Commun. Comput. Phys.
doi: 10.4208/cicp.OA-2020-0085

Vol. **28**, No. 5, pp. 1746-1767
November 2020

# Frequency Principle: Fourier Analysis Sheds Light on Deep Neural Networks

Zhi-Qin John Xu[1,*], Yaoyu Zhang[2], Tao Luo[1], Yanyang Xiao[3] and Zheng Ma[1]

[1] *School of Mathematical Sciences and Institute of Natural Sciences, Shanghai Jiao Tong University, Shanghai 200240, China.*
[2] *School of Mathematics, Institute for Advanced Study, Princeton, NJ 08540, USA.*
[3] *Brain Cognition and Brain Disease Institutes of Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China.*

Received 9 May 2020; Accepted (in revised version) 16 July 2020

**Abstract.** We study the training process of Deep Neural Networks (DNNs) from the Fourier analysis perspective. We demonstrate a very universal Frequency Principle (F-Principle) — DNNs often fit target functions from low to high frequencies — on high-dimensional benchmark datasets such as MNIST/CIFAR10 and deep neural networks such as VGG16. This F-Principle of DNNs is opposite to the behavior of Jacobi method, a conventional iterative numerical scheme, which exhibits faster convergence for higher frequencies for various scientific computing problems. With theories under an idealized setting, we illustrate that this F-Principle results from the smoothness/regularity of the commonly used activation functions. The F-Principle implies an implicit bias that DNNs tend to fit training data by a low-frequency function. This understanding provides an explanation of good generalization of DNNs on most real datasets and bad generalization of DNNs on parity function or a randomized dataset.

**AMS subject classifications**: 68Q32, 65N06, 68T01

**Key words**: Deep learning, training behavior, generalization, Jacobi iteration, Fourier analysis.

## 1 Introduction

Understanding the training process of Deep Neural Networks (DNNs) is a fundamental problem in the area of deep learning. We find a common behavior of the gradient-based training process of DNNs, that is, a Frequency Principle (F-Principle):

*Corresponding author. *Email addresses:* xuzhiqin@sjtu.edu.cn (Z.-Q. J. Xu), yaoyu@ias.edu (Y. Zhang), luotao41@sjtu.edu.cn (T. Luo), yy.xiao@siat.ac.cn (Y. Xiao), Wudy_MZ@sjtu.edu.cn (Z. Ma)

 

*DNNs often fit target functions from low to high frequencies during the training process.*

In another word, at the early stage of training, the low-frequencies are fitted and as iteration steps of training increase, the high-frequencies are fitted. For example, when a DNN is trained to fit $y=\sin(x)+\sin(2x)$, its output would be close to $\sin(x)$ at early stage and as training goes on, its output would be close to $\sin(x)+\sin(2x)$. Along with our previous works in [34], this paper is one of works that first discovery the F-Principle. In the same time, another group [†] independently found the F-Principle (or spectral bias) [26]. F-Principle was verified empirically in synthetic low-dimensional data with MSE loss during DNN training [26,34]. However, in deep learning, empirical phenomena could vary from one network structure to another, from one dataset to another and could exhibit significant difference between synthetic data and high-dimensional real data. Therefore, the universality of the F-Principle remains an important problem for further study. Especially for high-dimensional real problems, because the computational cost of high-dimensional Fourier transform is prohibitive in practice, it is of great challenge to demonstrate the F-Principle. On the other hand, the mechanism underlying the F-Principle and its implication to the application of DNNs, e.g., design of DNN-based PDE solver, as well as their generalization ability are also important open problems to be addressed.

In this work, we design two methods, i.e., projection and filtering methods, to show that the F-Principle exists in the training process of DNNs for high-dimensional benchmarks, i.e., MNIST [22], CIFAR10 [21]. The settings we have considered are i) different DNN architectures, e.g., fully-connected network, convolutional neural network (CNN), and VGG16 [28]; ii) different activation functions, e.g., tanh and rectified linear unit (ReLU); iii) different loss functions, e.g., cross entropy, mean squared error (MSE), and loss energy functional in variational problems. These results demonstrate the universality of the F-Principle.

To facilitate the designs and applications of DNN-based schemes, we characterize a stark difference between DNNs and the Jacobi method, a conventional numerical scheme exhibiting the opposite convergence behavior — faster convergence for higher frequencies. Numerical methods [5,7,31], such as well-known multigrid method [5,31], are developed to accelerate the convergence for low frequency. As the DNN-based schemes have potential to solve high-dimensional problems [8, 13–15, 17, 19, 30, 32], the low-frequency bias of DNN can be adopted to accelerate the convergence of low frequencies for computational problems.

We also intuitively explain with theories under an idealized setting how the smoothness/regularity of commonly used activation functions contributes to the F-Principle. Note that this mechanism is rigorously demonstrated for DNNs of general settings in a subsequent work [23]. Finally, we discuss that the F-Principle provides an understanding of good generalization of DNNs in many real datasets [37] and poor generalization in learning the parity function [25,27], that is, the F-Principle which implies that DNNs prefer low frequencies, is consistent with the property of low frequencies dominance in

[†]We acknowledge each other after communication with authors in [26]