CSIAM Trans. Appl. Math.
doi: 10.4208/csiam-am.SO-2021-0053

Vol. 3, No. 4, pp. 692-760
December 2022

Towards an Understanding of Residual Networks Using Neural Tangent Hierarchy (NTH)

Yuqing Li[1], Tao Luo[2,*] and Nung Kwan Yip[3]

[1] *School of Mathematical Sciences, CMA-Shanghai, Shanghai Jiao Tong University, Shanghai, 200240, P.R. China.*
[2] *School of Mathematical Sciences, CMA-Shanghai, Institute of Natural Sciences, MOE-LSC, and Qing Yuan Research Institute, Shanghai Jiao Tong University, Shanghai, 200240, P.R. China.*
[3] *Department of Mathematics, Purdue University, IN, 47907, USA.*

Received 7 December 2021; Accepted 11 April 2022

Abstract. Gradient descent yields zero training loss in polynomial time for deep neural networks despite non-convex nature of the objective function. The behavior of network in the infinite width limit trained by gradient descent can be described by the Neural Tangent Kernel (NTK) introduced in [25]. In this paper, we study dynamics of the NTK for finite width Deep Residual Network (ResNet) using the neural tangent hierarchy (NTH) proposed in [24]. For a ResNet with smooth and Lipschitz activation function, we reduce the requirement on the layer width m with respect to the number of training samples n from quartic to cubic. Our analysis suggests strongly that the particular skip-connection structure of ResNet is the main reason for its triumph over fully-connected network.

AMS subject classifications: 68U99, 90C26, 34A45

Key words: Residual networks, training process, neural tangent kernel, neural tangent hierarchy.

1 Introduction

Deep neural networks have achieved transcendent performance in a wide range of tasks such as speech recognition [9], computer vision [39], and natural language processing [8]. There are various methods to train neural networks, such as first-order gradient based methods like Gradient Descent (**GD**) and Stochastic Gradient Descent (**SGD**), which have been proven to achieve satisfactory results [20]. Experiments in [49] established

[*]Corresponding author. *Email addresses:* `liyuqing_551@sjtu.edu.cn` (Y. Li), `luotao41@sjtu.edu.cn` (T. Luo), `yipn@purdue.edu` (N. K. Yip)

that, even though with a random labeling of the training images, if one trains the state-of-the-art convolutional network for image classification using SGD, the network is still able to fit them well. There are numerous works trying to demystify such phenomenon theoretically. Du et al. [12, 15] proved that GD can obtain zero training loss for deep and shallow neural networks, and Zou et al. [52] analyzed the convergence of SGD on networks assembled with Rectified Linear Unit (**ReLU**) activation function. All these results are built upon the over-parameterized regime, and it is widely accepted that over-parameterization enables the neural network to fit all training data and bring no harm to the power of its generalization [49]. In particular, the deep neural networks that evaluated positions and selected moves for the well-known program AlphaGo are highly over-parameterized [41, 42].

Another advance is the outstanding performance of Deep Residual Network (**ResNet**), initially proposed by He et al. [22]. ResNet is arguably one of the most groundbreaking works in deep learning, in that it can train up to hundreds or even thousands of layers and still achieves compelling performance [23]. Recent works have shown that ResNet can utilize the features in transfer learning with better efficiency, and its residual link structure enables faster convergence of the training loss [45, 48]. Theoretically, Hardt and Ma [21] proved that for any residual linear networks with arbitrary depth, there are no spurious local optima. Du et al. [12] showed that in the scope of the convergence of GD via over-parameterization for different networks, training ResNet requires weaker conditions compared with fully-connected networks. Apart from that, the advantages of using residual connections remain to be discovered.

In this paper, we contribute to the further understanding of the above two aspects and make improvements in the analysis of their performance. We use the same ResNet structure as in [12]. (Details of the network structure are provided in Section 3.2.) The ResNet has L layers with width m. We will assume that the n data points are not parallel with each other. Such an assumption holds in general for a standard dataset [15]. We focus on the empirical risk minimization problem given by the quadratic loss and the activation function is 1-Lipschitz and analytic. We show that if $m=\Omega\left(n^3L^2\right)$, then the empirical risk $R_S(\boldsymbol{\theta}_t)$ under GD decays exponentially. More precisely,

$$R_S(\boldsymbol{\theta}_t)\leq R_S(\boldsymbol{\theta}_0)\exp\left(-\frac{\lambda t}{n}\right),$$

where λ is the least eigenvalue of $\boldsymbol{K}^{[L+1]}$, definition of which can be found in (4.2).

It is worth noticing that

- Given identical ResNet architectures, for the convergence of randomly initialized GD, our results improve upon [12] in the required number of width per layer from $m=\Omega(n^4L^2)$ to $m=\Omega(n^3L^2)$ (Corollary 4.1).

- For fully-connected network, the required amount of over-parameterization in [24] is $m=\Omega\left(n^3 2^{\mathcal{O}(L)}\right)$. We are able to reproduce the result of Du et al. [12], showing