CSIAM Trans. Appl. Math.
doi: 10.4208/csiam-am.SO-2020-0005

Vol. **2**, No. 3, pp. 484-507
September 2021

Theory of the Frequency Principle for General Deep Neural Networks

Tao Luo[1,*], Zheng Ma[1], Zhi-Qin John Xu[1] and Yaoyu Zhang[1,2]

[1] *School of Mathematical Sciences, Institute of Natural Sciences, MOE-LSC, and Qing Yuan Research Institute, Shanghai Jiao Tong University, Shanghai, 200240, China.*
[2] *Shanghai Center for Brain Science and Brain-Inspired Technology, Shanghai, 200031, China.*

Received 17 October 2020; Accepted 23 February 2021

Abstract. Along with fruitful applications of Deep Neural Networks (DNNs) to realistic problems, recently, empirical studies reported a universal phenomenon of Frequency Principle (F-Principle), that is, a DNN tends to learn a target function from low to high frequencies during the training. The F-Principle has been very useful in providing both qualitative and quantitative understandings of DNNs. In this paper, we rigorously investigate the F-Principle for the training dynamics of a general DNN at three stages: initial stage, intermediate stage, and final stage. For each stage, a theorem is provided in terms of proper quantities characterizing the F-Principle. Our results are general in the sense that they work for multilayer networks with general activation functions, population densities of data, and a large class of loss functions. Our work lays a theoretical foundation of the F-Principle for a better understanding of the training process of DNNs.

AMS subject classifications: 68Q32, 68T07, 37N40

Key words: Frequency principle, Deep Neural Networks, dynamical system, training process.

1 Introduction

Deep learning has achieved great success as in many fields [15], e.g., speech recognition [1], object recognition [10], natural language processing [35] and computer game control [21]. It has also been adopted into algorithms to solve scientific computing problems [8, 11,12,14]. In principle, the universal approximation theorem states that a commonly-used Deep Neural Network (DNN) of sufficiently large width can approximate any function to

*Corresponding author. *Email addresses:* luotao41@sjtu.edu.cn (T. Luo), zhengma@sjtu.edu.cn (Z. Ma), xuzhiqin@sjtu.edu.cn (Z.-Q. J. Xu), zhyy.sjtu@sjtu.edu.cn (Y. Zhang)

a desired precision [7]. However, it remains a mystery that how a DNN finds a minimum corresponding to such an approximation through the gradient-based training process. To understand the learning behavior of DNNs for the approximation problem, recent works model the gradient flow of parameters in a two-layer ReLU neural networks by a partial differential equation (PDE) in the mean-field limit [19, 25, 26]. However, it is not clear whether this PDE approach, which describes a neural network of one hidden layer of infinite width, can be extended to general DNNs of multiple hidden layers and limited neuron number. For further discussion on the mathematical understanding of DNNs, we refer the readers to a review article [9].

In this work, we take another approach that uses Fourier analysis to study the learning behavior of DNNs based on the phenomenon of *Frequency Principle (F-Principle), i.e., a DNN tends to learn a target function from low to high frequencies during the training* [23, 31, 32, 36]. Empirically, the F-Principle can be widely observed in general DNNs for both benchmark and synthetic data [31, 32]. Conceptually, it provides a qualitative explanation of the success and failure of DNNs [32]. E et al., (2019) [30] propose a continuous viewpoint for studying machine learning and suggest that the F-Principle underlying the gradient flows may be a main reason behind the success of modern machine learning. Empirically, the F-Principle provides us a perspective for quantifying the training process via the convergence of each frequency component [13, 22, 29, 33]. For example, it is used as an important phenomenon to pursue fundamentally different learning trajectories of meta-learning [22] and provides an understanding of why increasing the depth of a neural network may accelerate the training [33]. The F-Principle also provides important theoretical insights to design DNN-based algorithms [2, 3, 5, 16, 17, 20, 27, 28]. For example, Blind et al. [3] designs a loss function with explicit higher priority for high frequencies to significantly accelerate the simulation of fluid dynamics through DNN approach; MscaleDNN [16, 17, 28] is developed to accelerate the fitting of high frequency functions by shifting or rescaling high frequencies to lower ones. These works have signified the importance of the F-Principle. Theoretically, Xu et al. [32] propose a theorem for the characterization of the initial training stage of a two-layer tanh network, which is also adopted in the analysis of DNNs with ReLU activation function [23]. Another series of works [4, 6, 24, 34, 36] attempt to understand the F-Principle in very wide neural networks, which can be well approximated by the first-order expansion with respect to the network parameters (the linear neural tangent kernel (NTK) regime). The studies [6, 24, 34] from the perspective of eigen-decomposition of DNN dynamics in spatial domain require assumptions of very large network width and infinite samples. To study the F-Principle with finite samples, Zhang et al. [36] and Luo et al. [18] study the dynamics in the frequency domain and further obtain an effective model of linear F-Principle dynamics, which accurately predicts the learning results of two-layer ReLU neural networks of large widths, leads to an apriori estimate of the generalization error bound. However, the explanation of DNN's F-Principle beyond the NTK regime (non-linear regime) is still missing.

Following the same direction as in [32], in this work, we propose a theoretical frame-