Journal of Machine Learning **ISSN**: 2790-2048(e), 2790-203X(p)

On the Existence of Optimal Shallow Feedforward Networks with ReLU Activation

Steffen Dereich * [1] and Sebastian Kassing † [2]

[1]Institute for Mathematical Stochastics, Faculty of Mathematics and Computer Science, University of Münster, Germany.
[2]Faculty of Mathematics, University of Bielefeld, Germany.

Abstract. We prove existence of global minima in the loss landscape for the approximation of continuous target functions using shallow feedforward artificial neural networks with ReLU activation. This property is one of the fundamental artifacts separating ReLU from other commonly used activation functions. We propose a kind of closure of the search space so that in the extended space minimizers exist. In a second step, we show under mild assumptions that the newly added functions in the extension perform worse than appropriate representable ReLU networks. This then implies that the optimal response in the extended target space is indeed the response of a ReLU network.

Keywords:
Neural Networks,
Shallow Networks,
Best Approximation,
ReLU Activation,
Approximatively Compact.

Article Info.:
Volume: 3
Number: 1
Pages: 1 - 22
Date: March/2024
doi.org/10.4208/jml.230903

Article History:
Received: 03/09/2023
Accepted: 24/01/2024

Communicated by:
Arnulf Jentzen

1 Introduction

Modern machine learning algorithms are commonly based on the optimization of artificial neural networks (ANNs) through gradient based algorithms. The overwhelming success of these methods in practical applications has encouraged many scientists to build the mathematical foundations of machine learning and, in particular, to identify universal structures in the training dynamics that might provide an explanation for the mind-blowing observations practitioners make. One key component of ANNs is the activation function. Among the various activation functions that have been proposed, the rectified linear unit (ReLU), which is defined as the maximum between zero and the input value, has emerged as the most widely used and most effective activation function. There are several reasons why ReLU has become such a popular choice, e.g. it is easy to implement, computational efficient and overcomes the vanishing gradient problem, which is a common issue with other activation functions when training ANNs. In this work, we point out and prove a more subtle feature of the ReLU function that separates ReLU from several other common activation functions and might be one of the key reasons for its popularity in practice: the existence of global minima in the optimization landscape.

*Corresponding author. steffen.dereich@uni-muenster.de
†skassing@math.uni-bielefeld.de

A popular line of research studies the optimization procedure (also called training) for ANNs using gradient descent (GD) type methods. Since the error function in a typical machine learning optimization task is non-linear, non-convex and even non-coercive it remains an open problem to rigorously prove (or disprove) convergence of GD even in the simple scenario of optimizing a shallow ANN, i.e. an ANN with only one hidden layer. Existing theoretical convergence results often assume the process to stay bounded, i.e. for every realization there exists a compact set such that the process does not leave this set during training, see, e.g. [3, 11, 15] for results concerning gradient flows, [1, 2] for results concerning deterministic gradient methods, [5, 7, 23, 28] for results concerning stochastic gradient methods and [8] for results concerning gradient based diffusion processes. Many results go back to classical works by Łojasiewicz concerning gradient inequalities for analytic target functions and direct consequences for the convergence of gradient flow trajectories under the assumption of staying bounded [20–22].

In this context, it seems natural to ask for the existence of ANNs that solve the minimization task within the search space. More explicitly, if there does not exist a global minimum in the optimization landscape then every sequence that approaches the minimal loss value diverges to infinity. This might lead to slow convergence or even rule out convergence of the loss value, which is the property that practitioners are most interested in. Therefore, it seems reasonable to choose a network architecture, activation function and loss function such that there exist global optima in the optimization landscape.

Overparametrized networks in the setting of empirical risk minimization (more ReLU neurons than data points to fit) are able to perfectly interpolate the data (see, e.g. [12, Lemma 27.3]) such that there exists a network configuration achieving zero error and, thus, a global minimum in the search space. For shallow feedforward ANNs using ReLU activation it has been shown that also in the underparametrized regime there exists a global minimum if the ANN has a one-dimensional output [18], whereas there are pathological counterexamples in higher dimensions [19]. However, for general measures μ not necessarily consisting of a finite number of Dirac measures, the literature on the existence of global minima is very limited. There exist positive results for the approximation of functions in the space $L^p([0,1]^d)$ with shallow feedforward ANNs using heavyside activation [16], the approximation of Lipschitz continuous target functions with shallow feedforward ANNs using ReLU activation and the standard mean square error in the case where the input and output dimension is one-dimensional [15], and the approximation of multi-dimensional, real-valued continuous target functions with shallow residual ANNs using ReLU activation [6]. On the other hand, for several common (smooth) activations such as the standard logistic activation, softplus, arctan, hyperbolic tangent and softsign there, generally, do not exist minimizers in the optimization landscape for smooth target functions (or even polynomials), see [13, 24]. This phenomenon can also be observed in empirical risk minimization for the hyperbolic tangent activation. As shown in [19], in the underparametrized setting, there exist input data such that for all output data from a set of positive Lebesgue measure there does not exist minimizers in the optimization landscape.

In this article, we prove, for the first time, existence results for shallow feedforward ReLU ANNs with multi-dimensional input space for the population loss. Interestingly, minimizers exist under very mild assumptions on the optimization problem. This exis-