Commun. Comput. Phys.
doi: 10.4208/cicp.OA-2023-0264

Vol. **36**, No. 1, pp. 71-103
July 2024

Convergence Analysis for Over-Parameterized Deep Learning

Yuling Jiao[1], Xiliang Lu[1], Peiying Wu[1] and Jerry Zhijian Yang[1,*]

[1] *School of Mathematics and Statistics, Wuhan University, Wuhan 430072, P.R. China.*

Received 7 October 2023; Accepted (in revised version) 22 April 2024

Abstract. The success of deep learning in various applications has generated a growing interest in understanding its theoretical foundations. This paper presents a theoretical framework that explains why over-parameterized neural networks can perform well. Our analysis begins from the perspective of approximation theory and argues that over-parameterized deep neural networks with bounded norms can effectively approximate the target. Additionally, we demonstrate that the metric entropy of such networks is independent of the number of network parameters. We utilize these findings to derive consistency results for over-parameterized deep regression and the deep Ritz method, respectively. Furthermore, we prove convergence rates when the target has higher regularity, which, to our knowledge, represents the first convergence rate for over-parameterized deep learning.

AMS subject classifications: 65M15, 65N15, 65Y20

Key words: Over-parameterization, convergence rate, approximation, generalization.

1 Introduction

The success of deep learning in various applications has spurred a growing interest in understanding its theoretical foundations. One of the most crucial questions is why over-parameterized neural networks can perform well. The current literature [40] suggests that the generalization error of neural networks generally increases with the increasing complexity of the network function space, making it theoretically difficult for over-parameterized neural networks to converge in terms of generalization error. However, in practice, training over-parameterized deep neural networks is widely used since it makes model training more computationally convenient. Moreover, recent studies have shown

*Corresponding author. *Email addresses:* yulingjiaomath@whu.edu.cn (Y. Jiao), xllv.math@whu.edu.cn (X. Lu), peiyingwu@whu.edu.cn (P. Wu), zjyang.math@whu.edu.cn (J. Z. Yang)

that (stochastic) gradient descent with randomized initialization and small step-size converges linearly in over-parameterized regimes, even though the optimization problem in deep learning is highly non-convex, see [2,12,13,25,34,55] and the references therein. All of these indicate a conflict between existing theory and practice, and a new perspective is urgently needed to resolve this dilemma.

To address this dilemma, significant effort has been devoted to developing over-parameterized deep learning theory [4,6,10]. Belkin et al. proposed the double descent curve in [6] to describe the limitations of classical analysis, but did not provide explanations. Currently, the main perspective on understanding over-parameterization for linear and kernel models is benign overfitting due to the double descent phenomenon for estimators interpolating data with minimum norm [3,4,6–9,33,43,50]. However, [29] provides a negative result that the empirical risk minimization estimator can be inconsistent in nonparametric least squares regression with over-parameterized deep neural networks. In this work, we introduce a new theoretical framework based on function space theory and establish the consistency of norm-bounded over-parameterized deep learning. We demonstrate that the complexity of a neural network can be controlled by the metric entropy of the balls in certain metric space, which is independent of the number of parameters. This provides a novel perspective for understanding the good generalization ability of over-parameterized neural networks. We illustrate our approach with two representative examples: the regression model and the deep Ritz method. The main contributions of this work are summarized as follows.

- We establish a new bound for the approximation error of *ReLU* deep neural networks in the Sobolev space, which may be of independent interest.
- We provide a unified consistency analysis of over-parameterized regression models and deep Ritz methods, which offers a novel perspective for understanding over-parameterized deep learning.
- Our framework is applicable to various activation functions, including *ReLU* and *Sigmoidal* functions. By exploring the smoothness of the target and network, we drive improved convergence rate.

The paper is organized as follows. In Section 2, we give some notations and mathematical background used in this paper. Section 3 provides a brief overview of our main results. In Section 4, we present our proof framework. In Section 5, we summarize our findings and conclude the paper. Some technical detailed proofs are given in Section A.

2 Notations and background

In this section, we provide all the notations we need in this paper. For $k\in\mathbb{R}$, we define $\mathbb{R}_{>k}:=\{x\in\mathbb{R}|x>k\}$ and $\mathbb{N}_0:=\{x\in\mathbb{N}|x\geq 0\}$. If $x\in\mathbb{R}$, $\lfloor x\rfloor:=\max\{k\in\mathbb{Z}:k\leq x\}$ denotes the largest integer strictly smaller than x. $\mathfrak{C}\in\mathbb{R}$ is a positive constant number, and $\mathfrak{C}(d)$