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 3 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 3 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 5 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 $\phi =$* 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 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 3 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 $x, x' \in \mathbb{S}^{d-1}$*

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

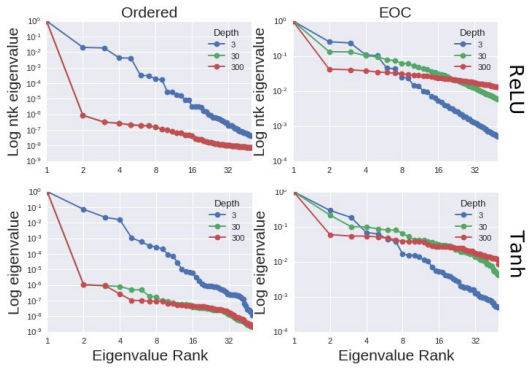

Figure 1: Normalized eigenvalues of K_L on the 2D sphere for an FFNN with different initializations, activations, and depths. (Red and Green lines are identical in the upper left figure.)

$(Y_{k,j})_{k\geq 0, j\in[1:N(d,k)]}$ *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.