Commun. Comput. Phys.
doi: 10.4208/cicp.OA-2019-0168

Vol. **27**, No. 2, pp. 379-411
February 2020

Better Approximations of High Dimensional Smooth Functions by Deep Neural Networks with Rectified Power Units

Bo Li[2,1,†], Shanshan Tang[3,†,‡] and Haijun Yu[1,2,*]

[1] *NCMIS & LSEC, Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.*
[2] *School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China.*
[3] *China Justice Big Data Institute, Beijing 100043, China.*

Received 30 September 2019; Accepted (in revised version) 24 October 2019

Abstract. Deep neural networks with rectified linear units (ReLU) are getting more and more popular due to their universal representation power and successful applications. Some theoretical progress regarding the approximation power of deep ReLU network for functions in Sobolev space and Korobov space have recently been made by [D. Yarotsky, Neural Network, 94:103-114, 2017] and [H. Montanelli and Q. Du, SIAM J Math. Data Sci., 1:78-92, 2019], etc. In this paper, we show that deep networks with rectified power units (RePU) can give better approximations for smooth functions than deep ReLU networks. Our analysis bases on classical polynomial approximation theory and some efficient algorithms proposed in this paper to convert polynomials into deep RePU networks of optimal size with no approximation error. Comparing to the results on ReLU networks, the sizes of RePU networks required to approximate functions in Sobolev space and Korobov space with an error tolerance ε, by our constructive proofs, are in general $\mathcal{O}(\log\frac{1}{\varepsilon})$ times smaller than the sizes of corresponding ReLU networks constructed in most of the existing literature. Comparing to the classical results of Mhaskar [Mhaskar, Adv. Comput. Math. 1:61-80, 1993], our constructions use less number of activation functions and numerically more stable, they can be served as good initials of deep RePU networks and further trained to break the limit of linear approximation theory. The functions represented by RePU networks are smooth functions, so they naturally fit in the places where derivatives are involved in the loss function.

[†] The first two authors contributed equally. Author list is alphabetical.

[‡] The work of this author is partially done during her Ph.D. study in Academy of Mathematics and Systems Science, Chinese Academy of Sciences.

[*] Corresponding author. *Email addresses:* libo1171309228@lsec.cc.ac.cn (B. Li), tangshanshan@lsec.cc.ac.cn (S. Tang), hyu@lsec.cc.ac.cn (H. Yu)

AMS subject classifications: 65D15, 65M12, 65M15

Key words: Deep neural network, high dimensional approximation, sparse grids, rectified linear unit, rectified power unit, rectified quadratic unit.

1 Introduction

Artificial neural network(ANN), whose origin may date back to the 1940s [1], is one of the most powerful tools in the field of machine learning. Especially, it became dominant in a lot of applications after the seminal works by Hinton et al. [2] and Bengio et al. [3] on efficient training of deep neural networks (DNNs), which pack up multi-layers of units with some nonlinear activation function. Since then, DNNs have greatly boosted the developments in different areas including image classification, speech recognition, computational chemistry and numerical solutions of high-dimensional partial differential equations and scientific problems, etc., see e.g. [4–12] to name a few.

The success of DNNs relies on two facts: 1) DNN is a powerful tool for general function approximation; 2) Efficient training methods are available to find minimizers with good generalization ability. In this paper, we focus on the first fact. It is known that artificial neural networks can approximate any C^0 and L^1 functions with any given error tolerance, using only one hidden layer (see e.g. [13,14]). However, it was realized recently that deep networks have better representation power(see e.g. [15–17]) than shallow networks. One of the commonly used activation functions with DNN is the so called rectified linear unit (ReLU) [18], which is defined as $\sigma(x)=\max(0,x)$. Telgarsky [16] gave a simple and elegant construction showing that for any k, there exist k-layer, $\mathcal{O}(1)$ wide ReLU networks on one-dimensional data, which can express a sawtooth function on $[0,1]$ with $\mathcal{O}(2^k)$ oscillations. Moreover, such a rapidly oscillating function cannot be approximated by poly(k)-wide ReLU networks with $o(k/\log(k))$ depth. Following this approach, several other works proved that deep ReLU networks have better approximation power than shallow ReLU networks [19–22]. In particular, for C^β-differentiable d-dimensional functions, Yarotsky [21] proved that the number of parameters needed to achieve an error tolerance of ε is $\mathcal{O}(\varepsilon^{-\frac{d}{\beta}}\log\frac{1}{\varepsilon})$. Petersen and Voigtlaender [22] proved that for a class of d-dimensional piecewise C^β continuous functions with the discontinuous interfaces being C^β continuous also, one can construct a ReLU neural network with $\mathcal{O}((1+\frac{\beta}{d})\log_2(2+\beta))$ layers, $\mathcal{O}(\varepsilon^{-\frac{2(d-1)}{\beta}})$ nonzero weights to achieve ε-approximation. The complexity bound is sharp. For analytic functions, E and Wang [23] proved that using ReLU networks with fixed width $d+4$, to achieve an error tolerance of ε, the depth of the network depends on $\log\frac{1}{\varepsilon}$ instead of ε itself. We also want to mention that the detailed relations between ReLU networks and linear finite elements have been studied by He et al. [24]. And recent work by Opschoor, Peterson and Schwab [25] reveals the connection between ReLU DNNs and high-order finite element methods.