Proceedings of Machine Learning Research vol 75:1–11, 2018 31st Annual Conference on Learning Theory

Optimal approximation of continuous functions by very deep ReLU networks

Dmitry Yarotsky D.YAROTSKY@SKOLTECH.RU
Skolkovo Institute of Science and Technology, Skolkovo Innovation Center, 3 Nobel st., Moscow 143026, Russia and *Institute for Information Transmission Problems, Bolshoy Karetny per. 19, build.1, Moscow 127051, Russia*

Editors: Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

We consider approximations of general continuous functions on finite-dimensional cubes by general deep ReLU neural networks and study the approximation rates with respect to the modulus of continuity of the function and the total number of weights W in the network. We establish the complete phase diagram of feasible approximation rates and show that it includes two distinct phases. One phase corresponds to slower approximations that can be achieved with constant-depth networks and continuous weight assignments. The other phase provides faster approximations at the cost of depths necessarily growing as a power law $L \sim W^\alpha, 0 < \alpha \leq 1$, and with necessarily discontinuous weight assignments. In particular, we prove that constant-width fully-connected networks of depth $L \sim W$ provide the fastest possible approximation rate $\|f - \widetilde{f}\|_\infty = O(\omega_f(O(W^{-2/\nu})))$ that cannot be achieved with less deep networks.[1]

1. Introduction

Expressiveness of deep neural networks with piecewise-linear (in particular, ReLU) activation functions has been a topic of much theoretical research in recent years. The topic has many aspects, with connections to combinatorics (Montufar et al., 2014; Telgarsky, 2016), topology (Bianchini and Scarselli, 2014), Vapnik-Chervonenkis dimension (Bartlett et al., 1998; Sakurai, 1999) and fat-shattering dimension (Kearns and Schapire, 1990; Anthony and Bartlett, 2009), hierarchical decompositions of functions (Mhaskar et al., 2016), information theory (Petersen and Voigtlaender, 2017), etc.

Here we adopt the perspective of classical approximation theory, in which the problem of expressiveness can be basically described as follows. Suppose that f is a multivariate function, say on the cube $[0, 1]^\nu$, and has some prescribed regularity properties; how efficiently can one approximate f by deep neural networks? The question has been studied in several recent publications. Depth-separation results for some explicit families of functions have been obtained in Safran and Shamir (2016); Telgarsky (2016). General upper and lower bounds on approximation rates for functions characterized by their degree of smoothness have been obtained in Liang and Srikant (2016); Yarotsky (2017). Hanin and Sellke (2017); Lu et al. (2017) establish the universal approximation property and convergence rates for deep and “narrow” (fixed-width) networks. Petersen and Voigtlaender (2017) establish con-

[1] Extended abstract. The full version with the proofs appears as [arXiv:1802.03620, v2]

vergence rates for approximations of discontinuous functions. Generalization capabilities of deep ReLU networks trained on finite noisy samples are studied in Schmidt-Hieber (2017).

In the present paper we consider and largely resolve the following question: what is the optimal rate of approximation of general continuous functions by deep ReLU networks, in terms of the number W of network weights and the modulus of continuity of the function? Specifically, for any W we seek a network architecture with W weights so that for any continuous $f : [0,1]^\nu \to \mathbb{R}$, as W increases, we would achieve the best convergence in the uniform norm $\|\cdot\|_\infty$ when using these architectures to approximate f.

In the slightly different but closely related context of approximation on balls in Sobolev spaces $\mathcal{W}^{d,\infty}([0,1]^\nu)$, this question of optimal convergence rate has been studied in Yarotsky (2017). That paper described ReLU network architectures with W weights ensuring approximation with error $O(W^{-d/\nu} \ln^{d/\nu} W)$ (Theorem 1). The construction was linear in the sense that the network weights depended on the approximated function linearly. Up to the logarithmic factor, this approximation rate matches the optimal rate over all parametric models under assumption of continuous parameter selection (DeVore et al. (1989)). It was also shown in Theorem 2 of Yarotsky (2017) that one can slightly (by a logarithmic factor) improve over this conditionally optimal rate by adjusting network architectures to the approximated function.

On the other hand, it was proved in Theorem 4 of Yarotsky (2017) that ReLU networks generally cannot provide approximation with accuracy better than $O(W^{-2d/\nu})$ – a bound with the power $\frac{2d}{\nu}$ twice as big as in the previously mentioned existence result. As was shown in the same theorem, this bound can be strengthened for shallow networks. However, without imposing depth constraints, there was a serious gap between the powers $\frac{2d}{\nu}$ and $\frac{d}{\nu}$ in the lower and upper accuracy bounds that was left open in that paper.

In the present paper we bridge this gap in the setting of continuous functions (which is slightly more general than the setting of the Sobolev space of Lipschitz functions, $\mathcal{W}^{1,\infty}([0,1]^\nu)$, i.e. the case $d = 1$). Our key insight is the close connection between approximation theory and VC dimension bounds. The lower bound on the approximation accuracy in Theorem 4 of Yarotsky (2017) was derived using the upper VCdim bound $O(W^2)$ from Goldberg and Jerrum (1995). More accurate upper and lower bounds involving the network depth L have been given in Bartlett et al. (1998); Sakurai (1999). The recent paper Bartlett et al. (2017) establishes nearly tight lower and upper VCdim bounds: $cWL\ln(W/L) \le \mathrm{VCdim}(W,L) \le CWL\ln W$, where $\mathrm{VCdim}(W,L)$ is the largest VC dimension of a piecewise linear network with W weights and L layers. The key element in the proof of the lower bound is the "bit extraction technique" (Bartlett et al. (1998)) providing a way to compress significant expressiveness in a single network weight. In the present paper we adapt this technique to the approximation theory setting.

Our main result is the complete phase diagram for the parameterized family of approximation rates involving the modulus of continuity ω_f of the function f and the number of weights W. We prove that using very deep networks one can approximate function f with error $O(\omega_f(O(W^{-2/\nu})))$, and this rate is optimal up to a logarithmic factor. In fact, the depth of the networks must necessarily grow almost linearly with W to achieve this rate, in sharp contrast to shallow networks that can provide approximation with error $O(\omega_f(O(W^{-1/\nu})))$. Moreover, whereas the slower rate $O(\omega_f(O(W^{-1/\nu})))$ can be achieved using a continuous weight assignment in the network, the optimal $O(\omega_f(O(W^{-2/\nu})))$ rate

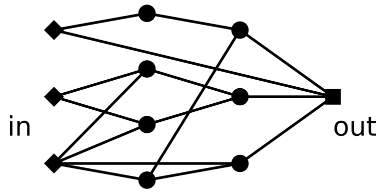

Figure 1: An example of a feed-forward network architecture of depth $L = 2$ with $W = 24$ weights.

necessarily requires a discontinuous weight assignment. All this allows us to regard these two kinds of approximations as being in different "phases". In addition, we explore the intermediate rates $O(\omega_f(O(W^{-p})))$ with $p \in (\frac{1}{\nu}, \frac{2}{\nu})$ and show that they are also in the discontinuous phase and require network depths $\sim W^{p\nu-1}$. We show that the optimal rate $O(\omega_f(O(W^{-2/\nu})))$ can be achieved with a deep constant-width fully-connected architecture, whereas the rates $O(\omega_f(O(W^{-p})))$ with $p \in (\frac{1}{\nu}, \frac{2}{\nu})$ and depth $O(W^{p\nu-1})$ can be achieved by stacking the deep constant-width architecture with a shallow parallel architecture. Apart from the bit extraction technique, we use the idea of the two-scales expansion from Theorem 2 in Yarotsky (2017) as an essential tool in the proofs of our results.

2. The results

We define the modulus of continuity ω_f of a function $f : [0,1]^\nu \to \mathbb{R}$ by

$$\omega_f(r) = \max\{|f(\mathbf{x}) - f(\mathbf{y})| : \mathbf{x}, \mathbf{y} \in [0,1]^\nu, |\mathbf{x} - \mathbf{y}| \le r\}, \tag{1}$$

where $|\mathbf{x}|$ is the euclidean norm of $\mathbf{x}$.

We approximate functions $f : [0,1]^\nu \to \mathbb{R}$ by usual feed-forward neural networks with the ReLU activation function $x \mapsto x_+ \equiv \max(0, x)$. The network has ν input units, some hidden units, and one output unit. The hidden units are assumed to be grouped in a sequence of layers so that the inputs of each unit is formed by outputs of some units from previous layers. The depth L of the network is the number of these hidden layers. A hidden unit computes a linear combination of its inputs followed by the activation function: $x_1, \ldots, x_s \mapsto (\sum_{k=1}^s w_k x_k + h)_+$, where w_k and h are the weights associated with this unit. The output unit acts similarly, but without the activation function: $x_1, \ldots, x_s \mapsto \sum_{k=1}^s w_k x_k + h$.

The network is determined by its architecture and weights. Clearly, the total number of weights, denoted by W, is equal to the total number of connections and computation units (not counting the input units). We don't impose any constraints on the network architecture (see Fig. 1 for an example of a valid architecture).

Throughout the paper, we consider the input dimension ν as fixed. Accordingly, by *constants* we will generally mean values that may depend on ν.

We are interested in relating the approximation errors to the complexity of the function f, measured by its modulus of continuity ω_f, and to the complexity of the approximating

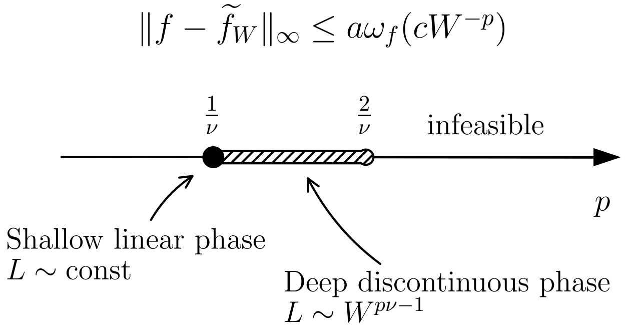

Figure 2: The phase diagram of convergence rates. At $p=\frac{1}{\nu}$ the rate is achieved by shallow networks with weights linearly (and continuously) depending on f. At $p\in(\frac{1}{\nu},\frac{2}{\nu}]$, the rate is achieved by deep networks with weights discontinuously depending on f. Rates with $p>\frac{2}{\nu}$ are infeasible.

network, measured by its total number of weights W. More precisely, we consider approximation rates in terms of the following procedure.

First, suppose that for each W we choose in some way a network architecture η_W with ν inputs and W weights. Then, for any $f:[0,1]^\nu\to\mathbb{R}$ we construct an approximation $\widetilde{f}_W:[0,1]^\nu\to\mathbb{R}$ to f by choosing in some way the values of the weights in the architecture η_W – in the sequel, we refer to this stage as the *weight assignment*. The question we ask is this: for which powers $p\in\mathbb{R}$ can we ensure, by the choice of the architecture and then the weights, that

$$\|f-\widetilde{f}_W\|_\infty \le a\omega_f(cW^{-p}),\quad \forall f\in C([0,1]^\nu),\tag{2}$$

with some constants a,c possibly depending on ν and p but not on W or f?

Clearly, if inequality (2) holds for some p, then it also holds for any smaller p. However, we expect that for smaller p the inequality can be in some sense easier to satisfy. In this paper we show that there is in fact a *qualitative* difference between different regions of p's.

Our findings are best summarized by the phase diagram shown in Fig. 2. We give an informal overview of the diagram before moving on to precise statements. The region of generally feasible rates is $p\le\frac{2}{\nu}$. This region includes two qualitatively distinct phases corresponding to $p=\frac{1}{\nu}$ and $p\in(\frac{1}{\nu},\frac{2}{\nu}]$. At $p=\frac{1}{\nu}$, the rate (2) can be achieved by fixed-depth networks whose weights depend linearly on the approximated function f. In contrast, at $p\in(\frac{1}{\nu},\frac{2}{\nu}]$ the rate can only be achieved by networks with growing depths $L\sim W^{p\nu-1}$ and whose weights depend *discontinuously* on the approximated function. In particular, at the rightmost feasible point $p=\frac{2}{\nu}$ the approximating architectures have $L\sim W$ and are thus necessarily extremely deep and narrow.

We now turn to precise statements. First we characterize the $p=\frac{1}{\nu}$ phase in which the approximation can be obtained using a standard piecewise linear interpolation. In the sequel, when writing $f=O(g)$ we mean that $|f|\le cg$ with some constant c that may depend on ν. For brevity, we will write $\widetilde{f}$ without the subscript W.

Proposition 1 *There exist network architectures η_W with W weights and, for each W, a weight assignment linear in f such that Eq. (2) is satisfied with $p=\frac{1}{\nu}$. The network*

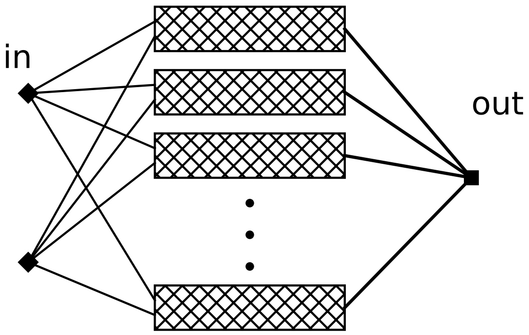

Figure 3: The parallel, constant-depth network architecture implementing piecewise linear interpolation and ensuring approximation rate (2) with $p = \frac{1}{\nu}$.

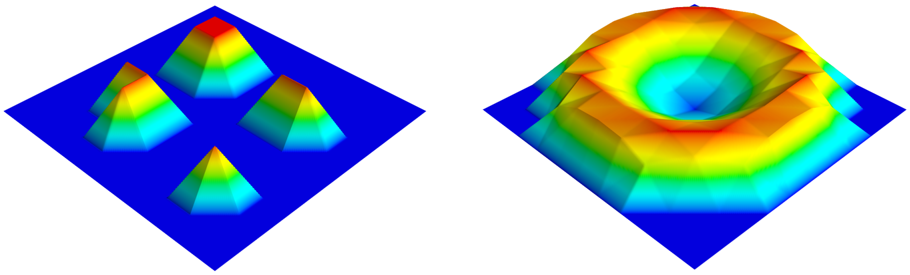

Figure 4: Approximation by linear combinations of "spike" functions in dimension $\nu = 2$. Left: a spike function and examples of sums of neighboring spike functions. Right: approximation of a radial function by a linear combination of spike functions.

architectures can be chosen as consisting of $O(W)$ parallel blocks each having the same architecture that only depends on ν (see Fig. 3). In particular, the depths of the networks depend on ν but not on W.

We explain now the idea of the proof. The approximating function $\widetilde{f}$ is constructed as a linear combination of "spike" functions sitting at the knots of the regular grid in $[0,1]^\nu$, with coefficients given by the values of f at these knots (see Fig. 4). For a grid of spacing $\frac{1}{N}$ with an appropriate N, the number of knots is $\sim N^\nu$ while the approximation error is $O(\omega_f(O(\frac{1}{N})))$. We implement each spike by a block in the network, and implement the whole approximation by summing blocks connected in parallel and weighted. Then the whole network has $O(N^\nu)$ weights and, by expressing N as $\sim W^{1/\nu}$, the approximation error is $O(\omega_f(O(W^{-1/\nu}))$, i.e. we obtain the rate (2) with $p = \frac{1}{\nu}$.

We note that the weights of the resulting network either do not depend on f at all or are given by $w = f(\mathbf{x})$ with some $\mathbf{x} \in [0,1]^\nu$. In particular, the weight assignment is continuous in f with respect to the standard topology of $C([0,1]^\nu)$.

We turn now to the region $p > \frac{1}{\nu}$. Several properties of this region are either direct consequences or slight modifications of existing results, and it is convenient to combine them in a single theorem.

Theorem 1

a) (Feasibility) Approximation rate (2) cannot be achieved with $p > \frac{2}{\nu}$.

b) (Inherent discontinuity) Approximation rate (2) cannot be achieved with $p > \frac{1}{\nu}$ if the weights of $\widetilde{f}$ are required to depend on f continuously with respect to the standard topology of $C([0,1]^\nu)$.

c) (Inherent depth) If approximation rate (2) is achieved with a particular $p \in (\frac{1}{\nu}, \frac{2}{\nu}]$, then the architectures η_W must have depths $L \geq dW^{p\nu-1}/\ln W$ with some possibly ν- and p-dependent constant $d > 0$.

Proof The proofs of these statements have the common element of considering the approximation for functions from the unit ball $F_{\nu,1}$ in the Sobolev space $\mathcal{W}^{1,\infty}([0,1]^\nu)$ of Lipschitz functions. Namely, suppose that the approximation rate (2) holds with some p. Then all $f \in F_{\nu,1}$ can be approximated by architectures η_W with accuracy

$$\epsilon_W = c_1 W^{-p} \tag{3}$$

with some constant c_1 independent of W. The three statements of the theorem are then obtained as follows.

a) This statement is a consequence of Theorem 4a) of Yarotsky (2017), which is in turn a consequence of the upper bound $O(W^2)$ for the VC dimension of a ReLU network (Goldberg and Jerrum (1995)). Precisely, Theorem 4a) implies that if an architecture η_W allows to approximate all $f \in F_{\nu,1}$ with accuracy ϵ_W, then $W \geq c_2 \epsilon_W^{-\nu/2}$ with some c_2. Comparing this with Eq. (3), we get $p \leq \frac{2}{\nu}$.

b) This statement is a consequence of the general bound of DeVore et al. (1989) on the efficiency of approximation of Sobolev balls with parametric models having parameters continuously depending on the approximated function. Namely, if the weights of the networks η_W depend on $f \in F_{\nu,1}$ continuously, then Theorem 4.2 of DeVore et al. (1989) implies that $\epsilon_W \geq c_2 W^{-1/\nu}$ with some constant c_2, which implies that $p \leq \frac{1}{\nu}$.

c) This statement can be obtained by combining arguments of Theorem 4 of Yarotsky (2017) with the recently established tight upper bound for the VC dimension of ReLU networks (Bartlett et al. (2017), Theorem 6) with given depth L and the number of weights W:

$$\mathrm{VCdim}(W, L) \leq CWL \ln W, \tag{4}$$

where C is a global constant.

Specifically, suppose that an architecture η_W allows to approximate all $f \in F_{\nu,1}$ with accuracy ϵ_W. Then, by considering suitable trial functions, one shows that if we threshold the network output, the resulting networks must have VC dimension $\mathrm{VCdim}(\eta_W) \geq c_2 \epsilon_W^{-\nu}$ (see Eq.(38) in Yarotsky (2017)). Hence, by Eq. (3), $\mathrm{VCdim}(\eta_W) \geq c_3 W^{p\nu}$. On the other hand, the upper bound (4) implies $\mathrm{VCdim}(\eta_W) \leq CWL \ln W$. We conclude that $c_3 W^{p\nu} \leq CWL \ln W$, i.e. $L \geq dW^{p\nu-1}/\ln W$ with some constant d. ■

Theorem 1 suggests the existence of an approximation phase drastically different from the phase $p = \frac{1}{\nu}$. This new phase would provide better approximation rates, up to $p = \frac{2}{\nu}$, at the cost of deeper networks and some complex, discontinuous weight assignment. The main contribution of the present paper is the proof that this phase indeed exists.

We describe some architectures that, as we will show, correspond to this phase. First we describe the architecture for $p = \frac{2}{\nu}$, i.e. for the fastest possible approximation rate. Consider the usual fully-connected architecture connecting neighboring layers and having a constant number of neurons in each layer, see Fig. 5. We refer to this constant number of neurons as the "width" H of the network. Such a network of width H and depth L has $W = L(H^2+H)+H^2+(\nu+1)H+1$ weights in total. We will be interested in the scenario of "narrow" networks where H is fixed and the network grows by increasing L; then W grows linearly with L. Below we will refer to the "narrow fully-connected architecture of width H having W weights": the depth L is supposed in this case to be determined from the above equality; we will assume without loss of generality that the equality is satisfied with an integer L. We will show that these narrow architectures provide the $p = \frac{2}{\nu}$ approximation rate if the width H is large enough (say, $H = 2\nu + 10$).

In the case $p \in (\frac{1}{\nu}, \frac{2}{\nu})$ we consider another kind of architectures obtained by stacking parallel shallow architectures (akin to those of Proposition 1) with the above narrow fully-connected architectures, see Fig. 6. The first, parallelized part of these architectures consists of blocks that only depend on ν (but not on W or p). The second, narrow fully-connected part will again have a fixed width, and we will take its depth to be $\sim W^{p\nu-1}$. All the remaining weights then go into the first parallel subnetwork, which in particular determines the number of blocks in it. Since the blocks are parallel and their architectures do not depend on W, the overall depth of the network is determined by the second, deep subnetwork and is $O(W^{p\nu-1})$. On the other hand, in terms of the number of weights, for $p < \frac{2}{\nu}$ most computation is performed by the first, parallel subnetwork (the deep subnetwork has $O(W^{p\nu-1})$ weights while the parallel one has an asymptotically larger number of weights, $W - O(W^{p\nu-1})$).

Clearly, these stacked architectures can be said to "interpolate" between the purely parallel architectures for $p = \frac{1}{\nu}$ and the purely serial architectures for $p = \frac{2}{\nu}$. Note that a parallel computation can be converted into a serial one at the cost of increasing the depth of the network. For $p < \frac{2}{\nu}$, rearranging the parallel subnetwork of the stacked architecture into a serial one would destroy the $O(W^{p\nu-1})$ bound on the depth of the full network, since the parallel subnetwork has $\sim W$ weights. However, for $p = \frac{2}{\nu}$ this rearrangement does not affect the $L \sim W$ asymptotic of the depth more than by a constant factor – that's why we don't include the parallel subnetwork into the full network in this case.

We state now our main result as the following theorem.

Theorem 2

a) For any $p \in (\frac{1}{\nu}, \frac{2}{\nu}]$, there exist a sequence of architectures η_W with depths $L = O(W^{p\nu-1})$ and respective weight assignments such that inequality (2) holds with this p.

b) For $p = \frac{2}{\nu}$, an example of such architectures is the narrow fully-connected architectures of constant width $2\nu + 10$.

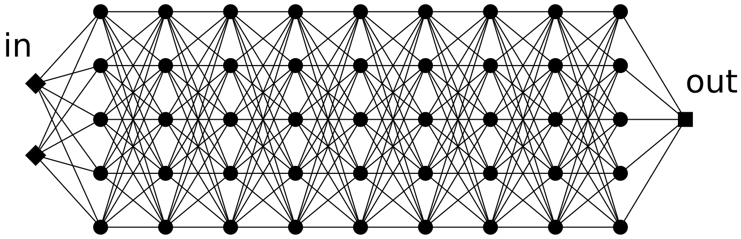

Figure 5: An example of "narrow" fully-connected network architecture having $\nu = 2$ inputs, depth $L = 9$ and width $H = 5$. These architectures provide the optimal approximation rate (2) with $p = \frac{2}{\nu}$ if H is sufficiently large and held constant while L is increased.

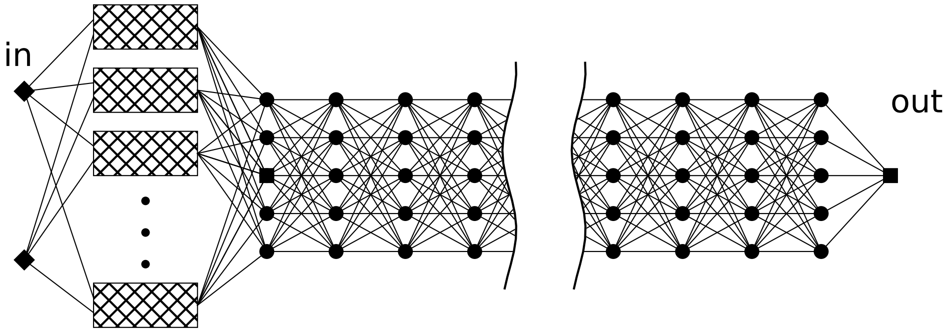

Figure 6: The "stacked" architectures for $p \in (\frac{1}{\nu}, \frac{2}{\nu})$, providing the optimal approximation rates (2) under the depth constraint $L = O(W^{p\nu-1})$.

c) For $p \in (\frac{1}{\nu}, \frac{2}{\nu})$, an example of such architectures are stacked architectures described above, with the narrow fully-connected subnetwork having width $3^\nu(2\nu+10)$ and depth $W^{p\nu-1}$.

Comparing this theorem with Theorem 1a) we see that the narrow fully-connected architectures provide the best possible approximation in the sense of Eq. (2). Moreover, for $p \in (\frac{1}{\nu}, \frac{2}{\nu})$ the upper bound on the network depth in Theorem 2c) matches the lower bound in Theorem 1c) up to a logarithmic factor. This proves that for $p \in (\frac{1}{\nu}, \frac{2}{\nu})$ our stacked architectures are also optimal (up to a logarithmic correction) if we additionally impose the asymptotic constraint $L = O(W^{p\nu-1})$ on the network depth.

We explain now the idea of the proof. Given a function f and some W, we first proceed as in Proposition 1 and construct its piecewise linear interpolation $\widetilde{f}_1$ on the length scale $\frac{1}{N}$ with $N \sim W^{1/\nu}$. This approximation has uniform error $O(\omega_f(O(W^{-1/\nu})))$. Then, we improve this approximation by constructing an additional approximation $\widetilde{f}_2$ for the discrepancy $f - \widetilde{f}_1$. This second approximation lives on a smaller length scale $\frac{1}{M}$ with $M \sim W^p$. In contrast to $\widetilde{f}_1$, the second approximation is inherently discrete: we consider a finite set of possible shapes of $f - \widetilde{f}_1$ in patches of linear size $\sim \frac{1}{N}$, and in each patch we use a special single network weight to encode the shape closest to $f - \widetilde{f}_1$. The second approximation is then fully determined by the collection of these special encoding weights found for all patches. We make the parallel subnetwork of the full network serve two purposes: in addition to computing the initial approximation $\widetilde{f}_1(\mathbf{x})$ as in Proposition 1, the subnetwork returns the position of $\mathbf{x}$ within its patch along with the weight that encodes the second approximation $\widetilde{f}_2$ within this patch. The remaining, deep narrow part of the network then serves to decode the second approximation within this patch from the special weight and compute the value $\widetilde{f}_2(\mathbf{x})$. Since the second approximation lives on the smaller length scale $\frac{1}{M}$, there are $Z = \exp(O((M/N)^\nu))$ possible approximations $\widetilde{f}_2$ within the patch that might need to be encoded in the special weight. It then takes a narrow network of depth $L \sim \ln Z$ to reconstruct the approximation from the special weight using the bit extraction technique of Bartlett et al. (1998). As $M \sim W^p$, we get $L \sim W^{p\nu-1}$. At the same time, the second approximation allows us to effectively improve the overall approximation scale from $\sim \frac{1}{N}$ down to $\sim \frac{1}{M}$, i.e. to $\sim W^{-p}$, while keeping the total number of weights in the network. This gives us the desired error bound $O(\omega_f(O(W^{-p})))$.

We remark that the discontinuity of the weight assignment in our construction is the result of the discreteness of the second approximation $\widetilde{f}_2$: whereas the variable weights in the network implementing the first approximation $\widetilde{f}_1$ are found by linearly projecting the approximated function to $\mathbb{R}$ (namely, by computing $f \mapsto f(\mathbf{x})$ at the knots $\mathbf{x}$), the variable weights for $\widetilde{f}_2$ are found by assigning to f one of the finitely many values encoding the possible approximate shapes in a patch. This operation is obviously discontinuous. While the discontinuity is present for all $p > \frac{1}{\nu}$, at smaller p it is “milder” in the sense of a smaller number of assignable values.

3. Discussion

We discuss now our result in the context of general approximation theory and practical machine learning. First, a theorem of Kainen et al. (1999) shows that in the optimal approximations by neural networks the weights generally discontinuously depend on the approximated function, so the discontinuity property that we have established is not surprizing. However, this theorem of Kainen et al. (1999) does not in any way quantify the accuracy gain that can be acquired by giving up the continuity of the weights. Our result does this in the practically important case of deep ReLU networks, and explicitly describes a relevant mechanism.

In general, many nonlinear approximation schemes involve some form of discontinuity, often explicit (e.g., using different expansion bases for different approximated functions (DeVore (1998)). At the same time, discontinuous selection of parameters in parametric models is often perceived as an undesirable phenomenon associated with unreliable approximation (DeVore et al. (1989); DeVore (1998)). We point out, however, that deep architectures considered in the present paper resemble some popular state-of-the-art practical networks for highly accurate image recognition – residual networks (He et al., 2016) and highway networks (Srivastava et al., 2015) that may have dozens or even hundreds of layers. While our model does not explicitly include shortcut connections as in ResNets, a very similar element is effectively present in the proof of Theorem 2 (in the form of channels reserved for passing forward the data). We expect, therefore, that our result may help better understand the properties of ResNet-like networks.

Quantized network weights have been previously considered from the information-theoretic point of view in Bölcskei et al. (2017); Petersen and Voigtlaender (2017). In the present paper we do not use quantized weights in the statement of the approximation problem, but they appear in the solution (namely, we use them to store small-scale descriptions of the approximated function). One can expect that weight quantization may play an important role in the future development of the theory of deep networks.

Acknowledgments

The author thanks Alexandr Kuleshov and the anonymous referees for helpful comments and suggestions. The research was supported by the Skoltech SBI Bazykin/Yarotsky project.

References

Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. Cambridge university press, 2009.

Peter L Bartlett, Vitaly Maiorov, and Ron Meir. Almost linear VC-dimension bounds for piecewise polynomial networks. *Neural computation*, 10(8):2159–2173, 1998.

Peter L Bartlett, Nick Harvey, Chris Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *arXiv preprint arXiv:1703.02930*, 2017.

Monica Bianchini and Franco Scarselli. On the complexity of neural network classifiers: A comparison between shallow and deep architectures. *IEEE transactions on neural networks and learning systems*, 25(8):1553–1565, 2014.

Helmut Bölcskei, Philipp Grohs, Gitta Kutyniok, and Philipp Petersen. Memory-optimal neural network approximation. In *Wavelets and Sparsity XVII*, volume 10394, page 103940Q. International Society for Optics and Photonics, 2017.

Ronald A DeVore. Nonlinear approximation. *Acta numerica*, 7:51–150, 1998.

Ronald A DeVore, Ralph Howard, and Charles Micchelli. Optimal nonlinear approximation. *Manuscripta mathematica*, 63(4):469–478, 1989.

Paul W Goldberg and Mark R Jerrum. Bounding the Vapnik-Chervonenkis dimension of concept classes parameterized by real numbers. *Machine Learning*, 18(2-3):131–148, 1995.

Boris Hanin and Mark Sellke. Approximating Continuous Functions by ReLU Nets of Minimal Width. *arXiv preprint arXiv:1710.11278*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Paul C Kainen, Věra Kůrková, and Andrew Vogt. Approximation by neural networks is not continuous. *Neurocomputing*, 29(1):47–56, 1999.

Michael J Kearns and Robert E Schapire. Efficient distribution-free learning of probabilistic concepts. In *Foundations of Computer Science, 1990. Proceedings., 31st Annual Symposium on*, pages 382–391. IEEE, 1990.

Shiyu Liang and R. Srikant. Why deep neural networks? *arXiv preprint arXiv:1610.04161*, 2016.

Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The expressive power of neural networks: A view from the width. In *Advances in Neural Information Processing Systems*, pages 6232–6240, 2017.

Hrushikesh Mhaskar, Qianli Liao, and Tomaso Poggio. Learning real and boolean functions: When is deep better than shallow. *arXiv preprint arXiv:1603.00988*, 2016.

Guido F Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Advances in neural information processing systems*, pages 2924–2932, 2014.

Philipp Petersen and Felix Voigtlaender. Optimal approximation of piecewise smooth functions using deep relu neural networks. *arXiv preprint arXiv:1709.05289*, 2017.

Itay Safran and Ohad Shamir. Depth separation in relu networks for approximating smooth non-linear functions. *arXiv preprint arXiv:1610.09887*, 2016.

Akito Sakurai. Tight Bounds for the VC-Dimension of Piecewise Polynomial Networks. In *Advances in Neural Information Processing Systems*, pages 323–329, 1999.

J. Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *arXiv preprint arXiv:1708.06633*, 2017.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015.

Matus Telgarsky. Benefits of depth in neural networks. *arXiv preprint arXiv:1602.04485*, 2016.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017.