CSIAM Trans. Appl. Math.
doi: 10.4208/csiam-am.SO-2020-0006

Vol. **3**, No. 2, pp. 299-317
June 2022

Implicit Bias in Understanding Deep Learning for Solving PDEs Beyond Ritz-Galerkin Method

Jihong Wang[1], Zhi-Qin John Xu[2,*], Jiwei Zhang[3,*] and Yaoyu Zhang[2]

[1] *Beijing Computational Science Research Center, Beijing 100193, P.R. China.*
[2] *School of Mathematical Sciences, Institute of Natural Sciences, MOE-LSC and Qing Yuan Research Institute, Shanghai Jiao Tong University, Shanghai 200240, P.R. China.*
[3] *School of Mathematics and Statistics, and Hubei Key Laboratory of Computational Science, Wuhan University, Wuhan 430072, P.R. China.*

Received 18 November 2020; Accepted 28 September 2021

Abstract. This paper aims at studying the difference between Ritz-Galerkin (R-G) method and deep neural network (DNN) method in solving partial differential equations (PDEs) to better understand deep learning. To this end, we consider solving a particular Poisson problem, where the information of the right-hand side of the equation f is only available at n sample points, that is, f is known at finite sample points. Through both theoretical and numerical studies, we show that solution of the R-G method converges to a piecewise linear function for the one dimensional problem or functions of lower regularity for high dimensional problems. With the same setting, DNNs however learn a relative smooth solution regardless of the dimension, this is, DNNs implicitly bias towards functions with more low-frequency components among all functions that can fit the equation at available data points. This bias is explained by the recent study of frequency principle. In addition to the similarity between the traditional numerical methods and DNNs in the approximation perspective, our work shows that the implicit bias in the learning process, which is different from traditional numerical methods, could help better understand the characteristics of DNNs.

AMS subject classifications: 35Q68, 65N30, 65N35

Key words: Deep learning, Ritz-Galerkin method, partial differential equations, F-Principle.

1 Introduction

Deep neural networks (DNNs) become increasingly important in scientific computing fields [5–7,10–13,16,17,22,26,31]. A major potential advantage over traditional numerical

*Corresponding author. *Email addresses:* jhwang@csrc.ac.cn (J. Wang), xuzhiqin@sjtu.edu.cn (Z. Xu), jiweizhang@whu.edu.cn (J. Zhang), zhyy.sjtu@sjtu.edu.cn (Y. Zhang)

methods is that DNNs could overcome the curse of dimensionality in high-dimensional problems. With traditional numerical methods, several studies have made progress on the understanding of the algorithm characteristics of DNNs. For example, by exploring ReLU DNN representation of continuous piecewise linear function in FEM, the work [13] theoretically establishes that a ReLU DNN can accurately represent any linear finite element functions. In the aspect of the convergence behavior, the works [32, 33] show a Frequency Principle (F-Principle) that DNNs often learn low-frequency components first while most of the conventional methods (e.g., Jacobi method) exhibit the opposite convergence behavior—higher-frequency components are learned faster. These understandings could lead to a better use of DNNs in practice, such as DNN-based algorithms are proposed based on the F-Principle to fast eliminate high-frequency error [3, 17].

As the DNN-based algorithms are increasingly important in solving PDEs, it is important to study the property of the DNN solution. The aim of this paper is to investigate the different behaviors between DNNs and Ritz-Galerkin (R-G) method (as a traditional numerical method). To this end, we utilize an example to show their stark difference, that is, solving PDEs only with a few given sample points. We denote n by the sample number and m by the basis number in the Ritz-Galerkin method or the neuron number in DNNs. In traditional PDE models, we consider the situation where the source functions in the equation are completely known, i.e. the sample number n can go to infinity. But in practical applications, such as signal processing, statistical mechanics, chemical and biophysical dynamic systems, we often encounter the problems that only a few sample values can be obtained. It is interesting to ask what effect R-G methods would have on solving this particular problem, and what the solution would be obtained by the DNN method. On the other hand, DNN is well-known often over-parameterized in real applications. For a fair comparison, the R-G method is also set as over-parameterized when the number of basis functions goes to infinity.

In this paper, we show that R-G method considers the discrete sampling points as linear combinations of Dirac delta functions, while DNN method always uses a relatively smooth function to interpolate the discrete sampling points. And we incorporate the F-Principle to show how DNN method is different from the R-G method, that is, for all functions that can fit the training data, DNNs implicitly bias towards functions with more low-frequency components. In addition to the similarity between the traditional numerical methods and DNNs in the approximation perspective [13], our work shows that the implicit bias in the learning process, which is different from traditional numerical methods, could help better understand the characteristics of DNNs.

The rest of the paper is organized as follows. In Section 2, we briefly introduce the R-G method and the DNN method. In Sections 3 and 4, we present the difference between the two methods in solving PDEs numerically, and provide some theoretical analysis. We end the paper with the conclusion in Section 5.