Journal of Computational Mathematics
Vol.39, No.6, 2021, 801–815.

http://www.global-sci.org/jcm
doi:10.4208/jcm.2007-m2019-0239

DEEP RELU NETWORKS OVERCOME THE CURSE OF DIMENSIONALITY FOR GENERALIZED BANDLIMITED FUNCTIONS*

Hadrien Montanelli
Centre de Mathématiques Appliquées, École Polytechnique, Palaiseau, France
Email: hadrien.montanelli@polytechnique.edu
Haizhao Yang[1)]
Department of Mathematics, Purdue University, Indiana, United States
Email: haizhao@purdue.edu
Qiang Du
Department of Applied Physics and Applied Mathematics, Columbia University, New York, United States
Email: qd2125@columbia.edu

Abstract

We prove a theorem concerning the approximation of generalized bandlimited multivariate functions by deep ReLU networks for which the curse of the dimensionality is overcome. Our theorem is based on a result by Maurey and on the ability of deep ReLU networks to approximate Chebyshev polynomials and analytic functions efficiently.

Mathematics subject classification: 68T01, 33F05, 41A10.
Key words: Machine learning, Deep ReLU networks, Curse of dimensionality, Approximation theory, Bandlimited functions, Chebyshev polynomials.

1. Introduction

The curse of dimensionality is a inevitable issue in high-dimensional scientific computing. Standard numerical algorithms whose cost is exponential in the dimension d are prohibitive when d is large. As a mesh-free function parametrization tool, neural networks are believed to be a suitable approach to conquer the curse of dimensionality. In this paper, we show that deep ReLU networks overcome the curse of dimensionality for *generalized bandlimited functions*, which we shall define at the end of the introduction. Let us first quickly review what networks are.

Shallow networks are approximations $\widetilde{f}_W$ of multivariate functions $f : \mathbb{R}^d \to \mathbb{R}$ of the form

$$\widetilde{f}_W(\boldsymbol{x}) = \sum_{i=1}^{W} \alpha_i \sigma(\boldsymbol{w}_i \cdot \boldsymbol{x} + \theta_i), \tag{1.1}$$

for some *activation function* $\sigma : \mathbb{R} \to \mathbb{R}$, weights $\alpha_i, \theta_i \in \mathbb{R}$, $\boldsymbol{w}_i \in \mathbb{R}^d$ and integer $W \geq 1$. Each operation $\sigma(\boldsymbol{w}_i \cdot \boldsymbol{x} + \theta_i)$ is called a *unit* and the W units in (1.1) form a *hidden layer*; this is a special form of nonlinear approximation [1, 2]. *Deep networks* are compositions of shallow networks and have several hidden layers, and each unit of each layer performs an operation

* Received October 25, 2019 / Revised version received May 25, 2020 / Accepted July 15, 2020 / Published online October 12, 2021 /
[1)] Corresponding author

of the form $\sigma(\boldsymbol{w} \cdot \boldsymbol{x} + \theta)$. Following Yarotsky [3], we allow connections between units in non-neighboring layers. We define the *depth* L of a network as the number of hidden layers and the *size* W as the total number of units. In practice, networks with depth $L = \mathcal{O}(1)$ are considered shallow, while deep networks have typically $L \gg 1$ layers.

Before the revolution of deep learning [4], most research concerned shallow networks with sigmoid activation functions. Nowadays, networks using the *REctifier Linear Unit* (*ReLU*) activation function $\sigma(x) = \max(0, x)$ have become the most popular tool, partly because sigmoid activation functions lead to severe gradient degeneracy during the optimization process. It was also shown in [5] that deep ReLU networks produce sparsity that helps a wide range of machine learning applications; smooth activation functions, including smoothed ReLU functions, do not. This is why we focus on ReLU networks in this paper.

The theory of approximating functions using shallow networks goes back to 1989 when Cybenko showed that any continuous functions can be approximated by shallow networks [6], while Hornik, Stinchcombe and White proved a similar result for Borel measurable functions [7]. In the 1990s, the attention shifted to the *approximation power*[1)] of shallow networks [8–11]. Of particular interest was the absence of the curse of dimensionality in the approximation of functions with fast decaying Fourier coefficients [12].

Fast forward to the 2010s and the success of deep networks, one of the most important theoretical problems is to determine why and when deep networks can lessen or break the curse of dimensionality, especially for ReLU networks. One may focus on a particular set of functions which have a very special structure (such as compositional or polynomial), and show that for this particular set deep networks overcome the curse of dimensionality [13–21]. Alternatively, one may consider a function space that is more generic for multivariate approximation in high dimensions, such as Korobov spaces [22], and prove convergence results for which the curse of dimensionality is lessened [23].

In this paper, we may consider *generalized bandlimited* functions $f : B = [0, 1]^d \to \mathbb{R}$ of the form

$$f(\boldsymbol{x}) = \int_{\mathbb{R}^d} F(\boldsymbol{w}) K(\boldsymbol{w} \cdot \boldsymbol{x}) d\boldsymbol{w}, \quad \operatorname{supp} F \subset [-M, M]^d, \quad M \geq 1, \tag{1.2}$$

for some square-integrable function $F : [-M, M]^d \to \mathbb{C}$ and analytic kernel $K : \mathbb{R} \to \mathbb{C}$. This class of functions contains several examples of Reproducing Kernel Hilbert Spaces (RKHSs), including the space of bandlimited functions. The latter are ubiquitous in science and engineering. In information theory, bandlimited signals are often used for analysis and representation after sampling. In scientific computing, after discretization, functions are bandlimited by the Nyquist–Shannon sampling theorem. Studying the approximation power of ReLU networks for bandlimited functions is particularly important for neural network-based scientific computing in high dimensions. In Section 3, we shall show that for any measure μ such functions can be approximated to accuracy ϵ in the $L^2(B, \mu)$-norm by deep ReLU networks of depth $L = \mathcal{O}\left(\log_2^2 \frac{1}{\epsilon}\right)$ and size $W = \mathcal{O}\left(\frac{1}{\epsilon^2} \log_2^2 \frac{1}{\epsilon}\right)$.

We review some properties of deep ReLU networks in Section 2, providing new proofs of existing results (Propositions 2.2 and 2.3), as well as presenting new results (Propositions 2.4 and 2.5, Theorem 2.1). In Section 3, we recall an existing theorem (Theorem 3.1), before proving our main theorem (Theorem 3.2).

[1)] For a real-valued function f in $\mathbb{R}^d$ whose smoothness is characterized by some integer $m \geq 1$, and for some prescribed accuracy $\epsilon > 0$, one shows that there exists a shallow network $\widetilde{f}_W$ of size $W = W(d, m)$ that satisfies $\|f - \widetilde{f}_W\| \leq \epsilon$ for some norm $\|\cdot\|$.