CS7020: Mini Project Report
Visualising NNGP, NTK and SGD-NNs for Function Regression

Varun Sundar
EE16B068

November 24, 2019

1 Introduction

Recent work in the theoretical advances of deep learning has focused attention towards behaviour of neural networks in the setting of infinite hidden nodes, also called the "infinite width" setting. Neural Tangent Kernels (NTK) [1, 2] and Neural Networks as a Gaussian Process (NNGP) [3–5] are two main perspectives in the infinite width case. The former shows that as the number of hidden nodes go to infinity, a deep network can be viewed as a kernel machine, allowing for kernel based ridge regression (KRR). On the other hand, the latter show that a randomly initialised neural network is a Gaussian Process, allowing for exact Bayesian Inference by determining its kernel function.

There are many interesting questions that arise at this point:

- Are NNGP and NTK formulations equivalent? The answer to this has been no so far. The kernels do not match.
- Even if the kernels do not match, are the two connected in some intricate fashion? What does literature on the relation between Gaussian Processes and Kernel Machines [6] tell us in this regard?
- Do empirically successful neural networks (trained by gradient descent methods) behave like kernel machines? The answer to this so far is no; methods such as NTK are inferior to SGD trained networks (SGD-NN) by a significant margin, such as 7% on CIFAR-10. Recent improvements to the NTK such as the Extended-CNTK [7] do attempt to bridge this gap.

2 Overview

We build a visualisation framework on top of the `neural-tangents` package based on `streamlit`, which allows for interactive presentation with parameters that can be toggled freely. This code-base compares three methods for the task of function regression: (a) NNGPs, (b) NTKs and (c) SGD-NNs. It is well known that kernel ridge regression can be obtained as the MAP estimate of a Gaussian Process with the same kernel (covariance) function. We include a short proof of this in the presentation. By using this, we can arrive at two methods to evaluate (b):

one using the equivalent Gaussian Process (which corresponds to training the kernel machine till convergence) and one by performing gradient descent using the kernel machine. Finally, in order to get a sense of the randomness induced by initialisation in a finite network trained by SGD, we use a finite ensemble of SGD-NNs for (c).

This visualisation also allows tinkering with the numerous parameters, including:

- Number of test and training points.
- Network architecture: depth, and width (for the finite case), architecture (*erf, ReLU* or linear).
- Function for generating training points, its range and periodicity, and the noise added, $y_i = f(x_i) + \sigma$.
- Variance for initialising network weights and biases.
- Number of training steps for KRR and SGD-NN.

We also include a brief section on extending these comparison to residual architectures. For all these experiments, we heavily utilise the `neural-tangents` library, which follows a `jax` like format for specifying network architectures.

References

[1] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 8580–8589, USA, 2018. Curran Associates Inc.

[2] Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *CoRR*, abs/1904.11955, 2019.

[3] Jaehoon Lee, Yasaman Bahri, Roman Novak, Sam Schoenholz, Jeffrey Pennington, and Jascha Sohl-dickstein. Deep neural networks as gaussian processes. 2018.

[4] Lechao Xiao, Yasaman Bahri, Jascha Sohl-Dickstein, Samuel S. Schoenholz, and Jeffrey Pennington. Dynamical isometry and a mean field theory of cnns: How to train 10,000-layer vanilla convolutional neural networks, 2018.

[5] Roman Novak, Lechao Xiao, Yasaman Bahri, Jaehoon Lee, Greg Yang, Jiri Hron, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Bayesian deep convolutional networks with many channels are gaussian processes. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.

[6] Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences, 2018.

[7] Zhiyuan Li, Ruosong Wang, Dingli Yu, Simon S. Du, Wei Hu, Ruslan Salakhutdinov, and Sanjeev Arora. Enhanced convolutional neural tangent kernels, 2019.