Understanding the Generalization Power of Overfitted NTK Models: 3-layer vs. 2-layer (Extended Abstract)

Peizhong Ju
Department of ECE
The Ohio State University
Columbus, USA
ju.171@osu.edu

Xiaojun Lin
School of ECE
Purdue University
West Lafayette, USA
linx@purdue.edu

Ness B. Shroff
Department of ECE and CSE
The Ohio State University
Columbus, USA
shroff.11@osu.edu

2022 58th Annual Allerton Conference on Communication, Control, and Computing (Allerton) | 979-8-3503-9998-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ALLERTON49937.2022.9929374

I. Introduction

Neural tangent kernel (NTK) models [1] have been recently used as an important intermediate step to understand the exceptional generalization power of overparameterized deep neural networks (DNNs). Compared to linear models with simple Gaussian or Fourier features, NTK models can capture the non-linear features inherent in neural networks. Indeed, the work in [2] has shown that, for a 2-layer NTK model, the generalization error of an overfitted solution decreases as the number of neurons increases. Further, this descent behavior is qualitatively different from that of linear models with simple Gaussian and Fourier features, and closer to that of an actual neural network.

However, the study in [2] is restricted to 2-layer networks. In this work, we study NTK models with 3 layers. Specifically, the input x of dimension d passes through the first layer of p_1 ReLU neurons that are fully connected with the second layer of p_2 ReLU neurons, followed by a linear summation at the third layer, to produce the output. We then study the generalization error of the min ℓ_2-norm solution that fits the training data. We aim to answer the following questions. First, does the interaction of the two hidden layers change the descent behavior in any way? Second, do 3-layer NTK models have any performance advantage over 2-layer NTK models?

II. Main Result and Insights

To answer these questions, we study the generalization performance of the overfitted min-ℓ_2-norm solutions $\hat{f}^{\ell_2}(\cdot)$ for 3-layer NTK models where the middle (i.e., second) layer is trained. Define a set of "learnable" ground-truth functions $\mathcal{F}_{(3)}^{\ell_2} = \left\{f_g(\boldsymbol{x}) = \int_{\mathcal{S}^{d-1}} K^{\text{Three}}(\boldsymbol{x}^T\boldsymbol{z})g(\boldsymbol{z})d\mu(\boldsymbol{z})\right\}$ where $K^{\text{Three}}(\cdot)$ denotes the limiting kernel of 3-layer NTK as the number of neurons approaches infinity. The function $g(\cdot)$ can be an arbitrary function, whose magnitude corresponds to the complexity of the ground-truth function, and the integral is over the uniform distribution μ on the unit-sphere $\mathcal{S}^{d-1}$ in d-dimension space. We then have the following result (detailed version in [3]).

Theorem 1 (an informal version): For any ground-truth function $f(x) = f_g(x) \in \mathcal{F}_{(3)}^{\ell_2}$, when d is fixed and p_1, p_2 are much larger than n, (with high probability) we have

$$|\hat{f}^{\ell_2}(\boldsymbol{x}) - f(\boldsymbol{x})| = \underbrace{O\left(\frac{\|g\|_\infty}{\sqrt{n}}\right)}_{\text{Term A}} + \Bigg(\underbrace{O\left(\frac{\|g\|_1}{\sqrt{p_2}}\right)}_{\text{Term B}} + \underbrace{O\left(\|g\|_1 \sqrt[4]{\frac{\log p_1}{p_1}}\right)}_{\text{Term C}} + \underbrace{\frac{\|\boldsymbol{\epsilon}\|_2}{\sqrt{n}}}_{\text{Term D}}\Bigg) \cdot \underbrace{O\left(n^{\frac{2}{d-1}+\frac{1}{2}} \cdot \sqrt{\log n}\right)}_{\text{Term E}}.$$

Similar to 2-layer NTK [2, 4, 5], our upper bound suggests that the generalization error also decreases to an error floor as the number of neurons p_1 and p_2 increase (as can be seen from Terms B and C). When there is no noise (i.e., no Term D), the error floor (Term A) further decreases as the number of samples n increases. Further, the product of Term D and Term E (i.e., noise term) will not explode when the number of neurons goes to infinity, which is also similar to that for 2-layer NTK. However, our upper bound also reveals new insights that are different from the results for 2-layer NTK, as follows.

Different descent speed: Our upper bound decreases with the number of neuron p_1 in the first hidden-layer at the slower speed of $\sqrt[4]{(\log p_1)/p_1}$ (see Term C), and decreases with the number of neurons p_2 in the second hidden-layer at the faster speed of $1/\sqrt{p_2}$ (see Term B). This difference in descent speed is due to the composition of the two layers.

Size of the Learnable Set: We then show that, even if we only train the middle-layer weights, the learnable set $\mathcal{F}_{(3)}^{\ell_2}$ (i.e., the set of ground-truth functions for which the above upper bound holds) of the 3-layer NTK model without bias contains all finite-degree polynomials, which is strictly larger than that of the 2-layer NTK without bias and is at least as large as the 2-layer NTK with bias. In particular, the learnable set of 2-layer NTK without bias contains only linear functions

This work has been supported in part by NSF grants: 2112471 (also partly funded by DHS), CNS-2106933, CNS-1901057, CNS-2113893, and the Bilsland Dissertation Fellowship at Purdue University, and a grant from the Army Research Office: W911NF-21-1-0244.

and even-power polynomials, and does not contain other odd-power polynomials [2, 4]. In contrast, the learnable sets of both 3-layer NTK (with or without bias) and 2-layer with bias contain polynomials of all powers [5].

Sensitivity to the Choices of Bias: Even though a similar learnable set can be attained by 2-layer NTK with bias, its actual generalization performance can still be quite sensitive to the choice of bias, especially when the input dimension d is large. One type of bias setting commonly used in literature [5, 6] is that the bias has a similar magnitude as each element of the input vector, which we refer to as "normal bias." However, as the dimension d increases, this normal bias diminishes to zero. As a result, the generalization performance will degrade to that of a 2-layer NTK without bias. To avoid this negative impact, it is important to use another type of bias that has a similar magnitude as the norm of the whole input vector, which we refer to as "balanced bias." In contrast, for 3-layer NTK, different bias settings do not have an obvious effect on the generalization performance. In summary, compared with 2-layer NTK, the use of an extra non-linear layer in 3-layer NTK appears to significantly reduce the impact due to the choice of bias, and therefore makes the learning more robust.

References

[1] A. Jacot, F. Gabriel, and C. Hongler, "Neural tangent kernel: Convergence and generalization in neural networks," in *Advances in neural information processing systems*, 2018, pp. 8571–8580.

[2] P. Ju, X. Lin, and N. Shroff, "On the generalization power of overfitted two-layer neural tangent kernel models," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5137–5147.

[3] P. Ju, X. Lin, and N. B. Shroff, "On the generalization power of the overfitted three-layer neural tangent kernel model," *arXiv preprint arXiv:2206.02047*, 2022.

[4] S. Arora, S. Du, W. Hu, Z. Li, and R. Wang, "Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks," in *International Conference on Machine Learning*, 2019, pp. 322–332.

[5] S. Satpathi and R. Srikant, "The dynamics of gradient descent for overparametrized neural networks," in *Learning for Dynamics and Control*. PMLR, 2021, pp. 373–384.

[6] B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari, "Linearized two-layers neural networks in high dimension," *The Annals of Statistics*, vol. 49, no. 2, pp. 1029–1054, 2021.