Neural Tangent Kernel Beyond the Infinite-Width Limit: Effects of Depth and Initialization

Abstract—Neural Tangent Kernel (NTK) is widely used to analyze overparameterized neural networks due to the famous result by Jacot et al. (2018) [1]: in the infinite-width limit, the NTK is deterministic and constant during training. However, this result cannot explain the behavior of deep networks, since it generally does not hold if the network’s depth and width tend to infinity simultaneously. In our recent paper, we study the NTK of fully-connected ReLU networks with depth comparable to width. We prove that the NTK properties depend significantly on the depth-to-width ratio and the distribution of parameters at initialization. In fact, our results indicate the importance of the three phases in the initialization hyperparameters’ space identified in Poole et al. (2016) [2]: ordered, chaotic and the edge of chaos (EOC). We derive exact expressions for the NTK dispersion in the infinite-depth-and-width limit in all three phases and conclude that the NTK variability grows exponentially with depth at the EOC and in the chaotic phase but not in the ordered phase. We also show that the NTK of deep networks may stay constant during training only in the ordered phase and discuss how the structure of the NTK matrix changes during training.

I. INTRODUCTION

Training dynamics and generalization capabilities of deep neural networks (DNNs) stand among the biggest open problems of deep learning theory. However, it is possible to fully address these challenges in the infinite-width limit of DNNs. The key notion in this limit is the Neural Tangent Kernel (NTK), which captures the first-order approximation of DNN’s evolution during gradient descent (GD) training. Consider the gradient flow dynamics of the DNN’s parameters:

\[ \dot{w} = -\nabla_w \mathcal{L}(\mathcal{D}) = -\sum_{(x_i, y_i) \in \mathcal{D}} \nabla_w f(x_i) \frac{\partial \mathcal{L}(\mathcal{D})}{\partial f(x_i)}, \]  

(1)

where \( w \) is the vector of all the trainable parameters, \( f(\cdot) \) is the DNN’s output function, \( \mathcal{L}(\cdot) \) is the loss function and \( \mathcal{D} \) is the dataset. Then the dynamics of the DNN’s output function is given by:

\[ \dot{f}(x) = \nabla_w f(x) \cdot \dot{w} = -\sum_{(x_i, y_i) \in \mathcal{D}} \Theta(x, x_i) \frac{\partial \mathcal{L}(\mathcal{D})}{\partial f(x_i)}, \]  

(2)

where \( \Theta(x, x_i) \) is called the NTK. Jacot et al. (2018) [1] showed that, in the infinite-width limit, the NTK is deterministic under proper random initialization and stays constant during training. Therefore, the dynamics in (2) is equivalent to kernel regression and has an analytical solution expressed in terms of the NTK. Hence, many recent papers used the NTK to explain properties of overparameterized DNNs theoretically [3], [4].

However, the extent to which the results in the infinite-width limit extrapolate to realistic DNNs remains largely an open question. Indeed, multiple authors have argued that the NTK regime and, in general, the infinite-width limit cannot explain the success of DNNs [5], [6]. The first argument in this direction is that no feature learning occurs if the NTK stays constant during training. Moreover, the infinite-width limit of the NTK becomes completely data-independent as depth increases [7], which suggests poor performance for deep networks in the NTK regime. Finally, numerous empirical results demonstrated that the performance of trained DNNs and the corresponding kernel methods often differs in practice [8], [9]. That is why it is essential to understand the statistical properties of the NTK and how they depend on the myriad of settings of a given DNN to assess if the infinite-width limit provides a reasonable approximation for this network. We contribute to this line of research by exploring the combined effect of two factors on the NTK: the network’s depth and initialization hyperparameters.

Network’s depth Most results on the NTK are derived in the setting where the network’s depth is kept constant while the width tends to infinity. This limit can only model very wide and shallow networks since the depth-to-width ratio tends to zero in it. Indeed, several recent papers demonstrated that infinite-width approximations often get worse as the depth increases [6], [10]. In particular, Hanin&Nica (2020) [11] first showed that the NTK of fully-connected ReLU DNNs may be random and change during training if depth and width are comparable. In our recent work, we expand on these results by precisely characterizing the variability of the NTK at initialization and generalizing to different initialization settings described below.

Initialization hyperparameters There are three phases in the initialization hyperparameter space where the properties of untrained infinitely-wide DNNs differ significantly: ordered, chaotic and the edge of chaos (EOC) [2]. In the ordered phase the gradient norms decrease with depth, whereas in the chaotic phase the gradient norms increase, and the edge of chaos is the initialization at the border between these two phases [12]. The results by Hanin&Nica (2020) [11] concerned the statistical properties of the NTK of wide and deep ReLU networks at the EOC. At the same time, several contributions demonstrated that the properties of the infinite-width NTK depend significantly on the phase of initialization [7], [13]. However, these results do not apply to networks with depth comparable to width since they assume infinite width before considering the effects of growing depth. We fill this gap by deriving statistical properties of the NTK for wide and deep ReLU networks in all three phases of initialization.

II. MAIN RESULTS

We study the variability of the NTK at initialization for fully-connected ReLU DNNs with depth comparable to width and varying initialization hyperparameters. Our contributions are as follows:

- We precisely characterize the dispersion of the diagonal elements \( \Theta(x, x) \) of the NTK (for arbitrary input \( x \)) in the infinite-depth-and-width limit and establish that the variability of the NTK grows exponentially with the depth-to-width ratio at the EOC and in the chaotic phase. Conversely, the variance of \( \Theta(x, x) \) tends to zero in the same limit in the ordered phase. We conclude that the NTK regime can approximate deep networks only in the ordered phase. Moreover, our results allow
Experiments:
- $\sigma_w^2 = 0.8$
- $\sigma_w^2 = 1.4$
- $\sigma_w^2 = 1.7$
- $\sigma_w^2 = 2.0$
- $\sigma_w^2 = 2.5$
- $\sigma_w^2 = 3.5$

Theory:
- Ordered
- EOC
- Chaotic

Fig. 1. $E[\Theta^2(x, x)] / E[\Theta(x, x)]$ ratio (dispersion) of ReLU DNNs with constant width $M \in \{100, 200, 500\}$ and varying hyperparameter $\sigma_w^2$, which characterizes variance of weights at initialization. Ordered phase corresponds to $\sigma_w^2 < 2$, chaotic phase – to $\sigma_w^2 > 2$. The EOC is the initialization with $\sigma_w^2 = 2$. The x-axis represents the depth-to-width ratio $\lambda := L/M$.

We provide non-asymptotic expressions for the first two moments of $\Theta(x, x)$ and illustrate finite-width effects that follow. We show that the variance of the finite-width NTK in the ordered phase gradually increases as the initialization approaches the EOC, i.e. the transition from order to chaos is gradual for finite-width DNNs. We also notice that the NTK dispersion depends on the architecture, i.e. on the varying widths of the fully-connected layers. Notably, the dispersion of $\Theta(x, x)$ decreases with depth in the ordered phase if the DNN increases dimensionality in consecutive layers. This enables us to conclude that such networks get more robust to random initialization with depth in the ordered phase.

We lower-bound the ratio of the expected non-diagonal elements of the NTK, i.e. $\Theta(x, \bar{x})$ with $x \neq \bar{x}$, and the diagonal elements $\Theta(x, x)$ in the infinite-depth-and-width limit. We also upper-bound the dispersion of the non-diagonal elements. In the ordered phase, our results allow to ensure that the whole NTK matrix is approximately deterministic and thus can be approximated by the infinite-width limit.

We provide extensive numerical experiments to verify our theoretical results. An excerpt in Figure 1 shows great agreement between theory and experiments for the dispersion of the diagonal elements of the NTK.

We study the training dynamics of the NTK for fully-connected ReLU DNNs with depth comparable to width and varying initialization hyperparameters. Our contributions are as follows:

- We show that the expected relative change of $\Theta(x, x)$ in the first GD step tends to infinity in the chaotic phase and to zero in the ordered phase in the infinite-depth-and-width limit. Combined with the result by Hanin&Nica (2020) [11], which states that the expected relative change of $\Theta(x, x)$ in the first GD step is exponential in the depth-to-width ratio at the EOC, we can conclude that the NTK of deep networks can stay approximately constant during GD training only in the ordered phase.

- We discuss how the structure of the NTK matrix changes during training outside of the NTK regime. The NTK matrix at initialization is label-agnostic and has an approximately diagonal structure with larger values on the main diagonal as compared to the non-diagonal ones. We speculate that the training process introduces label-awareness by changing the NTK structure to block-diagonal (with blocks of larger values corresponding to classes) and provide experiments to support this sentiment.

REFERENCES

[1] A. Jacot, C. Hongler, and F. Gabriel, “Neural tangent kernel: Convergence and generalization in neural networks,” in *Advances in Neural Information Processing Systems*, 2018, pp. 8580–8589.
[2] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, “Exponential expressivity in deep neural networks through transient chaos,” in *Advances in Neural Information Processing Systems*, 2016, pp. 3360–3368.
[3] K. Huang, Y. Wang, M. Tao, and T. Zhao, “Why do deep residual networks generalize better than deep feedforward networks? — A neural tangent kernel perspective,” in *Advances in Neural Information Processing Systems*, 2020.
[4] M. Geiger, A. Jacot, S. Spigler, F. Gabriel, L. Sagun, S. d’Ascoli, G. Biroli, C. Hongler, and M. Wyart, “Scaling description of generalization with number of parameters in deep learning,” *CoRR*, vol. abs/1901.01608, 2019. [Online]. Available: http://arxiv.org/abs/1901.01608
[5] L. Chizat, E. Oyallon, and F. R. Bach, “On lazy training in differentiable programming,” in *Advances in Neural Information Processing Systems*, 2019, pp. 2933–2943.
[6] M. B. Li, M. Nica, and D. M. Roy, “The future is log-gaussian: ResNets and their infinite-depth-and-width limit at initialization,” *CoRR*, vol. abs/2106.04013, 2021. [Online]. Available: https://arxiv.org/abs/2106.04013
[7] L. Xiao, J. Pennington, and S. Schoenholz, “Disentangling trainability and generalization in deep neural networks,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 10 462–10 472.
[8] S. Fort, G. K. Dziugaite, M. Paul, S. Kharaghani, D. M. Roy, and S. Ganguli, “Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel,” in *Advances in Neural Information Processing Systems*, 2020.
[9] J. Lee, S. S. Schoenholz, J. Pennington, B. Adlam, L. Xiao, R. Novak, and J. Sohl-Dickstein, “Finite versus infinite neural networks: an empirical study,” in *Advances in Neural Information Processing Systems*, 2020.
[10] A. G. de G. Matthews, J. Hron, M. Rowland, R. E. Turner, and Z. Ghahramani, “Gaussian process behaviour in wide deep neural networks,” in *International Conference on Learning Representations, ICLR*, 2018.
[11] B. Hanin and M. Nica, “Finite depth and width corrections to the neural tangent kernel,” in *International Conference on Learning Representations, ICLR*, 2020.
[12] S. S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein, “Deep information propagation,” in *International Conference on Learning Representations, ICLR*, 2017.
[13] S. Hayou, A. Doucet, and J. Rousseau, “Mean-field behaviour of neural tangent kernel for deep neural networks,” *CoRR*, vol. abs/1905.13654, 2019. [Online]. Available: https://arxiv.org/abs/1905.13654