{l-1}_{res}(x,x)\left(\frac{\sigma_w^2}{2}+1\right) + q^l(x,x)$$

Letting $\hat{K}^l_{res} = K^l_{res}/\left(1+\frac{\sigma_w^2}{l}\right)^{l-1}$ yields

$$\hat{K}^l_{res}(x,x) = \hat{K}^{l-1}_{res}(x,x) + \frac{\sigma_w^2}{d}\|x\|^2$$

Therefore, $\bar{K}^l_{res}(x,x) = \frac{\hat{K}^1_{res}(x,x)}{l} + (1-1/l)\frac{\sigma_w^2}{d}\|x\|^2$, the conclusion is straightforward since E is compact and $\hat{K}^1_{res}(x,x)$ is continuous which implies that it is uniformly bounded on E.

- Off-diagonal terms: the argument is similar to that of Theorem 1 with few key differences. From Appendix Lemma 7 we have that

$$\sup_{(x,x')\in B_\epsilon}\left|c^l(x,x') - 1 + \frac{\kappa_{\sigma_w}}{l^2} - \kappa'_{\sigma_w}\frac{\log(l)}{l^3}\right| = \mathcal{O}(l^{-3})$$

where $\kappa_{\sigma_w}, \kappa'_{\sigma_w} > 0$. Moreover, we have that

$$\sup_{(x,x')\in B_\epsilon}\left|f'(c^l(x,x')) - 1 + \frac{3(1+\frac{2}{\sigma_w^2})}{l} - \kappa''_{\sigma_w}\frac{\log(l)}{l^2}\right| = \mathcal{O}(l^{-2}).$$

Let $\alpha = \frac{\sigma_w^2}{2}$. We also have $\dot{q}^{l+1}(x,x') = \alpha f'(c^l(x,x'))$ where f is the ReLU correlation function given in fact 12. It follows that for all $(x,x') \in B_\epsilon$

$$1 + \dot{q}^{l+1}(x,x') = (1+\alpha)(1 - 3l^{-1} + \zeta\frac{\log(l)}{l^2} + \mathcal{O}(l^{-3}))$$

for some constant $\zeta \neq 0$ that does not depend on x, x'. The bound $\mathcal{O}$ does not depend on x, x' either. Now let $a_l = \frac{K^{l+1}_{res}(x,x')}{(1+\alpha)^l}$. Using the recursive formula of the NTK, we obtain

$$a_l = \lambda_l a_{l-1} + b_l$$

where $\lambda_l = 1 - 3l^{-1} + \zeta\frac{\log(l)}{l^2} + \mathcal{O}(l^{-3})$, $b_l = \frac{\sigma_w^2}{d}\sqrt{\|x\|\|x'\|}f(c^l(x,x')) = q(x,x') + \mathcal{O}(l^{-2})$ with $q(x,x') = \frac{\sigma_w^2}{d}\sqrt{\|x\|\|x'\|}$ and where we used the fact that $c^l(x,x') = 1 + \mathcal{O}(l^{-2})$ (Appendix Lemma 1) and the formula for ResNet variance terms given by equation (15). Observe that all bounds $\mathcal{O}$ are independent from the inputs (x,x'). Therefore, using Appendix Lemma 6, we have

$$\sup_{x,x'\in B_\epsilon}\left|K^{L+1}_{res}(x,x')/L(1+\alpha)^L - \bar{K}^\infty_{res}(x,x')\right| = \Theta(L^{-1}\log(L)),$$

which can also be written as

$$\sup_{x,x'\in B_\epsilon}\left|K^{L}_{res}(x,x')/(L-1)(1+\alpha)^{L-1} - \bar{K}^\infty_{res}(x,x')\right| = \Theta(L^{-1}\log(L)),$$

We conclude by observing that $K^L_{res}(x,x')/(L-1)(1+\alpha)^{L-1} = K^L_{res}(x,x')/L(1+\alpha)^{L-1} + \mathcal{O}(L^{-1})$ where $\mathcal{O}$ can be chosen to depend only on ϵ.

□

5.3 Spectral decomposition of the limiting NTK

5.3.1 Review on Spherical Harmonics

We start by giving a brief review of the theory of Spherical Harmonics MacRobert (1967). Let $\mathbb{S}^{d-1}$ be the unit sphere in $\mathbb{R}^d$ defined by $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$. For some $k \geq 1$, there exists a set $(Y_{k,j})_{1\leq j\leq N(d,k)}$ of Spherical Harmonics of degree k with $N(d,k) = \frac{2k+d-2}{k}\binom{k+d-3}{d-2}$.

The set of functions $(Y_{k,j})_{k\geq 1, j\in[1:N(d,k)]}$ form an orthonormal basis with respect to the uniform measure on the unit sphere $\mathbb{S}^{d-1}$.

For some function g, the Hecke-Funk formula is given by

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

where ν_{d-1} is the uniform measure on the unit sphere $\mathbb{S}^{d-1}$, 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

$$P_k^d(t) = \big(-\frac{1}{2}\big)^k \frac{\Gamma(\frac{d-1}{2})}{\Gamma(k+\frac{d-1}{2})} (1-t^2)^{\frac{3-d}{2}} \big(\frac{d}{dt}\big)^k (1-t^2)^{k+\frac{d-3}{2}}$$

$(P_k^d)_{k\geq 0}$ 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. Moreover, we have

$$\|P_k^d\|^2_{L^2([-1,1],(1-t^2)^{\frac{d-3}{2}} dt)} = \frac{(k+d-3)!}{(d-3)(k-d+3)!}$$

Using the Heck-Funk formula, we can easily conclude that any dot product kernel on the unit sphere $\mathbb{S}^{d-1}$, i.e. and kernel of the form $\kappa(x,x') = g(\langle x, x'\rangle)$ 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

$$\kappa(x,x') = \sum_{k\geq 0} \sum_{j=1}^{N(d,k)} \left[\int_{\mathbb{S}^{d-1}} g(\langle w, x'\rangle) Y_{k,j}(w) d\nu_{d-1}(w) \right] Y_{k,j}(x)$$

Using the Hecke-Funk formula yields

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

we conclude that

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

where $\mu_k = \frac{\Omega_{d-1}}{\Omega_d} \int_{-1}^{1} g(t) P_k^d(t) (1-t^2)^{(d-3)/2} dt$.

We use these result in the proof of the next theorem.

Proposition 2 (Spectral decomposition)**.** *Let κ^L be either, the NTK (K^L) for an FFNN with L layers initialized on the Ordered phase, The Average NTK (AK^L) for an FFNN with L layers initialized on the EOC, or the Normalized NTK ($\bar{K}^L_{res}$) for a ResNet with L layers (Fully Connected). Then, for all $L \geq 1$, there exists $(\mu_k^L)_{k\geq}$ such that for all $x, x' \in \mathbb{S}^{d-1}$*

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

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

Moreover, we have that $0 < \mu_0^\infty = \lim_{L\to\infty} \mu_0^L < \infty$, and for all $k \geq 1$, $\lim_{L\to\infty} \mu_k^L = 0$.

Proof. From the recursive formulas of the NTK for FFNN and ResNet architectures, it is straightforward that on the unit sphere $\mathbb{S}^{d-1}$, the kernel κ^L is zonal in the sense that it depends only on the scalar product, more precisely, for all $L \geq 1$, there exists a function g^L such that for all $x, x' \in \mathbb{S}^{d-1}$

$$\kappa^L(x,x') = g^L(\langle x, x'\rangle)$$

using the previous results on Spherical Harmonics, we have that for all $x, x' \in \mathbb{S}^{d-1}$

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

where $\mu_k^L = \frac{\Omega_{d-1}}{\Omega_d} \int_{-1}^{1} g^L(t) P_k^d(t) (1-t^2)^{(d-3)/2} dt$.

For $k = 0$, we have that for all $L \geq 1$, $\mu_0^L = \frac{\Omega_{d-1}}{\Omega_d} \int_{-1}^1 g^L(t)(1-t^2)^{(d-3)/2} dt$. By a simple dominated convergence argument, we have that $\lim_{L\to\infty} \mu_0^L = q\lambda \frac{\Omega_{d-1}}{\Omega_d} \int_{-1}^1 (1-t^2)^{(d-3)/2} dt > 0$, where q, λ are given in Theorems 1, 2 and Proposition 1 (where we take $q = 1$ for the Ordered/Chaotic phase initialization in Proposition 1). Using the same argument, we have that for $k \geq 1$, $\lim_{L\to\infty} \mu_k^L = q\lambda \frac{\Omega_{d-1}}{\Omega_d} \int_{-1}^1 P_k^d(t)(1-t^2)^{(d-3)/2} dt = q\lambda \frac{\Omega_{d-1}}{\Omega_d} \langle P_0^d, P_k^d \rangle_{L^2([-1,1],(1-t^2)^{\frac{d-3}{2}} dt)} = 0$.

□