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

Approximation Results for Gradient Flow Trained Neural Networks

Gerrit Welper * [1]

[1]Department of Mathematics, University of Central Florida, Orlando FL, USA.

Abstract. The paper contains approximation guarantees for neural networks that are trained with gradient flow, with error measured in the continuous $L_2(\mathbb{S}^{d-1})$-norm on the d-dimensional unit sphere and targets that are Sobolev smooth. The networks are fully connected of constant depth and increasing width. We show gradient flow convergence based on a neural tangent kernel (NTK) argument for the non-convex optimization of the second but last layer. Unlike standard NTK analysis, the continuous error norm implies an under-parametrized regime, possible by the natural smoothness assumption required for approximation. The typical over-parametrization re-enters the results in form of a loss in approximation rate relative to established approximation methods for Sobolev smooth functions.

Keywords:
Deep neural networks,
Approximation,
Gradient descent,
Neural tangent kernel.

Article Info.:
Volume: 3
Number: 2
Pages: 107 - 175
Date: /2024
doi.org/10.4208/jml.230924

Article History:
Received: 24/09/2023
Accepted: 25/03/2024

Communicated by:
Zhi-Qin John Xu

Contents

*Corresponding author. gerrit.welper@ucf.edu

1 Introduction

Direct approximation results for a large variety of methods, including neural networks, are typically of the form

$$\inf_{\theta} \|f_\theta - f\| \le n(\theta)^{-r}, \quad f \in K. \tag{1.1}$$

I.e. a target function f is approximated by an approximation method f_θ, parametrized by some degrees of freedom or weights θ up to a rate $n(\theta)^{-r}$ for some $n(\theta)$ that measures the richness of the approximation method as width, depth or number of weights for neural networks. Generally, the approximation rate can be arbitrarily slow unless the target f is contained in some compact set K, which depends on the approximation method and application and is typically a unit ball in a Sobolev, Besov, Barron or other normed smoothness space. Such results are well established for a variety of neural network architectures and compact sets K, however, these results rarely address how to practically compute the infimum in the formula above and instead use hand-picked weights.

On the other hand, the neural network optimization literature, typically considers discrete error norms (or losses)

$$\|f_\theta - f\|_* := \left(\frac{1}{n} \sum_{i=1}^{n} |f_\theta(x_i) - f(x_i)|^2 \right)^{\frac{1}{2}}$$

together with neural networks that are over-parametrized, i.e. for which the number of weights is larger than the number of samples n so that they can achieve zero training error