CSIAM Trans. Appl. Math.
doi: 10.4208/csiam-am.20-211

Vol. **1**, No. 3, pp. 387-440
September 2020

On the Banach Spaces Associated with Multi-Layer ReLU Networks: Function Representation, Approximation Theory and Gradient Descent Dynamics

Weinan E[1,2,3,*] and Stephan Wojtowytsch[2,*]

[1] *Department of Mathematics, Princeton University, Princeton, NJ 08544, USA.*
[2] *Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544, USA.*
[3] *Beijing Institute of Big Data Research, Beijing, P.R. China.*

Received 1 September 2020; Accepted 21 September 2020

Abstract. We develop Banach spaces for ReLU neural networks of finite depth L and infinite width. The spaces contain all finite fully connected L-layer networks and their L^2-limiting objects under bounds on the natural path-norm. Under this norm, the unit ball in the space for L-layer networks has low Rademacher complexity and thus favorable generalization properties. Functions in these spaces can be approximated by multi-layer neural networks with dimension-independent convergence rates.

The key to this work is a new way of representing functions in some form of expectations, motivated by multi-layer neural networks. This representation allows us to define a new class of continuous models for machine learning. We show that the gradient flow defined this way is the natural continuous analog of the gradient descent dynamics for the associated multi-layer neural networks. We show that the path-norm increases at most polynomially under this continuous gradient flow dynamics.

AMS subject classifications: 68T07, 46E15, 26B35, 35Q68, 34A12, 26B40

Key words: Barron space, multi-layer space, deep neural network, representations of functions, machine learning, infinitely wide network, ReLU activation, Banach space, path-norm, continuous gradient descent dynamics, index representation.

1 Introduction

It is well-known that neural networks can approximate any continuous function on a compact set arbitrarily well in the uniform topology as the number of trainable parameters increase [9, 26, 32]. However, the number and magnitude of the parameters required

[*] Corresponding author. *Email addresses:* `weinan@math.princeton.edu` (W. E), `stephanw@princeton.edu` (S. Wojtowytsch)

may make this result unfeasible for practical applications. Indeed it has been shown to be the case when two-layer neural networks are used to approximate general Lipschitz continuous functions [22]. It is therefore necessary to ask which functions can be approximated *well* by neural networks, by which we mean that as the number of parameters goes to infinity, the convergence rate should not suffer from the curse of dimensionality.

In classical approximation theory, the role of neural networks was taken by (piecewise) polynomials or Fourier series and the natural function spaces were Hölder spaces, (fractional) Sobolev spaces, or generalized versions thereof [33]. In the high-dimensional theories characteristic for machine learning, these spaces appear inappropriate (for example, approximation results of the kind discussed above do not hold for these spaces) and other concepts have emerged, such as reproducing kernel Hilbert spaces for random feature models [37], Barron spaces for two-layer neural networks [4,17–19,22,23,29], and the flow-induced space for residual neural network models [18].

In this article, we extend these ideas to networks with several hidden (infinitely wide) layers. The key is to find how functions in these spaces should be represented and what the right norm should be. Our most important results are:

1. There exists a class of Banach spaces associated with multi-layer neural networks which has low Rademacher complexity (i.e. multi-layer functions in these spaces are easily learnable).

2. The neural tree spaces introduced here are the appropriate function spaces for the corresponding multi-layer neural networks in terms of direct and inverse approximation theorems.

3. The gradient flow dynamics is well defined in a much simpler subspace of the corresponding neural tree space. Functions in this space admit an intuitive representation in terms of compositions of expectations. The path norm increases at most polynomially in time under the natural gradient flow dynamics. Since the path-norm controls the generalization gap, this slow increase suggests that gradient flow training does not lead to overfitting.

These results justify our choice of function representation and the norm.

Neural networks are parametrized by weight matrices which share indices only between adjacent layers. To understand the approximation power of neural networks, we rearrange the index structure of weights in a tree-like fashion and show that the approximation problem under path-norm bounds remains unchanged. This approach makes the problem more linear and easier to handle from the approximation perspective, but is unsuitable when describing training dynamics. To address this discrepancy, we introduce a subspace of the natural function spaces for very wide multi-layer neural networks (or neural trees) which automatically incorporates the structure of neural networks. For this subspace, we investigate the natural training dynamics and demonstrate that the path-norm increases at most polynomially during training.