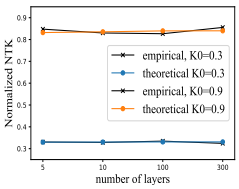
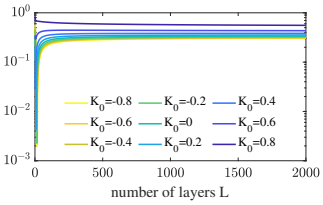
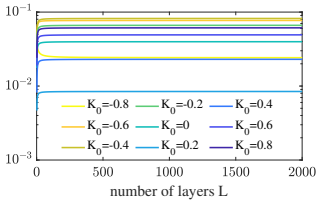
Re: Reviewer #1 For Thm. 4, we randomly initialize the ResNet with width=500, scaling factor $\gamma = 1$ and depth $L = 5, 10, 100, 300$, and then calculate the inner product of the Jacobians of the ResNet for two different inputs as in the definition of NTK (line 185-line187). We repeat the procedure for 500 times and plot the means of the empirical NTKs and the theoretical values in Fig. 1a which shows the two results match very well. For Thm. 5 and Thm. 6, Fig. 1b and Fig. 1c show that $\lim_{L\to\infty} |\overline{\Omega}_L(x, \tilde{x}) - 1/4| \cdot L/\log(\mathrm{L}) \approx$ constant and $\lim_{L\to\infty} |\overline{\Omega}_L(x, \tilde{x}) - \overline{\Omega}_1(x, \tilde{x})| \cdot L \approx$ constant with $x^\top \tilde{x} = K_0$ chosen at 9 points.



(a) Thm. 4. $m = 500$ and scaling $\gamma = 1$



(b) Thm. 5. y-axis is $|\overline{\Omega}_L(x, \tilde{x}) - 1/4| \cdot L/\log(\mathrm{L})$.



(c) Thm. 6. y-axis is $|\overline{\Omega}_L(x, \tilde{x}) - \overline{\Omega}_1(x, \tilde{x})| \cdot L$.

About ResNets' using convolutional layers (Conv) and batch norms(BN) in practice: we acknowledge that Conv and BN are indeed important and our techniques can be applied to them with some modification. We will add some discussions and leave them for our future investigation.

About the scaling factor: The scaling of the bottleneck weight effectively controls the norm of the activations of ResNets, and is commonly adopted in theoretical papers, e.g. [14] in our paper. In practice, the norm control is usually achieved by normalization techniques (batch norm, group norm, etc.).

Re: Reviewer #2 Thank you for the positive comments. About assumptions: we provide non-asymptotic results which bound the error between the finite-width NNs and the infinite-width NNs. Specifically, for ResNets with $\gamma = 1$, our theorem ensures good error control, where the width only depends logarithmically on the depth.

About generalizability: considering kernel ridge regression, one can show that the generalization error is "continuous" w.r.t. the kernel function under some integrable conditions. In this sense, the equivalence (in the view of NTK) means that the generalization of NTKs of sufficiently deep ResNets is close to the generalization of the NTK of 1-layer ResNet. The same applies to the poor generalization of deep FFNets.

Re: Reviewer #3. About our motivation: our increasing depth analysis is a common practice in theoretical research. This is because the infinite-depth behavior is very similar to the large-depth behavior of NNs. This is analogous to we using central limit theorem in practice, but CLT essentially characterizes the limiting distribution based on infinite many samples. Moreover, we highlight that existing results are not limited to 50 and 100 layers. For example, [10] (He et al. 2016) even show that an ultradeep ResNets of 1001 layers can still generalize very well.

We remind that these ultradeep ResNets are not often used in practice because of their massive sizes. If implementable, their generalization performance can be better than existing SOTA results. As more advanced computational hardware (faster GPU's) is developed, deeper ResNets will become more popular in practice.

About our experiments: some existing results on NTK-based methods aim at achieving better performance. Therefore, their experiments use more complex structures with additional tricks, e.g. NTKs of ConvNets with Global Average Pooling. However, the goal of our experiments is to justify the correctness of our theorems – the NTKs of deep FFNets do not generalize, while the NTKs of deep ResNets generalize well. Therefore, the results of our experiments sufficiently serve the purpose of our paper.

About degrading kernel: the degeneration of NTKs of deep FFNets is clearly and rigorously stated in line 237-line 250.

About missing references: We will discuss them in the future version, but please kindly notice that Yang's paper is made public **after** the submission deadline of NeurIPS 2020.

Re: Reviewer #4. Thank you for the positive comments and the references. The previous literatures are significant but have different aspects from ours. We will add more discussions in the next version.

About line 160: We are sorry for the misunderstanding here. Indeed, showing the equivalence of NNs trained by gradient flow and NTK kernel predictor is non-trivial. This result is rigorously proved in [22]. Thank you a lot for pointing out this issue. We will correct the statement in the next version.