The Curse of Depth in Kernel Regime

Soufiane Hayou[*]
Department of Statistics
University of Oxford
United Kingdom

Arnaud Doucet
Department of Statistics
University of Oxford
United Kingdom

Judith Rousseau
Department of Statistics
University of Oxford
United Kingdom

Abstract

Recent work by Jacot et al. (2018) has shown that training a neural network of any kind with gradient descent is strongly related to kernel gradient descent in function space with respect to the Neural Tangent Kernel (NTK). Empirical results in (Lee et al., 2019) demonstrated high performance of a linearized version of training using the so-called NTK regime. In this paper, we show that the *large depth* limit of this regime is unexpectedly trivial, and we fully characterize the convergence rate to this trivial regime.

1 Introduction

The Neural Tangent Kernel (Jacot et al., 2018), a.k.a the NTK, has been the main focus of a growing number of works aiming to understand the inductive bias of Deep Neural Networks (DNNs). To cite a few, Bietti and Mairal (2019); Karakida et al. (2018); Yang (2019); Arora et al. (2019); Bietti and Bach (2021). In the so-called NTK regime (infinite width), the whole training procedure is reduced to a linear model given by the first order Taylor expansion of the output function near its initialization value. It was shown in Lee et al. (2019), that such simple models could surprisingly achieve high performance. However, most experiments with NTK regime were conducted on shallow networks and have not sufficiently covered DNNs.

NTK regime (Infinite width) for DNNs. The infinite width limit of the NTK for different architectures was studied by Yang (2019), who introduced a tensor framework that allows the derivation of recursive formulas for the NTK.

Information propagation. In parallel, information propagation in infinite width DNNs has been studied in several works (Hayou et al., 2019; Lee et al., 2018; Schoenholz et al., 2017; Yang and Schoenholz, 2017a; Poole et al., 2016) where the authors identify a set of hyper-parameters known as the Edge of Chaos (**EOC**) and activation functions ensuring a deep propagation of the information carried by the input. This ensures that the network output does not 'forget' the input information as the depth grows. In this paper, we show that this has a direct impact on the NTK.

Contributions. There have been few attempts to understand the large depth limit of the NTK regime Xiao et al. (2020); Huang et al. (2020); Bietti and Bach (2021); however, none of these works have characterized the *limiting NTK* and more importantly the *exact convergence rate* of the NTK to this limiting kernel. The closest work is Huang et al. (2020) where the authors considered a scaled version of ResNet with ReLU and proved an upper bound on the convergence rate of order $\mathcal{O}(\frac{\text{polylog} L}{L})$. In this paper, we improve this result in many ways: we prove that the convergence to the limiting NTK happens with a rate of $\boldsymbol{\Theta(\log(L)L^{-1})}$ for different architectures and activation functions. Note that for NTK regime, the lower bound is more important as it ensures a sub-exponential convergence rate of the NTK to its trivial limiting kernel(e.g. constant). We also show that the large depth behaviour of the NTK is initialization-sensitive; in particular, we prove that for FFNNs, we obtain an exponential

[*]Correspondence to: `soufiane.hayou@yahoo.fr`

I (Still) Can't Believe It's Not Better Workshop at NeurIPS 2021.

convergence rate if the initialization is not on the EOC (this is a generalization of Xiao et al. (2020)), which is not the case with ResNet.

2 Neural Networks and Neural Tangent Kernel

2.1 Setup and notations

Consider a neural network model consisting of L layers of widths $(n_l)_{1\leq l\leq L}$, $n_0 = d$ is the input dimension, 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. The output f of the neural network is given by some mapping $s : \mathbb{R}^{n_L} \to \mathbb{R}^o$ 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 $f(x,\theta) = s(y^L(x)) \in \mathbb{R}^o$. We denote by θ_t the value of θ at training time (step) t and $f_t(x) = f(x,\theta_t)$. Let $\mathcal{D} = (x_i, z_i)_{1\leq i\leq N}$ be the dataset, and let $\mathcal{X} = (x_i)_{1\leq i\leq N}$, $\mathcal{Z} = (z_j)_{1\leq j\leq N}$ be the sequences of inputs and outputs respectively. We assume that there is no collinearity in the input dataset $\mathcal{X}$, i.e. for all $x, x' \in \mathcal{X}$ and $\alpha \in \mathbb{R}$, $x' \neq \alpha x$. We also assume that $\mathcal{X} \subset E$ where $E \subset \mathbb{R}^d$ is a compact set.
The NTK is defined as the $o \times o$ dimensional kernel satisfying for all $x, x' \in \mathbb{R}^d$

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

The NTK regime (Infinite width). For a fully connected feedforward neural network, Jacot et al. (2018) proved that $K^L_{\theta_t}$ converges pointwise to a kernel K^L(depends only on L) for all $t < T$ when $\min\{n_1, n_2, ..., n_L\} \to \infty$, where T is a constant. In this limit, for the quadratic loss, f_t is given by a simple linear model

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

where $\hat{K}^L = K^L(\mathcal{X},\mathcal{X})$ and $\gamma(x,\mathcal{X}) = K^L(x,\mathcal{X})(\hat{K}^L)^{-1}$. Hereafter, we refer to f_∞ by the "NTK regime". For the cross-entropy loss, Lee et al. (2019) introduced some approximations to obtain the NTK regime. These approximations are implemented in Novak et al. (2020).

Scale invariance. f_∞ is *scale invariant* in the sense that it does not change if we scale the kernel by some depth dependent scalar since for any $a_L > 0$,

$$\gamma(x,\mathcal{X}) = K^L(x,\mathcal{X})(\hat{K}^L)^{-1} = (K^L(x,\mathcal{X})/a_L)(\hat{K}^L/a_L)^{-1}. \tag{2}$$

Thus, *studying the NTK regime with kernel K^L is equivalent to studying the NTK regime with any scaled kernel* K^L/a_L. In Theorems 1 and 2, we study scaled kernels to mitigate an exploding kernel effect in the limit of large depth, as the NTK regime solution remains unchanged.

Generalization in the NTK regime. As observed in Du et al. (2019), the convergence rate (w.r.t time) of f_t to f_∞ (infinite training time) is given by the smallest eigenvalue of $\hat{K}^L$. If $\hat{K}_L$ becomes singular in the large depth limit, then the performance of NTK regime decreases and might even be trivial. Notice also that we can write $f_t(x) - f_0(x) = \sum_{i=1}^N a_i K^L(x_i, x)$ for some $a_1, ..., a_N \in \mathbb{R}$, i.e. the 'residual' $f_t - f_0$ belongs to the Reproducing Kernel Hilbert space of the K^L.

3 Asymptotic Neural Tangent Kernel regime

In this section, we characterize the behaviour of K^L as L goes to ∞. We prove that K^L converges to a kernel K^∞ (which is trivial) with an initialization-and-architecture-dependent convergence rate.

3.1 Deep NTK of a FeedForward Neural Network (FFNN)

Consider an FFNN of depth L, widths $(n_l)_{1\leq l\leq L}$, weights w^l and bias b^l. For some input $x \in \mathbb{R}^d$, the forward propagation using the NTK parameterization (similar to Jacot et al. (2018)) is given by

$$y^1_i(x) = \frac{\sigma_w}{\sqrt{d}} \sum_{j=1}^{d} w^1_{ij} x_j + \sigma_b b^1_i, \quad y^l_i(x) = \frac{\sigma_w}{\sqrt{n_{l-1}}} \sum_{j=1}^{n_{l-1}} w^l_{ij}\phi(y^{l-1}_j(x)) + \sigma_b b^l_i, \quad l \geq 2. \tag{3}$$

We initialize the model randomly with $w_{ij}^l, b_i^l \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, where $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution of mean μ and variance σ^2. In the limit of infinite width, the neurons $(y_i^l(.))_{i,l}$ converge to Gaussian processes (Neal, 1995; Lee et al., 2018; Matthews et al., 2018; Hayou et al., 2019; Schoenholz et al., 2017). Hereafter, we denote by $q^l(x,x')$ the covariance between $y_1^l(x)$ and $y_1^l(x')$ (y_1^l can be replaced by any y_i^l since $(y_i^l)_i$ are i.i.d. See appendix 4 for a comprehensive review of the signal propagation theory). We define the correlation $c^l(x,x')$. For the first layer, we have $q^1(x,x') = \sigma_b^2 + \frac{\sigma_w^2}{d} x \cdot x'$. For $\epsilon \in (0,1)$, we define the set B_ϵ by:

$$\text{FFNN} : B_\epsilon = \{(x,x') \in \mathbb{R}^d : c^1(x,x') \leq 1 - \epsilon\},$$

and we assume that there exists $\epsilon > 0$, such that for all $x \neq x' \in \mathcal{X}, (x,x') \in B_\epsilon$.

Edge of Chaos (EOC). Given an input x, we denote by $q^l(x)$ the variance of $y^l(x)$. The asymptotic behaviour of $q^l(x)$ was studied in Lee et al. (2018), Schoenholz et al. (2017), and Hayou et al. (2019). Under general regularity conditions, $q^l(x)$ converges to a point $q(\sigma_b, \sigma_w) > 0$ independent of x. The asymptotic behaviour of the correlation $c^l(x,x')$ between $y^l(x)$ and $y^l(x')$ for any two inputs x and x' is driven by the choice of (σ_b, σ_w); Schoenholz et al. (2017) showed that if $\sigma_w^2 \mathbb{E}[\phi'(\sqrt{q(\sigma_b,\sigma_w)}Z)^2] < 1$, where $Z \sim \mathcal{N}(0,1)$, then $c^l(x,x')$ converges to 1 exponentially quickly; this is called the ordered phase. If $\sigma_w^2 \mathbb{E}[\phi'(\sqrt{q(\sigma_b,\sigma_w)}Z)^2] > 1$ then $c^l(x,x')$ converges to $c < 1$, which is then referred to as the chaotic phase. The authors define the EOC as the set of parameters (σ_b, σ_w) that satisfy $\sigma_w^2 \mathbb{E}[\phi'(\sqrt{q(\sigma_b,\sigma_w)}Z)^2] = 1$. The EOC was studied in (Hayou et al., 2019) where the authors showed that the correlation converges to 1 with a polynomial rate (see Section 4 in the appendix). The following proposition establishes that any initialization on the Ordered or Chaotic phase, leads to a trivial limiting NTK as L becomes large.

Proposition 1 (NTK 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*

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

The proof of Proposition 1 relies on the asymptotic analysis of the second moment of the gradient. We refer the reader to Section 6 in the appendix for more details.
Proposition 1 show that K^L becomes trivial exponentially quickly w.r.t deph. In this case, the NTK regime yields trivial performance, i.e. no better than that of a random classifier. Empirically, we find that with depth $L = 30$, the NTK training fails when the network is initialized on the Ordered phase ??????????(Section **??**).
In the next theorem, we show that the NTK explodes in the limit of large depth when the network is initialized on the EOC. Leveraging our remark on the scale invariance property of the NTK (see Eq. (2)), we show that a scaled version of the kernel converges with a polynomial rate to the degenerate kernel, meaning that the infinite-depth NTK regime is also trivial in this case, although the convergence is much slower. The notation $g(x) = \Theta(m(x))$ means there exist two constants $A, B > 0$ such that $A\, m(x) \leq g(x) \leq B\, m(x)$.

Theorem 1 (NTK on the EOC)**.** *Let (σ_b, σ_w) be on the EOC and $\tilde{K}^L = K^L/L$. We have that*

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

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

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

- $\tilde{K}^\infty(x,x') = \frac{\sigma_w^2 \|x\|\, \|x'\|}{d}(1 - (1-\lambda)\mathbb{1}_{x \neq x'})$ *with* $\lambda = 1/4$, *for* $\phi = \text{ReLU}$.
- $\tilde{K}^\infty(x,x') = q(1 - (1-\lambda)\mathbb{1}_{x \neq x'})$ *where $q > 0$ is a constant and $\lambda = 1/3$, for $\phi =$ Tanh.*

We refer the reader to Section 1 in the appendix for the proof details. Theorem 1 shows that the EOC initialization yields a polynomial convergence rate (w.r.t L) of $\tilde{K}^L$ to $\tilde{K}^\infty$. This is important knowing that $\tilde{K}^\infty$ is trivial and brings hardly any information on x. Indeed, the convergence rate of $\tilde{K}^L$ to $\tilde{K}^\infty$ is $\Theta(\log(L)L^{-1})$. This means that as L grows, the kernel $\tilde{K}$ with EOC is still much further from the trivial kernel $\tilde{K}^\infty$ compared to the Ordered/Chaotic initialization. Thus, *the EOC alleviates the curse of depth for NTK regime*. However, as shown in table 1, NTK regime fails for very deep networks ($L = 300$).

3.2 Residual Neural Networks (ResNet)

Another important feature of DNNs, which is known to be highly influential, is their architecture. For residual networks, the next theorem shows that for any $\sigma_w > 0$, the NTK of a ResNet explodes (exponentially) as L grows. However, a normalized version $\bar{K}^L = K^L/\alpha_L$ of the NTK of a ResNet will always have a polynomial convergence rate to a limiting trivial kernel.

Table 1: Test accuracy on CIFAR10 dataset after 100 training epochs for $L \in \{3, 30\}$ and 160 epochs for $L = 300$. V-ResNet is a ResNet with Fully Connected blocks.

	NTK regime		SGD Training	
	EOC	Ordered	EOC	Ordered
L=3				
FFNN-ReLU	$48.13_{\pm 0.10}$	$48.45_{\pm 0.14}$	$55.13_{\pm 0.23}$	$54.10_{\pm 0.12}$
FFNN-Tanh	$48.32_{\pm 0.15}$	$48.10_{\pm 0.10}$	$56.13_{\pm 0.34}$	$54.10_{\pm 0.23}$
CNN-ReLU	$49.11_{\pm 0.16}$	$42.76_{\pm 3.32}$	$60.23_{\pm 0.45}$	$59.05_{\pm 0.15}$
V-ResNet	$47.82_{\pm 0.73}$	$48.01_{\pm 0.20}$	$54.40_{\pm 0.24}$	$54.28_{\pm 0.33}$
L=30				
FFNN-ReLU	$48.32_{\pm 0.10}$	—	$56.10_{\pm 0.41}$	—
FFNN-Tanh	$48.40_{\pm 0.12}$	—	$57.39_{\pm 0.08}$	—
CNN-ReLU	$48.42_{\pm 0.10}$	—	$75.39_{\pm 0.31}$	—
V-ResNet	—	—	$57.09_{\pm 0.47}$	$58.13_{\pm 0.18}$
L=300				
FFNN-ReLU	—	—	$30.25_{\pm 3.23}$	—
FFNN-Tanh	—	—	$58.25_{\pm 0.43}$	—
CNN-ReLU	—	—	$76.25_{\pm 0.21}$	—
V-ResNet	—	—	$58.87_{\pm 0.44}$	$59.25_{\pm 0.10}$

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

$$y^l(x) = y^{l-1}(x) + \mathcal{F}(w^l, y^{l-1}(x)), \quad l \geq 2, \tag{4}$$

where $\mathcal{F}$ is a dense layer (Eq. (3)) 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 $\epsilon \in (0,1)$

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

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

The proof techniques used in Theorem 2 are similar to those used in the proof of theorem 1. Details are provided in the appendix.

Theorem 2 shows that the NTK of a ReLU ResNet explodes exponentially w.r.t L. However, the normalized kernel $\bar{K}^L_{res} = K^L_{res}(x,x')/\alpha_L$ converges to a limiting kernel $\bar{K}^\infty_{res}$ at the exact polynomial rate $\Theta(\log(L)L^{-1})$ for all $\sigma_w > 0$. This suggests that ResNet act by default as an FFNN that is initialized on the EOC. However, $\bar{K}^L_{res}$ converges to a trivial kernel, which means that, even with ResNet, the performance of the NTK regime will decrease as we increase the depth, although it happens with a polynomial rate. Table 1 shows the performance of the NTK regime versus standard SGD training on CIFAR10. While the NTK regime fails with $L = 300$ for both Ordered/EOC initializations, it yields non-trivial performance when initialized on the EOC with $L = 30$, which is not the case with an Ordered phase. With ResNet, the performance is similar for both initializations which confirms the results of theorem 2. However, for all initializations schemes, NTK regime fails for $L = 300$ while standard SGD training succeeds.

We now leverage the previous results to obtain the asymptotic behaviour of the spectrum of the kernels studied in Theorems 1, 2 and Proposition 1, on the unit sphere $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$. On the sphere $\mathbb{S}^{d-1}$, all of these kernels (namely K^L for FFNN on the Ordered/Chaotic phase, $\tilde{K}^L$ for FFNN/CNN on the EOC, and $\bar{K}^L_{res}$ for ResNets) are dot-product kernels, i.e. for any of these kernels, denoted by κ_L, there exists a function g_L such that $\kappa^L(x,x') = g_L(x \cdot x')$ for all $x, x' \in \mathbb{S}^{d-1}$ (we refer the reader to appendix 4 for more details). This type of kernels is known to be diagonalizable on the sphere $\mathbb{S}^{d-1}$ and its eigenfunctions are the so-called Spherical Harmonics of $\mathbb{S}^{d-1}$. Many concurrent results have observed this fact Geifman et al. (2020); Cao et al. (2020); Bietti and Bach (2021). In the next proposition, we leverage the results of Section 3 to study the aforementioned kernels from a spectral perspective.

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

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

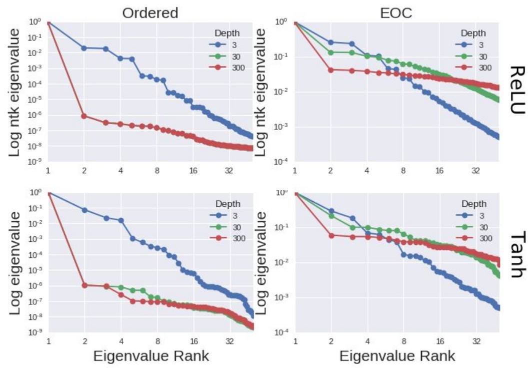

Figure 1: Normalized eigenvalues of K_L on the 2D sphere for an FFNN with different initializations, activations, and depths. (Red and Green lines are identical in the upper left figure.)

$(Y_{k,j})_{k\geq 0, j\in[1:N(d,k)]}$ *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* $k \geq 1$*,* $\lim_{L\to\infty} \mu_k^L = 0$*.*

The proof of Proposition 2 is based on a result from spectral theory analysis. The limiting eigenvalues are obtained by a simple application of the dominated convergence theorem.

Proposition 2 shows that in the limit of large L, the kernel κ^L becomes close to the trivial kernel $\kappa^\infty\,(x,x') \mapsto \mu_0^\infty Y_{0,0}(x)Y_{0,0}(x')$, where $Y_{0,0}$ is the constant function in the spherical harmonics class. Therefore, in the limit of infinite depth, the RKHS of the kernel κ^L is reduced to the space of constant functions, confirming that the NTK regime is trivial in this limit (recall that $f_\infty - f_0$ is in the RKHS of κ_L). Fig 1 illustrates this deterioration of the spectrum as the depth increases. Notice that with EOC, the deterioration happens with a much slower rate, which is expected from theorems 1 and 2.

4 Conclusion and Limitations

In this paper, we have shown that the infinite depth limit of the NTK regime is trivial and cannot explain the performance of DNNs. However, this convergence is initialization dependent. These findings add to a recent line of research which shows that the infinite width approximation of the NTK does not fully capture the training dynamics of DNNs (Chizat and Bach, 2018; Ghorbani et al., 2019; Huang and Yau, 2020; Hanin and Nica, 2019).

References

Arora, S., S. Du, W. Hu, , Z. Li, and R. Wand (2019). Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *ICML*.

Bietti, A. and F. Bach (2021). Deep equals shallow for reLU networks in kernel regimes. In *International Conference on Learning Representations*.

Bietti, A. and J. Mairal (2019). On the inductive bias of neural tangent kernels. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 32, pp. 12893–12904. Curran Associates, Inc.

Cao, Y., Z. Fang, Y. Wu, D. Zhou, and Q. Gu (2020). Towards understanding the spectral bias of deep learning. *arXiv prePrint 1912.01198*.

Chizat, L. and F. Bach (2018). A note on lazy training in supervised differentiable programming. *arXiv preprint arXiv:1812.07956*.

Du, S., J. Lee, H. Li, L. Wang, and X. Zhai (2019). Gradient descent finds global minima of deep neural networks. *ICML*.

Geifman, A., A. Yadav, Y. Kasten, M. Galun, D. Jacobs, and R. Basri (2020). On the similarity between the laplace and neural tangent kernels. *NeurIPS*.

Ghorbani, B., S. Mei, T. Misiakiewicz, and A. Montanari (2019). Linearized two-layers neural networks in high dimension. *arXiv preprint arXiv:1904.12191*.

Hanin, B. and M. Nica (2019). Finite depth and width corrections to the neural tangent kernel. *arXiv preprint arXiv:1909.05989*.

Hayou, S., A. Doucet, and J. Rousseau (2019). On the impact of the activation function on deep neural networks training. *ICML*.

Huang, J. and H. Yau (2020). Dynamics of deep neural networks and neural tangent hierarchy. *ICML*.

Huang, K., Y. Wang, M. Tao, and T. Zhao (2020). Why do deep residual networks generalize better than deep feedforward networks? – a neural tangent kernel perspective. *ArXiv preprint, arXiv:2002.06262*.

Jacot, A., F. Gabriel, and C. Hongler (2018). Neural tangent kernel: Convergence and generalization in neural networks. *32nd Conference on Neural Information Processing Systems*.

Karakida, R., S. Akaho, and S. Amari (2018). Universal statistics of Fisher information in deep neural networks: Mean field approach. *arXiv preprint arXiv:1806.01316*.

Lee, J., Y. Bahri, R. Novak, S. Schoenholz, J. Pennington, and J. Sohl-Dickstein (2018). Deep neural networks as Gaussian processes. *6th International Conference on Learning Representations*.

Lee, J., L. Xiao, S. Schoenholz, Y. Bahri, J. Sohl-Dickstein, and J. Pennington (2019). Wide neural networks of any depth evolve as linear models under gradient descent. *NeurIPS*.

Lillicrap, T., D. Cownden, D. Tweed, and C. Akerman (2016). Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications 7*(13276).

MacRobert, T. (1967). *Spherical harmonics: An elementary treatise on harmonic functions, with applications*. Pergamon Press.

Matthews, A., J. Hron, M. Rowland, R. Turner, and Z. Ghahramani (2018). Gaussian process behaviour in wide deep neural networks. *6th International Conference on Learning Representations*.

Neal, R. (1995). Bayesian learning for neural networks. *Springer Science & Business Media 118*.

Novak, R., L. Xiao, J. Hron, J. Lee, A. A. Alemi, J. Sohl-Dickstein, and S. S. Schoenholz (2020). Neural tangents: Fast and easy infinite neural networks in python. In *International Conference on Learning Representations*.

Poole, B., S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli (2016). Exponential expressivity in deep neural networks through transient chaos. *30th Conference on Neural Information Processing Systems*.

Schoenholz, S., J. Gilmer, S. Ganguli, and J. Sohl-Dickstein (2017). Deep information propagation. *5th International Conference on Learning Representations*.

Xiao, L., Y. Bahri, J. Sohl-Dickstein, S. S. Schoenholz, and P. Pennington (2018). Dynamical isometry and a mean field theory of cnns: How to train 10,000-layer vanilla convolutional neural networks. *ICML 2018*.

Xiao, L., J. Pennington, and S. Schoenholz (2020). Disentangling trainability and generalization in deep neural networks. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 10462–10472.

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

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

Yang, G. and S. Schoenholz (2017a). Mean field residual networks: On the edge of chaos. *Advances in Neural Information Processing Systems 30*, 2869–2869.

Yang, G. and S. Schoenholz (2017b). Mean field residual networks: On the edge of chaos. In *Advances in neural information processing systems*, pp. 7103–7114.

Appendix

0 Setup and notations

In the appendix, we prove more general results than those presented in the main text (we have chosen to present partial results for the sake of simplicity). The reader will find versions of Propositions 1, 2, and Theorems 1, 2, that include Convolutional Neural Networks.

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 Architectures

Let ϕ be the activation function. We consider the following architectures (FFNN and CNN)

- **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

$$\begin{aligned} y_i^1(x) &= \frac{\sigma_w}{\sqrt{d}} \sum_{j=1}^{d} w_{ij}^1 x_j + \sigma_b b_i^1 \\ y_i^l(x) &= \frac{\sigma_w}{\sqrt{n_{l-1}}} \sum_{j=1}^{n_{l-1}} w_{ij}^l \phi(y_j^{l-1}(x)) + \sigma_b b_i^l, \quad l \ge 2. \end{aligned} \tag{10}$$

- **Convolutional Neural Network (CNN/ConvNet)** Consider a 1D convolutional neural network of depth L, denoting by $[m : n]$ the set of integers $\{m, m+1, ..., n\}$ for $n \le m$, the forward propagation is given by

$$\begin{aligned} y_{i,\alpha}^1(x) &= \frac{\sigma_w}{\sqrt{v_1}} \sum_{j=1}^{n_0} \sum_{\beta \in ker_1} w_{i,j,\beta}^1 x_{j,\alpha+\beta} + \sigma_b b_i^1 \\ y_{i,\alpha}^l(x) &= \frac{\sigma_w}{\sqrt{v_l}} \sum_{j=1}^{n_{l-1}} \sum_{\beta \in ker_l} w_{i,j,\beta}^l \phi(y_{j,\alpha+\beta}^{l-1}(x)) + \sigma_b b_i^l, \end{aligned} \tag{11}$$

  where $i \in [1 : n_l]$ is the channel number, $\alpha \in [0 : M-1]$ is the neuron location in the channel, n_l is the number of channels in the l^{th} layer, and M is the number of neurons in each channel, $ker_l = [-k : k]$ is a filter with size $2k+1$ and $v_l = n_{l-1}(2k+1)$. Here, $w^l \in \mathbb{R}^{n_l \times n_{l-1} \times (2k+1)}$. We assume periodic boundary conditions, which results in having $y_{i,\alpha}^l = y_{i,\alpha+M}^l = y_{i,\alpha-M}^l$ and similarly for $l = 0$, $x_{i,\alpha+M_0} = x_{i,\alpha} = x_{i,\alpha-M_0}$. For the sake of simplification, we consider only the case of 1D CNN, the generalization to a mD CNN for $m \in \mathbb{N}$ is straightforward.

1 Proof techniques

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

Lemmas 1, 2, 3, 4. 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. However, for CNNs, the dynamics are a bit trickier since they involve convolution operators; We use some results from the theory of Circulant Matrices for this purpose.
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 through 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.

Convolutional Neural Networks. The infinite width approximation with 1D CNN yields a recursion for the kernel. However, the infinite width here means infinite number of channels, with a Monte Carlo error of $\mathcal{O}(1/\sqrt{n_{l-1}})$. The kernel in this case depends on the choice of the neurons in the channel and is given by

$$\kappa_{\alpha,\alpha'}^l(x, x') = \mathbb{E}[y_{i,\alpha}^l(x) y_{i,\alpha'}^l(x')] = \sigma_b^2 + \frac{\sigma_w^2}{2k+1} \sum_{\beta \in ker} \mathbb{E}[\phi(y_{1,\alpha+\beta}^{l-1}(x))\phi(y_{1,\alpha'+\beta}^{l-1}(x'))]$$

so that

$$\kappa_{\alpha,\alpha'}^l(x, x') = \sigma_b^2 + \frac{\sigma_w^2}{2k+1} \sum_{\beta \in ker} F_\phi(\kappa_{\alpha+\beta,\alpha'+\beta}^{l-1}(x, x), \kappa_{\alpha+\beta,\alpha'+\beta}^{l-1}(x, x'), \kappa_{\alpha+\beta,\alpha'+\beta}^{l-1}(x', x')).$$

The convolutional kernel $\kappa_{\alpha,\alpha'}^l$ has the 'self-averaging' property; i.e. it is an average over the kernels corresponding to different combination of neurons in the previous layer. However, it is easy to simplify the analysis in this case by studying the average kernel per channel defined by $\hat{\kappa}^l = \frac{1}{N^2} \sum_{\alpha,\alpha'} \kappa_{\alpha,\alpha'}^l$. Indeed, by summing terms in the previous equation and using the fact that we use circular padding, we obtain

$$\hat{\kappa}^l(x, x') = \sigma_b^2 + \sigma_w^2 \frac{1}{N^2} \sum_{\alpha,\alpha'} F_\phi(\kappa_{\alpha,\alpha'}^{l-1}(x, x), \kappa_{\alpha,\alpha'}^{l-1}(x, x'), \kappa_{\alpha,\alpha'}^{l-1}(x', x')).$$

This expression is similar in nature to that of FFNN. We will use this observation in the proofs.

Note that our analysis only requires the approximation that, in the infinite width limit, for any two inputs x, x', the variables $y_i^l(x)$ and $y_i^l(x')$ are Gaussian with covariance $\kappa^l(x, x')$ for FFNN, and $y_{i,\alpha}^l(x)$ and $y_{i,\alpha'}^l(x')$ are Gaussian with covariance $\kappa_{\alpha,\alpha'}^l(x, x')$ for CNN. We do not need the much stronger approximation that the process $y_i^l(x)$ ($y_{i,\alpha}^l(x)$ for CNN) is a Gaussian process.

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

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

whereas for ResNet with CNN-type layers, we have that

$$\begin{aligned}\kappa_{\alpha,\alpha'}^l(x, x') &= \kappa_{\alpha,\alpha'}^{l-1}(x, x') + \sigma_b^2 \\ &\quad + \frac{\sigma_w^2}{2k+1} \sum_{\beta \in ker} F_\phi(\kappa_{\alpha+\beta,\alpha'+\beta}^{l-1}(x, x), \kappa_{\alpha+\beta,\alpha'+\beta}^{l-1}(x, x'), \kappa_{\alpha+\beta,\alpha'+\beta}^{l-1}(x', x')).\end{aligned}$$

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 Discussion on Assumption 1

Assumption 1. *We assume that for all $x, x' \in \mathcal{X}$, $q_{\alpha,\alpha'}^1(x, x')$ is independent of α, α'.*

Assumption 1 implies that, there exists some function $e : (x, x') \mapsto e(x, x')$ such that for all α, α', x, x'

$$\sum_j \sum_{\beta \in ker_0} x_{j,\alpha+\beta} x'_{j,\alpha'+\beta} = e(x, x')$$

This system has N^2M^2 equations and $N \times 2n_0 \times M$ variables. Therefore, in the case $n_0 >> 1$, the set of solutions S is large. By using Assumption 1, we restrict our analysis to this case. Hereafter, for all CNN analysis, for some function G and set E, *taking the supremum* $\sup_{(x,x') \in E} G(x, x')$ *should be interpreted as* $\sup_{(x,x') \in E \cap \mathcal{X}^2} G(x, x')$.

Another justification to assumption 1 can be attributed a self-averaging property of the dynamics of the correlation inside a CNN. We refer the reader to the proof of Appendix lemma 3 for more details.

4 Warmup: Results from the Mean-Field theory of DNNs

4.1 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 $\dot{q}^l(x,x') = \sigma_w^2 \mathbb{E}[\phi'(y_1^{l-1}(x))\phi'(y_1^{l-1}(x'))]$.
For CNN layers, we use similar notation across channels. Let $q_\alpha^l(x)$ be the variance of $y_{1,\alpha,}^l(x)$ (the choice of the index 1 is not important here either since, in the limit of infinite number of channels, the random variables $(y_{i,\alpha}^l(x))_{i\in[1:N_l]}$ are iid). Let $q_{\alpha,\alpha'}^l(x,x')$ the covariance between $y_{1,\alpha}^l(x)$ and $y_{1,\alpha'}^l(x')$, and $c_{\alpha,\alpha'}^l(x,x')$ the corresponding correlation. We also define the pseudo-covariance $\hat{q}_{\alpha,\alpha'}^l(x,x') = \sigma_b^2 + \sigma_w^2 \mathbb{E}[\phi(y_{1,\alpha}^{l-1}(x))\phi(y_{1,\alpha'}^{l-1}(x'))]$ and $\dot{q}_{\alpha,\alpha'}^l(x,x') = \sigma_w^2 \mathbb{E}[\phi(y_{1,\alpha}^{l-1}(x))\phi(y_{1,\alpha'}^{l-1}(x'))]$.
The Gradient covariance is defined by $\tilde{q}_{\alpha,\alpha'}^l(x,x') = \mathbb{E}\left[\frac{\partial \mathcal{L}}{\partial y_{1,\alpha}^l}(x)\frac{\partial \mathcal{L}}{\partial y_{1,\alpha'}^l}(x')\right]$.

4.1.1 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

$$q^l(x,x') = \sigma_b^2 + \sigma_w^2 \mathbb{E}[\phi(y_i^{l-1}(x))\phi(y_i^{l-1}(x'))] \tag{12}$$

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 CNN. The only difference with FFNN is that the independence is across channels and not neurons. Simple calculus yields

$$q_{\alpha,\alpha'}^l(x,x') = \mathbb{E}[y_{i,\alpha}^l(x)y_{i,\alpha'}^l(x')] = \sigma_b^2 + \frac{\sigma_w^2}{2k+1}\sum_{\beta\in ker}\mathbb{E}[\phi(y_{1,\alpha+\beta}^{l-1}(x))\phi(y_{1,\alpha'+\beta}^{l-1}(x'))]$$

Observe that

$$q_{\alpha,\alpha'}^l(x,x') = \frac{1}{2k+1}\sum_{\beta inker}\hat{q}_{\alpha+\beta,\alpha'+\beta}^l(x,x') \tag{13}$$

With ReLU, we have

$$q_{\alpha,\alpha'}^l(x,x') = \sigma_b^2 + \frac{\sigma_w^2}{2k+1}\sum_{\beta\in ker}\sqrt{q_{\alpha+\beta}^l(x)}\sqrt{q_{\alpha'+\beta}^l(x')}f(c_{\alpha+\beta,\alpha'+\beta}^{l-1}(x,x')).$$

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

$$q^l(x,x') = q^{l-1}(x,x') + \sigma_b^2 + \frac{\sigma_w^2}{2}\sqrt{q^l(x)}\sqrt{q^l(x')}f(c^{l-1}) \tag{14}$$

with CNN layers, we have instead

$$q_{\alpha,\alpha'}^l(x,x') = q_{\alpha,\alpha'}^{l-1}(x,x') + \sigma_b^2 + \frac{\sigma_w^2}{2k+1}\sum_{\beta\in ker}\sqrt{q_{\alpha+\beta}^l(x)}\sqrt{q_{\alpha'+\beta}^l(x')}f(c_{\alpha+\beta,\alpha'+\beta}^{l-1}(x,x')) \tag{15}$$

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

Gradient Covariance back-propagation for CNN. We have that

$$\frac{\partial \mathcal{L}}{\partial W^l_{i,j,\beta}} = \sum_\alpha \frac{\partial \mathcal{L}}{\partial y^l_{i,\alpha}} \phi(y^{l-1}_{j,\alpha+\beta})$$

Moreover,

$$\frac{\partial \mathcal{L}}{\partial y^l_{i,\alpha}} = \sum_{j=1}^{n} \sum_{\beta \in ker} \frac{\partial \mathcal{L}}{\partial y^{l+1}_{j,\alpha-\beta}} W^{l+1}_{i,j,\beta} \phi'(y^l_{i,\alpha}).$$

Using the Gradient Independence 2.2, and taking the average over the number of channels we have that

$$\mathbb{E}\left[\frac{\partial \mathcal{L}}{\partial y^l_{i,\alpha}}^2\right] = \frac{\sigma_w^2 \mathbb{E}\left[\phi'(\sqrt{q^l_\alpha(x)}Z)^2\right]}{2k+1} \sum_{\beta \in ker} \mathbb{E}\left[\frac{\partial \mathcal{L}}{\partial y^{l+1}_{i,\alpha-\beta}}^2\right].$$

We can get similar recursion to that of the FFNN case by summing over α and using the periodic boundary condition, this yields

$$\sum_\alpha \mathbb{E}\left[\frac{\partial \mathcal{L}}{\partial y^l_{i,\alpha}}^2\right] = \chi(q^l_\alpha(x)) \sum_\alpha \mathbb{E}\left[\frac{\partial \mathcal{L}}{\partial y^{l+1}_{i,\alpha}}^2\right].$$

4.1.3 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 $w^l_{ij}, b^l_i \overset{iid}{\sim} \mathcal{N}(0, 1)$, 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 calculations are done in the limit of infinitely wide networks.

4.2 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 $(\sigma_b, \sigma_w) \in \mathrm{Ord}$, 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 (12).

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 $(\sigma_b, \sigma_w) \in \mathrm{Ch}$, 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 12 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*

$$f(c) = c + s(1-c)^{3/2} + b(1-c)^{5/2} + O((1-c)^{7/2}) \tag{16}$$

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 $5/2$.

Results for CNN with Tanh activation function.

Fact 15. *For any choice of $\sigma_b, \sigma_w \in \mathbb{R}^+$, there exist $q, \lambda > 0$ such that for all $l \geq 1$, $\sup_{\alpha,\alpha'} \sup_{x\in\mathbb{R}^d} |q^l_{\alpha,\alpha'}(x,x) - q| \leq e^{-\lambda l}$. (Equation (2.5) in Xiao et al. (2018) and variance convergence result in Schoenholz et al. (2017)).*

The behaviour of the correlation $c^l_{\alpha,\alpha'}(x,x')$ was studied in Xiao et al. (2018) only in the case $x' = x$. We give a comprehensive analysis of the asymptotic behaviour of $c^l_{\alpha,\alpha'}(x,x')$ in the next section.

General results on the correlation function.

Fact 16. *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 17. *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 5 (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 5. 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. □

4.3 Large depth behaviour of the correlation in CNNs

For CNNs, the infinite width will always mean the limit of infinite number of channels. Recall that, by definition, $\hat{q}^l_{\alpha,\alpha'}(x,x') = \sigma_b^2 + \sigma_w^2 \mathbb{E}[\phi(y^{l-1}_{1,\alpha}(x))\phi(y^{l-1}_{1,\alpha'}(x'))]$ and $q^l_{\alpha,\alpha'}(x,x') = \mathbb{E}[y^l_{i,\alpha}(x)y^l_{i,\alpha'}(x')]$.

Unlike FFNN, neurons in the same channel are correlated since they share the same filters. Let x, x' be two inputs and α, α' two nodes in the same channel i. Using Central Limit Theorem in the limit of large n_l (number of channels), we have

$$q^l_{\alpha,\alpha'}(x,x') = \mathbb{E}[y^l_{i,\alpha}(x)y^l_{i,\alpha'}(x')] = \frac{\sigma_w^2}{2k+1}\sum_{\beta\in ker}\mathbb{E}[\phi(y^{l-1}_{1,\alpha+\beta}(x))\phi(y^{l-1}_{1,\alpha'+\beta}(x'))] + \sigma_b^2$$

Let $c^l_{\alpha,\alpha'}(x,x')$ be the corresponding correlation. Since $q^l_{\alpha,\alpha}(x,x)$ converges exponentially to q which depends neither on x nor on α, the mean-field correlation as in Schoenholz et al. (2017); Hayou et al. (2019) is given by

$$c^l_{\alpha,\alpha'}(x,x') = \frac{1}{2k+1}\sum_{\beta\in ker} f(c^{l-1}_{\alpha+\beta,\alpha'+\beta}(x,x'))$$

where $f(c) = \frac{\sigma_w^2\mathbb{E}[\phi(\sqrt{q}Z_1)\phi(\sqrt{q}(cZ_1+\sqrt{1-c^2}Z_2))]+\sigma_b^2}{q}$ and Z_1, Z_2 are independent standard normal variables. The dynamics of $c^l_{\alpha,\alpha'}$ become similar to those of c^l in an FFNN under assumption 1. We show this in the proof of Appendix Lemma 3.

In Xiao et al. (2018), authors studied only the limiting behaviour of correlations $c^l_{\alpha,\alpha'}(x,x)$ (same input x), however, they do not study $c^l_{\alpha,\alpha'}(x,x')$ when $x \neq x'$. We do this in the following Lemma, which will prove also useful for the main results of the paper.

Appendix Lemma 3 (Asymptotic behaviour of the correlation in CNN with Tanh)**.** *We consider a CNN with Tanh activation function. Let $(\sigma_b,\sigma_w)\in(\mathbb{R}^+)^2$ and $\epsilon\in(0,1)$. Let $B_\epsilon = \{(x,x')\in\mathbb{R}^d : \sup_{\alpha,\alpha'} c^1_{\alpha,\alpha'}(x,x') < 1-\epsilon\}$. The following statements hold*

1. *If (σ_b,σ_w) are on the Ordered phase, then there exists $\beta>0$ such that*
$$\sup_{(x,x')\in\mathbb{R}^d}\sup_{\alpha,\alpha'}|c^l_{\alpha,\alpha'}(x,x')-1| = \mathcal{O}(e^{-\beta l})$$

2. *If (σ_b,σ_w) are on the Chaotic phase, then for all $\epsilon>0$ there exists $\beta>0$ and $c\in(0,1)$ such that*
$$\sup_{(x,x')\in B_\epsilon}\sup_{\alpha,\alpha'}|c^l_{\alpha,\alpha'}(x,x')-c| = \mathcal{O}(e^{-\beta l})$$

3. *Under Assumption 1, if $(\sigma_b,\sigma_w)\in EOC$, then we have*
$$\sup_{(x,x')\in B_\epsilon}\sup_{\alpha,\alpha'}\left|c^l_{\alpha,\alpha'}(x,x') - 1 + \frac{\kappa}{l} - \kappa(1-\kappa^2\zeta)\frac{\log(l)}{l^3}\right| = \mathcal{O}(l^{-3})$$
where $\kappa = \frac{2}{f''(1)} > 0$, $\zeta = \frac{f^3(1)}{6} > 0$, and f is the correlation function given in Fact 5.

We prove statements 1 and 2 for general inputs, i.e. without using Assumption 1. The third statement requires Assumption 1.

Proof. Let $(x,x')\in\mathbb{R}^d$. Without using assumption 1, we have that

$$c^l_{\alpha,\alpha'}(x,x') = \frac{1}{2k+1}\sum_{\beta\in ker} f(c^{l-1}_{\alpha+\beta,\alpha'+\beta}(x,x'))$$

Writing this in matrix form yields

$$C_l = \frac{1}{2k+1}Uf(C_{l-1})$$

where $C_l = ((c^l_{\alpha,\alpha+\beta}(x,x'))_{\alpha\in[0:N-1]})_{\beta\in[0:N-1]}$ is a vector in $\mathbb{R}^{N^2}$, U is a convolution matrix and f is applied element-wise. As an example, for $k=1$, U is given by

$$U = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 & 1\\ 1 & 1 & 1 & 0 & \ddots & 0\\ 0 & 1 & 1 & 1 & \ddots & 0\\ 0 & 0 & 1 & 1 & \ddots & 0\\ & \ddots & \ddots & \ddots & \ddots & \\ 1 & 0 & \dots & 0 & 1 & 1 \end{bmatrix}$$

For general k, U is a Circulant symmetric matrix with eigenvalues $\lambda_1 > \lambda_2 \geq \lambda_3 ... \geq \lambda_{N^2}$. The largest eigenvalue of U is given by $\lambda_1 = 2k+1$ and its equivalent eigenspace is generated by the vector $e_1 = \frac{1}{N}(1,1,...,1) \in \mathbb{R}^{N^2}$. This yields

$$(1+2k)^{-l}U^l = e_1 e_1^T + O(e^{-\beta l})$$

where $\beta = \log(\frac{\lambda_1}{\lambda_2})$.
This provides another justification to Assumption 1; as l grows, and assuming that $C_l \to e_1$ (which we show in the remainder of this proof), C_l exhibits a self-averaging property since $C_l \approx \frac{1}{2k+1}UC_{l-1}$. This system concentrates around the average value of the entries of C_l as l grows. Since the variances converge to a constant q as l goes to infinity (fact 15), this approximation implies that the entries of C_l become almost equal as l goes to infinity, thus making assumption 1 almost satisfied in deep layers. Let us now prove the statements.

1. Let (σ_b, σ_w) be in the Ordered phase, $(x, x') \in \mathbb{R}^d$ and $c^l_m = \min_{\alpha,\alpha'} c^l_{\alpha,\alpha'}(x,x')$. Using the fact that f is non-decreasing, we have that $c^l_{\alpha,\alpha'}(x,x') \geq \frac{1}{2k+1}\sum_{\beta\in ker} c^{l-1}_{\alpha+\beta,\alpha'+\beta}(x,x')) \geq c^{l-1}_m$. Taking the minimum again over α, α', we have $c^l_m \geq c^{l-1}_m$, therefore c^l_m is non-decreasing and converges to the unique fixed point of f which is $c = 1$. This proves that $\sup_{\alpha,\alpha'} |c^l_{\alpha,\alpha'}(x,x') - 1| \to 0$. Moreover, the convergence rate is exponential using the fact that (fact 16) $0 < f'(1) < 1$. To see this, observe that

   $$\sup_{\alpha,\alpha'} |1 - c^l_{\alpha,\alpha'}(x,x')| \leq \left(\sup_{\zeta \in [c^{l-1}_m, 1]} f'(\zeta) \right) \times \sup_{\alpha,\alpha'} |1 - c^l_{\alpha,\alpha'}(x,x')|$$

   Knowing that $\sup_{\zeta\in[c^{l-1}_m,1]} f'(\zeta) \to f'(1) < 1$, we conclude. Moreover, the convergence is uniform in (x, x') since the convergence rate depends only on $f'(1)$.

2. Let $\epsilon \in (0,1)$. In the chaotic phase, the only difference is the limit $c = c_1 < 1$ and the Supremum is taken over B_ϵ to avoid points where $c^1(x,x') = 1$. In the Chaotic phase (fact 16), f has two fixed points, 1 is an unstable fixed point and $c_1 \in (0,1)$ which is the unique stable fixed point. We conclude by following the same argument.

3. Let $\epsilon \in (0,1)$ and $(\sigma_b, \sigma_w) \in$ EOC. Using the same argument of monotony as in the previous cases and that f has 1 as unique fixed point, we have that $\lim_{l\to\infty} \sup_{x,x'} \sup_{\alpha,\alpha'} |1 - c^l_{\alpha,\alpha'}(x,x')| = 0$. From fact 7, the Taylor expansion of f near 1 is given by

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

   where $\alpha = \frac{f''(1)}{2}$ and $\zeta = \frac{f^{(3)}(1)}{6}$. Using fact 6, we know that $f^{(k)}(1) = \sigma_w^2 q^{k-1}\mathbb{E}[\phi^{(k)}(\sqrt{q}Z)^2]$. Therefore, we have $\alpha > 0$, and $\zeta < 0$.
   Under assumption 1, it is straightforward that for all α, α', and $l \geq 1$

   $$c^l_{\alpha,\alpha'}(x,x') = c^l(x,x')$$

   i.e. $c^l_{\alpha,\alpha'}$ are equal for all α, α'. The dynamics of $c^l(x,x')$ are exactly the dynamics of the correlation in an FFNN. We conclude using Appendix Lemma 2.

□

It is straightforward that the previous Appendix Lemma extend to ReLU activation, with slightly different dynamics. In this case, we use Appendix Lemma 1 to conclude for the third statement.

Appendix Lemma 4 (Asymptotic behaviour of the correlation in CNN with ReLU-like activation functions)**.** *We consider a CNN with ReLU activation. Let $(\sigma_b, \sigma_w) \in (\mathbb{R}^+)^2$. Let $(\sigma_b, \sigma_w) \in (\mathbb{R}^+)^2$ and $\epsilon \in (0,1)$. The following statements hold*

1. *If (σ_b, σ_w) are on the Ordered phase, then there exists $\beta > 0$ such that*

   $$\sup_{(x,x')\in\mathbb{R}^d} \sup_{\alpha,\alpha'} |c^l_{\alpha,\alpha'}(x,x') - 1| = \mathcal{O}(e^{-\beta l})$$

2. *If (σ_b, σ_w) are on the Chaotic phase, then there exists $\beta > 0$ and $c \in (0,1)$ such that*

   $$\sup_{(x,x')\in B_\epsilon} \sup_{\alpha,\alpha'} |c^l_{\alpha,\alpha'}(x,x') - c| = \mathcal{O}(e^{-\beta l})$$

3. *Under Assumption 1, if* $(\sigma_b, \sigma_w) \in EOC$*, then*

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

where $\kappa, \kappa' > 0$ *are universal constants.*

Proof. The proof is similar to the case of Tanh in Appendix Lemma 3. The only difference is that we use Appendix Lemma 1 to conclude for the third statement. □

5 A technical tool 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 5 (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 5 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 5.

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

6.1 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

$$\begin{aligned}(\partial_{w^{l+1}} y_i^{l+1}(x))(\partial_{w^{l+1}} y_i^{l+1}(x'))^t + (\partial_{b^{l+1}} y_i^{l+1}(x))(\partial_{b^{l+1}} y_i^{l+1}(x'))^t &= \frac{\sigma_w^2}{n_l}\sum_j \phi(y_j^l(x))\phi(y_j^l(x')) + \sigma_b^2 \\ &\underset{n_l\to\infty}{\to} \sigma_w^2\mathbb{E}[\phi(y_i^l(x))\phi(y_i^l(x'))] + \sigma_b^2 = q^{l+1}(x,x').\end{aligned}$$

which ends the proof.

□

We now provide the recursive formula satisfied by the NTK of a CNN, namely Lemma 2 of the paper.

Lemma 2 (Infinite width dynamics of the NTK of a CNN)**.** *Consider a CNN of the form (**??**), then we have that for all $x, x' \in \mathbb{R}^d$, $i, i' \leq n_1$ and $\alpha, \alpha' \in [0 : M-1]$*

$$K^1_{(i,\alpha),(i',\alpha')}(x,x') = \delta_{ii'}\big(\frac{\sigma_w^2}{n_0(2k+1)}[x,x']_{\alpha,\alpha'} + \sigma_b^2\big)$$

For $l \geq 2$, as $n_1, n_2, ..., n_{l-1} \to \infty$ recursively, we have for all $i, i' \leq n_l$, $\alpha, \alpha' \in [0 : M-1]$, $K^l_{(i,\alpha),(i',\alpha')}(x,x') = \delta_{ii'}K^l_{\alpha,\alpha'}(x,x')$, where $K^l_{\alpha,\alpha'}$ is given by the recursive formula

$$K^l_{\alpha,\alpha'} = \frac{1}{2k+1}\sum_{\beta\in ker_l}\Psi^{l-1}_{\alpha+\beta,\alpha'+\beta}$$

where $\Psi^{l-1}_{\alpha,\alpha'} = \dot{q}^l_{\alpha,\alpha'}K^{l-1}_{\alpha,\alpha'} + \hat{q}^l_{\alpha,\alpha'}$, and $\hat{q}^l_{\alpha,\alpha}, \dot{q}^l_{\alpha,\alpha'}$ are defined in Lemma 1, with $y^{l-1}_{1,\alpha}(x), y^{l-1}_{1,\alpha'}(x')$ in place of $y^{l-1}_1(x), y^{l-1}_1(x')$.

Proof. Let x, x' be two inputs. We have that

$$\begin{aligned}y^1_{i,\alpha}(x) &= \frac{\sigma_w}{\sqrt{v_1}}\sum_{j=1}^{n_0}\sum_{\beta\in ker_1} w^1_{i,j,\beta}x_{j,\alpha+\beta} + \sigma_b b^1_i \\ y^l_{i,\alpha}(x) &= \frac{\sigma_w}{\sqrt{v_l}}\sum_{j=1}^{n_{l-1}}\sum_{\beta\in ker_l} w^l_{i,j,\beta}\phi(y^{l-1}_{j,\alpha+\beta}(x)) + \sigma_b b^l_i\end{aligned}$$

therefore

$$\begin{aligned}K^1_{(i,\alpha),(i',\alpha')}(x,x') &= \sum_r\left(\sum_j\sum_\beta \frac{\partial y^1_{i,\alpha}(x)}{\partial w^1_{r,j,\beta}}\frac{\partial y^1_{i',\alpha'}(x)}{\partial w^1_{r,j,\beta}}\right) + \frac{\partial y^1_{i,\alpha}(x)}{\partial b^1_r}\frac{\partial y^1_{i',\alpha'}(x)}{\partial b^1_r} \\ &= \delta_{ii'}\left(\frac{\sigma_w^2}{n_0(2k+1)}\sum_j\sum_\beta x_{j,\alpha+\beta}x_{j,\alpha'+\beta} + \sigma_b^2\right)\end{aligned}$$

Assume the result is true for $l-1$, let us prove it for l. Let $\theta_{1:l-1}$ be model weights and bias in the layers 1 to $l-1$. Let $\partial_{\theta_{1:l-1}} y^l_{i,\alpha}(x) = \frac{\partial y^l_{i,\alpha}(x)}{\partial \theta_{1:l-1}}$. We have that

$$\partial_{\theta_{1:l-1}} y^l_{i,\alpha}(x) = \frac{\sigma_w}{\sqrt{n_{l-1}(2k+1)}} \sum_j \sum_\beta w^l_{i,j,\beta} \phi'(y^{l-1}_{j,\alpha+\beta}) \partial_{\theta_{1:l-1}} y^{l-1}_{i,\alpha+\beta}(x)$$

this yields

$$\partial_{\theta_{1:l-1}} y^l_{i,\alpha}(x) \partial_{\theta_{1:l-1}} y^l_{i',\alpha'}(x)^T = \\ \frac{\sigma_w^2}{n_{l-1}(2k+1)} \sum_{j,j'} \sum_{\beta,\beta'} w^l_{i,j,\beta} w^l_{i',j',\beta'} \phi'(y^{l-1}_{j,\alpha+\beta}) \phi'(y^{l-1}_{j',\alpha'+\beta}) \partial_{\theta_{1:l-1}} y^{l-1}_{j,\alpha+\beta}(x) \partial_{\theta_{1:l-1}} y^{l-1}_{j',\alpha'+\beta}(x)^T$$

as $n_1, n_2, ..., n_{l-2} \to \infty$ and using the induction hypothesis, we have

$$\partial_{\theta_{1:l-1}} y^l_{i,\alpha}(x) \partial_{\theta_{1:l-1}} y^l_{i',\alpha'}(x)^T \to \\ \frac{\sigma_w^2}{n_{l-1}(2k+1)} \sum_j \sum_{\beta,\beta'} w^l_{i,j,\beta} w^l_{i',j,\beta'} \phi'(y^{l-1}_{j,\alpha+\beta}) \phi'(y^{l-1}_{j,\alpha'+\beta}) K^{l-1}_{(j,\alpha+\beta),(j,\alpha'+\beta)}(x,x')$$

note that $K^{l-1}_{(j,\alpha+\beta),(j,\alpha'+\beta)}(x,x') = K^{l-1}_{(1,\alpha+\beta),(1,\alpha'+\beta)}(x,x')$ for all j since the variables are iid across the channel index j.

Now letting $n_{l-1} \to \infty$, we have that

$$\partial_{\theta_{1:l-1}} y^l_{i,\alpha}(x) \partial_{\theta_{1:l-1}} y^l_{i',\alpha'}(x)^T \to \\ \delta_{ii'}\big(\frac{1}{(2k+1)} \sum_{\beta,\beta'} \dot{q}^l_{\alpha+\beta,\alpha'+\beta} K^{l-1}_{(1,\alpha+\beta),(1,\alpha'+\beta)}(x,x')\big)$$

We conclude using the fact that

$$\partial_{\theta_l} y^l_{i,\alpha}(x) \partial_{\theta_l} y^l_{i',\alpha'}(x)^T \to \delta_{ii'}\big(\frac{\sigma_w^2}{2k+1} \sum_\beta \mathbb{E}[\phi(y^{l-1}_{\alpha+\beta}(x))\phi(y^{l-1}_{\alpha'+\beta}(x'))] + \sigma_b^2\big)$$

□

To alleviate notations, we use hereafter the notation K^L for both the NTK of FFNN and CNN. For FFNN, it represents the recursive kernel K^L given by lemma 1, whereas for CNN, it represents the recursive kernel $K^L_{\alpha,\alpha'}$ for any α, α', which means all results that follow are true for any α, α'.
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 6. *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, 17, 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

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

We conclude using Appendix Lemma 6.

Under Assumption 1, the proof is similar for CNN, using Appendix Lemmas 3 and 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 7. *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 8. *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 7, 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 7 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-like, 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. We start by proving the results for an FFNN architecture, the proof machinery remains the same for CNN under assumption:cnn.

Case 1: FFNN. 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 4). By definition, we have that $\dot{q}^l(x,x) = f'(c^{l-1}(x,x'))$. Let us first prove the result for ReLU.

- $\phi =$*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 16, 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 8 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 8. 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').

Case 2: CNN. Under Assumption 1, the NTK of a CNN is the same as that of an FFNN. Therefore, the results on the NTK of FFNN are all valid to the NTK of CNN $K^l_{\alpha,\alpha'}$ for any α, α'.

□

6.2 Proofs of the results of Section 3.2 on ResNets

In this section, we provide proofs for lemmas 3 and 4 together with Theorem **??** and proposition **??** on ResNets.

Lemma 3 in the paper gives the recursive formula for the mean-field NTK of a ResNet with Fully Connected blocks.

Lemma 3 (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 $x, x' \in \mathbb{R}^d$*

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

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 $i, i' \in [1 : n_l]$, $K_{ii'}^{res,l}(x, x') = \delta_{ii'} K_{res}^l(x, x')$, where $K_{res}^l(x, x')$ 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*

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

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

$$K_{res}^1(x, x') = \sum_j \frac{\partial y_1^1(x)}{\partial w_{1j}^1} \frac{\partial y_1^1(x)}{\partial w_{1j}^1} + \frac{\partial y_1^1(x)}{\partial b_1^1} \frac{\partial y_1^1(x)}{\partial b_1^1} = \frac{\sigma_w^2}{d} x \cdot x' + \sigma_b^2.$$

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

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

Therefore, we obtain

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

where

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

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

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

$$\to K_{res}^l(x, x') + \frac{\sigma_w^2}{n_l} \sum_j^{n_l} (w_{ij}^{l+1})^2 \phi'(y_j^l(x)) \phi'(y_j^l(x')) K_{res}^l(x, x') + I',$$

where $I' = \frac{\sigma_w^2}{n_l} w_{ii}^{l+1} (\phi'(y_i^l(x)) + \phi'(y_i^l(x'))) K_{res}^l(x, x')$.

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

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

Moreover, we have that

$$\begin{aligned}(\partial_{w^{l+1}} y_i^{l+1}(x))(\partial_{w^{l+1}} y_i^{l+1}(x'))^t + (\partial_{b^{l+1}} y_i^{l+1}(x))(\partial_{b^{l+1}} y_i^{l+1}(x'))^t &= \frac{\sigma_w^2}{n_l} \sum_j \phi(y_j^l(x)) \phi(y_j^l(x')) + \sigma_b^2 \\ &\underset{n_l \to \infty}{\to} \sigma_w^2 \mathbb{E}[\phi(y_i^l(x)) \phi(y_i^l(x'))] + \sigma_b^2 = q^{l+1}(x, x').\end{aligned}$$

□

Now we proof the recursive formula for ResNets with Convolutional layers.

Lemma 4 (NTK of a ResNet with Convolutional layers in the infinite width limit)**.** *Let $K^{res,1}$ be the exact NTK for the ResNet with 1 layer. Then*
• For the first layer (without residual connections), we have for all $x, x' \in \mathbb{R}^d$

$$K^{res,1}_{(i,\alpha),(i',\alpha')}(x,x') = \delta_{ii'}\Big(\frac{\sigma_w^2}{n_0(2k+1)}[x,x']_{\alpha,\alpha'} + \sigma_b^2\Big)$$

• For $l \geq 2$, as $n_1, n_2, ..., n_{l-1} \to \infty$ recursively, we have for all $i, i' \in [1 : n_l]$, $\alpha, \alpha' \in [0 : M-1]$, $K^{res,l}_{(i,\alpha),(i',\alpha')}(x,x') = \delta_{ii'}K^{res,l}_{\alpha,\alpha'}(x,x')$, where $K^{res,l}_{\alpha,\alpha'}$ is given by the recursive formula for all $x, x' \in \mathbb{R}^d$, using the same notations as in lemma 2,

$$K^{res,l}_{\alpha,\alpha'} = K^{res,l-1}_{\alpha,\alpha'} + \frac{1}{2k+1}\sum_{\beta}\Psi^{l-1}_{\alpha+\beta,\alpha'+\beta}.$$

where $\Psi^l_{\alpha,\alpha'} = \dot{q}^l_{\alpha,\alpha'}K^{res,l}_{\alpha,\alpha'} + \hat{q}^l_{\alpha,\alpha'}$.

Proof. Let x, x' be two inputs. We have that

$$\begin{aligned} K^1_{(i,\alpha),(i',\alpha')}(x,x') &= \sum_j \left(\sum_\beta \frac{\partial y^1_{i,\alpha}(x)}{\partial w^1_{i,j,\beta}}\frac{\partial y^1_{i',\alpha'}(x)}{\partial w^1_{i',j,\beta}} + \frac{\partial y^1_{i,\alpha}(x)}{\partial b^1_j}\frac{\partial y^1_{i',\alpha'}(x)}{\partial b^1_j}\right) \\ &= \delta_{ii'}\left(\frac{\sigma_w^2}{n_0(2k+1)}\sum_j\sum_\beta x_{j,\alpha+\beta}x_{j,\alpha'+\beta} + \sigma_b^2\right). \end{aligned}$$

Assume the result is true for $l-1$, let us prove it for l. Let $\theta_{1:l-1}$ be model weights and bias in the layers 1 to $l-1$. Let $\partial_{\theta_{1:l-1}}y^l_{i,\alpha}(x) = \frac{\partial y^l_{i,\alpha}(x)}{\partial\theta_{1:l-1}}$. We have that

$$\partial_{\theta_{1:l-1}}y^l_{i,\alpha}(x) = \partial_{\theta_{1:l-1}}y^{l-1}_{i,\alpha}(x) + \frac{\sigma_w}{\sqrt{n_{l-1}(2k+1)}}\sum_j\sum_\beta w^l_{i,j,\beta}\phi'(y^{l-1}_{j,\alpha+\beta})\partial_{\theta_{1:l-1}}y^{l-1}_{i,\alpha+\beta}(x)$$

this yields

$$\begin{aligned} &\partial_{\theta_{1:l-1}}y^l_{i,\alpha}(x)\partial_{\theta_{1:l-1}}y^l_{i',\alpha'}(x)^T = \partial_{\theta_{1:l-1}}y^{l-1}_{i,\alpha}(x)\partial_{\theta_{1:l-1}}y^{l-1}_{i',\alpha'}(x)^T + \\ &\frac{\sigma_w^2}{n_{l-1}(2k+1)}\sum_{j,j'}\sum_{\beta,\beta'} w^l_{i,j,\beta}w^l_{i',j',\beta'}\phi'(y^{l-1}_{j,\alpha+\beta})\phi'(y^{l-1}_{j',\alpha'+\beta})\partial_{\theta_{1:l-1}}y^{l-1}_{j,\alpha+\beta}(x)\partial_{\theta_{1:l-1}}y^{l-1}_{j',\alpha'+\beta}(x)^T + I, \end{aligned}$$

where

$$I = \frac{\sigma_w}{\sqrt{n_{l-1}(2k+1)}}\sum_{j,\beta} w^l_{i,j,\beta}\phi'(y^{l-1}_{j,\alpha+\beta})(\partial_{\theta_{1:l-1}}y^{l-1}_{i,\alpha}(x)\partial_{\theta_{1:l-1}}y^{l-1}_{i,\alpha+\beta}(x)^T + \partial_{\theta_{1:l-1}}y^{l-1}_{i,\alpha+\beta}(x)\partial_{\theta_{1:l-1}}y^{l-1}_{i,\alpha}(x)^T).$$

As $n_1, n_2, ..., n_{l-2} \to \infty$ and using the induction hypothesis, we have

$$\begin{aligned} &\partial_{\theta_{1:l-1}}y^l_{i,\alpha}(x)\partial_{\theta_{1:l-1}}y^l_{i',\alpha'}(x)^T \to \delta_{ii'}K^{l-1}_{\alpha,\alpha'}(x,x') + \\ &\frac{\sigma_w^2}{n_{l-1}(2k+1)}\sum_j\sum_{\beta,\beta'} w^l_{i,j,\beta}w^l_{i',j,\beta'}\phi'(y^{l-1}_{j,\alpha+\beta})\phi'(y^{l-1}_{j,\alpha'+\beta})K^{l-1}_{(j,\alpha+\beta),(j,\alpha'+\beta)}(x,x'). \end{aligned}$$

Note that $K^{l-1}_{(j,\alpha+\beta),(j,\alpha'+\beta)}(x,x') = K^{l-1}_{(1,\alpha+\beta),(1,\alpha'+\beta)}(x,x')$ for all j since the variables are iid across the channel index j. Now letting $n_{l-1} \to \infty$, we have that

$$\begin{aligned} &\partial_{\theta_{1:l-1}}y^l_{i,\alpha}(x)\partial_{\theta_{1:l-1}}y^l_{i',\alpha'}(x)^T \to \\ &\delta_{ii'}K^{l-1}_{\alpha,\alpha'}(x,x') + \delta_{ii'}\Big(\frac{1}{(2k+1)}\sum_{\beta,\beta'} f'(c^{l-1}_{\alpha+\beta,\alpha'+\beta}(x,x'))K^{l-1}_{(1,\alpha+\beta),(1,\alpha'+\beta)}(x,x')\Big), \end{aligned}$$

where $f'(c^{l-1}_{\alpha+\beta,\alpha'+\beta}(x,x')) = \sigma_w^2\mathbb{E}[\phi'(y^{l-1}_{j,\alpha+\beta})\phi'(y^{l-1}_{j,\alpha'+\beta})]$.
We conclude using the fact that

$$\partial_{\theta_l}y^l_{i,\alpha}(x)\partial_{\theta_l}y^l_{i',\alpha'}(x)^T \to \delta_{ii'}\Big(\frac{\sigma_w^2}{2k+1}\sum_\beta \mathbb{E}[\phi(y^{l-1}_{\alpha+\beta}(x))\phi(y^{l-1}_{\alpha'+\beta}(x'))] + \sigma_b^2\Big).$$

□

Before moving to the main theorem on ResNets, We first prove a Lemma on the asymptotic behaviour of c^l for ResNet.

Appendix Lemma 9 (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. Moreover, this result holds also for CNNs where the supremum should be replaced by $\sup_{(x,x')\in B_\epsilon} \sup_{\alpha,\alpha'}$.

Proof. We first prove the result for ResNet with fully connected layers, then we generalize it to convolutional layers. 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
$$q^l(x,x) = q^{l-1}(x,x) + \sigma_w^2/2q^{l-1}(x,x) = (1+\sigma_w^2/2)^{l-1}q^1(x,x)$$
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.
- Under Assumption 1, results of FFNN hold for CNN.

□

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*

$$y^l(x) = y^{l-1}(x) + \mathcal{F}(w^l, y^{l-1}(x)), \quad l \geq 2, \tag{17}$$

where $\mathcal{F}$ is either a convolutional or 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 $\alpha_L = L(1+\frac{\sigma_w^2}{2})^{L-1}$. If the layers are convolutional assume Assumption 1 holds. 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. We start by proving the result for a ResNet architecture with fully-connected layers. As in the previous proofs, the result easily extends to a ResNet with convolutional layers under assumption:cnn.

Case 1: ResNet with fully-connected layers. 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

$$q^l(x,x) = (1+\sigma_w^2/2)^{l-1} \frac{\sigma_w^2}{d} \|x\|^2 \tag{18}$$

Recall that the recursive formula of NTK of a ResNet with fully-connected layers is given by (Appendix Lemma 3)

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

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 9 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 (18). Observe that all bounds $\mathcal{O}$ are independent from the inputs (x,x'). Therefore, using Appendix Lemma 8, 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.

Case 2: ResNet with Convolutional layers. Under Assumption 1, the dynamics of the correlation and NTK are exactly the same for FFNN, hence all results on FFNN apply to CNN.

□

6.2.1 Beyond ResNet: Scaled ResNet

The term α_L in the residual NTK might cause numerical stability issues for NTK training, and the triviality of the limiting kernel yields a trivial NTK regime solution (recall that $f_t - f_0$ belongs to the RKHS of the NTK; see section 2). It turns out that we can improve the performance of NTK training of ResNets with a simple scaling of the ResNet blocks.

Theorem 3 (Scaled ResNet)**.** *Consider a ResNet satisfying*

$$y^l(x) = y^{l-1}(x) + \frac{1}{\sqrt{l}}\mathcal{F}(w^l, y^{l-1}(x)), \quad l \geq 2, \tag{19}$$

*where $\mathcal{F}$ is either a convolutional or dense layer ((3) and (**??**)) with ReLU activation. Then the results of Theorem 2 apply with $\alpha_L = L^{1+\sigma_w^2/2}$ and the convergence rate $\Theta(\log(L)^{-1})$.*

Theorem 3 shows that scaling the residual blocks by $1/\sqrt{l}$ has two important effects on the NTK: first, it stabilizes the NTK which only grows as $L^{1+\frac{\sigma_w^2}{2}}$ instead of $L(1+\frac{\sigma_w^2}{2})^{L-1}$; second, it drastically slows down the convergence rate to the limiting (trivial) $\bar{K}^\infty_{res}$. Both properties are highly desirable for NTK training. The second property in particular means that with the scaling, we can 'NTK train' deeper ResNets compared to the non-scaled ResNet. A more aggressive scaling was studied in Huang et al. (2020), where authors scale the blocks with $1/L$ instead of our scaling $1/\sqrt{l}$, and show that it also stabilizes the NTK of ResNet. This is the main topic of the next chapter. We particularly show that a suitable scaling ensures that the limiting NTK is universal, i.e. we can approximate any continuous function on some compact set K with a function from the Reproducing Kernel Hilbert Space of the limiting NTK. This is a desirable property since the second term in the solution of the NTK regime lives in the RKHS of the NTK.

Now let us prove the Scaled Resnet result. Before that, we prove the following Lemma

Appendix Lemma 10. *Consider a Residual Neural Network with the following forward propagation equations*

$$y^l(x) = y^{l-1}(x) + \frac{1}{\sqrt{l}}\mathcal{F}(w^l, y^{l-1}(x)), \quad l \geq 2. \tag{20}$$

*where $\mathcal{F}$ is either a convolutional or dense layer (equations 3 and **??**) with ReLU activation. Then there exists $\zeta, \nabla > 0$ such that for all $\epsilon \in (0,1)$*

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

where the bound 'o' depends only on ϵ.
For CNN, under Assumption 1, the result holds and the supremum is taken also over α, α', i.e.

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

Proof. We first start with the dense layer case. Let $\epsilon \in (0,1)$ and $(x,x') \in B_\epsilon$ be two inputs and denote by $c^l := c^l(x,x')$. Following the same machinery as in the proof of Appendix Lemma 9, we have that

$$c^l = \frac{1}{1+\alpha_l}c^{l-1} + \frac{\alpha_l}{1+\alpha_l}f(c^{l-1})$$

where $\alpha_l = \frac{\sigma_w^2}{2l}$. Using fact 12, it is straightforward that $f' \geq 0$, hence f is non-decreasing. Therefore, $c^l \geq c^{l-1}$ and c^l converges to a fixed point c. Let us prove that $c = 1$. By contradiction, suppose $c < 1$ so that $f(c) - c > 0$ (f has a unique fixed point which is 1). This yields

$$c^l - c = c^{l-1} - c + \frac{f(c)-c}{l} + \mathcal{O}(\frac{c^l - c}{l}) + \mathcal{O}(l^{-2})$$

by summing, this leads to $c^l - c \sim (f(c) - c)\log(l)$ which is absurd since $f(c) \neq c$ (f has only 1 as a fixed point). We conclude that $c = 1$. Using the non-decreasing nature of f, it is easy to conclude that the convergence is uniform over B_ϵ.

Now let us find the asymptotic expansion of $1-c^l$. Recall the Taylor expansion of f near 1 given in fact 14

$$f(c) \underset{x\to 1-}{=} c + s(1-c)^{3/2} + b(1-c)^{5/2} + \mathcal{O}((1-c)^{7/2}) \tag{21}$$

where $s = \frac{2\sqrt{2}}{3\pi}$ and $b = \frac{\sqrt{2}}{30\pi}$. Letting $\gamma_l = 1 - c^l$, we obtain

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

which yields

$$\gamma_l^{-1/2} = \gamma_{l-1}^{-1/2} + \frac{s}{2}\delta_l + \frac{3}{8}s^2\delta_l^2 \gamma_{l-1}^{1/2} + \frac{b}{2}\delta_l \gamma_{l-1} + O(\delta_l \gamma_{l-1}^{3/2}). \tag{22}$$

therefore, we have that

$$\gamma_l^{-1/2} \sim \frac{s\sigma_w^2}{4}\log(l)$$

and $1 - c^l \sim \frac{\zeta}{\log(l)^2}$ where $\zeta = 16/s^2\sigma_w^4$.

we can further expand the asymptotic approximation to have

$$1 - c^l = \frac{\zeta}{\log(l)^2} - \frac{\nabla}{\log(l)^3} + o(\frac{1}{\log(l)^3})$$

where $\nabla > 0$. the 'o' holds uniformly for $(x, x') \in B_\epsilon$ as in the proof of Appendix Lemma 1.

This result holds for a ResNet with CNN layers under Assumption 1 since the dynamics are the same in this case.

□

Proof of Theorem 3.

Proof. We use the same techniques as in the non scaled case. Let us prove the result for fully connected layers, the proof for convolutional layers follows the same analysis. Let $\epsilon \in (0,1)$ and $x, x' \in B_\epsilon$ be two inputs. We first prove the result for the diagonal term $K_{res}^L(x,x)$ then $K_{res}^L(x,x')$.

- We have that $\dot{q}^l(x,x) = \frac{\sigma_w^2}{2l} f(1) = \frac{\sigma_w^2}{2l}$. Moreover, we have $q^l(x,x) = q^{l-1}(x,x) + \sigma_w^2/2l \times q^{l-1}(x,x) = [\prod_{k=1}^l (1+\sigma_w^2/2k)] \frac{\sigma_w^2}{d}\|x\|^2$. Recall that

  $$K_{res}^l(x,x) = K_{res}^{l-1}(x,x)(1 + \frac{\sigma_w^2}{2l}) + q^l(x,x)$$

  letting $k_l' = \frac{K_{res}^l(x,x)}{\prod_{k=1}^l (1+\sigma_w^2/2k)}$,we have that

  $$k_l' = k_{l-1}' + \frac{\sigma_w^2}{d}\|x\|$$

  using the fact that $\prod_{k=1}^l (1+\sigma_w^2/2k) = \Theta(l^{\sigma_w^2/2})$, we conclude for $K_{res}^l(x,x)$.

- Recall that

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

  Let $c^l := c^l(x,x')$. From Appendix Lemma 10 we have that

  $$1 - c^l = \frac{\zeta}{\log(l)^2} - \frac{\nabla}{\log(l)^3} + o(\frac{1}{\log(l)^3})$$

  $\zeta = \frac{16}{s^2\sigma_w^4}$ and $\nabla > 0$. Using the Taylor expansion of f' as in Appendix Lemma 1, it follows that

  $$f'(c^l(x,x')) = 1 - \frac{6}{\sigma_w^2}\log(l)^{-1} + \zeta'\log(l)^{-2} + \mathcal{O}(\log(l)^{-3})$$

  where $\zeta' = \frac{\nabla}{\sqrt{2\pi\zeta}}$. We obtain

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

where $\zeta'' = \frac{\sigma_w^2}{2}\zeta'$. Letting $a_l = \frac{K_{res}^{l+1}(x,x')}{\prod_{k=1}^{l}(1+\sigma_w^2/2k)}$, we obtain

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

where $\lambda_l = 1 - l^{-1} - 3l^{-1}\log(l)^{-1} + \mathcal{O}(l^{-1}\log(l)^{-2})$, $b_l = \sqrt{q^1(x,x)}\sqrt{q^1(x',x')}f(c^l(x,x')) = q(x,x') + \mathcal{O}(\log(l)^{-2})$ with $q = \sqrt{q^1(x,x)}\sqrt{q^1(x',x')}$ and where we used the fact that $c^l = 1 + \mathcal{O}(\log(l)^{-2})$ (Appendix Lemma 10).

Now we proceed in the same way as in the proof of Appendix Lemma 8. Let $x_l = \frac{a_l}{l} - q$, then there exists $M_1, M_2 > 0$ such that

$$x_{l-1}(1-\frac{1}{l}) - M_1 l^{-1}\log(l)^{-1} \leq x_l \leq x_{l-1}(1-\frac{1}{l}) - M_2 l^{-1}\log(l)^{-1}$$

therefore, there exists l_0 independent of (x,x') such that for all $l \geq l_0$

$$x_l \leq x_{l_0}\prod_{k=l_0}^{l}(1-\frac{1}{k}) - M_2\sum_{k=l_0}^{l}\prod_{j=k+1}^{l}(1-\frac{1}{j})k^{-1}\log(k)^{-1}$$

and

$$x_l \geq x_{l_0}\prod_{k=l_0}^{l}(1-\frac{1}{k}) - M_1\sum_{k=l_0}^{l}\prod_{j=k+1}^{l}(1-\frac{1}{j})k^{-1}\log(k)^{-1}$$

after simplification, we have that

$$\sum_{k=l_0}^{l}\prod_{j=k+1}^{l}(1-\frac{1}{j})k^{-1}\log(k)^{-1} = \Theta\left(\frac{1}{l}\int^{l}\frac{1}{\log(t)}dt\right) = \Theta(\log(l)^{-1})$$

where we have used the asymptotic approximation of the Logarithmic Intergal function $\text{Li}(x) = \int^{t}\frac{1}{\log(t)} \sim_{x\to\infty} \frac{x}{\log(x)}$

we conclude that $\alpha_L = L \times \prod_{k=1}^{l}(1+\sigma_w^2/2k) \sim L^{1+\frac{\sigma_w^2}{2}}$ and the convergence rate of the NTK is now $\Theta(\log(L)^{-1})$ which is better than $\Theta(L^{-1})$. The convergence is uniform over the set B_ϵ.

In the limit of large L, the matrix NTK of the scaled resnet has the following form

$$A\hat{K}_{res}^{l} = qU + \log(L)^{-1}\Theta(M_L)$$

where U is the matrix of ones, and M_L has all elements but the diagonal equal to 1 and the diagonal terms are $\mathcal{O}(L^{-1}\log(L)) \to 0$. Therefore, M_L is inversible for large L which makes $\hat{K}_{res}^{l}$ also inversible. Moreover, observe that the convergence rate for scaled resnet is $\log(L)^{-1}$ which means that for the same depth L, the NTK remains far more expressive for scaled resnet compared to standard resnet, this is particularly important for the generalization.

□

Does Scaled ResNet outperforms ResNet with SGD? We train standard ResNet with depths 32, 50, and 104 on CIFAR100 with SGD. We use a decaying learning rate schedule; we start with 0.1 and divide by 10 after $n_e/2$ epochs, where n_e is the total number of epochs; we scale again, by 10, after $n_e/4$ epochs. We use a batch size of 128, and we train the model with 160 epochs. Theorem 3 shows that the NTK of Scaled ResNet is more stable compared to the NTK of standard ResNet. Although this result is limited to NTK training, we investigate the impact of scaling on SGD training. Table 2 displays test accuracy for standard ResNet and scaled ResNet after 10 and 160 epochs; Scaled ResNet outperforms ResNet and converges faster. However, it is not clear whether this is linked to the NTK, or caused by something else. We leave this for future work.

Table 2: Test accuracy on CIFAR100 for ResNet.

		Epoch 10	Epoch 160
ResNet32	standard	**54.18±1.21**	72.49±0.18
	scaled	53.89±2.32	**74.07±0.22**
ResNet50	standard	51.09±1.73	73.63±1.51
	scaled	**55.39±1.52**	**75.02±0.44**
ResNet104	standard	47.02±3.23	74.77±0.29
	scaled	**56.38±2.54**	**76.14±0.98**

6.3 Spectral decomposition of the limiting NTK

6.3.1 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

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

$(P_k^d)_{k \geq 0}$ 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, CNN 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 $k \geq 1$, $\lim_{L \to \infty} \mu_k^L = q\lambda \frac{\Omega_{d-1}}{\Omega_d} \int_{-1}^{1} P_k^d(t)(1-t^2)^{(d-3)/2} dt = q\lambda \frac{\Omega_{d-1}}{\Omega_d} \langle P_0^d, P_k^d \rangle_{L^2([-1,1],(1-t^2)^{\frac{d-3}{2}} dt)} = 0$.

□

7 Further experimental results

• **Results for L between 30 and 300**: In our experiments, we observed degeneracy of the NTK regime when $L \sim 300$ (hence our choice of $L = 300$ in the paper). The Figure included here shows the percentage drop in performance of the NTK regime for a 100x100 FFNN with ReLU on MNIST for the Ordered phase/EOC Init.

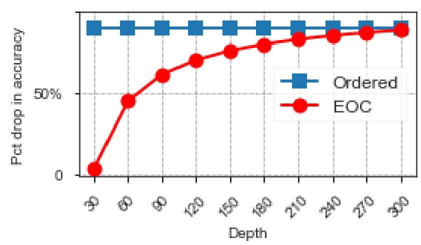

Figure 2: Deterioration of NTK regime

References

Arora, S., S. Du, W. Hu, , Z. Li, and R. Wand (2019). Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *ICML*.

Bietti, A. and F. Bach (2021). Deep equals shallow for reLU networks in kernel regimes. In *International Conference on Learning Representations*.

Bietti, A. and J. Mairal (2019). On the inductive bias of neural tangent kernels. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 32, pp. 12893–12904. Curran Associates, Inc.

Cao, Y., Z. Fang, Y. Wu, D. Zhou, and Q. Gu (2020). Towards understanding the spectral bias of deep learning. *arXiv prePrint 1912.01198*.

Chizat, L. and F. Bach (2018). A note on lazy training in supervised differentiable programming. *arXiv preprint arXiv:1812.07956*.

Du, S., J. Lee, H. Li, L. Wang, and X. Zhai (2019). Gradient descent finds global minima of deep neural networks. *ICML*.

Geifman, A., A. Yadav, Y. Kasten, M. Galun, D. Jacobs, and R. Basri (2020). On the similarity between the laplace and neural tangent kernels. *NeurIPS*.

Ghorbani, B., S. Mei, T. Misiakiewicz, and A. Montanari (2019). Linearized two-layers neural networks in high dimension. *arXiv preprint arXiv:1904.12191*.

Hanin, B. and M. Nica (2019). Finite depth and width corrections to the neural tangent kernel. *arXiv preprint arXiv:1909.05989*.

Hayou, S., A. Doucet, and J. Rousseau (2019). On the impact of the activation function on deep neural networks training. *ICML*.

Huang, J. and H. Yau (2020). Dynamics of deep neural networks and neural tangent hierarchy. *ICML*.

Huang, K., Y. Wang, M. Tao, and T. Zhao (2020). Why do deep residual networks generalize better than deep feedforward networks? – a neural tangent kernel perspective. *ArXiv preprint, arXiv:2002.06262*.

Jacot, A., F. Gabriel, and C. Hongler (2018). Neural tangent kernel: Convergence and generalization in neural networks. *32nd Conference on Neural Information Processing Systems*.

Karakida, R., S. Akaho, and S. Amari (2018). Universal statistics of Fisher information in deep neural networks: Mean field approach. *arXiv preprint arXiv:1806.01316*.

Lee, J., Y. Bahri, R. Novak, S. Schoenholz, J. Pennington, and J. Sohl-Dickstein (2018). Deep neural networks as Gaussian processes. *6th International Conference on Learning Representations*.

Lee, J., L. Xiao, S. Schoenholz, Y. Bahri, J. Sohl-Dickstein, and J. Pennington (2019). Wide neural networks of any depth evolve as linear models under gradient descent. *NeurIPS*.

Lillicrap, T., D. Cownden, D. Tweed, and C. Akerman (2016). Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications 7*(13276).

MacRobert, T. (1967). *Spherical harmonics: An elementary treatise on harmonic functions, with applications*. Pergamon Press.

Matthews, A., J. Hron, M. Rowland, R. Turner, and Z. Ghahramani (2018). Gaussian process behaviour in wide deep neural networks. *6th International Conference on Learning Representations*.

Neal, R. (1995). Bayesian learning for neural networks. *Springer Science & Business Media 118*.

Novak, R., L. Xiao, J. Hron, J. Lee, A. A. Alemi, J. Sohl-Dickstein, and S. S. Schoenholz (2020). Neural tangents: Fast and easy infinite neural networks in python. In *International Conference on Learning Representations*.

Poole, B., S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli (2016). Exponential expressivity in deep neural networks through transient chaos. *30th Conference on Neural Information Processing Systems*.

Schoenholz, S., J. Gilmer, S. Ganguli, and J. Sohl-Dickstein (2017). Deep information propagation. *5th International Conference on Learning Representations*.

Xiao, L., Y. Bahri, J. Sohl-Dickstein, S. S. Schoenholz, and P. Pennington (2018). Dynamical isometry and a mean field theory of cnns: How to train 10,000-layer vanilla convolutional neural networks. *ICML 2018*.

Xiao, L., J. Pennington, and S. Schoenholz (2020). Disentangling trainability and generalization in deep neural networks. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 10462–10472.

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

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

Yang, G. and S. Schoenholz (2017a). Mean field residual networks: On the edge of chaos. *Advances in Neural Information Processing Systems 30*, 2869–2869.

Yang, G. and S. Schoenholz (2017b). Mean field residual networks: On the edge of chaos. In *Advances in neural information processing systems*, pp. 7103–7114.