INTERNATIONAL JOURNAL OF
NUMERICAL ANALYSIS AND MODELING
Volume 21, Number 5, Pages 674–696

© 2024 Institute for Scientific
Computing and Information
doi: 10.4208/ijnam2024-1027

HOW CAN DEEP NEURAL NETWORKS FAIL EVEN WITH GLOBAL OPTIMA?

QINGGUANG GUAN

Abstract. Fully connected deep neural networks are successfully applied to classification and function approximation problems. By minimizing the cost function, i.e., finding the proper weights and biases, models can be built for accurate predictions. The ideal optimization process can achieve global optima. However, do global optima always perform well? If not, how bad can it be? In this work, we aim to: 1) extend the expressive power of shallow neural networks to networks of any depth using a simple trick, 2) construct extremely overfitting deep neural networks that, despite having global optima, still fail to perform well on classification and function approximation problems. Different types of activation functions are considered, including ReLU, Parametric ReLU, and Sigmoid functions. Extensive theoretical analysis has been conducted, ranging from one-dimensional models to models of any dimensionality. Numerical results illustrate our theoretical findings.

Key words. Deep neural network, global optima, binary classification, function approximation, overfitting.

1. Introduction

Fully connected deep neural networks are the fundamental components of modern deep learning architectures, serving as the building blocks for various models like convolutional neural networks [12], transformers [21], and numerous others. The effectiveness of deep neural networks lies in their ability to approximate complex functions, making them essential tools for tasks ranging from image recognition to natural language processing. However, along with their expressive power, deep neural networks also exhibit a phenomenon known as overfitting, where they may fit the training data very well instead of capturing the underlying patterns. This underscores the importance of understanding both the approximation capabilities and the limitations of deep neural networks. Since neural network models are obtained through training, which involves optimizing a cost function. The ultimate goal is to find the global optima, which represent configurations of the network parameters that minimize the discrepancy between the predicted outputs and the actual targets. However, achieving global optima does not guarantee optimal performance, as the network may still suffer from overfitting or other issues. Therefore, it is crucial to thoroughly examine the properties of global optima to understand how they affect the performance of the model.

In this paper, we will focus on the regression problem formulated as scalar-valued function approximation. Let the target be a scalar-valued function $g(\mathbf{x})$ (in the case of binary classification, $g(\mathbf{x})$ has values 1 and -1). The variable is $\mathbf{x} \in \mathbb{R}^d$, where d is a positive integer. The training set is defined as

$$\left\{\mathbf{x}_l, y_l\right\}_{l=1}^{\mathbb{L}},$$

where $y_l = g(\mathbf{x}_l)$, and $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{\mathbb{L}}$ are samples drawn from a uniform distribution in a d-dimensional cube $[0, 1]^d$. Thus, the input layer has d neurons, and the output

Received by the editors on May 27, 2024 and, accepted on August 3, 2024.
2000 *Mathematics Subject Classification.* 68T07, 65Z05, 65D99.

layer has one neuron. Suppose there are $\mathbb{K}$ hidden layers in the network. We define the output as a function $f_{\mathbb{K}}(\mathbf{x})$, omitting the parameters of weights and biases in the function definition. The cost functions are Median Absolute Error (MAE error) defined in equation (1) or Mean Squared Error (MSE error) defined in equation (2):

$$C_{mae}(W,B) = \frac{1}{\mathbb{L}}\sum_{l=1}^{\mathbb{L}} \Big| g(\mathbf{x}_l) - f_{\mathbb{K}}(\mathbf{x}_l) \Big|, \tag{1}$$

$$C_{mse}(W,B) = \frac{1}{\mathbb{L}}\sum_{l=1}^{\mathbb{L}} \Big(g(\mathbf{x}_l) - f_{\mathbb{K}}(\mathbf{x}_l) \Big)^2, \tag{2}$$

where W, B are weights and biases of $f_{\mathbb{K}}(\mathbf{x})$. From (1) and (2), we know $C_{mae}(W,B) \geq 0$ and $C_{mse}(W,B) \geq 0$. If there exist W^* and B^* such that $C_{mae}(W^*,B^*) = 0$ or $C_{mse}(W^*,B^*) = 0$, then (W^*,B^*) is a global minimizer for the corresponding cost function.

For properly designed binary classification and function approximation problems, we can construct neural networks of any depth that fit the training data perfectly, achieving global optima and zero training loss. However, those neural networks have the worst generalization error. The extreme case is that the model only works on the training set; for any data not in the training set, the output of the model is meaningless.

The paper is organized as follows: In Section 2, we propose a simple trick to extend the universal approximation of shallow networks to deep neural networks of any depth. Various activation functions are considered. In Section 3, we construct examples of binary classification and function approximation in one, two, and high dimensions for networks with ReLU activation functions. Section 4 is devoted to deep neural networks with Parametric ReLU activation functions. The constructions are slightly different compared to the ReLU function. In Section 5, we only consider function approximation problems for networks with Sigmoid activation functions. Conclusions are drawn in Section 6.

2. A Simple Trick to Extend the Expressivity of Shallow Neural Networks to Any Depth

The approximation properties of shallow neural networks have been extensively studied, including universal approximation [3, 17, 9, 1, 4, 13], and higher order estimations [19, 20]. However, extending these existing results to any depth is either too complicated or requires many neurons in the subsequent hidden layers, see [8, 22, 6, 18, 10]. Before presenting examples that can cause deep neural networks (DNNs) to fail, we employ a very simple trick to extend the expressive power of shallow neural networks to networks of any depth, the minimum width of the following attached hidden layers can be as small as one. The activation functions are assumed to be ReLU-like [15, 2, 11], which have a linear part x, if $x \geq 0$; or C^2 continuous with bounded second order derivatives, such as Sigmoid, Tanh, Softplus [5], Gaussian and RBFs [17].

Theorem 2.1. *Suppose a bounded scalar valued function $f(\mathbf{x}), \mathbf{x} \in \mathbb{R}^d, d \geq 1$ can be approximated by a fully connected neural network with $\mathbb{K}$ hidden layers, $\mathbb{K} \geq 1$, then after attaching $\mathbb{N}$ extra hidden layers with any width ≥ 1, the function can still be approximated by the deep neural network with $\mathbb{K} + \mathbb{N}$ hidden layers. $\mathbb{N}$ can be any positive integer.*