

Deep Network Approximation: Achieving Arbitrary Accuracy with Fixed Number of Neurons*

Zuowei Shen[†] Haizhao Yang[‡] Shijun Zhang[§]

Abstract

This paper develops simple feed-forward neural networks that achieve the universal approximation property for all continuous functions with a fixed finite number of neurons. These neural networks are simple because they are designed with a simple and computable continuous activation function σ leveraging a triangular-wave function and a softsign function. We prove that σ -activated networks with width $36d(2d+1)$ and depth 11 can approximate any continuous function on a d -dimensional hypercube within an arbitrarily small error. Hence, for supervised learning and its related regression problems, the hypothesis space generated by these networks with a size not smaller than $36d(2d+1) \times 11$ is dense in the space of continuous functions. Furthermore, classification functions arising from image and signal classification are in the hypothesis space generated by σ -activated networks with width $36d(2d+1)$ and depth 12, when there exist pairwise disjoint closed bounded subsets of \mathbb{R}^d such that the samples of the same class are located in the same subset.

Key words. Universal Approximation Theorem; Fixed-Size Neural Network; Periodic Function; Continuous Function; Classification Function.

1 Introduction

Deep neural networks have been widely used in data science and artificial intelligence. Their tremendous successes in various applications have motivated extensive research to establish the theoretical foundation of deep learning. Understanding the approximation capacity of deep neural networks is one of the keys to reveal the power of deep learning. The most basic layers of deep neural networks are nonlinear functions as the composition of an affine linear transform and a nonlinear activation function. The composition of these simple nonlinear functions can generate a complicated deep neural network with powerful approximation capacity, which is the key difference to classic approximation tools. In this paper, we show that the hypothesis space of deep neural networks generated from the composition of 11 such simple nonlinear functions is

*Submitted to the editors DATE.

[†]Department of Mathematics, National University of Singapore (matzuows@nus.edu.sg).

[‡]Department of Mathematics, Purdue University (haizhao@purdue.edu).

[§]Department of Mathematics, National University of Singapore (zhangshijun@u.nus.edu).

32 dense in the continuous function space $C([a, b]^d)$, when the affine linear transforms are
 33 parameterized with $\mathcal{O}(d^2)$ parameters in total and the nonlinear activation function is
 34 constructed from a simple triangular-wave function and a softsign function.

35 1.1 Main results

36 One of the key elements of a neural network is its activation functions. Searching
 37 for simple activation functions enabling powerful approximation capacity of neural net-
 38 works is an important mathematical problem probably originated in the Kolmogorov
 39 superposition theorem (KST) [20] for Hilbert’s 13-th problem, where a two-hidden-layer
 40 neural network with $\mathcal{O}(d)$ neurons and complicated activation functions depending on
 41 the target functions are constructed to represent an arbitrary function in $C([0, 1]^d)$.
 42 Since then, whether simple and computable activation functions independent of the tar-
 43 get function exist to make the space of neural networks with $\mathcal{O}(d)$ neurons dense in
 44 $C([0, 1]^d)$ or even equal to $C([0, 1]^d)$ has been an open problem. A function $\varrho : \mathbb{R} \rightarrow \mathbb{R}$
 45 is said to be a universal activation function (UAF) if the function space generated by
 46 ϱ -activated networks with $C_{\varrho, d}$ neurons is dense in $C([0, 1]^d)$, where $C_{\varrho, d}$ is a constant
 47 determined by ϱ and d . That is, if ϱ is a UAF, then ϱ -activated networks with $C_{\varrho, d}$
 48 neurons can approximate any continuous function within an arbitrary error on $[0, 1]^d$ by
 49 only adjusting the parameters.

50 In this paper, we first construct a simple and computable example of UAFs. As
 51 a typical and simple UAF, this activation function is called the elementary universal
 52 activation function (EUAF), and the corresponding networks are called EUAF networks.
 53 Then, we prove that the function space generated by EUAF networks with $\mathcal{O}(d^2)$ neurons
 54 is dense in $C([a, b]^d)$. Furthermore, it is shown that EUAF networks with $\mathcal{O}(d^2)$ neurons
 55 can exactly represent d -dimensional classification functions.

56 While a good activation function should be simple and numerically implementable,
 57 the neural network activated by it should be able to approximate continuous functions well
 58 with a manageable size. Considering these requirements and motivated by previous works
 59 [32, 33, 40], the activation function to be chosen should have appropriate nonlinearity,
 60 periodicity, and the capacity to reproduce step functions. It is challenging to find a single
 61 activation function with all these properties. Here, we propose an activation function
 62 with all required properties by using two simple functions σ_1 and σ_2 defined below.

63 Let σ_1 be the continuous triangular-wave function with period 2, i.e.,

$$64 \quad \sigma_1(x) := |x| \quad \text{for any } x \in [-1, 1] \quad (1.1)$$

65 and $\sigma_1(x + 2) = \sigma_1(x)$ for any $x \in \mathbb{R}$. Alternatively, σ_1 can also be written as:

$$66 \quad \sigma_1(x) = \left| x - 2 \left\lfloor \frac{x+1}{2} \right\rfloor \right| \quad \text{for any } x \in \mathbb{R}, \quad \text{where } \lfloor \cdot \rfloor \text{ is the floor function.}$$

67 Clearly, σ_1 is periodic and $x - \sigma_1(x)$ is a continuous variant of the floor function as
 68 desired.

69 To introduce high nonlinearity, let σ_2 be the softsign activation function commonly
 70 used in machine learning [21, 36]:

$$71 \quad \sigma_2(x) := \frac{x}{|x| + 1} \quad \text{for any } x \in \mathbb{R}. \quad (1.2)$$

72 Then the activation function σ is defined as:

$$73 \quad \sigma(x) := \begin{cases} \sigma_1(x) & \text{for } x \in [0, \infty), \\ \sigma_2(x) & \text{for } x \in (-\infty, 0). \end{cases} \quad (1.3)$$

74 See an illustration of σ in Figure 1. This activation function σ is the EUAF used to
75 construct powerful neural networks in this paper.

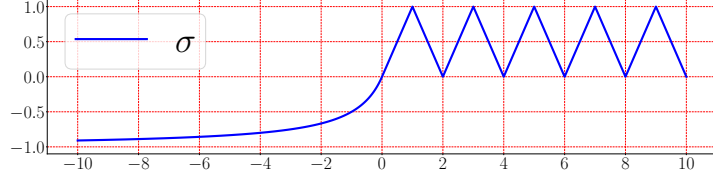


Figure 1: An illustration of σ on $[-10, 10]$.

76 The periodicity of the triangular-wave function σ_1 and the nonlinearity of the soft-
77 sign function σ_2 play crucial roles in the proof of our main results. Observing that σ_1 is
78 an even function and σ_2 is an odd function, i.e., $\sigma(x) = \sigma_1(x) = \sigma_1(-x)$ for any $x \geq 0$ and
79 $-\sigma(-x) = -\sigma_2(-x) = \sigma_2(x)$ for any $x \geq 0$. This implies that $\sigma(x)$ and $-\sigma(-x)$ with $x \geq 0$
80 have both required periodicity and nonlinearity features and play the same roles as $\sigma_1(x)$
81 and $\sigma_2(x)$, respectively. These requirements lead to our choice of σ as the activation
82 function. If allowed to be more complicated, one can design many other UAFs satisfy-
83 ing stronger requirements for various applications. For example, the idea of designing a
84 C^s UAF is given in Section 5.1 and a sigmoidal UAF (see Figure 14) is constructed in
85 Section 5.2.

86 With the activation function σ in hand, let us introduce the network (architecture)
87 using σ as the activation function, called σ -activated network (architecture). To be
88 precise, a σ -activated network with a (vector) input $\mathbf{x} \in \mathbb{R}^d$, an output $\Phi(\mathbf{x}, \boldsymbol{\theta}) \in \mathbb{R}$, and
89 $L \in \mathbb{N}^+$ hidden layers can be briefly described as follows:

$$90 \quad \mathbf{x} = \tilde{\mathbf{h}}_0 \xrightarrow[\mathcal{L}_0]{\mathbf{A}_0, \mathbf{b}_0} \mathbf{h}_1 \xrightarrow{\sigma} \tilde{\mathbf{h}}_1 \cdots \xrightarrow[\mathcal{L}_{L-1}]{\mathbf{A}_{L-1}, \mathbf{b}_{L-1}} \mathbf{h}_L \xrightarrow{\sigma} \tilde{\mathbf{h}}_L \xrightarrow[\mathcal{L}_L]{\mathbf{A}_L, \mathbf{b}_L} \mathbf{h}_{L+1} = \Phi(\mathbf{x}, \boldsymbol{\theta}), \quad (1.4)$$

91 where $N_0 = d \in \mathbb{N}^+$, $N_1, N_2, \dots, N_L \in \mathbb{N}^+$, $N_{L+1} = 1$, $\mathbf{A}_i \in \mathbb{R}^{N_{i+1} \times N_i}$ and $\mathbf{b}_i \in \mathbb{R}^{N_{i+1}}$ are the
92 weight matrix and the bias vector in the i -th affine linear transform \mathcal{L}_i , respectively, i.e.,

$$93 \quad \mathbf{h}_{i+1} = \mathbf{A}_i \cdot \tilde{\mathbf{h}}_i + \mathbf{b}_i =: \mathcal{L}_i(\tilde{\mathbf{h}}_i) \quad \text{for } i = 0, 1, \dots, L$$

94 and

$$95 \quad \tilde{h}_{i,j} = \sigma(h_{i,j}) \quad \text{for } j = 1, 2, \dots, N_i \text{ and } i = 1, 2, \dots, L.$$

96 $\boldsymbol{\theta}$ is a fattened vector consisting of all parameters in $\mathbf{A}_0, \mathbf{b}_0, \dots, \mathbf{A}_L, \mathbf{b}_L$. $\tilde{h}_{i,j}$ and $h_{i,j}$ are
97 the j -th entry of $\tilde{\mathbf{h}}_i$ and \mathbf{h}_i , respectively, for $j = 1, 2, \dots, N_i$ and $i = 1, 2, \dots, L$. If σ is
98 applied to a vector entrywisely, i.e., given any $k \in \mathbb{N}^+$,

$$99 \quad \sigma(\mathbf{y}) = [\sigma(y_1), \dots, \sigma(y_k)]^T \quad \text{for any } \mathbf{y} = [y_1, \dots, y_k]^T \in \mathbb{R}^k.$$

100 then Φ can be represented in a form of function compositions as follows:

$$101 \quad \Phi(\mathbf{x}, \boldsymbol{\theta}) = \mathcal{L}_L \circ \sigma \circ \mathcal{L}_{L-1} \circ \sigma \circ \cdots \circ \sigma \circ \mathcal{L}_1 \circ \sigma \circ \mathcal{L}_0(\mathbf{x}) \quad \text{for any } \mathbf{x} \in \mathbb{R}^d.$$

Given $N, L \in \mathbb{N}^+$, let $\Phi_{N,L}(\mathbf{x}, \boldsymbol{\theta})$ denote the σ -activated network architecture $\Phi(\mathbf{x}, \boldsymbol{\theta})$ in Equation (1.4) with $N_1 = N_2 = \dots = N_L = N$. Let

$$W = W_{d,N,L} = d \times N + N + (N \times N + N) \times (L - 1) + 1 \times N + 1 = \mathcal{O}(dN + N^2L)$$

be the total number of parameters in $\Phi_{N,L}(\mathbf{x}, \boldsymbol{\theta})$, i.e., $\boldsymbol{\theta} \in \mathbb{R}^W$.

Define the hypothesis space $\mathcal{H}_d(N, L)$ as the function space generated by EUAF networks with width N and depth L , i.e.,

$$\mathcal{H}_d(N, L) := \left\{ \phi : \phi(\mathbf{x}) = \Phi_{N,L}(\mathbf{x}, \boldsymbol{\theta}) \text{ for any } \mathbf{x} \in \mathbb{R}^d, \quad \boldsymbol{\theta} \in \mathbb{R}^W \right\}. \quad (1.5)$$

Let $C([a, b]^d)$ be the space of all continuous functions $f : [a, b]^d \rightarrow \mathbb{R}$ with the maximum norm. Our first main result, Theorem 1.1 below, shows that σ -activated networks with a fixed size $\mathcal{O}(d^2)$ enjoy the universal approximation property by only adjusting their parameters.

Theorem 1.1. *Let $f \in C([a, b]^d)$ be a continuous function and $\mathcal{H}_d(N, L)$ be the hypothesis space defined in (1.5) with $N = 36d(2d + 1)$ and $L = 11$. Then, for an arbitrary $\varepsilon > 0$, there exists $\phi \in \mathcal{H}_d(N, L)$ such that*

$$|\phi(\mathbf{x}) - f(\mathbf{x})| < \varepsilon \quad \text{for any } \mathbf{x} \in [a, b]^d.$$

Remark. The network realizing ϕ in Theorem 1.1 has

$$d \times N + N + (N \times N + N) \times (L - 1) + N \times 1 + 1 \sim d^4$$

parameters, where $N = 36d(2d + 1)$ and $L = 11$. However, as shown in our constructive proof of Theorem 1.1, it is enough to adjust $5437(d + 1)(2d + 1) = \mathcal{O}(d^2) \ll d^4$ parameters and set all the others to 0.

Since for an arbitrary $M > 0$, $2M\sigma(\frac{x+M}{2M}) - M = x$ for all $x \in [-M, M]$, we can manually add hidden layers to EUAF networks without changing the output. This leads to the following immediate corollary of Theorem 1.1.

Corollary 1.2. *Assume $N \geq 36d(2d + 1)$ and $L \geq 11$, then the hypothesis space $\mathcal{H}_d(N, L)$ defined in (1.5) is dense in $C([a, b]^d)$.*

One can ask whether the arbitrary error $\varepsilon > 0$ in Theorem 1.1 can be further reduced to 0. This is not true in general, but it is true for a class of interesting functions widely used in image classifications. Given any pairwise disjoint closed bounded subsets $E_1, E_2, \dots, E_J \subseteq \mathbb{R}^d$, define “the classification function space” of these subsets as

$$\mathcal{C}_d(E_1, E_2, \dots, E_J) := \left\{ f : f = \sum_{j=1}^J r_j \cdot \mathbf{1}_{E_j} \text{ for any } r_1, r_2, \dots, r_J \in \mathbb{Q} \right\},$$

where $\mathbf{1}_{E_n}$ is the indicator function of E_j for each j . Our second main result, Theorem 1.3 below, shows that each element of $\mathcal{C}_d(E_1, E_2, \dots, E_J)$ can be exactly represented by a σ -activated network with $\mathcal{O}(d^2)$ neurons in $\bigcup_{j=1}^J E_j$.

Theorem 1.3. Let $E_1, E_2, \dots, E_J \subseteq \mathbb{R}^d$ be pairwise disjoint closed bounded subsets and $\mathcal{H}_d(N, L)$ be the hypothesis space defined in (1.5) with $N = 36d(2d+1)$ and $L = 12$. Then, for $f \in \mathcal{C}_d(E_1, E_2, \dots, E_J)$, there exists $\phi \in \mathcal{H}_d(N, L)$ such that

$$\phi(\mathbf{x}) = f(\mathbf{x}) \quad \text{for any } \mathbf{x} \in \bigcup_{j=1}^J E_j.$$

Remark. The network realizing ϕ in Theorem 1.3 has

$$d \times N + N + (N \times N + N) \times (L - 1) + N \times 1 + 1 \sim d^4$$

parameters, where $N = 36d(2d+1)$ and $L = 12$. However, as shown in our constructive proof of Theorem 1.3, it is enough to adjust $5509(d+1)(2d+1) = \mathcal{O}(d^2) \ll d^4$ parameters and set all the others to 0.

For a general function space \mathcal{F} , define $\mathcal{F}|_E := \{f|_E : f \in \mathcal{F}\}$, where $f|_E$ is the function achieved via limiting f on E . Then, we have a corollary of Theorem 1.3 as follows.

Corollary 1.4. Let $E_1, E_2, \dots, E_J \subseteq \mathbb{R}^d$ be pairwise disjoint closed bounded subsets and $\mathcal{H}_d(N, L)$ be the hypothesis space defined in (1.5). Assume $N \geq 36d(2d+1)$ and $L \geq 12$, then

$$\mathcal{C}_d(E_1, E_2, \dots, E_J)|_E \subseteq \mathcal{H}_d(N, L)|_E,$$

where $E = \bigcup_{j=1}^J E_j$.

One of the most successful applications of deep learning is the image and signal classifications. In supervised classification problems, given a few samples and their labels (usually integers), the goal of the task is to learn how to assign a label to a new sample. For example, in binary classification via deep learning, a neural network is trained based on given samples (and labels) to approximate a classification function mapping one class of samples to 0 and the other class of samples to 1. Theorem 1.3 (or Corollary 1.4) implies that the classification function can be exactly realized by an EUAF network with a size depending only on the dimension of the problem domain via adjusting its parameters. This means that the best approximation error of EUAF networks to classification functions in the classification problem is 0.

Remark that, in the worst scenario, there might exist complicated high-dimensional functions such that, the parameters of the EUAF network in Theorem 1.3 (or Corollary 1.4) require high computer precision for storage, and the precision might be exponentially large in the problem dimension. We refer this as the curse of memory, which may make Theorem 1.3 and Corollary 1.4 less interesting in real applications, though the number of parameters can be very small. The key question to be addressed is how rare the curse of memory would happen in real applications. If the target functions in real applications typically have no curse of memory with a high probability, then EUAF networks would be very useful in real applications. As future work, we will explore the statistical characterization of high-dimensional functions for the curse of memory of EUAF networks.

1.2 Related work

In recent years, there has been an increasing amount of literature on the approximation power of neural networks as a special case of nonlinear approximation [5, 7, 8]. In the early works of approximation theory for neural networks, the universal approximation theorem [6, 16, 17] without approximation errors showed that there exists a sufficiently large neural network approximating a target function in a certain function space within any given error $\varepsilon > 0$. There are also other versions of the universal approximation theorem. For example, it was shown in [24] that the ReLU-activated residual neural networks with one neuron per hidden layer and a sufficiently large depth are a universal approximator. The universal approximation property for general residual neural networks was proved in [22] via a dynamical system approach. In all papers discussed above, the network size goes to infinity when the target approximation error approaches 0. However, our result in Theorem 1.1 implies that EUAF networks with a fixed size ($\mathcal{O}(d^2)$ neurons in total) can achieve an arbitrary small error for approximating $f \in C([a, b]^d)$.

The approximation errors in terms of the total number of parameters of ReLU networks are well studied for basic function spaces with (nearly) optimal approximation errors, e.g., (nearly) optimal asymptotic errors for continuous functions [38], C^s functions [40], piecewise smooth functions [30], functions that can be optimally approximated by affine systems [2], and Sobolev spaces [37]. Approximation errors in terms of width and depth would be more useful than those in terms of the total number of nonzero parameters in practice, because width and depth are two essential hyper-parameters in every numerical algorithm instead of the number of nonzero parameters. This motivated the works on the (nearly) optimal non-asymptotic errors in terms of width and depth with explicit pre-factors for approximating continuous functions in [31, 34, 41] and for C^s functions in [25, 41]. As the errors are optimal, there are two possible directions to improve the approximation error in order to reduce the effect of the curse of dimensionality. The first one is to consider smaller target function spaces, e.g., analytic functions [3, 11], Barron spaces [1, 10, 13, 35], and band-limited functions [4, 27].

Another direction is to design advanced activation functions, where one can use multiple activation functions, to enhance the power of neural networks, especially to conquer the curse of dimensionality in network approximation. There have been several papers designing activation functions to achieve good approximation errors. The results in [40] imply that (sin, ReLU)-activated neural networks (i.e., the activation function of a neuron can be chosen from either sin or ReLU) with W parameters can approximate Lipschitz continuous functions with an asymptotic approximation error $\mathcal{O}(e^{-c_d \sqrt{W}})$, where c_d is a constant depending on d and might cause the curse of dimensionality, though the approximation error is root-exponentially small in W . In [32], it was shown that (Floor, ReLU)-activated neural networks with width $\mathcal{O}(N)$ and depth $\mathcal{O}(L)$ admit a quantitative approximation error $\mathcal{O}(\sqrt{d}N^{-\sqrt{L}})$ for Lipschitz continuous functions, conquering the curse of dimensionality in approximation with a root-exponentially small error in depth L .^① In [33], it was shown that, even if the depth is as small as 3, neural

^①Although there is no curse of dimensionality in network approximation, the construction requires exponentially many data samples of the target function and computer memory. Hence, there would be a curse of dimensionality in inferring a target function from its finite samples when standard learning techniques are applied on a computer.

networks with width N and $\mathcal{O}(d + N)$ nonzero parameters can approximate Lipschitz continuous functions with an exponentially small error $\mathcal{O}(\sqrt{d}2^{-N})$, if the floor function $\lfloor x \rfloor$, the exponential function 2^x , and the step function $\mathbb{1}_{\{x \geq 0\}}$ are used as activation functions. Corollary 1.2 implies that the hypothesis space of EUAF networks activated by a single activation function with $\mathcal{O}(d^2)$ neurons is dense in $C([a, b]^d)$. Particularly, all continuous functions can be arbitrarily approximated by fixed-size EUAF networks with width N and depth L on a d -dimensional hypercube, whenever $N \geq 36d(2d + 1)$ and $L \geq 11$.

There is another research line for the approximation error of neural networks: apply KST [20] or its variants to explore new activation functions for a fixed-size network to achieve an arbitrary error. The original KST shows that any multivariate function $f \in C([0, 1]^d)$ can be represented as $f(\mathbf{x}) = \sum_{i=0}^{2^d} g_i(\sum_{j=1}^d h_{i,j}(x_j))$ for any $\mathbf{x} = [x_1, \dots, x_d]^T \in [0, 1]^d$, where g_i and $h_{i,j}$ are univariate continuous functions. In fact, the composition architecture of KST can be regarded as a special neural network with (complicated) activation functions depending on the target function, which results in the failure of KST in practice. To alleviate this issue, a single activation function independent of the target function is designed in [26] to construct networks with a fixed size ($\mathcal{O}(d)$ neurons) to achieve an arbitrary error for approximating functions in $C([-1, 1]^d)$. However, the activation function in [26] has no closed form and is hardly computable. See Section 2.2 for a detailed discussion of [26]. The computability issue of activation functions was addressed recently in [39]. It was shown in [39] that, for an arbitrary $\varepsilon > 0$ and any function f in $C([0, 1]^d)$, there exists a network of size only depending on d constructed with multiple activation functions either (\sin & \arcsin) or ($\lfloor \cdot \rfloor$ & a non-polynomial analytic function) to approximate f within an error ε . To the best of our knowledge, there is no explicit characterization of the size dependence on d in [39]. For example, a very important question is whether the dependence can be mild, e.g., only a polynomial of d , or has to be severe, e.g., exponentially in d . The results of current paper provide positive answers to all the issues discussed above: we show that EUAF networks with a single simple and computable activation function, width $36d(2d + 1)$, and depth 11 can approximate functions in $C([a, b]^d)$ within an arbitrary pre-specified error $\varepsilon > 0$.

In summary, the aim of this paper is to design a simple and computable activation function σ to construct fixed-size neural networks with the universal approximation property. The network sizes of the width and depth have an explicit characterization that only depends on the dimension d . The fixed-size neural network is designed to approximate any continuous functions on a hypercube within an arbitrary error by only adjusting $\mathcal{O}(d^2)$ network parameters. Moreover, we prove that an arbitrary classification function can be exactly represented by such a fixed-size network architecture via only adjusting $\mathcal{O}(d^2)$ network parameters. The main contribution of this paper is to develop a rigorous mathematical analysis for the universal approximation property of fixed-size neural networks. Some of the mathematical analysis and ideas developed here may be applied to understand other neural networks.

Finally, the EUAF network is simple and computable in the sense that the output and subgradient of EUAF networks can be efficiently evaluated. Therefore, EUAF can be directly applied to existing deep learning software in the same way as other popular activation functions (such as ReLU or sigmoid). As opposed to the computability of

our EUAF, the powerful activation function proposed in [26] is not computable in the sense that there is no numerical algorithm to evaluate the output and gradient of the corresponding network. In real applications, the parameters of the EUAF network are learned from the samples of the target function, which involves sophisticated numerical optimization. We refer to the learnability of network parameters as the existence of a numerical optimization algorithm that can identify network parameters to achieve a target approximation error. Obviously, the computability of the EUAF networks does not imply learnability. As we shall see in the next subsection, the learnability involves approximation, optimization, and generalization error analyses. The result in this paper show that there exist computable EUAF networks achieving an arbitrarily small approximation error. Whether or not EUAF networks can be useful in real applications also depends on optimization and generalization, which is out of the scope of this paper. Optimization and generalization error analyses of practical deep neural networks including EUAF networks has been a challenging problem for years. To the best of our knowledge, there is no complete error analysis to address the learnability of neural networks with nonlinear activation functions, even for a small function space like Barron spaces [1, 10, 13, 35]. We will discuss the full error analysis for the learnability in the next subsection.

1.3 Error analysis

The error analysis of deep learning generally includes approximation, generalization, and optimization errors. Our results in this paper only deal with the approximation error. Here, we give a brief discussion on these three errors to illustrate the importance of controlling approximation errors in the applications of deep neural networks. One may find more details in [25, 32]. Let $\Phi(\mathbf{x}, \boldsymbol{\theta})$ denote a function in $\mathbf{x} \in \mathbb{R}^d$ generated by a network architecture parameterized with $\boldsymbol{\theta} \in \mathbb{R}^W$. Given a target function f , the final goal is to find the expected risk minimizer

$$\boldsymbol{\theta}_{\mathcal{D}} \in \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^W} R_{\mathcal{D}}(\boldsymbol{\theta}), \quad \text{where } R_{\mathcal{D}}(\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{x} \sim U(\mathcal{X})} [\ell(\Phi(\mathbf{x}, \boldsymbol{\theta}), f(\mathbf{x}))]$$

with a loss function $\ell(\cdot, \cdot)$ and an unknown data distribution $U(\mathcal{X})$.

Theorem 1.1 implies $\inf_{\boldsymbol{\theta} \in \mathbb{R}^W} \|\Phi(\cdot, \boldsymbol{\theta}) - f(\cdot)\|_{L^\infty([a, b]^d)} = 0$ for all $f \in C([a, b]^d)$ with $\mathcal{X} = [a, b]^d$. However, $\boldsymbol{\theta}_{\mathcal{D}}$ may not be always achievable. When $\boldsymbol{\theta}_{\mathcal{D}}$ is achievable, $\mathbb{E}_{\mathbf{x} \sim U(\mathcal{X})} [\ell(\Phi(\mathbf{x}, \boldsymbol{\theta}_{\mathcal{D}}), f(\mathbf{x}))] = R_{\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}}) = 0$. When $\boldsymbol{\theta}_{\mathcal{D}}$ is not attainable, for any pre-specified $\eta > 0$, one could identify $\boldsymbol{\theta}_{\mathcal{D}, \eta} \in \mathbb{R}^W$ as the parameter set satisfying

$$R_{\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D}, \eta}) \leq \inf_{\boldsymbol{\theta} \in \mathbb{R}^W} R_{\mathcal{D}}(\boldsymbol{\theta}) + \eta/2. \quad (1.6)$$

In practice, for given samples $\{(\mathbf{x}_i, f(\mathbf{x}_i))\}_{i=1}^n$, the goal of supervised learning is to identify the empirical risk minimizer

$$\boldsymbol{\theta}_{\mathcal{S}} \in \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^W} R_{\mathcal{S}}(\boldsymbol{\theta}), \quad \text{where } R_{\mathcal{S}}(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^n \ell(\Phi(\mathbf{x}_i, \boldsymbol{\theta}), f(\mathbf{x}_i)).$$

Similarly, when $\boldsymbol{\theta}_{\mathcal{S}}$ is not attainable, our goal is to identify $\boldsymbol{\theta}_{\mathcal{S}, \eta}$ instead of $\boldsymbol{\theta}_{\mathcal{S}}$ for any pre-specified $\eta > 0$, where $\boldsymbol{\theta}_{\mathcal{S}, \eta} \in \mathbb{R}^W$ satisfies

$$R_{\mathcal{S}}(\boldsymbol{\theta}_{\mathcal{S}, \eta}) \leq \inf_{\boldsymbol{\theta} \in \mathbb{R}^W} R_{\mathcal{S}}(\boldsymbol{\theta}) + \eta/2. \quad (1.7)$$

In practical implementation, only a numerical minimizer $\boldsymbol{\theta}_{\mathcal{N}}$ of $R_{\mathcal{S}}(\boldsymbol{\theta})$ can be achieved via a numerical optimization method. The discrepancy between the learned function $\Phi(\mathbf{x}, \boldsymbol{\theta}_{\mathcal{N}})$ and the target function f is measured by $R_{\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{N}})$, which is bounded by

$$\begin{aligned}
R_{\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{N}}) &= \underbrace{[R_{\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{N}}) - R_{\mathcal{S}}(\boldsymbol{\theta}_{\mathcal{N}})]}_{\text{GE}} + \underbrace{[R_{\mathcal{S}}(\boldsymbol{\theta}_{\mathcal{N}}) - R_{\mathcal{S}}(\boldsymbol{\theta}_{\mathcal{S},\eta})]}_{\text{OE}} + \underbrace{[R_{\mathcal{S}}(\boldsymbol{\theta}_{\mathcal{S},\eta}) - R_{\mathcal{S}}(\boldsymbol{\theta}_{\mathcal{D},\eta})]}_{\leq \eta/2 \text{ by (1.7)}} + \underbrace{[R_{\mathcal{S}}(\boldsymbol{\theta}_{\mathcal{D},\eta}) - R_{\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D},\eta})]}_{\text{GE}} + \underbrace{R_{\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D},\eta})}_{\leq \inf_{\boldsymbol{\theta} \in \mathbb{R}^W} R_{\mathcal{D}}(\boldsymbol{\theta}) + \eta/2 \text{ by (1.6)}} \\
&\leq \underbrace{\eta}_{\text{Perturbation}} + \underbrace{\inf_{\boldsymbol{\theta} \in \mathbb{R}^W} R_{\mathcal{D}}(\boldsymbol{\theta})}_{\text{Approximation error}} + \underbrace{[R_{\mathcal{S}}(\boldsymbol{\theta}_{\mathcal{N}}) - \inf_{\boldsymbol{\theta} \in \mathbb{R}^W} R_{\mathcal{S}}(\boldsymbol{\theta})]}_{\text{Optimization error (OE)}} + \underbrace{[R_{\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{N}}) - R_{\mathcal{S}}(\boldsymbol{\theta}_{\mathcal{N}})] + [R_{\mathcal{S}}(\boldsymbol{\theta}_{\mathcal{D},\eta}) - R_{\mathcal{D}}(\boldsymbol{\theta}_{\mathcal{D},\eta})]}_{\text{Generalization error (GE)}}.
\end{aligned}$$

The pre-specified hyper-parameter η can be arbitrarily small and Theorem 1.1 guarantees $\inf_{\boldsymbol{\theta} \in \mathbb{R}^W} R_{\mathcal{D}}(\boldsymbol{\theta}) = 0$. Therefore, the error analysis of deep learning can be reduced to the analysis of the optimization and generalization errors, which depends on data samples, optimization algorithms, etc. One could refer to [9, 10, 12, 15, 18, 19, 23, 28, 29] for the analysis of the generalization and optimization errors. The computability of the activation function implies that we can numerically implement the optimization algorithm to find a minimizer of the empirical risk. Since our theory provides an arbitrarily small approximation error, the learnability of the best approximation is reduced to achieving small generalization and optimization errors, which depends on the given data, the empirical risk model, and the optimization algorithm.

The rest of this paper is organized as follows. In Section 2, we first prove Theorem 1.1 by assuming Theorem 2.1 is true, and then prove Theorem 1.3. Next, Theorem 2.1 is proved in Section 3 based on Proposition 2.2, the proof of which can be found in Section 4. Then, several UAFs with better properties are proposed in Section 5. Finally, Section 6 concludes this paper with a short discussion.

2 Proof of main theorems

In this section, we first prove Theorem 1.1 by assuming Theorem 2.1 is true, and then prove Theorem 1.3. Notation throughout this paper are summarized in Section 2.1.

2.1 Notation

Let us summarize all basic notation used in this paper as follows.

- Let \mathbb{R} , \mathbb{Q} , and \mathbb{Z} denote the set of real numbers, rational numbers, and integers, respectively.
- Let \mathbb{N} and \mathbb{N}^+ denote the set of natural numbers and positive natural numbers, respectively. That is, $\mathbb{N}^+ = \{1, 2, 3, \dots\}$ and $\mathbb{N} = \mathbb{N}^+ \cup \{0\}$.
- For any $x \in \mathbb{R}$, let $\lfloor x \rfloor := \max\{n : n \leq x, n \in \mathbb{Z}\}$ and $\lceil x \rceil := \min\{n : n \geq x, n \in \mathbb{Z}\}$.
- Let $\mathbb{1}_S$ be the indicator (characteristic) function of a set S , i.e., $\mathbb{1}_S$ is equal to 1 on S and 0 outside S .
- The set difference of two sets A and B is denoted by $A \setminus B := \{x : x \in A, x \notin B\}$.

• Matrices are denoted by bold uppercase letters. For instance, $\mathbf{A} \in \mathbb{R}^{m \times n}$ is a real matrix of size $m \times n$, and \mathbf{A}^T denotes the transpose of \mathbf{A} . Vectors are denoted as bold lowercase letters. For example, $\mathbf{v} = [v_1, \dots, v_d]^T = \begin{bmatrix} v_1 \\ \vdots \\ v_d \end{bmatrix} \in \mathbb{R}^d$ is a column vector. Besides, “[” and “]” are used to partition matrices (vectors) into blocks, e.g., $\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$.

• For any $p \in [1, \infty)$, the p -norm (or ℓ^p -norm) of a vector $\mathbf{x} = [x_1, x_2, \dots, x_d]^T \in \mathbb{R}^d$ is defined by

$$\|\mathbf{x}\|_p = \|\mathbf{x}\|_{\ell^p} := (|x_1|^p + |x_2|^p + \dots + |x_d|^p)^{1/p}.$$

In the case $p = \infty$,

$$\|\mathbf{x}\|_\infty = \|\mathbf{x}\|_{\ell^\infty} := \max \{|x_i| : i = 1, 2, \dots, d\}.$$

• For any $a_1, a_2, \dots, a_J \in \mathbb{R}$, we say a_1, a_2, \dots, a_J are **rationally independent** if they are linearly independent over the rational numbers \mathbb{Q} . That is, if there exist $\lambda_1, \lambda_2, \dots, \lambda_J \in \mathbb{Q}$ such that $\sum_{j=1}^J \lambda_j \cdot a_j = 0$, then $\lambda_1 = \lambda_2 = \dots = \lambda_J = 0$. For a simple example, 1, $\sqrt{2}$, and $\sqrt{3}$ are rationally independent.

• An **algebraic** number is any complex number (including real numbers) that is a root of a polynomial equation with rational coefficients, i.e., α is an algebraic number if and only if there exist $\lambda_0, \lambda_1, \dots, \lambda_J \in \mathbb{Q}$ with $\sum_{j=0}^J \lambda_j \alpha^j = 0$.^② Denote the set of all algebraic numbers by \mathbb{A} . A complex number is called **transcendental** if it is not in \mathbb{A} . The set \mathbb{A} is countable, and, therefore, almost all numbers are transcendental. The best known transcendental numbers are π (the ratio of a circle’s circumference to its diameter) and e (the natural logarithmic base).

• The expression “a network (architecture) with width N and depth L ” means

- The maximum width of this network (architecture) for all **hidden** layers is no more than N .
- The number of **hidden** layers of this network (architecture) is no more than L .

2.2 Key ideas of proving Theorem 1.1

The proof of Theorem 1.1 has two main steps: 1) prove the one-dimensional case; 2) reduce the d -dimensional approximation to the one-dimensional case via KST [20]. In fact, in the case of $d = 1$, the size of the network in Theorem 1.1 can be further reduced as shown in Theorem 2.1 below. Theorem 2.1 is actually an enhanced version of Theorem 1.1, and, therefore, implies Theorem 1.1 in the case $d = 1$.

Theorem 2.1. *Let $f \in C([a, b])$ be a continuous function. Then, for an arbitrary $\varepsilon > 0$, there exists a function ϕ generated by an EUAF network with width 36 and depth 5 such that*

$$|\phi(x) - f(x)| < \varepsilon \quad \text{for any } x \in [a, b] \subseteq \mathbb{R}.$$

^②For simplicity, we denote $1 = x^0$ for any $x \in \mathbb{R}$, including the case 0^0 .

The detailed proof of Theorem 2.1 can be found in Section 3. The main ideas of proving Theorem 2.1 are developed from some ideas of our early works [32, 33]. Roughly speaking, we eventually convert a function approximation problem to a point-fitting problem via the composition architecture of neural networks in the following three steps.

- Divide $[0, 1)$ into small intervals $\mathcal{I}_k = [\frac{k-1}{K}, \frac{k}{K})$ with a left endpoint x_k for $k \in \{1, 2, \dots, K\}$, where K is an integer determined by the given error and the target function f .
- Construct a sub-network to generate a function ϕ_1 mapping the whole interval \mathcal{I}_k to k for each k . The floor function $\lfloor \cdot \rfloor$ is a good choice to implement this step. Precisely, we can define $\phi_1(x) = \lfloor Kx \rfloor$. The floor function is not continuous and has zero-derivative almost everywhere. As we shall see later, σ_1 (or σ) can be a continuous alternative to implement this step, but the construction is more complicated.
- The final step is to design another sub-network to generate a function ϕ_2 mapping k approximately to $f(x_k)$ for each k . Then $\phi_2 \circ \phi_1(x) = \phi_2(k) \approx f(x_k) \approx f(x)$ for any $x \in \mathcal{I}_k$ and $k \in \{1, 2, \dots, K\}$, which implies $\phi_2 \circ \phi_1 \approx f$ on $[0, 1)$. After the above two steps, we simplify the approximation problem to a point-fitting problem, where k is approximately mapped to $f(k)$. This step is the bottleneck of the construction in our previous papers [32, 33]. Roughly speaking, the final approximation error is essentially determined by how many points we can fit using a neural network.

For the second step, the capacity to generate step functions with sufficiently many “steps” via a sub-network with a limited number of neurons plays an important role. The reproduced step functions can be considered as a continuous version of the floor function $(\lfloor \cdot \rfloor)$ in [32, 33], which is a perfect step function with infinite “steps” that improves the approximation power of networks as shown in [32, 33]. The key ingredient in the third step of the proof of Theorem 2.1 is essentially a point-fitting problem with arbitrarily many points. This requires the following proposition motivated by the well-known fact that an irrational winding on the torus is dense. See Figure 2 for illustrations of such a fact. Here, we propose a new point-fitting technique that can fit arbitrarily many points within an arbitrary error using neural networks.

Proposition 2.2. *For any $K \in \mathbb{N}^+$, the following point set*

$$\left\{ \left[\sigma_1\left(\frac{w}{\pi+1}\right), \sigma_1\left(\frac{w}{\pi+2}\right), \dots, \sigma_1\left(\frac{w}{\pi+K}\right) \right]^T : w \in \mathbb{R} \right\} \subseteq [0, 1]^K$$

is dense in $[0, 1]^K$, where π is the ratio of the circumference of a circle to its diameter.

The proof of this proposition can be found in Section 4. This proposition implies that for any given sample points $(k, y_k) \in \mathbb{R}^2$ with $y_k \in [0, 1]$ for $k = 1, 2, \dots, K$ and any $K \in \mathbb{N}^+$, there exists $w_0 \in \mathbb{R}$ such that the function $x \mapsto \sigma_1(\frac{w_0}{\pi+x})$ can fit the points $(k, y_k) \in \mathbb{R}^2$ for $k = 1, 2, \dots, K$ within an arbitrary pre-specified error $\varepsilon > 0$. To put it another way, for any $\varepsilon > 0$, there exists $w_0 \in \mathbb{R}$ such that $|\sigma_1(\frac{w_0}{\pi+k}) - y_k| < \varepsilon$ for all k .

As we shall see later in the proof of Proposition 2.2, the key point is the periodicity of the outer function σ_1 . Of course, the inner function $x \mapsto \frac{w_0}{\pi+x}$ is also necessary since it

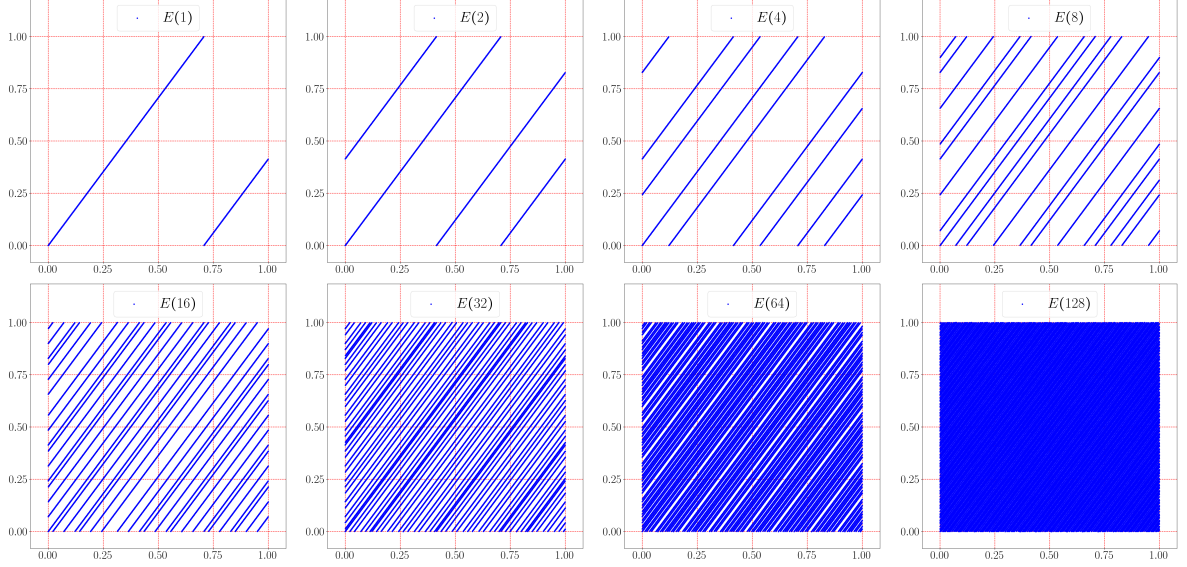


Figure 2: Illustrations of the denseness of $E(\infty)$ in $[0, 1]^2$, where $E(r)$ is a winding of an “irrational” direction $[1, \sqrt{2}]^T$ on $[0, r)$, i.e., $E(r) = \{[\tau(t), \tau(\sqrt{2}t)]^T : t \in [0, r)\}$ with $\tau(t) = t - \lfloor t \rfloor$.

helps to adjust sample points for $x = 1, 2, \dots, K$. In fact, the inner function $x \mapsto \frac{w_0}{\pi+x}$ can be regarded as a variant of σ_2 via scaling and shifting. The periodicity has been explored to improve neural network approximation in the literature, e.g. the sin function in [40] is periodic and the floor function ($\lfloor \cdot \rfloor$) in [32, 33] is implicitly periodic because $x - \lfloor x \rfloor$ is periodic. Remark that a similar result holds if we replace σ_1 by a non-trivial periodic function and replace the sample locations $x = 1, 2, \dots, K$ by distinct rational numbers $r_1, r_2, \dots, r_K \in \mathbb{Q}$. See Section 4 for a further discussion.

Theorem 2.1 essentially proves Theorem 1.1 for the univariate case. To prove the general case, we need KST [20] given below to reduce a multivariate problem to a one-dimensional case.

Theorem 2.3 (Kolmogorov superposition theorem (KST) [20]). *There exist continuous functions $h_{i,j} \in C([0, 1])$ for $i = 0, 1, \dots, 2d$ and $j = 1, 2, \dots, d$ such that any continuous function $f \in C([0, 1]^d)$ can be represented as*

$$f(\mathbf{x}) = \sum_{i=0}^{2d} g_i \left(\sum_{j=1}^d h_{i,j}(x_j) \right) \quad \text{for any } \mathbf{x} = [x_1, x_2, \dots, x_d]^T \in [0, 1]^d,$$

where $g_i : \mathbb{R} \rightarrow \mathbb{R}$ is a continuous function for each $i \in \{0, 1, \dots, 2d\}$.

KST [20] is often used to reduce a multidimensional problem to a one-dimensional one. In fact, the compositional representation in KST can be regarded as a special neural network with (complicated) activation functions depending on the target function, which makes KST useless in practical computation. To avoid this dependency, an activation function was designed in [26] to construct neural network representations with $\mathcal{O}(d)$ neurons that can approximate functions in $C([-1, 1]^d)$ within an arbitrary error. Let us briefly summarize the main ideas in [26]: 1) Identify a dense and countable subset $\{u_k\}_{k=1}^\infty$

of $C([-1, 1])$, e.g., polynomials with rational coefficients. 2) Construct an activation function ϱ to encode all $u_k(x)$ for $x \in [-1, 1]$. In fact, for each k , $u_k|_{[-1, 1]}$ is “stored” in ϱ on $[4k, 4k + 2]$, and the values of ϱ on $[4k + 2, 4k + 4]$ are properly assigned to make ϱ a smooth and monotonically increasing function. That is, let $\varrho(x + 4k + 1) = a_k + b_k x + c_k u_k(x)$ for any $x \in [-1, 1]$ with carefully chosen constants a_k , b_k , and $c_k \neq 0$ such that $\varrho(x)$ can be a sigmoid function. 3) For any $g \in C([-1, 1])$, there exists a one-hidden-layer ϱ -activated network with width 3 approximating g within an arbitrary error δ , i.e., there exists k such that $g \stackrel{\delta}{\approx} u_k =: \frac{\varrho(x + 4k + 1) - a_k - b_k x}{c_k}$. 4) Replace the inner and outer functions in KST with these one-hidden-layer networks to achieve a two-hidden-layer ϱ -activated network with width $\mathcal{O}(d)$ to approximate $f \in C([0, 1]^d)$ within an arbitrary error ε . As we can see, the key point of the construction in [26] is to encode a dense and countable subset of the target function space in an activation function.

We note that both [26] and this paper use KST to reduce dimension. However, the activation function of [26] is complicated without any close form and there is no efficient numerical algorithm to evaluate it. After encoding a dense subset of continuous function into a single but complicated activation function, one only needs to construct affine linear transformations to select appropriate functions of this dense subset from this complicated activation function to construct approximation. Hence, such a complicated activation function simplifies the proof of the denseness, since the denseness is encoded in the activation function. As a contrast, we design a simple activation function with efficient numerical implementation (see Figure 1 for an illustration) achieving the universal approximation property with fixed-size networks, because simple and implementable activation functions are a basic requirement for a neural network to be used in applications. However, the proof of the denseness of a neural network generated by such a simple activation function becomes difficult. A sophisticated analysis will be developed in the rest of this paper to overcome the difficulties.

We start with proving Theorem 1.1 by assuming Theorem 2.1, whose proof will be given in Section 3.

2.3 Proof of Theorem 1.1

The detailed proof of Theorem 1.1 converts the above ideas to implementations using neural networks with fixed sizes. The whole construction procedure can be divided into three steps.

- (1) Apply KST to reduce dimension, i.e., represent $f \in C([a, b]^d)$ by the compositions and combinations of univariate continuous functions.
- (2) Apply Theorem 2.1 to design sub-networks to approximate the univariate continuous functions in the previous step within the desired error.
- (3) Integrate the sub-networks to form the final network and estimate its size.

Step 1: Apply KST to reduce dimension.

To apply KST, we define a linear function $\mathcal{L}_1(t) = (b - a)t - a$ for any $t \in [0, 1]$. Clearly, \mathcal{L}_1 is a bijection from $[0, 1]$ to $[a, b]$. Define

$$\tilde{f}(\mathbf{y}) := f(\mathcal{L}_1(y_1), \mathcal{L}_1(y_2), \dots, \mathcal{L}_1(y_d)) \quad \text{for any } \mathbf{y} = [y_1, y_2, \dots, y_d]^T \in [0, 1]^d.$$

470 Then $\tilde{f}: [0, 1]^d \rightarrow \mathbb{R}$ is a continuous function since $f \in C([a, b]^d)$. By Theorem 2.3, there
 471 exists $\tilde{h}_{i,j} \in C([0, 1])$ and $\tilde{g}_i \in C(\mathbb{R})$ for $i = 0, 1, \dots, 2d$ and $j = 1, 2, \dots, d$ such that

$$472 \quad \tilde{f}(\mathbf{y}) = \sum_{i=0}^{2d} \tilde{g}_i \left(\sum_{j=1}^d \tilde{h}_{i,j}(y_j) \right) \quad \text{for any } \mathbf{y} = [y_1, y_2, \dots, y_d]^T \in [0, 1]^d.$$

473 Let $\tilde{\mathcal{L}}_1$ be the inverse of \mathcal{L}_1 , i.e., define $\tilde{\mathcal{L}}_1(t) = (t - a)/(b - a)$ for any $t \in [a, b]$. Then, for
 474 any $x_j \in [a, b]$, there exists a unique $y_j \in [0, 1]$ such that $\mathcal{L}_1(y_j) = x_j$ and $y_j = \tilde{\mathcal{L}}_1(x_j)$ for
 475 any $j = 1, 2, \dots, d$, which implies

$$\begin{aligned} f(\mathbf{x}) &= f(x_1, x_2, \dots, x_d) = f(\mathcal{L}_1(y_1), \mathcal{L}_1(y_2), \dots, \mathcal{L}_1(y_d)) = \tilde{f}(\mathbf{y}) \\ 476 \quad &= \sum_{i=0}^{2d} \tilde{g}_i \left(\sum_{j=1}^d \tilde{h}_{i,j}(y_j) \right) = \sum_{i=0}^{2d} \tilde{g}_i \left(\sum_{j=1}^d \tilde{h}_{i,j}(\tilde{\mathcal{L}}_1(x_j)) \right) = \sum_{i=0}^{2d} \tilde{g}_i \left(\sum_{j=1}^d \tilde{h}_{i,j} \circ \tilde{\mathcal{L}}_1(x_j) \right). \end{aligned}$$

477 It follows that

$$478 \quad f(\mathbf{x}) = \sum_{i=0}^{2d} \tilde{g}_i \left(\sum_{j=1}^d \tilde{h}_{i,j} \circ \tilde{\mathcal{L}}_1(x_j) \right) = \sum_{i=0}^{2d} \tilde{g}_i \circ \hat{h}_i(\mathbf{x}) \quad \text{for any } \mathbf{x} \in [a, b]^d,$$

479 where

$$480 \quad \hat{h}_i(\mathbf{x}) = \sum_{j=1}^d \tilde{h}_{i,j} \circ \tilde{\mathcal{L}}_1(x_j) \quad \text{for any } \mathbf{x} = [x_1, x_2, \dots, x_d]^T \in [a, b]^d. \quad (2.1)$$

481 Denote

$$482 \quad M = \max_{i \in \{0, 1, \dots, 2d\}} \|\tilde{h}_i\|_{L^\infty([a, b]^d)} + 1 > 0.$$

483 Define $\mathcal{L}_2(t) = (t + 2M)/4M$ and $\tilde{\mathcal{L}}_2(t) = 4Mt - 2M$ for any $t \in \mathbb{R}$. Then \mathcal{L}_2 is a bijection
 484 from $[-M, M]$ to $[\frac{1}{4}, \frac{3}{4}]$ and $\tilde{\mathcal{L}}_2$ is the inverse of \mathcal{L}_2 . Clearly, $\tilde{\mathcal{L}}_2 \circ \mathcal{L}_2(t) = t$ for any
 485 $t \in [-M, M]$, which implies $\hat{h}_i(\mathbf{x}) = \tilde{\mathcal{L}}_2 \circ \mathcal{L}_2 \circ \hat{h}_i(\mathbf{x})$ for any $\mathbf{x} \in [a, b]^d$. Therefore, for any
 486 $\mathbf{x} \in [a, b]^d$, we have

$$487 \quad f(\mathbf{x}) = \sum_{i=0}^{2d} \tilde{g}_i \circ \hat{h}_i(\mathbf{x}) = \sum_{i=0}^{2d} \tilde{g}_i \circ \tilde{\mathcal{L}}_2 \circ \mathcal{L}_2 \circ \hat{h}_i(\mathbf{x}) = \sum_{i=0}^{2d} g_i \circ h_i(\mathbf{x}),$$

488 where

$$489 \quad g_i = \tilde{g}_i \circ \tilde{\mathcal{L}}_2 \quad \text{and} \quad h_i = \mathcal{L}_2 \circ \hat{h}_i \quad \text{for } i = 0, 1, \dots, 2d. \quad (2.2)$$

490 Clearly, $\mathcal{L}_2(t) \in [\frac{1}{4}, \frac{3}{4}]$ for any $t \in [-M, M]$, which implies

$$491 \quad h_i(\mathbf{x}) = \mathcal{L}_2 \circ \hat{h}_i(\mathbf{x}) \in [\frac{1}{4}, \frac{3}{4}] \quad \text{for any } \mathbf{x} \in [a, b]^d \text{ and } i = 0, 1, \dots, 2d.$$

492 **Step 2:** Design sub-networks to approximate g_i and h_i .

493 Next, we represent g_i and h_i by sub-networks. Obviously, $g_i = \tilde{g}_i \circ \tilde{\mathcal{L}}_2$ is continuous
 494 on \mathbb{R} , and, therefore, uniformly continuous on $[0, 1]$ for each i . Thus, for $i = 0, 1, \dots, 2d$,
 495 there exists $\delta_i > 0$ such that

$$496 \quad |g_i(z_1) - g_i(z_2)| < \varepsilon / (4d + 2) \quad \text{for any } z_1, z_2 \in [0, 1] \text{ with } |z_1 - z_2| < \delta_i.$$

Set $\delta = \min(\{\delta_i : i = 0, 1, \dots, 2d\} \cup \{\frac{1}{4}\})$. Then, for $i = 0, 1, \dots, 2d$, we have

$$|g_i(z_1) - g_i(z_2)| < \varepsilon/(4d+2) \quad \text{for any } z_1, z_2 \in [0, 1] \text{ with } |z_1 - z_2| < \delta. \quad (2.3)$$

For each $i \in \{0, 1, \dots, 2d\}$, by Theorem 2.1, there exists a function ϕ_i generated by an EUAF network with width 36 and depth 5 such that

$$|g_i(z) - \phi_i(z)| < \varepsilon/(4d+2) \quad \text{for any } z \in [0, 1]. \quad (2.4)$$

Fix $i \in \{0, 1, \dots, 2d\}$, we will design an EUAF network to generate a function $\psi_i : [a, b]^d \rightarrow \mathbb{R}$ satisfying

$$|h_i(\mathbf{x}) - \psi_i(\mathbf{x})| < \delta \quad \text{for any } \mathbf{x} \in [a, b]^d.$$

For any $\mathbf{x} = [x_1, x_2, \dots, x_d]^T \in [a, b]^d$, by Equations (2.1) and (2.2), we have

$$\begin{aligned} h_i(\mathbf{x}) &= \mathcal{L}_2 \circ \widehat{h}_i(\mathbf{x}) = \mathcal{L}_2 \left(\sum_{j=1}^d \widetilde{h}_{i,j} \circ \widetilde{\mathcal{L}}_1(x_j) \right) = \frac{\left(\sum_{j=1}^d \widetilde{h}_{i,j} \circ \widetilde{\mathcal{L}}_1(x_j) \right) + 2M}{4M} \\ &= \sum_{j=1}^d \left(\frac{\widetilde{h}_{i,j} \circ \widetilde{\mathcal{L}}_1(x_j)}{4M} + \frac{1}{2d} \right) =: \sum_{j=1}^d h_{i,j}(x_j), \end{aligned}$$

where

$$h_{i,j}(t) := \frac{\widetilde{h}_{i,j} \circ \widetilde{\mathcal{L}}_1(t)}{4M} + \frac{1}{2d} \quad \text{for any } t \in [a, b] \text{ and } j = 1, 2, \dots, d.$$

For each $j \in \{1, 2, \dots, d\}$, by Theorem 2.1, there exists a function $\psi_{i,j}$ generated by an EUAF network with width 36 and depth 5 such that

$$|h_{i,j}(t) - \psi_{i,j}(t)| < \delta/d \quad \text{for any } t \in [a, b].$$

Define $\psi_i(\mathbf{x}) := \sum_{j=1}^d \psi_{i,j}(x_j)$ for any $\mathbf{x} = [x_1, x_2, \dots, x_d]^T \in [a, b]^d$. Then, for any $\mathbf{x} = [x_1, x_2, \dots, x_d]^T \in [a, b]^d$, we have

$$|h_i(\mathbf{x}) - \psi_i(\mathbf{x})| = \left| \sum_{j=1}^d h_{i,j}(x_j) - \sum_{j=1}^d \psi_{i,j}(x_j) \right| = \sum_{j=1}^d |h_{i,j}(x_j) - \psi_{i,j}(x_j)| < \sum_{j=1}^d \delta/d = \delta.$$

Step 3: Integrate sub-networks.

Finally, we build an integrated network with the desired size to approximate the target function f . The desired function ϕ can be defined as

$$\phi(\mathbf{x}) := \sum_{i=0}^{2d} \phi_i \circ \psi_i(\mathbf{x}) = \sum_{i=0}^{2d} \phi_i \left(\sum_{j=1}^d \psi_{i,j}(x_j) \right) \quad \text{for any } \mathbf{x} = [x_1, x_2, \dots, x_d]^T \in [a, b]^d.$$

Let us first estimate the approximation error and then determine the size of the target network realizing ϕ . See Figure 3 for an illustration of the target network realizing ϕ for the case $d = 2$.

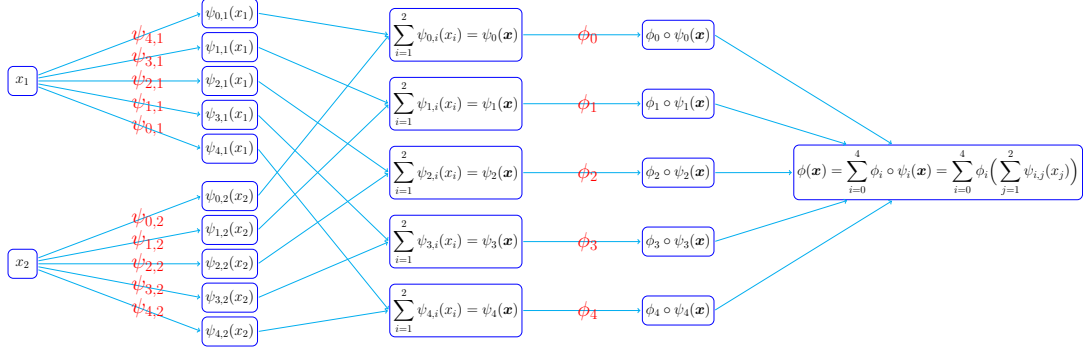


Figure 3: An illustration of the target network realizing ϕ for any $\mathbf{x} \in [a, b]^d$ in the case of $d = 2$. This network contains $(2d + 1)d + (2d + 1) = (d + 1)(2d + 1)$ sub-networks that realize $\psi_{i,j}$ and ϕ_i for $i = 0, 1, \dots, 2d$ and $j = 1, 2, \dots, d$.

Fix $\mathbf{x} \in [a, b]^d$ and $i \in \{0, 1, \dots, 2d\}$. Recall that $h_i(\mathbf{x}) \in [\frac{1}{4}, \frac{3}{4}]$ and $|h_i(\mathbf{x}) - \psi_i(\mathbf{x})| < \delta \leq \frac{1}{4}$, which implies $\psi_i(\mathbf{x}) \in [0, 1]$. Then by Equation (2.3) (set $z_1 = h_i(\mathbf{x})$ and $z_2 = \psi_i(\mathbf{x})$ therein), we have

$$|g_i \circ h_i(\mathbf{x}) - g_i \circ \psi_i(\mathbf{x})| = |g_i(h_i(\mathbf{x})) - g_i(\psi_i(\mathbf{x}))| < \varepsilon / (4d + 2).$$

By Equation (2.4) (set $z = \psi_i(\mathbf{x}) \in [0, 1]$ therein), we have

$$|g_i \circ \psi_i(\mathbf{x}) - \phi_i \circ \psi_i(\mathbf{x})| = |g_i(\psi_i(\mathbf{x})) - \phi_i(\psi_i(\mathbf{x}))| < \varepsilon / (4d + 2).$$

Therefore, for any $\mathbf{x} \in [a, b]^d$, we have

$$\begin{aligned} |f(\mathbf{x}) - \phi(\mathbf{x})| &= \left| \sum_{i=0}^{2d} g_i \circ h_i(\mathbf{x}) - \sum_{i=0}^{2d} \phi_i \circ \psi_i(\mathbf{x}) \right| = \sum_{i=0}^{2d} |g_i \circ h_i(\mathbf{x}) - \phi_i \circ \psi_i(\mathbf{x})| \\ &\leq \sum_{i=0}^{2d} \left(|g_i \circ h_i(\mathbf{x}) - g_i \circ \psi_i(\mathbf{x})| + |g_i \circ \psi_i(\mathbf{x}) - \phi_i \circ \psi_i(\mathbf{x})| \right) \\ &< \sum_{i=0}^{2d} \left(\varepsilon / (4d + 2) + \varepsilon / (4d + 2) \right) = \varepsilon. \end{aligned}$$

It remains to show ϕ can be generated by an EUAF network with the desired size. Recall that, for each $i \in \{0, 1, \dots, 2d\}$ and each $j \in \{1, 2, \dots, d\}$, $\psi_{i,j}$ can be generated by an EUAF network with width 36, depth 5, and, therefore, at most

$$(36 + 36) + (36 \times 36 + 36) \times 4 + (36 + 1) = 5437$$

nonzero parameters. Hence, for each $i \in \{0, 1, \dots, 2d\}$, ψ_i , given by $\psi_i(\mathbf{x}) = \sum_{j=1}^d \psi_{i,j}(x_j)$, can be generated by an EUAF network with width $36d$, depth 5, and at most $5437d$ nonzero parameters.

Since $\psi_i(\mathbf{x}) \in [0, 1]$ for any $\mathbf{x} \in [a, b]^d$ and $i = 0, 1, \dots, 2d$, we have $\sigma(\psi_i(\mathbf{x})) = \psi_i(\mathbf{x})$ for any $\mathbf{x} \in [a, b]^d$. Hence, $\phi_i \circ \psi_i$ can be generated by an EUAF network as visualized in Figure 4.

Recall that ϕ_i can be generated by an EUAF network with width 36 and depth 5. Hence, the network generating ϕ_i has at most 5437 nonzero parameters. As we can see

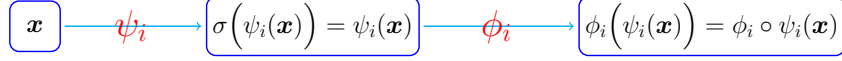


Figure 4: An illustration of the target EUAF network generating $\phi_i \circ \psi_i(\mathbf{x})$ for any $\mathbf{x} \in [a, b]^d$ and $i = 0, 1, \dots, 2d$.

from Figure 4, $\phi_i \circ \psi_i$ can be generated by an EUAF network with width $36d$, depth $5 + 1 + 5 = 11$, and at most $5437d + 5437 = 5437(d + 1)$ nonzero parameters. This means $\phi = \sum_{i=0}^{2d} \phi_i \circ \psi_i$ can be generated by an EUAF network with width $36d(2d + 1)$, depth 11, and, therefore, at most $5437(d + 1)(2d + 1)$ nonzero parameters as desired. So we finish the proof.

2.4 Proof of Theorem 1.3

The proof of Theorem 1.3 relies on a basic result of real analysis given in the following lemma.

Lemma 2.4. *Suppose $A, B \subseteq \mathbb{R}^d$ are two disjoint bounded closed sets. Then there exists a continuous function $f \in C(\mathbb{R}^d)$ such that $f(\mathbf{x}) = 1$ for any $\mathbf{x} \in A$ and $f(\mathbf{y}) = 0$ for any $\mathbf{y} \in B$.*

Proof. Define $\text{dist}(\mathbf{x}, A) = \inf\{\|\mathbf{x} - \mathbf{y}\|_2 : \mathbf{y} \in A\}$ and $\text{dist}(\mathbf{x}, B) = \inf\{\|\mathbf{x} - \mathbf{y}\|_2 : \mathbf{y} \in B\}$ for any $\mathbf{x} \in \mathbb{R}^d$. It is easy to verify that $\text{dist}(\mathbf{x}, A)$ and $\text{dist}(\mathbf{x}, B)$ are continuous in $\mathbf{x} \in \mathbb{R}^d$. Since $A, B \subseteq \mathbb{R}^d$ are two disjoint bounded closed subsets, we have $\text{dist}(\mathbf{x}, A) + \text{dist}(\mathbf{x}, B) > 0$ for any $\mathbf{x} \in \mathbb{R}^d$. Finally, define

$$f(\mathbf{x}) := \frac{\text{dist}(\mathbf{x}, B)}{\text{dist}(\mathbf{x}, A) + \text{dist}(\mathbf{x}, B)} \quad \text{for any } \mathbf{x} \in \mathbb{R}^d.$$

Then f meets the requirements. So we finish the proof. \square

With Lemma 2.4, we can prove Theorem 1.3.

Proof of Theorem 1.3. For any $f = \sum_{j=1}^J r_j \cdot \mathbb{1}_{E_j} \in \mathcal{C}_d(E_1, E_2, \dots, E_J)$, our goal is to construct a function ϕ generated by a σ -activated network such that $\phi(\mathbf{x}) = f(\mathbf{x})$ for any $\mathbf{x} \in \bigcup_{j=1}^J E_j$, where E_1, E_2, \dots, E_J are pairwise disjoint bounded closed subsets of \mathbb{R}^d . Set $E := \bigcup_{j=1}^J E_j$ and choose $a, b \in \mathbb{R}$ properly such that $E \subseteq [a, b]^d$.

For each $j \in \{1, 2, \dots, J\}$, E_j and $\tilde{E}_j := E \setminus E_j$ are two disjoint bounded closed subsets. Then, for each j , by Lemma 2.4, there exists $g_j \in C(\mathbb{R}^d)$ such that $g_j(\mathbf{x}) = 1$ for any $\mathbf{x} \in E_j$ and $g_j(\mathbf{y}) = 0$ for any $\mathbf{y} \in \tilde{E}_j$. By defining $g := \sum_{j=1}^J r_j \cdot g_j \in C(\mathbb{R}^d)$, we have $g(\mathbf{x}) = \sum_{j=1}^J r_j \cdot \mathbb{1}_{E_j}(\mathbf{x}) = f(\mathbf{x})$ for any $\mathbf{x} \in E = \bigcup_{j=1}^J E_j$.

Since r_1, r_2, \dots, r_J are rational numbers and $g : [a, b]^d \rightarrow \mathbb{R}$ is continuous, there exist $n_1, n_2 \in \mathbb{Z}$ such that

- $n_1 \cdot r_j + n_2 \in \mathbb{N}^+$ for $j = 1, 2, \dots, J$;
- $n_1 \cdot g(\mathbf{x}) + n_2 \geq 0$ for any $\mathbf{x} \in [a, b]^d$.

By applying Theorem 1.1 to $2(n_1 \cdot g + n_2) + 1$, there exists a function ϕ_1 generated by an EUAF network with width $36d(2d+1)$, depth 11, and at most $5437(d+1)(2d+1)$ nonzero parameters such that

$$\left| 2(n_1 \cdot g(\mathbf{x}) + n_2) + 1 - \phi_1(\mathbf{x}) \right| \leq 1/2 \quad \text{for any } \mathbf{x} \in [a, b]^d. \quad (2.5)$$

It follows that

$$\left| 2\left(n_1 \cdot \sum_{j=1}^J r_j \cdot \mathbf{1}_{E_j}(\mathbf{x}) + n_2\right) + 1 - \phi_1(\mathbf{x}) \right| \leq 1/2 \quad \text{for any } \mathbf{x} \in E = \bigcup_{j=1}^J E_j.$$

Since E_1, E_2, \dots, E_J are pairwise disjoint, we have

$$\left| 2(n_1 \cdot r_j + n_2) + 1 - \phi_1(\mathbf{x}) \right| \leq 1/2 \quad \text{for any } \mathbf{x} \in E_j \text{ and each } j \in \{1, 2, \dots, J\}. \quad (2.6)$$

Define $\phi_2(x) = x + 1/2 - \sigma(x + 3/2)$ for any $x \in \mathbb{R}$. See Figure 5 for an illustration. It is

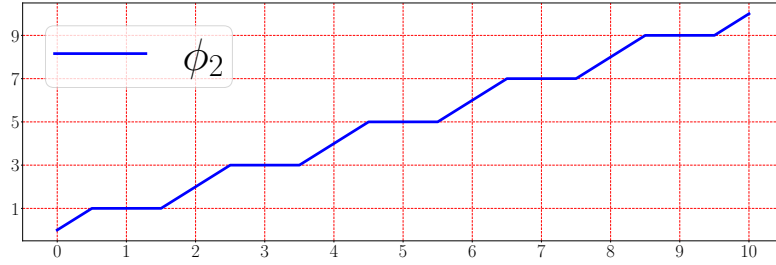


Figure 5: An illustration of ϕ_2 on $[0, 10]$.

easy to verify that

$$\phi_2(y) = 2k + 1 \quad \text{for any } y \text{ and } k \in \mathbb{N}^+ \text{ with } |2k + 1 - y| \leq 1/2. \quad (2.7)$$

Therefore, by Equations (2.6) and (2.7) (set $y = \phi_1(\mathbf{x})$ and $k = n_1 \cdot r_j + n_2$ therein), we have $\phi_2 \circ \phi_1(\mathbf{x}) = 2(n_1 \cdot r_j + n_2) + 1$ for any $\mathbf{x} \in E_j$ and any $j \in \{1, 2, \dots, J\}$, which implies

$$\frac{\phi_2 \circ \phi_1(\mathbf{x}) - 2n_2 - 1}{2n_1} = r_j \quad \text{for any } \mathbf{x} \in E_j \text{ and any } j \in \{1, 2, \dots, J\}.$$

Define

$$\phi(\mathbf{x}) := \frac{\phi_2 \circ \phi_1(\mathbf{x}) - 2n_2 - 1}{2n_1} \quad \text{for any } \mathbf{x} \in [a, b]^d.$$

Clearly, we have $\phi(\mathbf{x}) = r_j$ for any $\mathbf{x} \in E_j$ and each $j \in \{1, 2, \dots, J\}$, which implies $\phi(\mathbf{x}) = \sum_{j=1}^J r_j \cdot \mathbf{1}_{E_j}(\mathbf{x}) = f(\mathbf{x})$ for any $\mathbf{x} \in E = \bigcup_{j=1}^J E_j$ as desired.

Set $M = 2\|n_1 g + n_2\|_{L^\infty([a, b]^d)} + 3/2 > 0$. By Equation (2.5) and the fact $n_1 \cdot g(\mathbf{x}) + n_2 \geq 0$ for any $\mathbf{x} \in [a, b]^d$, we have

$$\phi_1(\mathbf{x}) \in [1/2, 2\|n_1 g + n_2\|_{L^\infty([a, b]^d)} + 1 + 1/2] \subseteq [0, M] \quad \text{for any } \mathbf{x} \in [a, b]^d.$$

Then, for any $\mathbf{x} \in [a, b]^d$, we have

$$\phi_2 \circ \phi_1(\mathbf{x}) = \phi_1(\mathbf{x}) + 1/2 - \sigma(\phi_1(\mathbf{x}) + 3/2) = M\sigma(\phi_1(\mathbf{x})/M) + 1/2 - \sigma(\phi_1(\mathbf{x}) + 3/2).$$

595 It follows that

$$596 \quad \phi(\mathbf{x}) = \frac{\phi_2 \circ \phi_1(\mathbf{x}) - 2n_2 - 1}{2n_1} = \frac{M\sigma(\phi_1(\mathbf{x})/M) - \sigma(\phi_1(\mathbf{x}) + 3/2) - 2n_2 - 1/2}{2n_1},$$

597 for any $\mathbf{x} \in [a, b]^d$. The network realizing ϕ has just one more hidden layer with 2
 598 neurons, compared to the network realizing ϕ_1 . Recall that ϕ_1 can be generated by an
 599 EUAF network with width $36d(2d+1)$, depth 11, and at most $5437(d+1)(2d+1)$ nonzero
 600 parameters. Therefore, ϕ , limited on $[a, b]^d$, can be generated by an EUAF network with
 601 width $36d(2d+1)$, depth 12, and at most

$$602 \quad 5437(d+1)(2d+1) + \underbrace{36d(2d+1) \times 2 + 2 + 2 + 1}_{\text{all possible new parameters}} \leq 5509(d+1)(2d+1)$$

603 nonzero parameters. So we finish the proof. \square

604 3 Proof of Theorem 2.1

605 To prove Theorem 2.1, we need to introduce two auxiliary theorems, Theorems 3.1
 606 and 3.2, which serve as two important intermediate steps.

607 **Theorem 3.1.** *Let $f \in C([0, 1])$ be a continuous function. Given any $\varepsilon > 0$, if K is a*
 608 *positive integer satisfying*

$$609 \quad |f(x_1) - f(x_2)| < \varepsilon/2 \quad \text{for any } x_1, x_2 \in [0, 1] \text{ with } |x_1 - x_2| < 1/K, \quad (3.1)$$

610 *then there exists a function ϕ generated by an EUAF network with width 2 and depth 3*
 611 *such that $\|\phi\|_{L^\infty([0,1])} \leq \|f\|_{L^\infty([0,1])} + 1$ and*

$$612 \quad |\phi(x) - f(x)| < \varepsilon \quad \text{for any } x \in \bigcup_{k=0}^{K-1} \left[\frac{2k}{2K}, \frac{2k+1}{2K} \right].$$

613 **Theorem 3.2.** *Let $f \in C([0, 1])$ be a continuous function. Then, for an arbitrary $\varepsilon > 0$,*
 614 *there exists a function ϕ generated by an EUAF network with width 36 and depth 5 such*
 615 *that^③*

$$616 \quad |\phi(x) - f(x)| < \varepsilon \quad \text{for any } x \in [0, \frac{9}{10}].$$

617 To prove Theorem 3.1, we only need to care about the approximation on “half” of
 618 $[0, 1]$. It is not necessary to care about the approximation on the other “half” of $[0, 1]$.
 619 Such an idea is similar to the “trifling region” in [25, 41]. As we shall see later, the proof
 620 of Theorem 3.1 can eventually be converted to a point-fitting problem, which can be
 621 solved by applying Proposition 2.2.

622 The key idea to prove Theorem 3.2 is to apply Theorem 3.1 to several horizontally
 623 shifted variants of the target function. Then a good approximation can be constructed
 624 via the combinations and multiplications of these variants, similar to the idea of [25, 41]

^③Theorem 3.2 still holds via replacing $\frac{9}{10}$ by any number in $[0, 1)$. In fact, it is true for $[0, \frac{1}{K}]$, and K can be arbitrarily large.

to obtain an error estimation with the L^∞ -norm from a result with the L^p -norm for $p \in [1, \infty)$.

The proofs of Theorems 3.1 and 3.2 will be presented in Sections 3.1 and 3.2, respectively. Let us first prove Theorem 2.1 by assuming Theorem 3.2 is true.

Proof of Theorem 2.1. Define a linear function \mathcal{L} by $\mathcal{L}(x) = a + \frac{10(b-a)}{9}x$ for any $x \in [0, \frac{9}{10}]$. Then \mathcal{L} is a bijection from $[0, \frac{9}{10}]$ to $[a, b]$. It follows that $f \circ \mathcal{L}$ is a continuous function on $[0, \frac{9}{10}]$. By Theorem 3.2, there exists a function $\tilde{\phi}$ generated by an EUAF network with width 36 and depth 5 such that

$$|f \circ \mathcal{L}(x) - \tilde{\phi}(x)| < \varepsilon \quad \text{for any } x \in [0, \frac{9}{10}].$$

Define $\tilde{\mathcal{L}}(y) = \frac{9(y-a)}{10(b-a)}$ for any $y \in [a, b]$. Clearly, it is the inverse of \mathcal{L} , i.e., $\mathcal{L} \circ \tilde{\mathcal{L}}(y) = y$ for any $y \in [a, b]$. Therefore, for any $y \in [a, b]$, we have $x = \tilde{\mathcal{L}}(y) \in [0, \frac{9}{10}]$, which implies

$$\begin{aligned} |f(y) - \tilde{\phi} \circ \tilde{\mathcal{L}}(y)| &= |f \circ \mathcal{L} \circ \tilde{\mathcal{L}}(y) - \tilde{\phi} \circ \tilde{\mathcal{L}}(y)| \\ &= |f \circ \mathcal{L}(\tilde{\mathcal{L}}(y)) - \tilde{\phi}(\tilde{\mathcal{L}}(y))| \leq |f \circ \mathcal{L}(x) - \tilde{\phi}(x)| < \varepsilon. \end{aligned}$$

By defining $\phi := \tilde{\phi} \circ \tilde{\mathcal{L}}$, we have $|f(y) - \phi(y)| < \varepsilon$ for any $y \in [a, b]$ as desired. Note that $\tilde{\phi}$ can be realized by an EUAF network with width 36 and depth 5. We can compose $\tilde{\mathcal{L}}$ and the affine linear map of the network $\tilde{\phi}$ that connects the input layer and the first hidden layer. Therefore, $\phi = \tilde{\phi} \circ \tilde{\mathcal{L}}$ can also be realized by an EUAF network with width 36 and depth 5. So we finish the proof. \square

3.1 Proof of Theorem 3.1

Partition $[0, 1]$ into $2K$ small intervals \mathcal{I}_k and $\tilde{\mathcal{I}}_k$ for $k = 1, 2, \dots, K$, i.e.,

$$\mathcal{I}_k = \left[\frac{2k-2}{2K}, \frac{2k-1}{2K} \right] \quad \text{and} \quad \tilde{\mathcal{I}}_k = \left[\frac{2k-1}{2K}, \frac{2k}{2K} \right].$$

Clearly, $[0, 1] = \bigcup_{k=1}^K (\mathcal{I}_k \cup \tilde{\mathcal{I}}_k)$. Let x_k be the right endpoint of \mathcal{I}_k , i.e., $x_k = \frac{2k-1}{2K}$ for $k = 1, 2, \dots, K$. See an illustration of \mathcal{I}_k , $\tilde{\mathcal{I}}_k$, and x_k in Figure 6 for the case $K = 5$.

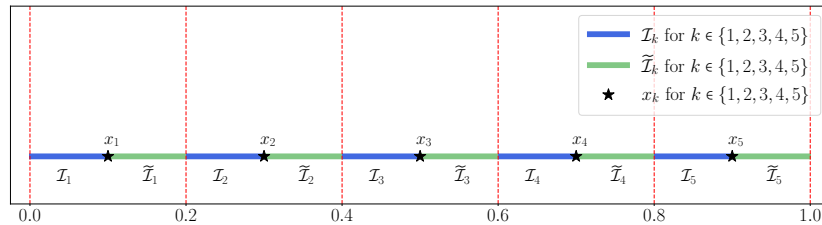


Figure 6: An illustration of \mathcal{I}_k and $\tilde{\mathcal{I}}_k$ for $k \in \{1, 2, \dots, K\}$ with $K = 5$.

Our goal is to construct a function ϕ generated by an EUAF network with the desired size to approximate f well on \mathcal{I}_k for $k = 1, 2, \dots, K$. It is not necessary to care about the values of ϕ on $\tilde{\mathcal{I}}_k$ for all k . In other words, we only need to care about the approximation on a “half” of $[0, 1]$, which is the key for our proof.

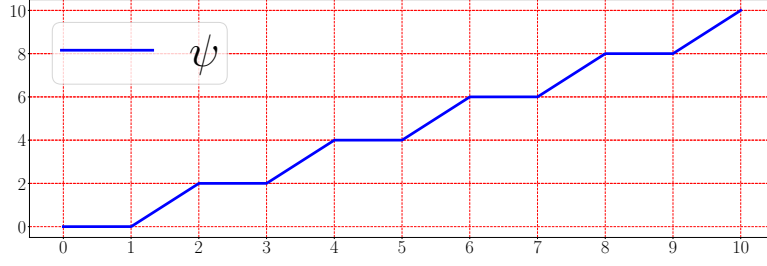


Figure 7: An illustration of ψ on $[0, 10]$.

Define $\psi(x) = x - \sigma(x)$ for any $x \in \mathbb{R}$, where σ is defined in Equation (1.3). See Figure 7 for an illustration of ψ .

It is easy to verify that

$$\psi(y) = 2k - 2 \quad \text{for any } y \in [2k - 2, 2k - 1] \text{ and each } k \in \{1, 2, \dots, K\}.$$

It follows that

$$\psi(2Kx)/2 + 1 = k \quad \text{for any } x \in [\frac{2k-2}{2K}, \frac{2k-1}{2K}] = \mathcal{I}_k \text{ and each } k \in \{1, 2, \dots, K\}.$$

Recall that x_k is the right endpoint of \mathcal{I}_k for $k = 1, 2, \dots, K$. Set $M = \|f\|_{L^\infty([0,1])} + 1$ and define

$$\xi_k := \frac{f(x_k) + M}{2M} \in [0, 1] \quad \text{for } k = 1, 2, \dots, K.$$

Then $[\xi_1, \xi_2, \dots, \xi_K]^T$ is in $[0, 1]^K$. By Proposition 2.2, there exists $w_0 \in \mathbb{R}$ such that

$$\left| \sigma_1\left(\frac{w_0}{\pi+k}\right) - \xi_k \right| < \varepsilon/(4M) \quad \text{for } k = 1, 2, \dots, K.$$

Let m_0 be an integer larger than $|w_0|$, e.g., set $m_0 = \lfloor |w_0| \rfloor + 1$. It is easy to verify that

$$\frac{w_0}{\pi+k} + 2m_0 \geq 0 \quad \text{for any } x \in [0, 1].$$

Since $\sigma(x) = \sigma_1(x)$ for $x \geq 0$ and σ_1 is periodic with period 2, we have

$$\left| \sigma\left(\frac{w_0}{\pi+k} + 2m_0\right) - \xi_k \right| = \left| \sigma_1\left(\frac{w_0}{\pi+k} + 2m_0\right) - \xi_k \right| = \left| \sigma_1\left(\frac{w_0}{\pi+k}\right) - \xi_k \right| < \varepsilon/(4M),$$

for $k = 1, 2, \dots, K$. It follows that

$$\begin{aligned} \left| 2M\sigma\left(\frac{w_0}{\pi+k} + 2m_0\right) - M - f(x_k) \right| &= \left| 2M\sigma\left(\frac{w_0}{\pi+k} + 2m_0\right) - M - (2M\xi_k - M) \right| \\ &= 2M \left| \sigma\left(\frac{w_0}{\pi+k} + 2m_0\right) - \xi_k \right| < 2M \frac{\varepsilon}{4M} = \varepsilon/2, \end{aligned} \tag{3.2}$$

for $k = 1, 2, \dots, K$.

The desired ϕ is defined as

$$\phi(x) := 2M\sigma\left(\frac{w_0}{\pi+\psi(2Kx)/2+1} + 2m_0\right) - M \quad \text{for any } x \in [0, 1].$$

Recall that $m_0 \geq |w_0|$ and $\psi(x) \geq 0$ for any $x \geq 0$, which implies $\frac{w_0}{\pi+\psi(2Kx)/2+1} + 2m_0 \geq 0$ for any $x \in [0, 1]$. Thus, $\|\phi\|_{L^\infty([0,1])} \leq M = \|f\|_{L^\infty([0,1])} + 1$ since $0 \leq \sigma(y) \leq 1$ for any $y \geq 0$.

For any $x \in \mathcal{I}_k$ and each $k \in \{1, 2, \dots, K\}$, we have $\psi(2Kx)/2 + 1 = k$, which implies

$$\phi(x) = 2M\sigma\left(\frac{w_0}{\pi+\psi(2Kx)/2+1} + 2m_0\right) - M = 2M\sigma\left(\frac{w_0}{\pi+k} + 2m_0\right) - M.$$

For any $x \in \mathcal{I}_k$ and each $k \in \{1, 2, \dots, K\}$, we have $|x_k - x| < 1/K$, which implies $|f(x_k) - f(x)| < \varepsilon/2$ by Equation (3.1). Therefore, by Equation (3.2), we have

$$\begin{aligned} |\phi(x) - f(x)| &= \left| 2M\sigma\left(\frac{w_0}{\pi+k} + 2m_0\right) - M - f(x) \right| \\ &\leq \left| 2M\sigma\left(\frac{w_0}{\pi+k} + 2m_0\right) - M - f(x_k) \right| + |f(x_k) - f(x)| < \varepsilon/2 + \varepsilon/2 = \varepsilon, \end{aligned}$$

for any $x \in \mathcal{I}_k$ and each $k \in \{1, 2, \dots, K\}$. It follows that

$$|\phi(x) - f(x)| < \varepsilon \quad \text{for any } x \in \bigcup_{j=1}^K \mathcal{I}_j = \bigcup_{j=1}^K \left[\frac{2j-2}{2K}, \frac{2j-1}{2K} \right] = \bigcup_{k=0}^{K-1} \left[\frac{2k}{2K}, \frac{2k+1}{2K} \right].$$

It remains to show that ϕ can be generated by an EUAF network with the desired size. Observe that

$$\sigma(y) + 1 = \frac{y}{|y| + 1} + 1 = \frac{y}{-y + 1} + 1 = \frac{1}{-y + 1} \quad \text{for any } y \leq 0.$$

By setting $y = -\pi - \psi(2Kx)/2 \leq 0$ for any $x \in [0, 1]$, we have

$$\begin{aligned} \frac{1}{\pi+\psi(2Kx)/2+1} &= \frac{1}{-y+1} = \sigma(y) + 1 = \sigma(-\pi - \psi(2Kx)/2) + 1 \\ &= \sigma(-\pi - (2Kx - \sigma(2Kx))/2) + 1 \\ &= \sigma(-\pi - Kx + \sigma(2Kx)/2) + 1, \end{aligned}$$

where the large equality comes from $\psi(z) = z - \sigma(z)$ for any $z \in \mathbb{R}$. Therefore, we get

$$\begin{aligned} \phi(x) &= 2M\sigma\left(\frac{w_0}{\pi+\psi(2Kx)/2+1} + 2m_0\right) - M \\ &= 2M\sigma\left(w_0\sigma\left(-\pi - Kx + \sigma(2Kx)/2\right) + w_0 + 2m_0\right) - M. \end{aligned} \tag{3.3}$$

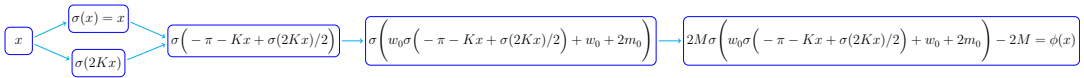


Figure 8: An illustration of the target EUAF network realizing $\phi(x)$ for $x \in [0, 1]$ based on Equation (3.3).

Thus, the desired EUAF network realizing ϕ is shown in Figure 8. Clearly, the network in Figure 8 has width 2 and depth 3 as desired. It is easy to verify that the network architecture of ϕ is independent of the target function f and the desired error ε . That is, we can fix the architecture and only adjust parameters to achieve the desired approximation error. So we finish the proof.

3.2 Proof of Theorem 3.2

The key idea of proving Theorem 3.2 is to apply Theorem 3.1 to several horizontally shifted variants of the target function. Then a good approximation can be expected via combinations and multiplications of these variants. Thus, we need to reproduce $f(x, y) = xy$ locally via an EUAF network as shown in the following lemma.

Lemma 3.3. *For any $M > 0$, there exists a function ϕ generated by an EUAF network with width 9 and depth 2 such that*

$$\phi(x, y) = xy \quad \text{for any } x, y \in [-M, M].$$

The proof of this lemma is given in Section 3.3. Now let us first prove Theorem 3.2 by assuming this lemma is true.

Proof of Theorem 3.2. Set $\tilde{\varepsilon} = \varepsilon/4$ and extend f from $[0, 1]$ to $[-1, 1]$ by defining $f(x) = f(0)$ for $x \in [-1, 0)$. Then f is continuous on $[-1, 1]$, and, therefore, uniformly continuous. Thus, there exists $K = K(f, \varepsilon) \in \mathbb{N}^+$ with $K \geq 10$ such that

$$|f(x_1) - f(x_2)| < \tilde{\varepsilon}/2 \quad \text{for any } x_1, x_2 \in [-1, 1] \text{ with } |x_1 - x_2| < 1/K.$$

For $i = 1, 2, 3, 4$, define

$$f_i(x) := f\left(x - \frac{i}{4K}\right) \quad \text{for any } x \in [0, 1].$$

For each $i \in \{1, 2, 3, 4\}$ and any $x_1, x_2 \in [0, 1]$ with $|x_1 - x_2| < 1/K$, we have $x_1 - \frac{i}{4K}, x_2 - \frac{i}{4K} \in [-1, 1]$ and $\left|(x_1 - \frac{i}{4K}) - (x_2 - \frac{i}{4K})\right| = |x_1 - x_2| < 1/K$, which implies

$$|f_i(x_1) - f_i(x_2)| = |f(x_1 - \frac{i}{4K}) - f(x_2 - \frac{i}{4K})| < \tilde{\varepsilon}/2.$$

That is, for $i = 1, 2, 3, 4$, we have

$$|f_i(x_1) - f_i(x_2)| < \tilde{\varepsilon}/2 \quad \text{for any } x_1, x_2 \in [0, 1] \text{ with } |x_1 - x_2| < 1/K.$$

For each $i \in \{1, 2, 3, 4\}$, by Theorem 3.1, there exist a function ϕ_i generated by an EUAF network with width 2 and depth 3 such that $\|\phi_i\|_{L^\infty([0,1])} \leq \|f_i\|_{L^\infty([0,1])} + 1 \leq \|f\|_{L^\infty([-1,1])} + 1$ and

$$|\phi_i(x) - f_i(x)| < \tilde{\varepsilon} = \varepsilon/4 \quad \text{for any } x \in \bigcup_{k=0}^{K-1} \left[\frac{2k}{2K}, \frac{2k+1}{2K}\right].$$

Define

$$\psi(x) = \sigma(x + 1 - \sigma(x + 1)) \quad \text{for any } x \in \mathbb{R}.$$

See an illustration of ψ on $[0, 2K]$ for $K = 5$ in Figure 9.

Clearly, $0 \leq \psi(2Kx) \leq 1$ for any $x \in [0, 1]$, which results in

$$\left|(\phi_i(x) - f_i(x))\psi(2Kx)\right| \leq |\phi_i(x) - f_i(x)| < \varepsilon/4 \quad \text{for any } x \in \bigcup_{k=0}^{K-1} \left[\frac{2k}{2K}, \frac{2k+1}{2K}\right].$$

Observe that $\psi(y) = 0$ for $y \in \bigcup_{k=0}^{K-1} [2k+1, 2k+2]$, which implies

$$\psi(2Kx) = 0 \quad \text{for any } x \in \bigcup_{k=0}^{K-1} \left[\frac{2k+1}{2K}, \frac{2k+2}{2K}\right] \supseteq [0, 1] \setminus \bigcup_{k=0}^{K-1} \left[\frac{2k}{2K}, \frac{2k+1}{2K}\right].$$

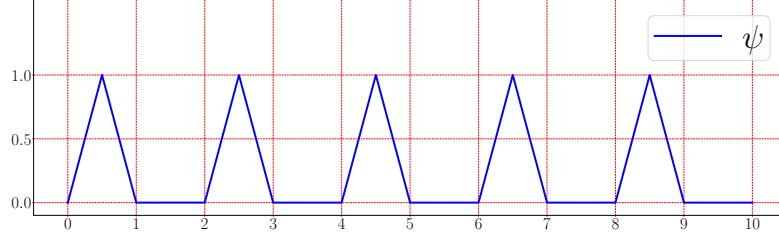


Figure 9: An illustration of ψ on $[0, 2K]$ for $K = 5$.

It follows that

$$\left| (\phi_i(x) - f_i(x))\psi(2Kx) \right| < \varepsilon/4 \quad \text{for any } x \in [0, 1] \text{ and } i = 1, 2, 3, 4. \quad (3.4)$$

For each $i \in \{1, 2, 3, 4\}$ and any $z \in [0, \frac{9}{10}] \subseteq [0, 1 - \frac{i}{4K}]$, we have $y_i = z + \frac{i}{4K} \in [\frac{i}{4K}, 1] \subseteq [0, 1]$. Therefore, by bringing $y_i \in [0, 1]$ into Equation (3.4) (set $x = y_i$ therein), we have

$$\begin{aligned} \varepsilon/4 &> \left| (\phi_i(y_i) - f_i(y_i))\psi(2Ky_i) \right| = \left| \phi_i(y_i)\psi(2Ky_i) - f_i(y_i)\psi(2Ky_i) \right| \\ &= \left| \phi_i\left(z + \frac{i}{4K}\right)\psi\left(2K\left(z + \frac{i}{4K}\right)\right) - f_i\left(z + \frac{i}{4K}\right)\psi\left(2K\left(z + \frac{i}{4K}\right)\right) \right| \\ &= \left| \phi_i\left(z + \frac{i}{4K}\right)\psi\left(2Kz + \frac{i}{2}\right) - f\left(z\right)\psi\left(2Kz + \frac{i}{2}\right) \right|, \end{aligned} \quad (3.5)$$

where the last equality comes from the fact that $f_i(x) = f(x - \frac{i}{4K})$ for any $x \in [0, 1] \supseteq [\frac{i}{4K}, 1]$. The desired ϕ is defined as

$$\phi(x) := \sum_{i=1}^4 \phi_i\left(x + \frac{i}{4K}\right)\psi\left(2Kx + \frac{i}{2}\right) \quad \text{for any } x \in [0, \frac{9}{10}].$$

It is easy to verify that $\sum_{i=1}^4 \psi\left(x + \frac{i}{2}\right) = 1$ for any $x \geq 0$ based on the definition of ψ . See Figure 10 for illustrations. It follows that $\sum_{i=1}^4 \psi\left(2Kz + \frac{i}{2}\right) = 1$ for any $z \in [0, \frac{9}{10}]$.

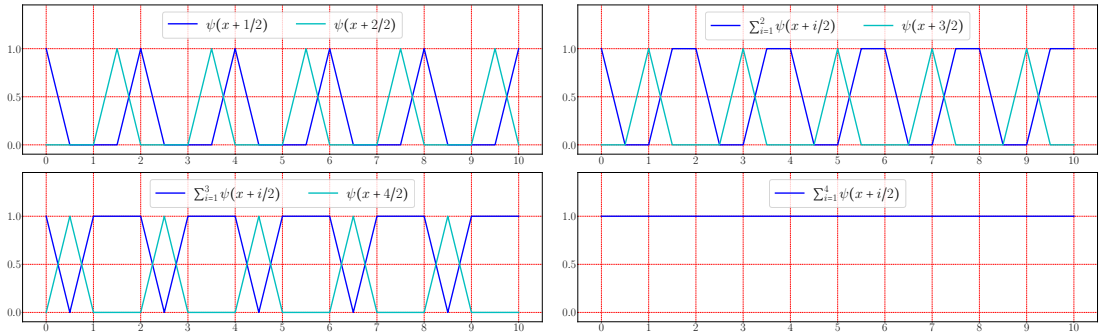


Figure 10: Illustrations of $\sum_{i=1}^4 \psi(x + i/2) = 1$ for any $x \in [0, 10]$.

Hence, by Equation (3.5), we have

$$\begin{aligned} |\phi(z) - f(z)| &= \left| \sum_{i=1}^4 \phi_i\left(z + \frac{i}{4K}\right)\psi\left(2Kz + \frac{i}{2}\right) - f(z) \sum_{i=1}^4 \psi\left(2Kz + \frac{i}{2}\right) \right| \\ &\leq \sum_{i=1}^4 \left| \phi_i\left(z + \frac{i}{4K}\right)\psi\left(2Kz + \frac{i}{2}\right) - f(z)\psi\left(2Kz + \frac{i}{2}\right) \right| < 4 \frac{\varepsilon}{4} = \varepsilon. \end{aligned}$$

737 That is, $|\phi(x) - f(x)| < \varepsilon$ for any $x \in [0, \frac{9}{10}]$ as desired. It remains to show that ϕ , limited
738 on $[0, \frac{9}{10}]$, can be generated by an EUAF network with the desired size.
739 Note that $x + 1 = (2K + 1)\sigma(\frac{x+1}{2K+1})$ for any $x \in [0, 2K]$, which implies

$$740 \quad \psi(x) = \sigma(x + 1 - \sigma(x + 1)) = \sigma((2K + 1)\sigma(\frac{x+1}{2K+1}) - \sigma(x + 1)).$$

741 This means ψ , limited on $[0, 2K]$, can be generated by an EUAF network with width
742 2 and depth 2. Since $0 \leq 2Kx + \frac{i}{2} \leq 2K\frac{9}{10} + 2 = 2K(\frac{9}{10} + \frac{1}{K}) \leq 2K$ for any $x \in [0, \frac{9}{10}]$,
743 $\psi(2K \cdot + \frac{i}{2})$, limited on $[0, \frac{9}{10}]$, can also be generated by an EUAF network with width 2
744 and depth 2.

745 Note that ϕ_i , limited on $[0, 1]$, can also be generated by an EUAF network with
746 width 2 and depth 3. Clearly, $x + \frac{i}{4K} \in [0, 1]$ for any $x \in [0, \frac{9}{10}]$, and, therefore, $\phi_i(\cdot + \frac{i}{4K})$,
747 limited on $[0, \frac{9}{10}]$, can also be generated by an EUAF network with width 2 and depth
748 3.

749 Recall that $\|\phi_i\|_{L^\infty([0,1])} \leq \|f\|_{L^\infty([-1,1])} + 1 =: M$. Thus, $|\phi_i(x + \frac{i}{4K})| \leq M$ and $|\psi(2Kx + \frac{i}{2})| \leq 1 \leq M$ for any $x \in [0, \frac{9}{10}]$ and $i = 1, 2, 3, 4$. By Lemma 3.3, there exists a function Γ
750 generated by an EUAF network with width 9 and depth 2 such that
751

$$752 \quad \Gamma(x, y) = xy \quad \text{for any } x, y \in [-M, M].$$

753 It follows that

$$754 \quad \Gamma\left(\phi_i\left(x + \frac{i}{4K}\right), \psi\left(2Kx + \frac{i}{2}\right)\right) = \phi_i\left(x + \frac{i}{4K}\right)\psi\left(2Kx + \frac{i}{2}\right) \quad \text{for } i = 1, 2, 3, 4.$$

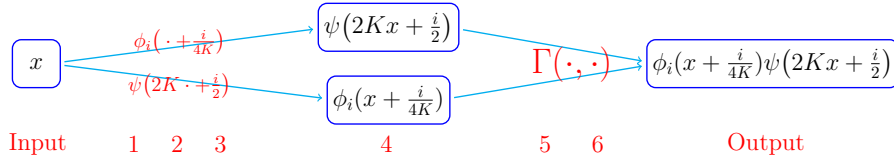


Figure 11: An illustration of the target EUAF network realizing each component of $\phi(x)$, $\phi_i(x + \frac{i}{4K})\psi(2Kx + \frac{i}{2})$, for any $x \in [0, \frac{9}{10}]$ and each $i \in \{1, 2, 3, 4\}$. The networks realizing $\phi_i(\cdot + \frac{i}{4K})$ and $\psi(2K \cdot + \frac{i}{2})$ can be placed in parallel since we can manually add a hidden layer to ψ since $\sigma \circ \psi(2Kx + \frac{i}{2}) = \psi(2Kx + \frac{i}{2})$ for any $x \in [0, \frac{9}{10}]$.

755 Therefore, each component of $\phi(x)$, $\phi_i(x + \frac{i}{4K})\psi(2Kx + \frac{i}{2})$ for some $i \in \{1, 2, 3, 4\}$,
756 can be generated by the network in Figure 11 for any $x \in [0, \frac{9}{10}]$. Clearly, such a network
757 has width 9 and depth 6. Since the 4-th hidden layer of the network in Figure 11 uses
758 identity as activation function for each neuron in this hidden layer, we can reduce the
759 depth by 1 via composing two adjacent affine linear maps to generate a new one. Thus,
760 the network in Figure 11 can be interpreted as an EUAF network with width 9 and
761 depth 5.

762 Note that ϕ is the sum of its four components, namely,

$$763 \quad \phi(x) = \sum_{i=1}^4 \phi_i\left(x + \frac{i}{4K}\right)\psi\left(2Kx + \frac{i}{2}\right) \quad \text{for any } x \in [0, \frac{9}{10}].$$

Therefore, ϕ , limited on $[0, \frac{9}{10}]$, can be generated by an EUAF network with width $9 \times 4 = 36$ and depth 5 as desired. It is easy to verify that the designed network architecture is independent of the target function f and the desired error ε . That is, we can fix the architecture and only adjust parameters to achieve an arbitrarily desired approximation error. So we finish the proof. \square

3.3 Proof of Lemma 3.3

The key idea of proving Lemma 3.3 is the polarization identity $2xy = (x+y)^2 - x^2 - y^2$. Thus, we need to reproduce x^2 locally by an EUAF network as shown in the following lemma.

Lemma 3.4. *There exists a function ϕ generated by an EUAF network with width 3 and depth 2 such that*

$$\phi(x) = x^2 \quad \text{for any } x \in [-1, 1].$$

Proof. Observe that

$$\sigma(y) + 1 = \frac{y}{|y| + 1} + 1 = \frac{y}{-y + 1} + 1 = \frac{1}{-y + 1} \quad \text{for any } y \leq 0.$$

For any $x \in [-1, 1]$, we have $-x - 1 \leq 0$ and $-x - 2 \leq 0$, which implies

$$\begin{aligned} \sigma(-x - 1) - \sigma(-x - 2) &= \left(\sigma(-x - 1) + 1 \right) - \left(\sigma(-x - 2) + 1 \right) \\ &= \frac{1}{-(-x - 1) + 1} - \frac{1}{-(-x - 2) + 1} = \frac{1}{x + 2} - \frac{1}{x + 3} = \frac{1}{(x + 2)(x + 3)}. \end{aligned}$$

It follows from $1 - \frac{12}{(x+2)(x+3)} \leq 0$ for any $x \in [-1, 1]$ that

$$\sigma\left(1 - \frac{12}{(x + 2)(x + 3)}\right) + 1 = \frac{1}{-\left(1 - \frac{12}{(x+2)(x+3)}\right) + 1} = \frac{x^2 + 5x + 6}{12},$$

implying

$$\begin{aligned} x^2 &= 12\sigma\left(1 - \frac{12}{(x + 2)(x + 3)}\right) + 12 - (5x + 6) \\ &= 12\sigma\left(1 - 12(\sigma(-x - 1) - \sigma(-x - 2))\right) + 11\frac{6 - 5x}{11} \\ &= 12\sigma\left(1 - 12\sigma(-x - 1) + 12\sigma(-x - 2)\right) + 11\sigma\left(\frac{6 - 5x}{11}\right) := \phi(x), \end{aligned}$$

where the equality $\frac{6-5x}{11} = \sigma\left(\frac{6-5x}{11}\right)$ comes from two facts: $\frac{6-5x}{11} \in [0, 1]$ since $x \in [-1, 1]$ and $\sigma(z) = z$ for any $z \in [0, 1]$.

Then, x^2 can be generated by the network shown in Figure 12 for any $x \in [-1, 1]$. The target network has width 3 and depth 2. So we finish the proof. \square

With Lemma 3.4 at hand, we are ready to prove Lemma 3.3.

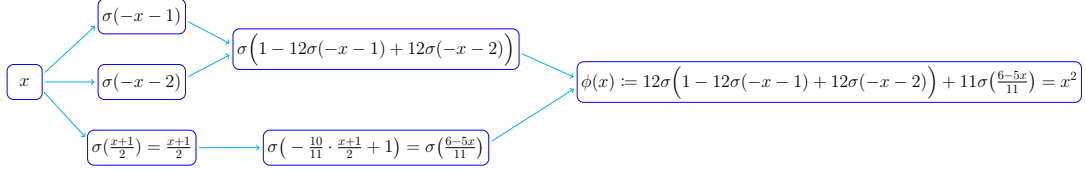


Figure 12: An illustration of the target EUAF network realizing $\phi(x) = x^2$ for $x \in [-1, 1]$.

789 *Proof of Lemma 3.3.* By Lemma 3.4, there exists a function $\tilde{\phi}$ generated by an EUAF
 790 network such that $\tilde{\phi}(t) = t^2$ for any $t \in [-1, 1]$. Thus, for any $x, y \in [-M, M]$, we have

$$\begin{aligned}
 xy &= 2M^2 \left(\left(\frac{x+y}{2M} \right)^2 - \left(\frac{x}{2M} \right)^2 - \left(\frac{y}{2M} \right)^2 \right) \\
 &= 2M^2 \left(\tilde{\phi}\left(\frac{x+y}{2M}\right) - \tilde{\phi}\left(\frac{x}{2M}\right) - \tilde{\phi}\left(\frac{y}{2M}\right) \right) =: \phi(x, y).
 \end{aligned}$$

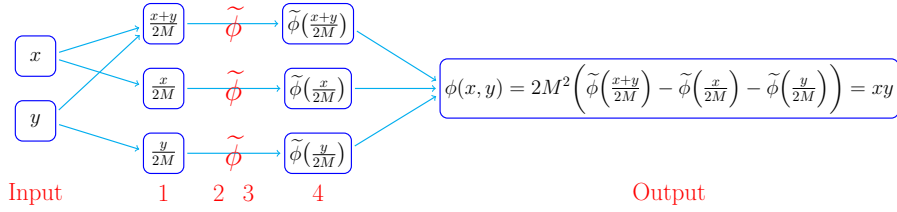


Figure 13: An illustration of the target network realizing $\phi(x, y) = xy$ for $x, y \in [-M, M]$. “ $\tilde{\phi}$ ” means the network realizing $\tilde{\phi}$, i.e., an EUAF network with width 3 and depth 2.

792 The target network realizing ϕ with width 9 and depth 4 is shown in Figure 13.
 793 Note that we can reduce the depth by one if the activation function of each neuron in
 794 a hidden layer is identity. In fact, we can eliminate this hidden layer by composing two
 795 adjacent affine linear maps to generate a new one. The 1-st and 4-th hidden layers in
 796 the network in Figure 13 use identity as an activation function. Thus, the network in
 797 Figure 13 can be interpreted as an EUAF network with width 9 and depth 2. So we
 798 finish the proof. \square

799 4 Proof of Proposition 2.2

800 We will prove Proposition 2.2 in this section. The proof includes two main steps.
 801 First, we show how to simply generate a set of rationally independent numbers in
 802 Lemma 4.1 below. Next, we prove that the target point set via a winding of the gener-
 803 ated rationally independent numbers is dense in a hypercube. Such proof relies on the
 804 fact that an irrational winding on the torus is dense (e.g., see Lemma 2 of [39]) as shown
 805 in Lemma 4.2 below in a hypercube.

806 **Lemma 4.1.** *Given any $K \in \mathbb{N}^+$, any transcendental number $\alpha \in \mathbb{R} \setminus \mathbb{A}$, and any pairwise*
 807 *distinct rational numbers $r_1, r_2, \dots, r_K \in \mathbb{Q}$, the set of numbers*

$$\left\{ \frac{1}{\alpha + r_k} : k = 1, 2, \dots, K \right\}$$

809 are rationally independent.

810 **Lemma 4.2.** *Given any rationally independent numbers a_1, a_2, \dots, a_K for any $K \in \mathbb{N}^+$*
 811 *and an arbitrary periodic function $g: \mathbb{R} \rightarrow \mathbb{R}$ with period T , i.e., $g(x+T) = g(x)$ for any*
 812 *$x \in \mathbb{R}$, assume there exist $x_1, x_2 \in \mathbb{R}$ with $0 < x_2 - x_1 < T$ such that g is continuous on*
 813 *$[x_1, x_2]$. Then the following set*

$$814 \quad \left\{ [g(wa_1), g(wa_2), \dots, g(wa_K)]^T : w \in \mathbb{R} \right\}$$

815 *is dense in $[M_1, M_2]^K$, where $M_1 = \min_{x \in [x_1, x_2]} g(x)$ and $M_2 = \max_{x \in [x_1, x_2]} g(x)$.*

816 The proofs of these two lemmas can be found in Sections 4.1 and 4.2, respectively.
 817 With these two lemmas at hand, the proof of Proposition 2.2 is straightforward. In fact,
 818 we can prove a more general result in Proposition 4.3 below, which implies Proposition 2.2
 819 immediately.

820 **Proposition 4.3.** *Given an arbitrary periodic function $g: \mathbb{R} \rightarrow \mathbb{R}$ with period T , i.e.,*
 821 *$g(x+T) = g(x)$ for any $x \in \mathbb{R}$, assume there exist $x_1, x_2 \in \mathbb{R}$ with $0 < x_2 - x_1 < T$ such that*
 822 *g is continuous on $[x_1, x_2]$. Then, for any $K \in \mathbb{N}^+$, any transcendental number $\alpha \in \mathbb{R} \setminus \mathbb{A}$,*
 823 *and any pairwise distinct rational numbers $r_1, r_2, \dots, r_K \in \mathbb{Q}$, the following set*

$$824 \quad \left\{ \left[g\left(\frac{w}{\alpha+r_1}\right), g\left(\frac{w}{\alpha+r_2}\right), \dots, g\left(\frac{w}{\alpha+r_K}\right) \right]^T : w \in \mathbb{R} \right\}$$

825 *is dense in $[M_1, M_2]^K$, where $M_1 = \min_{x \in [x_1, x_2]} g(x)$ and $M_2 = \max_{x \in [x_1, x_2]} g(x)$. In the case of*
 826 *$M_1 < M_2$, the following set*

$$827 \quad \left\{ \left[u \cdot g\left(\frac{w}{\alpha+r_1}\right) + v, u \cdot g\left(\frac{w}{\alpha+r_2}\right) + v, \dots, u \cdot g\left(\frac{w}{\alpha+r_K}\right) + v \right]^T : u, v, w \in \mathbb{R} \right\}$$

828 *is dense in \mathbb{R}^K .*

829 Clearly, Proposition 2.2 is a special case of Proposition 4.3 with $g = \sigma_1$, $\alpha = \pi$, $r_k = k$
 830 for $k = 1, 2, \dots, K$. The transcendence of π is well known (e.g., see Lindemann–Weierstrass
 831 theorem). By setting $x_1 = 0$ and $x_2 = 1$, we have $[M_1, M_2] = [0, 1]$ and σ_1 is continuous
 832 on $[0, 1]$, which means that the following set

$$833 \quad \left\{ \left[\sigma_1\left(\frac{w}{\pi+1}\right), \sigma_1\left(\frac{w}{\pi+2}\right), \dots, \sigma_1\left(\frac{w}{\pi+K}\right) \right]^T : w \in \mathbb{R} \right\}$$

834 *is dense in $[0, 1]^K$ as desired.*

835 Finally, let us prove Proposition 4.3 by assuming Lemmas 4.2 and 4.1 are true.

836 *Proof of Proposition 4.3.* By Lemma 4.1, the set of numbers

$$837 \quad \left\{ \frac{1}{\alpha+r_k} : k = 1, 2, \dots, K \right\}$$

838 are rationally independent. Denote $a_k = \frac{1}{\alpha+r_k}$ for $k = 1, 2, \dots, K$. Then, by Lemma 4.2,

$$839 \quad \begin{aligned} & \left\{ [g(wa_1), g(wa_2), \dots, g(wa_K)]^T : w \in \mathbb{R} \right\} \\ &= \left\{ \left[g\left(\frac{w}{\alpha+r_1}\right), g\left(\frac{w}{\alpha+r_2}\right), \dots, g\left(\frac{w}{\alpha+r_K}\right) \right]^T : w \in \mathbb{R} \right\} \end{aligned}$$

is dense in $[M_1, M_2]^K$. Now consider the case $M_1 < M_2$ for the latter result. For any $\varepsilon > 0$ and any $\mathbf{x} \in \mathbb{R}^K$, by setting $J = \|\mathbf{x}\|_\infty + 1 > 0$, we have $\frac{\mathbf{x}+J}{2J} \in [0, 1]^K$, and hence

$$\mathbf{y} := \frac{\mathbf{x}+J}{2J}(M_2 - M_1) + M_1 \in [M_1, M_2]^K.$$

By the former result, there exists $w_0 \in \mathbb{R}$ such that

$$\left\| \mathbf{y} - \left[g\left(\frac{w_0}{\alpha+r_1}\right), g\left(\frac{w_0}{\alpha+r_2}\right), \dots, g\left(\frac{w_0}{\alpha+r_K}\right) \right]^T \right\|_\infty < \frac{M_2-M_1}{2J} \varepsilon$$

It follows from $\mathbf{y} = \frac{\mathbf{x}+J}{2J}(M_2 - M_1) + M_1$ that $\mathbf{x} = \frac{2J}{M_2-M_1}\mathbf{y} + \frac{J(M_1+M_2)}{M_1-M_2} =: u_0\mathbf{y} + v_0$, where $u_0 = \frac{2J}{M_2-M_1}$ and $v_0 = \frac{J(M_1+M_2)}{M_1-M_2}$. Therefore,

$$\begin{aligned} & \left\| \mathbf{x} - \left[u_0 g\left(\frac{w_0}{\alpha+r_1}\right) + v_0, u_0 g\left(\frac{w_0}{\alpha+r_2}\right) + v_0, \dots, u_0 g\left(\frac{w_0}{\alpha+r_K}\right) + v_0 \right]^T \right\|_\infty \\ &= \left\| u_0\mathbf{y} + v_0 - \left[u_0 g\left(\frac{w_0}{\alpha+r_1}\right) + v_0, u_0 g\left(\frac{w_0}{\alpha+r_2}\right) + v_0, \dots, u_0 g\left(\frac{w_0}{\alpha+r_K}\right) + v_0 \right]^T \right\|_\infty < u_0 \frac{M_2-M_1}{2J} \varepsilon = \varepsilon. \end{aligned}$$

Since $\varepsilon > 0$ and $\mathbf{x} \in \mathbb{R}^K$ are arbitrary, the following set

$$\left\{ \left[u \cdot g\left(\frac{w}{\alpha+r_1}\right) + v, u \cdot g\left(\frac{w}{\alpha+r_2}\right) + v, \dots, u \cdot g\left(\frac{w}{\alpha+r_K}\right) + v \right]^T : u, v, w \in \mathbb{R} \right\}$$

is dense in \mathbb{R}^K . So we finish the proof. \square

4.1 Proof of Lemma 4.1

Before proving Lemma 4.1, let us have a brief discussion on related concepts. Recall that a complex number α is an algebraic number if and only if there exist $\lambda_0, \lambda_1, \dots, \lambda_J \in \mathbb{Q}$ with $\sum_{j=0}^J \lambda_j \alpha^j = 0$. The set of all algebraic numbers is denoted by \mathbb{A} . A complex number is called **transcendental** if it is not in \mathbb{A} . It is well known that the set \mathbb{A} is **countable**, and, therefore, almost all numbers are transcendental. Therefore, for almost all $\alpha \in \mathbb{R}$, the set of numbers $\left\{ \frac{1}{\alpha+k} : k = 1, 2, \dots, K \right\}$ are rationally independent. The best known transcendental numbers are π (the ratio of a circle's circumference to its diameter) and e (the natural logarithmic base). Thus, both sets of numbers $\left\{ \frac{1}{\pi+k} : k = 1, 2, \dots, K \right\}$ and $\left\{ \frac{1}{e+k} : k = 1, 2, \dots, K \right\}$ are rational independent.

In order to prove lemma 4.1, we need an auxiliary lemma below, characterizing some properties of coefficients of Lagrange basis polynomials. Recall that, for any given pairwise distinct numbers $x_1, x_2, \dots, x_K \in \mathbb{R}$, the Lagrange basis polynomials are

$$p_k(x) := \prod_{\substack{j \in \{1, 2, \dots, K\} \\ j \neq k}} \frac{x - x_j}{x_k - x_j} = \frac{x - x_1}{x_k - x_1} \dots \frac{x - x_{k-1}}{x_k - x_{k-1}} \frac{x - x_{k+1}}{x_k - x_{k+1}} \dots \frac{x - x_K}{x_k - x_K}, \quad (4.1)$$

for $k = 1, 2, \dots, K$. They are polynomials of degree $\leq K - 1$. Thus, the coefficients of these K Lagrange basis polynomials form a matrix

$$\mathbf{A} = (a_{i,j}) = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,K} \\ a_{2,1} & a_{2,2} & \dots & a_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ a_{K,1} & a_{K,2} & \dots & a_{K,K} \end{bmatrix} \in \mathbb{R}^{K \times K}, \quad (4.2)$$

868 which satisfies the following equality

$$869 \quad p_k(x) = \sum_{j=1}^K a_{k,j} x^{j-1} = a_{k,1} + a_{k,2}x + \cdots + a_{k,K}x^{K-1} \quad \text{for } k = 1, 2, \dots, K \text{ and any } x \in \mathbb{R}.$$

870 The lemma below essentially characterizes the linear independence of Lagrange basis
871 polynomials.

872 **Lemma 4.4.** *With the same setting just above, the matrix \mathbf{A} given in Equation (4.2) is*
873 *invertible.*

874 *Proof.* For any $\mathbf{y} = [y_1, y_2, \dots, y_K] \in \mathbb{R}^K$, by the definition of Lagrange basis polynomials
875 $p_k(x)$ for $k = 1, 2, \dots, K$ in Equation (4.1), $p(x) = \sum_{k=1}^K y_k p_k(x)$ is the target interpolation
876 polynomial for sample points $(x_1, y_1), (x_2, y_2), \dots, (x_K, y_K)$. That is, for any
877 $\ell \in \{1, 2, \dots, K\}$, we have

$$\begin{aligned} y_\ell &= p(x_\ell) = \sum_{k=1}^K y_k p_k(x_\ell) = \sum_{k=1}^K y_k \sum_{j=1}^K a_{k,j} x_\ell^{j-1} \\ 878 \quad &= [y_1, y_2, \dots, y_K] \cdot \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,K} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ a_{K,1} & a_{K,2} & \cdots & a_{K,K} \end{bmatrix} \cdot \begin{bmatrix} x_\ell^0 \\ x_\ell^1 \\ \vdots \\ x_\ell^{K-1} \end{bmatrix} = \mathbf{y}^T \mathbf{A} \begin{bmatrix} x_\ell^0 \\ x_\ell^1 \\ \vdots \\ x_\ell^{K-1} \end{bmatrix}. \end{aligned}$$

879 It follows that

$$880 \quad \mathbf{y}^T = [y_1, y_2, \dots, y_K] = \mathbf{y}^T \mathbf{A} \begin{bmatrix} x_1^0 & x_2^0 & \cdots & x_K^0 \\ x_1^1 & x_2^1 & \cdots & x_K^1 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{K-1} & x_2^{K-1} & \cdots & x_K^{K-1} \end{bmatrix}.$$

881 Since $\mathbf{y} \in \mathbb{R}^K$ is arbitrary, we have

$$882 \quad \mathbf{A} \begin{bmatrix} x_1^0 & x_2^0 & \cdots & x_K^0 \\ x_1^1 & x_2^1 & \cdots & x_K^1 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{K-1} & x_2^{K-1} & \cdots & x_K^{K-1} \end{bmatrix} = \mathbf{I}_K,$$

883 where $\mathbf{I}_K \in \mathbb{R}^{K \times K}$ is the identity matrix. Recall that x_1, x_2, \dots, x_K are pairwise distinct,
884 which implies the Vandermonde matrix

$$885 \quad \begin{bmatrix} x_1^0 & x_2^0 & \cdots & x_K^0 \\ x_1^1 & x_2^1 & \cdots & x_K^1 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{K-1} & x_2^{K-1} & \cdots & x_K^{K-1} \end{bmatrix}$$

886 is invertible. Thus, \mathbf{A} is also invertible. So we complete the proof. □

887 With Lemma 4.4 at hand, we are ready to prove Lemma 4.1.

888 *Proof of Lemma 4.1.* Let $x_k = -r_k \in \mathbb{Q}$ for $k = 1, 2, \dots, K$ and define the Lagrange basis
 889 polynomials as

$$890 \quad p_k(x) := \prod_{\substack{j \in \{1, 2, \dots, K\} \\ j \neq k}} \frac{x - x_j}{x_k - x_j} = w_k \prod_{\substack{j \in \{1, 2, \dots, K\} \\ j \neq k}} (x - x_j), \quad \text{where } w_k = \prod_{\substack{j \in \{1, 2, \dots, K\} \\ j \neq k}} \frac{1}{x_k - x_j} \neq 0,$$

891 for $k = 1, 2, \dots, K$. Note that w_k is rational and nonzero for any k , which is important for
 892 later proof. The coefficients of these K Lagrange basis polynomials form a matrix

$$893 \quad \mathbf{A} = (a_{i,j}) = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,K} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ a_{K,1} & a_{K,2} & \cdots & a_{K,K} \end{bmatrix} \in \mathbb{R}^{K \times K},$$

894 which satisfies the following equality

$$895 \quad p_k(x) = \sum_{j=1}^K a_{k,j} x^{j-1} = a_{k,1} + a_{k,2}x + \cdots + a_{k,K}x^{K-1} \quad \text{for } k = 1, 2, \dots, K \text{ and any } x \in \mathbb{R}.$$

896 Now assume there exist $\lambda_1, \lambda_2, \dots, \lambda_K \in \mathbb{Q}$ such that $\sum_{k=1}^K \lambda_k \cdot \frac{1}{\alpha + r_k} = 0$. Our goal is to
 897 prove $\lambda_1 = \lambda_2 = \cdots = \lambda_K = 0$. Clearly, we have

$$\begin{aligned} 0 &= \sum_{k=1}^K \lambda_k \cdot \frac{1}{\alpha + r_k} = \sum_{k=1}^K \frac{\lambda_k}{\alpha - x_k} = \prod_{j=1}^K (\alpha - x_j) \cdot \sum_{k=1}^K \frac{\lambda_k}{\alpha - x_k} = \sum_{k=1}^K \frac{\lambda_k}{w_k} \cdot w_k \prod_{\substack{j \in \{1, 2, \dots, K\} \\ j \neq k}} (\alpha - x_j) \\ 898 \quad &= \sum_{k=1}^K \frac{\lambda_k}{w_k} \cdot p_k(\alpha) = \sum_{k=1}^K \frac{\lambda_k}{w_k} \sum_{j=1}^K a_{k,j} \alpha^{j-1} = \sum_{j=1}^K \left(\underbrace{\sum_{k=1}^K \frac{\lambda_k}{w_k} a_{k,j}}_{=0 \text{ since } \alpha \in \mathbb{R} \setminus \mathbb{A}} \right) \cdot \alpha^{j-1}. \end{aligned}$$

899 Note that $\alpha \in \mathbb{R} \setminus \mathbb{A}$ is not an algebraic number and $\frac{\lambda_k}{w_k} \in \mathbb{Q}$ since $\lambda_k, w_k \in \mathbb{Q}$ for any k .
 900 Thus, the coefficients must be 0, namely,

$$901 \quad \sum_{k=1}^K \frac{\lambda_k}{w_k} a_{k,j} = 0 \quad \text{for } j = 1, 2, \dots, K.$$

902 It follows that

$$903 \quad \mathbf{0} = \left[\frac{\lambda_1}{w_1}, \frac{\lambda_2}{w_2}, \dots, \frac{\lambda_K}{w_K} \right] \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,K} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ a_{K,1} & a_{K,2} & \cdots & a_{K,K} \end{bmatrix} = \left[\frac{\lambda_1}{w_1}, \frac{\lambda_2}{w_2}, \dots, \frac{\lambda_K}{w_K} \right] \mathbf{A}.$$

904 By Lemma 4.4, \mathbf{A} is invertible. Thus, $\left[\frac{\lambda_1}{w_1}, \frac{\lambda_2}{w_2}, \dots, \frac{\lambda_K}{w_K} \right] = \mathbf{0}$, which implies $\lambda_1 = \lambda_2 = \cdots =$
 905 $\lambda_K = 0$. Hence, the set of numbers $\left\{ \frac{1}{\alpha + r_k} : k = 1, 2, \dots, K \right\}$ are rationally independent,
 906 which means we finish the proof. \square

4.2 Proof of Lemma 4.2

The proof of Lemma 4.2 is mainly based on the fact that an irrational winding is dense on the torus (e.g., see Lemma 2 of [39]). For completeness, we establish a lemma below and give its detailed proof.

Lemma 4.5. *Given any $K \in \mathbb{N}^+$ and an arbitrary set of rationally independent numbers $\{a_k : k = 1, 2, \dots, K\} \subseteq \mathbb{R}$, the following set*

$$\left\{ \left[\tau(wa_1), \tau(wa_2), \dots, \tau(wa_K) \right]^T : w \in \mathbb{R} \right\} \subseteq [0, 1)^K$$

is dense in $[0, 1]^K$, where $\tau(x) := x - \lfloor x \rfloor$ for any $x \in \mathbb{R}$.

The proof of Lemma 4.5 can be found later in this section. Now let us first prove Lemma 4.2 by assuming Lemma 4.5 is true.

Proof of Lemma 4.2. Define $\tilde{g}(x) := g(Tx)$ for any $x \in \mathbb{R}$. The continuity of g on $[x_1, x_2]$ implies \tilde{g} is continuous on $[\frac{x_1}{T}, \frac{x_2}{T}]$, and, therefore, uniformly continuous on $[\frac{x_1}{T}, \frac{x_2}{T}]$. For any $\varepsilon > 0$, there exists $\delta \in (0, \frac{x_2 - x_1}{T})$ such that

$$|\tilde{g}(u) - \tilde{g}(v)| < \varepsilon \quad \text{for any } u, v \in [\frac{x_1}{T}, \frac{x_2}{T}] \text{ with } |u - v| < \delta. \quad (4.3)$$

Given any $\xi = [\xi_1, \xi_2, \dots, \xi_K] \in [M_1, M_2]^K$, by the intermediate value theorem, there exists $z_1, z_2, \dots, z_K \in [x_1, x_2]$ such that $g(z_k) = \xi_k$ for any $k = 1, 2, \dots, K$.

For any $k = 1, 2, \dots, K$, set $y_k = z_k/T \in [\frac{x_1}{T}, \frac{x_2}{T}]$ and

$$\tilde{y}_k = y_k + \frac{\delta}{2} \cdot \mathbb{1}_{\{y_k \leq \frac{x_1}{T} + \frac{\delta}{2}\}} - \frac{\delta}{2} \cdot \mathbb{1}_{\{y_k \geq \frac{x_2}{T} - \frac{\delta}{2}\}}.$$

Then, for $k = 1, 2, \dots, K$, we have

$$\tilde{y}_k = y_k + \frac{\delta}{2} \cdot \mathbb{1}_{\{y_k \leq \frac{x_1}{T} + \frac{\delta}{2}\}} - \frac{\delta}{2} \cdot \mathbb{1}_{\{y_k \geq \frac{x_2}{T} - \frac{\delta}{2}\}} \in \left[\frac{x_1}{T} + \frac{\delta}{2}, \frac{x_2}{T} - \frac{\delta}{2} \right]$$

and

$$|\tilde{y}_k - y_k| \leq \left| \frac{\delta}{2} \cdot \mathbb{1}_{\{y_k \leq \frac{x_1}{T} + \frac{\delta}{2}\}} - \frac{\delta}{2} \cdot \mathbb{1}_{\{y_k \geq \frac{x_2}{T} - \frac{\delta}{2}\}} \right| \leq \delta/2.$$

Define $\tau(x) = x - \lfloor x \rfloor$ for any $x \in \mathbb{R}$. Clearly, $[\tau(\tilde{y}_1), \tau(\tilde{y}_2), \dots, \tau(\tilde{y}_K)]^T \in [0, 1]^K$. Then by Lemma 4.5, there exists $w_0 \in \mathbb{R}$ such that

$$|\tau(w_0 a_k) - \tau(\tilde{y}_k)| < \delta/2 \quad \text{for } k = 1, 2, \dots, K.$$

It follows that

$$\left| \tau(w_0 a_k) + \lfloor \tilde{y}_k \rfloor - \tilde{y}_k \right| = \left| \tau(w_0 a_k) - (\tilde{y}_k - \lfloor \tilde{y}_k \rfloor) \right| = \left| \tau(w_0 a_k) - \tau(\tilde{y}_k) \right| < \delta/2,$$

for $k = 1, 2, \dots, K$. Since $\tilde{y}_k \in [\frac{x_1}{T} + \frac{\delta}{2}, \frac{x_2}{T} - \frac{\delta}{2}]$, we have $\tau(w_0 a_k) + \lfloor \tilde{y}_k \rfloor \in [\frac{x_1}{T}, \frac{x_2}{T}]$. Besides,

$$\left| \tau(w_0 a_k) + \lfloor \tilde{y}_k \rfloor - y_k \right| \leq \left| \tau(w_0 a_k) + \lfloor \tilde{y}_k \rfloor - \tilde{y}_k \right| + |\tilde{y}_k - y_k| < \delta/2 + \delta/2 = \delta,$$

for $k = 1, 2, \dots, K$. Then, by Equation (4.3), we have

$$\left| \widetilde{g}(\tau(w_0 a_k) + \lfloor \widetilde{y}_k \rfloor) - \widetilde{g}(y_k) \right| < \varepsilon \quad \text{for } k = 1, 2, \dots, K.$$

By the definition of \widetilde{g} , it is periodic with period 1 since g is periodic with period T . This implies

$$\widetilde{g}(\tau(w_0 a_k) + \lfloor \widetilde{y}_k \rfloor) = \widetilde{g}(w_0 a_k - \lfloor w_0 a_k \rfloor + \lfloor \widetilde{y}_k \rfloor) = \widetilde{g}(w_0 a_k) = g(T \cdot w_0 a_k),$$

for $k = 1, 2, \dots, K$. Also, $\widetilde{g}(y_k) = g(T y_k) = g(z_k) = \xi_k$ for $k = 1, 2, \dots, K$. It follows that

$$\left| g(T \cdot w_0 a_k) - \xi_k \right| = \left| \widetilde{g}(\tau(w_0 a_k) + \lfloor \widetilde{y}_k \rfloor) - \widetilde{g}(y_k) \right| < \varepsilon \quad \text{for } k = 1, 2, \dots, K.$$

That is

$$\left\| [g(w_1 a_1), g(w_1 a_2), \dots, g(w_1 a_K)]^T - \boldsymbol{\xi} \right\|_\infty < \varepsilon,$$

where $w_1 = T \cdot w_0 \in \mathbb{R}$. Since $\boldsymbol{\xi} \in [M_1, M_2]^K$ and $\varepsilon > 0$ are arbitrary, the following set

$$\left\{ [g(w a_1), g(w a_2), \dots, g(w a_K)]^T : w \in \mathbb{R} \right\}$$

is dense in $[M_1, M_2]^K$ as desired. So we finish the proof. \square

Finally, let us present the detailed proof of Lemma 4.5.

Proof of Lemma 4.5. We prove this lemma by mathematical induction. First, we consider the case $K = 1$. Note that $a_1 \neq 0$ since it is rationally independent. Thus, we have $\{\tau(w a_1) : w \in \mathbb{R}\} = [0, 1)$, which implies $\{\tau(w a_1) : w \in \mathbb{R}\}$ is dense in $[0, 1]$.

Now assume this lemma holds for $K = J - 1 \in \mathbb{N}^+$. Our goal is to prove the case $K = J$. Given any $\varepsilon \in (0, 1/100)$ and arbitrary $\boldsymbol{\xi} = [\xi_1, \xi_2, \dots, \xi_J]^T \in [0, 1]^J$, our goal is to find proper $w \in \mathbb{R}$ such that

$$|\tau(w a_j) - \xi_j| < C\varepsilon \quad \text{for } j = 1, 2, \dots, J, \quad \text{where } C \text{ is an absolute constant.} \quad (4.4)$$

As we shall see later, we need an assumption that the given point is in $[6\varepsilon, 1 - 6\varepsilon]^J$. Thus, we set

$$\widetilde{\xi}_j = \xi_j + 6\varepsilon \cdot \mathbf{1}_{\{\xi_j \leq 6\varepsilon\}} - 6\varepsilon \cdot \mathbf{1}_{\{\xi_j \geq 1 - 6\varepsilon\}} \quad \text{for } j = 1, 2, \dots, J.$$

Then, we have

$$\widetilde{\xi}_j \in [6\varepsilon, 1 - 6\varepsilon] \quad \text{for } j = 1, 2, \dots, J \quad (4.5)$$

and

$$|\xi_j - \widetilde{\xi}_j| = |6\varepsilon \cdot \mathbf{1}_{\{\xi_j \leq 6\varepsilon\}} - 6\varepsilon \cdot \mathbf{1}_{\{\xi_j \geq 1 - 6\varepsilon\}}| \leq 6\varepsilon \quad \text{for } j = 1, 2, \dots, J. \quad (4.6)$$

Define

$$\widehat{\xi}_j := \tau(\widetilde{\xi}_j - \frac{\widetilde{\xi}_J}{a_J} a_j) \quad \text{for } j = 1, 2, \dots, J. \quad (4.7)$$

Then $\widehat{\xi}_J = 0$ and $\widehat{\xi}_j \in [0, 1)$ for $j = 1, 2, \dots, J - 1$. To approximate $[\widehat{\xi}_1, \widehat{\xi}_2, \dots, \widehat{\xi}_{J-1}]^T \in [0, 1)^{J-1}$, we only need to consider $J - 1$ indices, and, therefore, we can use the induction hypothesis to continue our proof.

Clearly, the rational independence of a_1, a_2, \dots, a_J implies none of them is equal to zero. Define

$$\mathbf{b}_n := \left[\tau\left(\frac{n}{a_J}a_1\right), \tau\left(\frac{n}{a_J}a_2\right), \dots, \tau\left(\frac{n}{a_J}a_{J-1}\right) \right]^T \in [0, 1)^{J-1}.$$

Then the bounded sequence $(\mathbf{b}_n)_{n=1}^\infty$ has a convergent subsequence by the Bolzano-Weierstrass theorem. Thus, there exist $n_1, n_2 \in \mathbb{N}^+$ with $n_1 < n_2$ such that $\|\mathbf{b}_{n_2} - \mathbf{b}_{n_1}\|_\infty < \varepsilon$. That is,

$$\left| \tau\left(\frac{n_2}{a_J}a_j\right) - \tau\left(\frac{n_1}{a_J}a_j\right) \right| < \varepsilon \quad \text{for } j = 1, 2, \dots, J-1.$$

Set $\widehat{n} = n_2 - n_1 \in \mathbb{N}^+$ and $k_j = \left\lfloor \frac{n_1}{a_J}a_j \right\rfloor - \left\lfloor \frac{n_2}{a_J}a_j \right\rfloor$ for $j = 1, 2, \dots, J-1$. Then, by defining

$$\widehat{a}_j := \frac{\widehat{n}}{a_J}a_j + k_j \quad \text{for } j = 1, 2, \dots, J-1,$$

we have

$$\begin{aligned} |\widehat{a}_j| &= \left| \frac{\widehat{n}}{a_J}a_j + k_j \right| = \left| \frac{n_2}{a_J}a_j - \frac{n_1}{a_J}a_j + \left\lfloor \frac{n_1}{a_J}a_j \right\rfloor - \left\lfloor \frac{n_2}{a_J}a_j \right\rfloor \right| \\ &= \left| \left(\frac{n_2}{a_J}a_j - \left\lfloor \frac{n_2}{a_J}a_j \right\rfloor \right) - \left(\frac{n_1}{a_J}a_j - \left\lfloor \frac{n_1}{a_J}a_j \right\rfloor \right) \right| = \left| \tau\left(\frac{n_2}{a_J}a_j\right) - \tau\left(\frac{n_1}{a_J}a_j\right) \right| < \varepsilon. \end{aligned} \quad (4.8)$$

It is easy to verify that $\widehat{a}_1, \widehat{a}_2, \dots, \widehat{a}_{J-1}$ are rationally independent. To see this, assume there exist $\lambda_1, \lambda_2, \dots, \lambda_{J-1} \in \mathbb{Q}$ such that

$$0 = \sum_{j=1}^{J-1} \lambda_j \widehat{a}_j = \sum_{j=1}^{J-1} \lambda_j \left(\frac{\widehat{n}}{a_J}a_j + k_j \right) = \sum_{j=1}^{J-1} \lambda_j \frac{\widehat{n}}{a_J}a_j + \sum_{j=1}^{J-1} \lambda_j k_j,$$

then

$$0 = \sum_{j=1}^{J-1} \lambda_j \widehat{n} a_j + \left(\sum_{j=1}^{J-1} \lambda_j k_j \right) a_J.$$

Since a_1, a_2, \dots, a_J are rationally independent, we have $\lambda_j \widehat{n} = 0$ for $j = 1, 2, \dots, J-1$. It follows from $\widehat{n} = n_2 - n_1 > 0$ that $\lambda_1 = \lambda_2 = \dots = \lambda_{J-1} = 0$. Thus, $\widehat{a}_1, \widehat{a}_2, \dots, \widehat{a}_{J-1}$ are rationally independent as desired.

By the induction hypothesis, the following set

$$\left\{ \left[\tau(s \cdot \widehat{a}_1), \tau(s \cdot \widehat{a}_2), \dots, \tau(s \cdot \widehat{a}_{J-1}) \right]^T : s \in \mathbb{R} \right\} \subseteq [0, 1)^{J-1}$$

is dense in $[0, 1)^{J-1}$. Recall that $\widehat{\xi}_j \in [0, 1]$ for $j = 1, \dots, J-1$, which implies

$$\widehat{\xi}_j + 3\varepsilon \cdot \mathbf{1}_{\{\widehat{\xi}_j \leq 3\varepsilon\}} - 3\varepsilon \cdot \mathbf{1}_{\{\widehat{\xi}_j \geq 1-3\varepsilon\}} \in [3\varepsilon, 1-3\varepsilon] \quad \text{for } j = 1, \dots, J-1.$$

Hence, there exists $s_0 \in \mathbb{R}$ such that

$$\left| \tau(s_0 \widehat{a}_j) - \left(\widehat{\xi}_j + 3\varepsilon \cdot \mathbf{1}_{\{\widehat{\xi}_j \leq 3\varepsilon\}} - 3\varepsilon \cdot \mathbf{1}_{\{\widehat{\xi}_j \geq 1-3\varepsilon\}} \right) \right| < \varepsilon \quad \text{for } j = 1, \dots, J-1.$$

It follows that

$$\tau(s_0 \widehat{a}_j) \in [2\varepsilon, 1-2\varepsilon] \quad \text{for } j = 1, \dots, J-1$$

and

$$\left| \tau(s_0 \widehat{a}_j) - \widehat{\xi}_j \right| < \varepsilon + \left| 3\varepsilon \cdot \mathbf{1}_{\{\widehat{\xi}_j \leq 3\varepsilon\}} - 3\varepsilon \cdot \mathbf{1}_{\{\widehat{\xi}_j \geq 1-3\varepsilon\}} \right| \leq 4\varepsilon \quad \text{for } j = 1, \dots, J-1. \quad (4.9)$$

997 To estimate $\tau(\lfloor s_0 \rfloor \widehat{a}_j) - \widehat{\xi}_j$, we need to bound $\tau(s_0 \widehat{a}_j) - \tau(\lfloor s_0 \rfloor \widehat{a}_j)$. To this end, we
 998 need an observation for any $x, y \in \mathbb{R}$ as follows.

$$999 \quad |x - y| < \varepsilon \quad \text{and} \quad \tau(x) \in [2\varepsilon, 1 - 2\varepsilon] \implies |\tau(x) - \tau(y)| < \varepsilon. \quad (4.10)$$

1000 In fact, $\tau(x) \in [2\varepsilon, 1 - 2\varepsilon]$ implies $\varepsilon \leq \tau(x) - \varepsilon \leq \tau(x) + \varepsilon \leq 1 - \varepsilon$, deducing

$$1001 \quad y \in [x - \varepsilon, x + \varepsilon] = \left[\underbrace{\lfloor x \rfloor + \tau(x) - \varepsilon}_{\geq \varepsilon}, \underbrace{\lfloor x \rfloor + \tau(x) + \varepsilon}_{\leq 1 - \varepsilon} \right] \subseteq [\lfloor x \rfloor + \varepsilon, \lfloor x \rfloor + 1 - \varepsilon] \subseteq [\lfloor x \rfloor, \lfloor x \rfloor + 1).$$

1002 Thus, $\lfloor y \rfloor = \lfloor x \rfloor$, which implies $|\tau(x) - \tau(y)| = |\tau(x) - \tau(y) + \lfloor x \rfloor - \lfloor y \rfloor| = |x - y| < \varepsilon$ as
 1003 desired.

1004 By Equation (4.8), we have

$$1005 \quad \left| s_0 \widehat{a}_j - \lfloor s_0 \rfloor \widehat{a}_j \right| \leq \left| s_0 - \lfloor s_0 \rfloor \right| \cdot |\widehat{a}_j| < \varepsilon \quad \text{for } j = 1, 2, \dots, J - 1.$$

1006 Recall that $\tau(s_0 \widehat{a}_j) \in [2\varepsilon, 1 - 2\varepsilon]$ for $j = 1, \dots, J - 1$. Then, for each $j \in \{1, 2, \dots, J - 1\}$, by
 1007 the observation above in Equation (4.10) (set $x = s_0 \widehat{a}_j$ and $y = \lfloor s_0 \rfloor \widehat{a}_j$ therein), we have
 1008 $|\tau(s_0 \widehat{a}_j) - \tau(\lfloor s_0 \rfloor \widehat{a}_j)| < \varepsilon$. Therefore, by Equations (4.7) and (4.9), we have

$$1009 \quad \begin{aligned} \left| \tau(\lfloor s_0 \rfloor \widehat{a}_j) - \tau(\widetilde{\xi}_j - \frac{\widetilde{\xi}_J}{a_J} a_j) \right| &= \left| \tau(\lfloor s_0 \rfloor \widehat{a}_j) - \widehat{\xi}_j \right| \\ &\leq \left| \tau(\lfloor s_0 \rfloor \widehat{a}_j) - \tau(s_0 \widehat{a}_j) \right| + \left| \tau(s_0 \widehat{a}_j) - \widehat{\xi}_j \right| < \varepsilon + 4\varepsilon = 5\varepsilon, \end{aligned}$$

1010 for $j = 1, 2, \dots, J - 1$. Recall the fact: For any $x, y \in \mathbb{R}$, it holds that $\tau(x) - \tau(y) =$
 1011 $x - \lfloor x \rfloor - (y - \lfloor y \rfloor) = x - y - z$, where $z = \lfloor x \rfloor - \lfloor y \rfloor \in \mathbb{Z}$.

1012 Therefore, for $j = 1, 2, \dots, J - 1$, there exists $z_j \in \mathbb{Z}$ such that

$$1013 \quad \tau(\lfloor s_0 \rfloor \widehat{a}_j) - \tau(\widetilde{\xi}_j - \frac{\widetilde{\xi}_J}{a_J} a_j) = \lfloor s_0 \rfloor \widehat{a}_j - (\widetilde{\xi}_j - \frac{\widetilde{\xi}_J}{a_J} a_j) - z_j = \lfloor s_0 \rfloor \widehat{a}_j + \frac{\widetilde{\xi}_J}{a_J} a_j - (z_j + \widetilde{\xi}_j),$$

1014 which implies

$$1015 \quad \left| \lfloor s_0 \rfloor \widehat{a}_j + \frac{\widetilde{\xi}_J}{a_J} a_j - (z_j + \widetilde{\xi}_j) \right| = \left| \tau(\lfloor s_0 \rfloor \widehat{a}_j) - \tau(\widetilde{\xi}_j - \frac{\widetilde{\xi}_J}{a_J} a_j) \right| < 5\varepsilon.$$

1016 It follows that

$$1017 \quad \lfloor s_0 \rfloor \widehat{a}_j + \frac{\widetilde{\xi}_J}{a_J} a_j \in \underbrace{[z_j + \widetilde{\xi}_j - 5\varepsilon, z_j + \widetilde{\xi}_j + 5\varepsilon]}_{\geq \varepsilon} \subseteq [z_j + \varepsilon, z_j + 1 - \varepsilon] \quad \text{for } j = 1, 2, \dots, J - 1,$$

1018 where the fact $\varepsilon \leq \widetilde{\xi}_j - 5\varepsilon \leq \widetilde{\xi}_j + 5\varepsilon \leq 1 - \varepsilon$ comes from Equation (4.5). Therefore,

$$1019 \quad \tau(\lfloor s_0 \rfloor \widehat{a}_j + \frac{\widetilde{\xi}_J}{a_J} a_j) = (\lfloor s_0 \rfloor \widehat{a}_j + \frac{\widetilde{\xi}_J}{a_J} a_j) - z_j \in [\widetilde{\xi}_j - 5\varepsilon, \widetilde{\xi}_j + 5\varepsilon] \quad \text{for } j = 1, 2, \dots, J - 1.$$

1020 For $j = 1, 2, \dots, J - 1$, we have

$$1021 \quad \lfloor s_0 \rfloor \widehat{a}_j + \frac{\widetilde{\xi}_J}{a_J} a_j = \lfloor s_0 \rfloor \left(\frac{\widehat{n}}{a_J} a_j + k_j \right) + \frac{\widetilde{\xi}_J}{a_J} a_j = \underbrace{\frac{\lfloor s_0 \rfloor \widehat{n} + \widetilde{\xi}_J}{a_J} a_j + k_j \lfloor s_0 \rfloor}_{\in \mathbb{Z}},$$

1022 which implies

$$1023 \quad \tau\left(\frac{\lfloor s_0 \rfloor \tilde{n} + \tilde{\xi}_J}{a_J} a_j\right) = \tau(\lfloor s_0 \rfloor \tilde{a}_j + \frac{\tilde{\xi}_J}{a_J} a_j) \in [\tilde{\xi}_j - 5\varepsilon, \tilde{\xi}_j + 5\varepsilon] \quad \text{for } j = 1, 2, \dots, J-1.$$

1024 By Equation (4.5), we have $\tilde{\xi}_J \in [6\varepsilon, 1 - 6\varepsilon]$, which implies

$$1025 \quad \tau\left(\frac{\lfloor s_0 \rfloor \tilde{n} + \tilde{\xi}_J}{a_J} a_J\right) = \tau(\lfloor s_0 \rfloor \tilde{n} + \tilde{\xi}_J) = \tilde{\xi}_J.$$

1026 Thus, for $j = 1, 2, \dots, J$, we have

$$1027 \quad \left| \tau\left(\frac{\lfloor s_0 \rfloor \tilde{n} + \tilde{\xi}_J}{a_J} a_j\right) - \tilde{\xi}_j \right| \leq 5\varepsilon.$$

1028 By Equation (4.6), we have $|\tilde{\xi}_j - \xi_j| < 6\varepsilon$ for $j = 1, 2, \dots, J$, which implies

$$1029 \quad \left| \tau\left(\frac{\lfloor s_0 \rfloor \tilde{n} + \tilde{\xi}_J}{a_J} a_j\right) - \xi_j \right| \leq \left| \tau\left(\frac{\lfloor s_0 \rfloor \tilde{n} + \tilde{\xi}_J}{a_J} a_j\right) - \tilde{\xi}_j \right| + |\tilde{\xi}_j - \xi_j| \leq 5\varepsilon + 6\varepsilon = 11\varepsilon.$$

1030 Therefore, $w_0 = \frac{\lfloor s_0 \rfloor \tilde{n} + \tilde{\xi}_J}{a_J}$ is the desired w in Equation (4.4). That is,

$$1031 \quad \left| \tau(w_0 a_j) - \xi_j \right| \leq 11\varepsilon \quad \text{for } j = 1, 2, \dots, J.$$

1032 Since $\boldsymbol{\xi} = [\xi_1, \xi_2, \dots, \xi_J]^T \in [0, 1]^J$ is arbitrary, the following set

$$1033 \quad \left\{ [\tau(w a_1), \tau(w a_2), \dots, \tau(w a_J)]^T : w \in \mathbb{R} \right\} \subseteq [0, 1]^J$$

1034 is dense in $[0, 1]^J$ as desired. We finish the process of mathematical induction, and,
1035 therefore, finish the proof by the principle of mathematical induction. \square

1036 We remark that the target parameter $w_0 = \frac{\lfloor s_0 \rfloor \tilde{n} + \tilde{\xi}_J}{a_J}$ designed in the above proof may
1037 not be bounded uniformly for any approximation error ε since \tilde{n} can be arbitrarily large
1038 depending on ε . Therefore, the network in Theorem 1.1 may require sufficient large
1039 parameters to achieve a target error ε .

1040 5 Other examples of UAFs

1041 This section aims at designing new UAFs with additional properties such as smooth
1042 or sigmoidal functions. As discussed in the introduction and shown in the proof of
1043 our main theorem, the construction of UAFs mainly relies on three properties: high
1044 nonlinearity, periodicity, and the capacity to reproduce step functions. The EUAF σ
1045 defined in Equation (1.3) is a simple and typical example of UAFs satisfying these three
1046 properties. Indeed, having these properties plays an important role in our proof and
1047 is a necessary but not sufficient condition for designing a UAF. In other words, these
1048 properties are important, but cannot guarantee the successful construction of UAFs.

1049 Here, we present another idea to design new UAFs, which mainly relies on the
1050 following observation: If a UAF ϱ can be approximated by a fixed-size network activated
1051 by a new function $\tilde{\varrho}$ within an arbitrary error on any bounded interval, then $\tilde{\varrho}$ is also a
1052 UAF. Such an observation is a direct result of the lemma below.

Lemma 5.1. Let $\varrho, \tilde{\varrho}: \mathbb{R} \rightarrow \mathbb{R}$ be two functions with $\varrho \in C(\mathbb{R})$. For an arbitrary given function $f \in [a, b]^d \rightarrow \mathbb{R}$ and $\varepsilon > 0$, suppose that the following two conditions hold:

- There exists a function ϕ_ϱ realized by a ϱ -activated network with width N and depth L such that

$$|\phi_\varrho(\mathbf{x}) - f(\mathbf{x})| < \varepsilon/2 \quad \text{for any } \mathbf{x} \in [a, b]^d.$$

- For any $M > 0$ and each $\delta \in (0, 1)$, there exists a function ϱ_δ realized by a $\tilde{\varrho}$ -activated network with width \tilde{N} and depth \tilde{L} such that

$$\varrho_\delta(t) \rightrightarrows \varrho(t) \quad \text{as } \delta \rightarrow 0^+ \quad \text{for any } t \in [-M, M],$$

where \rightrightarrows denotes the uniform convergence.

Then, there exists a function $\phi = \phi_{\tilde{\varrho}}$ generated by a $\tilde{\varrho}$ -activated network with width $N\tilde{N}$ and depth $L\tilde{L}$ such that

$$|\phi(\mathbf{x}) - f(\mathbf{x})| < \varepsilon \quad \text{for any } \mathbf{x} \in [a, b]^d.$$

The proof of Lemma 5.1 is placed in Section 5.3. Based on Lemma 5.1, we will propose two UAFs with better mathematical properties. That is, the idea of designing a C^s UAF is given in Section 5.1 and a sigmoidal UAF is constructed in detail in Section 5.2.

5.1 Smooth UAF

The smoothness of a function is one of the most desired properties in mathematical modeling and computation. The EUAF σ is continuous but not smooth. So we will show how to construct a C^s UAF based on an existing one. The key point is the fact that the integral of a continuous function is continuously differentiable.

Suppose ϱ is a continuous UAF. Define

$$\tilde{\varrho}(x) := \int_0^x \varrho(t) dt \quad \text{for any } x \in \mathbb{R}.$$

For any $M > 0$, it holds that

$$\frac{\tilde{\varrho}(x + \delta) - \tilde{\varrho}(x)}{\delta} = \frac{1}{\delta} \int_x^{x+\delta} \varrho(t) dt \rightrightarrows \varrho(x) \quad \text{as } \delta \rightarrow 0^+ \quad \text{for any } x \in [-M, M].$$

This means ϱ can be approximated by a one-hidden-layer $\tilde{\varrho}$ -activated network with width 2 arbitrarily well on any bounded interval. It follows that $\tilde{\varrho}$ is also a UAF. By repeated applications of the above idea, one could easily construct a C^s UAF.

In particular, set $\varrho_0 = \sigma$ and define $\varrho_1, \varrho_2, \dots, \varrho_s$ by induction as follows.

$$\varrho_{i+1}(x) := \int_0^x \varrho_i(t) dt \quad \text{for any } x \in \mathbb{R} \text{ and } i \in \{0, 1, \dots, s-1\}. \quad (5.1)$$

Then, ϱ_s is a C^s UAF as shown in the following theorem.

Theorem 5.2. Let $\varrho_s \in C^s(\mathbb{R})$ be the function defined in Equation (5.1) for any $s \in \mathbb{N}^+$. Then, for any $f \in C([a, b]^d)$ and $\varepsilon > 0$, there exists a function ϕ generated by a ϱ_s -activated network with width $72sd(2d+1)$ and depth 11 such that

$$|\phi(\mathbf{x}) - f(\mathbf{x})| < \varepsilon \quad \text{for any } \mathbf{x} \in [a, b]^d.$$

Proof. For any $i \in \{0, 1, \dots, s-1\}$ and any $M > 0$, it is easy to verify that

$$\frac{\varrho_{i+1}(x+\delta) - \varrho_{i+1}(x)}{\delta} = \frac{1}{\delta} \int_x^{x+\delta} \varrho_i(t) dt \rightrightarrows \varrho_i(x) \quad \text{as } \delta \rightarrow 0^+ \quad \text{for any } x \in [-M, M].$$

This means ϱ_i can be approximated by a one-hidden-layer ϱ_{i+1} -activated network with width 2 arbitrarily well on any bounded interval. By induction, one could easily prove that $\varrho_0 = \sigma$ can be approximated by a one-hidden-layer ϱ_s -activated network with width $2s$ arbitrarily well on any bounded interval. That is, for each $\delta \in (0, 1)$, there exists a function $\sigma_{s,\delta}$ realized by a ϱ_s -activated network with width $2s$ and depth 1 such that

$$\sigma_{s,\delta}(t) \rightrightarrows \sigma(t) \quad \text{as } \delta \rightarrow 0^+ \quad \text{for any } t \in [-M, M].$$

By Theorem 1.1, there exists a function ϕ_σ generated by a σ -activated network with width $36d(2d+1)$ and depth 11 such that

$$|\phi_\sigma(\mathbf{x}) - f(\mathbf{x})| < \varepsilon/2 \quad \text{for any } \mathbf{x} \in [a, b]^d.$$

Then, by Lemma 5.1, there exists another function $\phi = \phi_{\varrho_s}$ realized by a ϱ_s -activated network with width $2s \times 36d(2d+1) = 72sd(2d+1)$ and depth $1 \times 11 = 11$ such that

$$|\phi(\mathbf{x}) - f(\mathbf{x})| < \varepsilon \quad \text{for any } \mathbf{x} \in [a, b]^d.$$

So we finish the proof. □

5.2 Sigmoidal UAF

Many activation functions used in real applications are sigmoidal functions. Generally, we say a function $g : \mathbb{R} \rightarrow \mathbb{R}$ is sigmoidal (or sigmoid, e.g., see [14]) if it satisfies the following conditions.

- Bounded: $\lim_{x \rightarrow \infty} g(x) = 1$ and $\lim_{x \rightarrow -\infty} g(x) = -1$ (or 0).
- Differentiable: $g'(x)$ exists and continuous for all $x \in \mathbb{R}$.
- Increasing: $g'(x)$ is non-negative for all $x \in \mathbb{R}$.

Our goal is to construct a sigmoidal UAF. To this end, we need to design a new function $\tilde{\sigma}$ based on σ such that σ can be reproduced/approximated by a $\tilde{\sigma}$ -activated network with a fixed size. Making $\tilde{\sigma}$ bounded and increasing is not difficult. The key is to make $\tilde{\sigma}$ continuously differentiable, which can be true by the fact that the integral of a continuous function is continuously differentiable. To be exact, we can define $\tilde{\sigma}$ as follows.

- For $x \in (-\infty, 0]$, define $\tilde{\sigma}(x) := \sigma(x) = \frac{x}{-x+1}$.

- For $x \in (0, \infty)$, define

$$\tilde{\sigma}(x) := \int_0^x \frac{c\sigma(t) + 1}{(2t+1)^2} dt, \quad \text{where} \quad c = \frac{1}{2 \int_0^\infty \frac{\sigma(t)}{(2t+1)^2} dt} \approx 2.554.$$

Remark that there are many possible choices for the integrand in the above definition of $\sigma(x)$ for $x \in (0, \infty)$. Here, we just give a simple example. See an illustration of $\tilde{\sigma}$ in Figure 14.

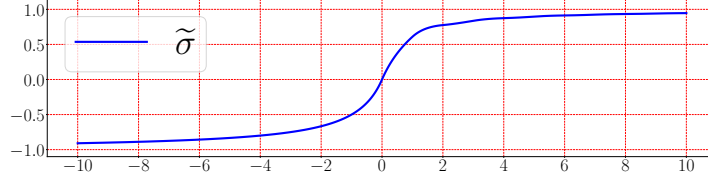


Figure 14: An illustration of $\tilde{\sigma}$ on $[-10, 10]$.

Then, $\tilde{\sigma}$ is a sigmoidal function as verified below.

- Clearly, $\lim_{x \rightarrow -\infty} \tilde{\sigma}(x) = \lim_{x \rightarrow -\infty} \frac{x}{-x+1} = -1$. Moreover,

$$\lim_{x \rightarrow \infty} \tilde{\sigma}(x) = \int_0^\infty \frac{c\sigma(t) + 1}{(2t+1)^2} dt = \frac{1}{2} + \int_0^\infty \frac{1}{(2t+1)^2} dt = 1.$$

- Obviously, $\tilde{\sigma}$ is continuously differentiable on $(-\infty, 0)$ and $(0, \infty)$. Meanwhile, we have $\tilde{\sigma}'(0) = 1$ and $\lim_{x \rightarrow 0} \tilde{\sigma}'(x) = 1$. Therefore, we have $\tilde{\sigma} \in C^1(\mathbb{R})$ as desired.
- For $x \in (-\infty, 0)$, $\tilde{\sigma}'(x) = \frac{1}{(-x+1)^2} > 0$. For $x = 0$, $\tilde{\sigma}'(x) = 1 > 0$. For $x \in (0, \infty)$, $\tilde{\sigma}'(x) = \frac{c\sigma(x)+1}{(2x+1)^2} > 0$. That is, $\tilde{\sigma}'(x) > 0$ for all $x \in \mathbb{R}$.

Based on Theorem 1.1 corresponding to σ , we establish a similar theorem for $\tilde{\sigma}$, Theorem 5.3 below, showing that fixed-size $\tilde{\sigma}$ -activated networks can also approximate continuous functions within an arbitrary error on a hypercube.

Theorem 5.3. *For any $f \in C([a, b]^d)$ and $\varepsilon > 0$, there exists a function ϕ generated by a $\tilde{\sigma}$ -activated network with width $1044d(2d+1)$ and depth 66 such that*

$$|\phi(\mathbf{x}) - f(\mathbf{x})| < \varepsilon \quad \text{for any } \mathbf{x} \in [a, b]^d.$$

To prove this theorem based on Theorem 1.1, we only need to show σ can be approximated by a fixed-size $\tilde{\sigma}$ -activated network within an arbitrary error on any pre-specified interval as presented in the following lemma.

Lemma 5.4. *For any $\varepsilon > 0$ and any $M > 0$, there exists a function ϕ realized by a $\tilde{\sigma}$ -activated network with width 29 and depth 6 such that*

$$|\phi(x) - \sigma(x)| < \varepsilon \quad \text{for any } x \in [-M, M].$$

The proof of Lemma 5.4 can be found later. By assuming Lemma 5.4 is true, we can give the proof of Theorem 5.3.

1143 *Proof of Theorem 5.3.* By Theorem 1.1, there exists a function ϕ_σ generated by a σ -
 1144 activated network with width $36d(2d+1)$ and depth 11 such that

$$1145 \quad |\phi_\sigma(\mathbf{x}) - f(\mathbf{x})| < \varepsilon/2 \quad \text{for any } \mathbf{x} \in [a, b]^d.$$

1146 By Lemma 5.4, for any $M > 0$ and each $\delta \in (0, 1)$, there exists a function σ_δ realized by
 1147 a $\tilde{\sigma}$ -activated network with width 29 and depth 6 such that

$$1148 \quad \sigma_\delta(t) \rightrightarrows \sigma(t) \quad \text{as } \delta \rightarrow 0^+ \quad \text{for any } t \in [-M, M].$$

1149 Then, by Lemma 5.1, there exists another function $\phi = \phi_{\tilde{\sigma}}$ realized by a $\tilde{\sigma}$ -activated
 1150 network with width $29 \times 36d(2d+1) = 1044d(2d+1)$ and depth $6 \times 11 = 66$ such that

$$1151 \quad |\phi(\mathbf{x}) - f(\mathbf{x})| < \varepsilon \quad \text{for any } \mathbf{x} \in [a, b]^d.$$

1152 So we finish the proof. □

1153 Finally, let us present the detailed proof of Lemma 5.4.

1154 *Proof of Lemma 5.4.* Since $1 = \tilde{\sigma}'(0) = \lim_{x \rightarrow 0} \frac{\tilde{\sigma}(x)}{x}$, it is easy to show: For any $\mathcal{E} > 0$ and
 1155 any $R > 0$, there exists a sufficiently small $w > 0$ such that

$$1156 \quad \|\tilde{\sigma}(wx)/w - x\|_{L^\infty([-R, R])} < \mathcal{E}.$$

1157 Thus, we may assume the identity map is allowed to be the activation function in $\tilde{\sigma}$ -
 1158 activated networks. Without loss of generality, we may assume $M \geq 2$ because $\widehat{M} =$
 1159 $\max\{2, M\}$ implies $[-M, M] \subseteq [-\widehat{M}, \widehat{M}]$.

1160 For simplicity, we denote $\mathcal{H}_{\tilde{\sigma}}(N, L)$ as the (hypothesis) space of functions generated
 1161 by $\tilde{\sigma}$ -activated networks with width N and depth L . Then the proof can be roughly
 1162 divided into three steps as follows.

1163 (1) Design $\Gamma \in \mathcal{H}_{\tilde{\sigma}}(9, 2)$ to reproduce xy on $[-4\widehat{M}, 4\widehat{M}]^2$, where $\widehat{M} = (M+1)^2$.

1164 (2) Design $\psi_\delta \in \mathcal{H}_{\tilde{\sigma}}(20, 4)$ based on the first step to approximate σ well on $[0, M]$.

1165 (3) Design $\phi \in \mathcal{H}_{\tilde{\sigma}}(29, 6)$ based on the previous two steps to approximate σ well on
 1166 $[-M, M]$.

1167 The details of the three steps can be found below.

1168 **Step 1:** Design $\Gamma \in \mathcal{H}_{\tilde{\sigma}}(9, 2)$ to reproduce xy on $[-4\widehat{M}, 4\widehat{M}]^2$.

1169 Observe that

$$1170 \quad \tilde{\sigma}(y) + 1 = \frac{y}{|y| + 1} + 1 = \frac{y}{-y + 1} + 1 = \frac{1}{-y + 1} \quad \text{for any } y \leq 0.$$

1171 For any $x \in [-4, 4]$, we have $-x - 4 \leq 0$ and $-x - 5 \leq 0$, implying

$$1172 \quad \begin{aligned} \tilde{\sigma}(-x - 4) - \tilde{\sigma}(-x - 5) &= (\tilde{\sigma}(-x - 4) + 1) - (\tilde{\sigma}(-x - 5) + 1) \\ &= \frac{1}{-(-x - 4) + 1} - \frac{1}{-(-x - 5) + 1} = \frac{1}{x + 5} - \frac{1}{x + 6} = \frac{1}{(x + 5)(x + 6)}. \end{aligned}$$

1173 It follows from $1 - \frac{90}{(x+5)(x+6)} \leq 0$ for any $x \in [-4, 4]$ that

$$1174 \quad \tilde{\sigma}\left(1 - \frac{90}{(x+5)(x+6)}\right) + 1 = \frac{1}{-\left(1 - \frac{90}{(x+5)(x+6)}\right) + 1} = \frac{x^2 + 11x + 30}{90},$$

1175 implying

$$\begin{aligned} x^2 &= 90\tilde{\sigma}\left(1 - \frac{90}{(x+5)(x+6)}\right) + 90 - (11x + 30) \\ 1176 \quad &= 90\tilde{\sigma}\left(1 - 90(\tilde{\sigma}(-x-4) - \tilde{\sigma}(-x-5))\right) - 11x + 60 \\ &= 90\tilde{\sigma}\left(1 - 90\tilde{\sigma}(-x-4) + 90\tilde{\sigma}(-x-5)\right) - 11x + 60. \end{aligned} \quad (5.2)$$

1177 Thus, x^2 can be realized by a $\tilde{\sigma}$ -activated network with width 3 and depth 2 on $[-4, 4]$.
 1178 Set $\widetilde{M} = (M+1)^2$. Then, for any $x, y \in [-4\widetilde{M}, 4\widetilde{M}]$, we have $\frac{x}{2\widetilde{M}}, \frac{y}{2\widetilde{M}}, \frac{x+y}{2\widetilde{M}} \in [-4, 4]$. Recall
 1179 the fact

$$1180 \quad xy = 2\widetilde{M}^2\left(\left(\frac{x+y}{2\widetilde{M}}\right)^2 - \left(\frac{x}{2\widetilde{M}}\right)^2 - \left(\frac{y}{2\widetilde{M}}\right)^2\right).$$

1181 Thus, xy can be realized by a $\tilde{\sigma}$ -activated network with width 9 and depth 2 for any $x, y \in$
 1182 $[-4\widetilde{M}, 4\widetilde{M}]$. That is, there exists $\Gamma \in \mathcal{H}_{\tilde{\sigma}}(9, 2)$ such that $\Gamma(x, y) = xy$ on $[-4\widetilde{M}, 4\widetilde{M}]^2$.

1183 **Step 2:** Design $\psi_{\delta} \in \mathcal{H}_{\tilde{\sigma}}(9, 4)$ to approximate σ well on $[0, M]$.

1184 Recall that x^2 can be realized by a $\tilde{\sigma}$ -activated network with width 3 and depth 2
 1185 on $[-4, 4]$. There exists $\psi_1 \in \mathcal{H}_{\tilde{\sigma}}(3, 2)$ such that

$$1186 \quad \psi_1(x) = \frac{(2x+1)^2}{(2M+1)^2} \quad \text{for any } x \in [-M, M].$$

1187 Define

$$1188 \quad \psi_{2,\delta}(x) := \frac{\tilde{\sigma}(x+\delta) - \tilde{\sigma}(x)}{\delta} \quad \text{for any } x \in \mathbb{R}.$$

1189 Then, we have $\psi_{2,\delta} \in \mathcal{H}_{\tilde{\sigma}}(2, 1)$ and

$$1190 \quad \psi_{2,\delta}(x) := \frac{\tilde{\sigma}(x+\delta) - \tilde{\sigma}(x)}{\delta} \Rightarrow \frac{d}{dx}\tilde{\sigma}(x) = \frac{c\sigma(x) + 1}{(2x+1)^2} \quad \text{as } \delta \rightarrow 0^+,$$

1191 for any $x \in [0, M]$ and

$$1192 \quad c = \frac{1}{2 \int_0^\infty \frac{\sigma(t)}{(2t+1)^2} dt} \approx 2.554.$$

1193 Define

$$1194 \quad \psi_{\delta}(x) := \frac{(2M+1)^2}{c} \Gamma\left(\psi_1(x), \psi_{2,\delta}(x)\right) - \frac{1}{c} \quad \text{for any } x \in \mathbb{R}.$$

1195 Since $\Gamma \in \mathcal{H}_{\tilde{\sigma}}(9, 2)$, $\psi_1 \in \mathcal{H}_{\tilde{\sigma}}(3, 2)$, and $\psi_{2,\delta} \in \mathcal{H}_{\tilde{\sigma}}(2, 1)$, we have $\psi_{\delta} \in \mathcal{H}_{\tilde{\sigma}}(9, 4)$.

1196 Clearly, for any $x \in [0, M]$, we have $\psi_1(x) = \frac{(x+1)^2}{(2M+1)^2} \in [0, 1]$ and $\psi_{2,\delta}(x) \approx \frac{c\sigma(x)+1}{(2x+1)^2} \in$
 1197 $[0, 3.6]$, implying $\psi_1(x), \psi_{2,\delta}(x) \in [-4, 4] \subseteq [-4\widetilde{M}, 4\widetilde{M}]^2$ for any small $\delta > 0$. Thus, for
 1198 any $x \in [0, M]$, as δ goes to 0^+ , we get

$$\begin{aligned} \psi_{\delta}(x) &= \frac{(2M+1)^2}{c} \Gamma\left(\psi_1(x), \psi_{2,\delta}(x)\right) - \frac{1}{c} = \frac{(2M+1)^2}{c} \cdot \psi_1(x) \cdot \psi_{2,\delta}(x) - \frac{1}{c} \\ 1199 \quad &\Rightarrow \frac{(2M+1)^2}{c} \cdot \frac{(2x+1)^2}{(2M+1)^2} \cdot \frac{c\sigma(x)+1}{(2x+1)^2} - \frac{1}{c} = \sigma(x). \end{aligned}$$

1200 That is, for any $x \in [0, M]$,

$$1201 \quad \psi_\delta(x) \rightrightarrows \sigma(x) \quad \text{as } \delta \rightarrow 0^+.$$

1202 **Step 3:** Design $\phi \in \mathcal{H}_{\tilde{\sigma}}(29, 6)$ to approximate σ well on $[-M, M]$.

1203 Note that $\tilde{\sigma}(x) = \sigma(x)$ for all $x \in [-M, 0]$ and $\psi_\delta(x)$ approximates $\sigma(x)$ well for
 1204 all $x \in [0, M]$. Then, $\tilde{\sigma}(x) \cdot \mathbb{1}_{\{x \in [-M, 0]\}} + \psi_\delta(x) \cdot \mathbb{1}_{\{x \in [0, M]\}}$ approximates $\sigma(x)$ well for
 1205 all $x \in [-M, M]$. To design ϕ approximating σ well on $[-M, M]$, we need to design a
 1206 $\tilde{\sigma}$ -activated network to approximate an indicator function $\mathbb{1}_{\{x \in [0, M]\}}$ well.

1207 It is impossible to approximate $\mathbb{1}_{\{x \in [0, M]\}}$ well by a $\tilde{\sigma}$ -activated network due to the
 1208 continuity of $\tilde{\sigma}$. However, we define a continuous function g to replace $\mathbb{1}_{\{x \in [0, M]\}}$. By the
 1209 continuity of $\tilde{\sigma}$ and σ , there exists $\eta_0 \in (0, 1)$ such that

$$1210 \quad |\tilde{\sigma}(x)| < \varepsilon/6 \quad \text{and} \quad |\sigma(x)| < \varepsilon/6 \quad \text{for any } x \in [0, \eta_0]. \quad (5.3)$$

1211 Then we define

$$1212 \quad g(x) := \frac{\text{ReLU}(x) - \text{ReLU}(x - \eta_0)}{\eta_0}, \quad \text{where } \text{ReLU}(x) = \max\{0, x\} \quad \text{for any } x \in \mathbb{R}.$$

1213 See Figure 15 for an illustration of g .

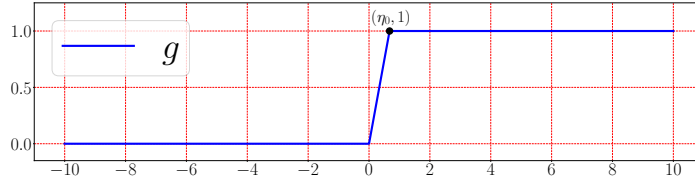


Figure 15: An illustration of g on $[-10, 10]$.

1214 We will construct a $\tilde{\sigma}$ -activated network to approximate g well. To this end, we
 1215 first design a $\tilde{\sigma}$ -activated network to approximate the ReLU function well. For $x \in$
 1216 $[-M - 1, M + 1]$, we have $\frac{x}{M+1} + 1 \in [0, 2] \subseteq [0, M]$, implying

$$1217 \quad 1 - \psi_\delta\left(\frac{x}{M+1} + 1\right) \rightrightarrows 1 - \sigma\left(\frac{x}{M+1} + 1\right) = \left|\frac{x}{M+1}\right| \quad \text{as } \delta \rightarrow 0^+,$$

1218 where the last equality comes from $1 - \sigma(y) = |y - 1|$ for any $y \in [0, 2]$. Note that
 1219 $\text{ReLU}(x) = \frac{x}{2} + \frac{|x|}{2} = \frac{x}{2} + \frac{M+1}{2} \cdot \left|\frac{x}{M+1}\right|$ for any $x \in [-M - 1, M + 1]$. Define

$$1220 \quad \tilde{g}_\delta(x) := \frac{x}{2} + \frac{M+1}{2} \left(1 - \psi_\delta\left(\frac{x}{M+1} + 1\right)\right) \quad \text{for any } x \in \mathbb{R}.$$

1221 Then, $\psi_\delta \in \mathcal{H}_{\tilde{\sigma}}(9, 4)$ implies $\tilde{g}_\delta \in \mathcal{H}_{\tilde{\sigma}}(10, 4)$. Moreover, for any $x \in [-M - 1, M + 1]$,

$$1222 \quad \tilde{g}_\delta(x) \rightrightarrows \frac{x}{2} + \frac{M+1}{2} \cdot \left|\frac{x}{M+1}\right| = \text{ReLU}(x) \quad \text{as } \delta \rightarrow 0^+.$$

1223 Define

$$1224 \quad g_\delta(x) := \frac{\tilde{g}_\delta(x) - \tilde{g}_\delta(x - \eta_0)}{\eta_0} \quad \text{for any } x \in \mathbb{R}.$$

Clearly, $\tilde{g}_\delta \in \mathcal{H}_{\tilde{\sigma}}(10, 4)$ implies $g_\delta \in \mathcal{H}_{\tilde{\sigma}}(20, 4)$. For any $x \in [-M, M]$, we have $x, x - \eta_0 \in [-M - 1, M + 1]$, implying

$$g_\delta(x) = \frac{\tilde{g}_\delta(x) - \tilde{g}_\delta(x - \eta_0)}{\eta_0} \rightrightarrows \frac{\text{ReLU}(x) - \text{ReLU}(x - \eta_0)}{\eta_0} = g(x) \quad \text{as } \delta \rightarrow 0^+.$$

Next, define

$$\phi_\delta(x) := \Gamma(\psi_\delta(x), g_\delta(x)) + \Gamma(\tilde{\sigma}(x), 1 - g_\delta(x)) \quad \text{for any } x \in \mathbb{R}.$$

Since $\Gamma \in \mathcal{H}_{\tilde{\sigma}}(9, 2)$, $\psi_\delta \in \mathcal{H}_{\tilde{\sigma}}(9, 4)$, and $g_\delta \in \mathcal{H}_{\tilde{\sigma}}(20, 4)$, we have $\phi_\delta \in \mathcal{H}_{\tilde{\sigma}}(29, 6)$.

Clearly, $\tilde{\sigma}(x)$, $g_\delta(x)$, and $1 - g_\delta(x)$ are all in $[-4\widetilde{M}, 4\widetilde{M}]$ for any small $\delta > 0$ and all $x \in [-M, M]$. We will show $\psi_\delta(x) \in [-4\widetilde{M}, 4\widetilde{M}]$ for any small $\delta > 0$ and all $x \in [-M, M]$ via two cases as follows.

• For $x \in [0, M]$, $\psi_\delta(x) \rightrightarrows \sigma(x)$ implies $\psi_\delta(x) \in [-4\widetilde{M}, 4\widetilde{M}]$ for any small $\delta > 0$.

• For $x \in [-M, 0)$, we have $\psi_1(x) = \frac{(x+1)^2}{(2M+1)^2} \in [0, 1]$ and

$$\psi_{2,\delta}(x) = \frac{\tilde{\sigma}(x+\delta) - \tilde{\sigma}(x)}{\delta} \rightrightarrows \frac{d}{dx}\tilde{\sigma}(x) = \frac{1}{(-x+1)^2} \quad \text{as } \delta \rightarrow 0^+.$$

Thus, for any $x \in [-M, 0)$, as δ goes to 0^+ , we get

$$\begin{aligned} \psi_\delta(x) &= \frac{(2M+1)^2}{c} \Gamma(\psi_1(x), \psi_{2,\delta}(x)) - \frac{1}{c} = \frac{(2M+1)^2}{c} \cdot \psi_1(x) \cdot \psi_{2,\delta}(x) - \frac{1}{c} \\ &\rightrightarrows \frac{(2M+1)^2}{c} \cdot \frac{(2x+1)^2}{(2M+1)^2} \cdot \frac{1}{(-x+1)^2} - \frac{1}{c} = \frac{(2x+1)^2 - 1}{c(-x+1)^2}. \end{aligned}$$

Since $\widetilde{M} = (M+1)^2$, we have $\frac{(2x+1)^2 - 1}{c(-x+1)^2} \in [0, 4\widetilde{M} - 1]$ for all $x \in [-M, 0)$, implying $\psi_\delta(x) \in [-4\widetilde{M}, 4\widetilde{M}]$ for any small $\delta > 0$.

Thus, for any $x \in [\eta_0, M]$, we have $1 - g(x) = 0$, implying

$$\phi_\delta(x) = \psi_\delta(x) \cdot g_\delta(x) + \tilde{\sigma}(x) \cdot (1 - g_\delta(x)) \rightrightarrows \sigma(x) \cdot g(x) + 0 = \sigma(x) \quad \text{as } \delta \rightarrow 0^+.$$

Similarly, for any $x \in [-M, 0]$, we have $g(x) = 0$, implying

$$\phi_\delta(x) = \psi_\delta(x) \cdot g_\delta(x) + \tilde{\sigma}(x) \cdot (1 - g_\delta(x)) \rightrightarrows 0 + \tilde{\sigma}(x) \cdot (1 - g(x)) = \sigma(x) \quad \text{as } \delta \rightarrow 0^+.$$

Therefore, there exists a small $\delta_0 > 0$ such that

$$|\phi_{\delta_0}(x) - \sigma(x)| < \varepsilon \quad \text{for any } x \in [-M, 0] \cup [\eta_0, M],$$

$$\|g_{\delta_0}\|_{L^\infty([0, \eta_0])} \leq 2, \quad \|1 - g_{\delta_0}\|_{L^\infty([0, \eta_0])} \leq 2, \quad \text{and}$$

$$\|\psi_{\delta_0}\|_{L^\infty([0, \eta_0])} \leq \|\sigma\|_{L^\infty([0, \eta_0])} + \varepsilon/12,$$

where the above inequality comes from $\psi_\delta(x)$ uniformly converges to $\sigma(x)$ for any $x \in [0, \eta_0] \subseteq [0, M]$.

Clearly, for $x \in [0, \eta_0]$, by Equation (5.3), we have

$$\begin{aligned}
|\phi_{\delta_0}(x) - \sigma(x)| &\leq |\phi_{\delta_0}(x)| + |\sigma(x)| < \left| \psi_{\delta_0}(x) \cdot g_{\delta_0}(x) + \tilde{\sigma}(x) \cdot (1 - g_{\delta_0}(x)) \right| + \varepsilon/6 \\
&\leq |\psi_{\delta_0}(x)| \cdot |g_{\delta_0}(x)| + |\tilde{\sigma}(x)| \cdot |1 - g_{\delta_0}(x)| + \varepsilon/6 \\
&\leq \left(\|\sigma\|_{L^\infty([0, \eta_0])} + \frac{\varepsilon}{12} \right) \cdot 2 + \frac{\varepsilon}{6} \cdot 2 + \frac{\varepsilon}{6} \\
&\leq \left(\frac{\varepsilon}{6} + \frac{\varepsilon}{12} \right) \cdot 2 + \frac{\varepsilon}{6} \cdot 2 + \frac{\varepsilon}{6} = \varepsilon.
\end{aligned}$$

By setting $\phi = \phi_{\delta_0}$, we have $\phi = \phi_{\delta_0} \in \mathcal{H}_{\tilde{\sigma}}(29, 6)$ and

$$|\phi(x) - \sigma(x)| = |\phi_{\delta_0}(x) - \sigma(x)| < \varepsilon \quad \text{for any } x \in [-M, M].$$

So we finish the proof. \square

5.3 Proof of Lemma 5.1

Let the activation function be applied to a vector elementwisely. Then, ϕ_{ϱ} can be represented in a form of function compositions as follows:

$$\phi_{\varrho}(\mathbf{x}) = \mathcal{L}_L \circ \varrho \circ \mathcal{L}_{L-1} \circ \varrho \circ \cdots \circ \varrho \circ \mathcal{L}_1 \circ \varrho \circ \mathcal{L}_0(\mathbf{x}) \quad \text{for any } \mathbf{x} \in \mathbb{R}^d,$$

where $N_0 = d$, $N_1, N_2, \dots, N_L \in \mathbb{N}^+$, $N_{L+1} = 1$, $\mathbf{A}_{\ell} \in \mathbb{R}^{N_{\ell+1} \times N_{\ell}}$ and $\mathbf{b}_{\ell} \in \mathbb{R}^{N_{\ell+1}}$ are the weight matrix and the bias vector in the ℓ -th affine linear transform $\mathcal{L}_{\ell} : \mathbf{y} \mapsto \mathbf{A}_{\ell} \mathbf{y} + \mathbf{b}_{\ell}$ for each $\ell \in \{0, 1, \dots, L\}$. Define

$$\phi_{\varrho_{\delta}}(\mathbf{x}) := \mathcal{L}_L \circ \varrho_{\delta} \circ \mathcal{L}_{L-1} \circ \varrho_{\delta} \circ \cdots \circ \varrho_{\delta} \circ \mathcal{L}_1 \circ \varrho_{\delta} \circ \mathcal{L}_0(\mathbf{x}) \quad \text{for any } \mathbf{x} \in \mathbb{R}^d.$$

Recall that ϱ_{δ} can be realized by a $\tilde{\varrho}$ -activated network with width \tilde{N} and depth \tilde{L} . Thus, $\phi_{\varrho_{\delta}}$ can be realized by a $\tilde{\varrho}$ -activated network with width $N\tilde{N}$ and depth $L\tilde{L}$.

We will prove

$$\phi_{\varrho_{\delta}}(\mathbf{x}) \rightrightarrows \phi_{\varrho}(\mathbf{x}) \quad \text{as } \delta \rightarrow 0^+ \quad \text{for any } \mathbf{x} \in [a, b]^d.$$

For any $\mathbf{x} \in \mathbb{R}^d$ and each $\ell \in \{1, 2, \dots, L+1\}$, define

$$\mathbf{h}_{\ell}(\mathbf{x}) := \mathcal{L}_{\ell-1} \circ \varrho \circ \mathcal{L}_{\ell-2} \circ \varrho \circ \cdots \circ \varrho \circ \mathcal{L}_1 \circ \varrho \circ \mathcal{L}_0(\mathbf{x})$$

and

$$\mathbf{h}_{\ell, \delta}(\mathbf{x}) := \mathcal{L}_{\ell-1} \circ \varrho_{\delta} \circ \mathcal{L}_{\ell-2} \circ \varrho_{\delta} \circ \cdots \circ \varrho_{\delta} \circ \mathcal{L}_1 \circ \varrho_{\delta} \circ \mathcal{L}_0(\mathbf{x}).$$

Note that \mathbf{h}_{ℓ} and $\mathbf{h}_{\ell, \delta}$ are two maps from \mathbb{R}^d to $\mathbb{R}^{N_{\ell}}$ for each ℓ .

We will prove by induction that

$$\mathbf{h}_{\ell, \delta}(\mathbf{x}) \rightrightarrows \mathbf{h}_{\ell}(\mathbf{x}) \quad \text{as } \delta \rightarrow 0^+ \tag{5.4}$$

for any $\mathbf{x} \in [a, b]^d$ and each $\ell \in \{1, 2, \dots, L+1\}$.

First, we consider the case $\ell = 1$. Clearly,

$$\mathbf{h}_{1, \delta}(\mathbf{x}) = \mathcal{L}_0(\mathbf{x}) = \mathbf{h}_1(\mathbf{x}) \quad \text{as } \delta \rightarrow 0^+ \quad \text{for any } \mathbf{x} \in [a, b]^d.$$

1278 This means Equation (5.4) holds for $\ell = 1$.

1279 Next, suppose Equation (5.4) holds for $\ell = i \in \{1, 2, \dots, L\}$. Our goal is to prove that
 1280 it also holds for $\ell = i + 1$. Define

$$1281 \quad M := \sup \left\{ \|\mathbf{h}_j(\mathbf{x})\|_{\ell^\infty} + 1 : \mathbf{x} \in [a, b]^d, \quad j = 1, 2, \dots, L + 1 \right\},$$

1282 where the continuity of ϱ guarantees the above supremum is finite. By the induction
 1283 hypothesis, we have

$$1284 \quad \mathbf{h}_{i,\delta}(\mathbf{x}) \rightrightarrows \mathbf{h}_i(\mathbf{x}) \quad \text{as } \delta \rightarrow 0^+ \quad \text{for any } \mathbf{x} \in [a, b]^d.$$

1285 Clearly, for any $\mathbf{x} \in [a, b]^d$, we have $\|\mathbf{h}_i(\mathbf{x})\|_{\ell^\infty} \leq M$ and $\|\mathbf{h}_{i,\delta}(\mathbf{x})\|_{\ell^\infty} \leq \|\mathbf{h}_i(\mathbf{x})\|_{\ell^\infty} + 1 \leq$
 1286 M for any small $\delta > 0$.

1287 Recall the fact $\varrho_\delta(t) \rightrightarrows \varrho(t)$ as $\delta \rightarrow 0^+$ for any $t \in [-M, M]$. Then

$$1288 \quad \varrho_\delta \circ \mathbf{h}_{i,\delta}(\mathbf{x}) - \varrho \circ \mathbf{h}_{i,\delta}(\mathbf{x}) \rightrightarrows \mathbf{0} \quad \text{as } \delta \rightarrow 0^+.$$

1289 The continuity of ϱ implies the uniform continuity of ϱ on $[-M, M]$, deducing

$$1290 \quad \varrho \circ \mathbf{h}_{i,\delta}(\mathbf{x}) - \varrho \circ \mathbf{h}_i(\mathbf{x}) \rightrightarrows \mathbf{0} \quad \text{as } \delta \rightarrow 0^+ \quad \text{for any } \mathbf{x} \in [a, b]^d.$$

1291 Therefore, for any $\mathbf{x} \in [a, b]^d$, as $\delta \rightarrow 0^+$, we have

$$1292 \quad \varrho_\delta \circ \mathbf{h}_{i,\delta}(\mathbf{x}) - \varrho \circ \mathbf{h}_i(\mathbf{x}) = \underbrace{\varrho_\delta \circ \mathbf{h}_{i,\delta}(\mathbf{x}) - \varrho \circ \mathbf{h}_{i,\delta}(\mathbf{x})}_{\rightrightarrows \mathbf{0}} + \underbrace{\varrho \circ \mathbf{h}_{i,\delta}(\mathbf{x}) - \varrho \circ \mathbf{h}_i(\mathbf{x})}_{\rightrightarrows \mathbf{0}} \rightrightarrows \mathbf{0},$$

1293 implying

$$1294 \quad \mathbf{h}_{i+1,\delta}(\mathbf{x}) = \mathcal{L}_i \circ \varrho_\delta \circ \mathbf{h}_{i,\delta}(\mathbf{x}) \rightrightarrows \mathcal{L}_i \circ \varrho \circ \mathbf{h}_i(\mathbf{x}) = \mathbf{h}_{i+1}(\mathbf{x}).$$

1295 This means Equation (5.4) holds for $\ell = i + 1$. So we complete the inductive step.

1296 By the principle of induction, we have

$$1297 \quad \phi_{\varrho_\delta}(\mathbf{x}) = \mathbf{h}_{L+1,\delta}(\mathbf{x}) \rightrightarrows \mathbf{h}_{L+1}(\mathbf{x}) = \phi_\varrho(\mathbf{x}) \quad \text{as } \delta \rightarrow 0^+ \quad \text{for any } \mathbf{x} \in [a, b]^d.$$

1298 There exists a small $\delta_0 > 0$ such that

$$1299 \quad |\phi_{\varrho_{\delta_0}}(\mathbf{x}) - \phi_\varrho(\mathbf{x})| < \varepsilon/2 \quad \text{for any } \mathbf{x} \in [a, b]^d.$$

1300 By setting $\phi = \phi_{\varrho_{\delta_0}}$, we have

$$1301 \quad |\phi(\mathbf{x}) - f(\mathbf{x})| \leq |\phi_{\varrho_{\delta_0}}(\mathbf{x}) - \phi_\varrho(\mathbf{x})| + |\phi_\varrho(\mathbf{x}) - f(\mathbf{x})| < \varepsilon/2 + \varepsilon/2 = \varepsilon$$

1302 for any $\mathbf{x} \in [a, b]^d$. Moreover, $\phi = \phi_{\varrho_{\delta_0}}$ can be generated by a $\tilde{\varrho}$ -activated network with
 1303 width $N\tilde{N}$ and depth $L\tilde{L}$. So we finish the proof.

6 Conclusion

This paper studies the super approximation power of deep feed-forward neural networks with a fixed size. It is proved by construction that there exists an EUAF network architecture with d input neurons, a maximum width $36d(2d+1)$, 11 hidden layers, and at most $5437(d+1)(2d+1)$ nonzero parameters, achieving the universal approximation property by only adjusting its finitely many parameters. That is, without changing the network size, our EUAF network can approximate any continuous function $f : [a, b]^d \rightarrow \mathbb{R}$ within an arbitrarily small error $\varepsilon > 0$ with appropriate parameters depending on f , ε , d , a , and b . Moreover, augmenting this EUAF network using one more layer with 2 neurons can exactly realize a classification function $\sum_{j=1}^J r_j \cdot \mathbb{1}_{E_j}$ in $\bigcup_{j=1}^J E_j$ for any $J \in \mathbb{N}^+$, where r_1, r_2, \dots, r_J are distinct rational numbers, $\mathbb{1}_{E_j}$ is the indicator function of E_j for each j , and E_1, E_2, \dots, E_J are arbitrary pairwise disjoint closed bounded subsets of \mathbb{R}^d . While we are interested in the theoretical analysis here, it is interesting to explore the numerical implementation in various applications of the proposed EUAF neural network. Furthermore, it would be very interesting to investigate the generalization and optimization errors of the EUAF networks in deep learning.

Acknowledgments

Z. Shen is supported by Tan Chin Tuan Centennial Professorship. H. Yang was partially supported by the US National Science Foundation under award DMS-1945029. S. Zhang is supported by a Postdoctoral Fellowship under NUS ENDOWMENT FUND (EXP WBS) (01 651).

References

- [1] A. R. BARRON, *Universal approximation bounds for superpositions of a sigmoidal function*, IEEE Transactions on Information Theory, 39 (1993), pp. 930–945.
- [2] H. BÖLCSKEI, P. GROHS, G. KUTYNIOK, AND P. PETERSEN, *Optimal approximation with sparsely connected deep neural networks*, SIAM Journal on Mathematics of Data Science, 1 (2019), pp. 8–45.
- [3] A. BONITO, R. DEVORE, P. J. DIANE GUIGNARD, AND G. PETROVA, *Polynomial approximation of anisotropic analytic functions of several variables*, Constructive Approximation, 53 (2021), pp. 319–348.
- [4] L. CHEN AND C. WU, *A note on the expressive power of deep rectified linear unit networks in high-dimensional spaces*, Mathematical Methods in the Applied Sciences, 42 (2019), pp. 3400–3404.
- [5] A. COHEN, R. DEVORE, G. PETROVA, AND P. WOJTASZCZYK, *Optimal stable nonlinear approximation*, arXiv e-prints, (2020), p. arXiv:2009.09907.
- [6] G. CYBENKO, *Approximation by superpositions of a sigmoidal function*, MCSS, 2 (1989), pp. 303–314.

- [7] I. DAUBECHIES, R. DEVORE, S. FOUCART, B. HANIN, AND G. PETROVA, *Nonlinear approximation and (deep) ReLU networks*, Constructive Approximation, (2021).
- [8] R. A. DEVORE, *Nonlinear approximation*, Acta Numerica, 7 (1998), pp. 51–150.
- [9] W. E, C. MA, AND Q. WANG, *A priori estimates of the population risk for residual networks*, arXiv e-prints, (2019), p. arXiv:1903.02154.
- [10] W. E, C. MA, AND L. WU, *A priori estimates of the population risk for two-layer neural networks*, Communications in Mathematical Sciences, 17 (2019), pp. 1407–1425.
- [11] W. E AND Q. WANG, *Exponential convergence of the deep neural network approximation for analytic functions*, CoRR, abs/1807.00297 (2018).
- [12] W. E AND S. WOJTOWYTSCH, *A priori estimates for classification problems using neural networks*, arXiv e-prints, (2020), p. arXiv:2009.13500.
- [13] ———, *Representation formulas and pointwise properties for barron functions*, arXiv e-prints, (2020), p. arXiv:2006.05982.
- [14] J. HAN AND C. MORAGA, *The influence of the sigmoid function parameters on the speed of backpropagation learning*, in From Natural to Artificial Neural Computation, J. Mira and F. Sandoval, eds., Berlin, Heidelberg, 1995, Springer Berlin Heidelberg, pp. 195–201.
- [15] J. HE, X. JIA, J. XU, L. ZHANG, AND L. ZHAO, *Make ℓ_1 regularization effective in training sparse CNN*, Computational Optimization and Applications, 77 (2020), pp. 163–182.
- [16] K. HORNIK, *Approximation capabilities of multilayer feedforward networks*, Neural Networks, 4 (1991), pp. 251–257.
- [17] K. HORNIK, M. STINCHCOMBE, AND H. WHITE, *Multilayer feedforward networks are universal approximators*, Neural Networks, 2 (1989), pp. 359–366.
- [18] K. KAWAGUCHI, *Deep learning without poor local minima*, in Advances in Neural Information Processing Systems 29, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, eds., Curran Associates, Inc., 2016, pp. 586–594.
- [19] K. KAWAGUCHI AND Y. BENGIO, *Depth with nonlinearity creates no bad local minima in resnets*, Neural Networks, 118 (2019), pp. 167–174.
- [20] A. N. KOLMOGOROV, *On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition*, Doklady Akademii Nauk SSSR, 114 (1957), pp. 953–956.
- [21] P. LE AND W. ZUIDEMA, *Compositional distributional semantics with long short term memory*, in Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics, Denver, Colorado, June 2015, Association for Computational Linguistics, pp. 10–19.

- 1379 [22] Q. LI, T. LIN, AND Z. SHEN, *Deep learning via dynamical systems: An approxi-*
1380 *mation perspective*, Journal of European Mathematical Society, (to appear).
- 1381 [23] Q. LI, C. TAI, AND W. E, *Stochastic modified equations and dynamics of stochas-*
1382 *tic gradient algorithms I: Mathematical foundations*, Journal of Machine Learning
1383 Research, 20 (2019), pp. 1–47.
- 1384 [24] H. LIN AND S. JEGELKA, *Resnet with one-neuron hidden layers is a universal*
1385 *approximator*, in Advances in Neural Information Processing Systems, S. Bengio,
1386 H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds.,
1387 vol. 31, Curran Associates, Inc., 2018.
- 1388 [25] J. LU, Z. SHEN, H. YANG, AND S. ZHANG, *Deep network approximation for*
1389 *smooth functions*, SIAM Journal on Mathematical Analysis, (to appear).
- 1390 [26] V. MAIOROV AND A. PINKUS, *Lower bounds for approximation by MLP neural*
1391 *networks*, Neurocomputing, 25 (1999), pp. 81–91.
- 1392 [27] H. MONTANELLI, H. YANG, AND Q. DU, *Deep ReLU networks overcome the curse*
1393 *of dimensionality for bandlimited functions*, Journal of Computational Mathematics,
1394 (to appear).
- 1395 [28] B. NEYSHABUR, Z. LI, S. BHOJANAPALLI, Y. LECUN, AND N. SREBRO, *The*
1396 *role of over-parametrization in generalization of neural networks*, in International
1397 Conference on Learning Representations, 2019.
- 1398 [29] Q. N. NGUYEN AND M. HEIN, *The loss surface of deep and wide neural networks*,
1399 CoRR, abs/1704.08045 (2017).
- 1400 [30] P. PETERSEN AND F. VOIGTLAENDER, *Optimal approximation of piecewise smooth*
1401 *functions using deep ReLU neural networks*, Neural Networks, 108 (2018), pp. 296–
1402 330.
- 1403 [31] Z. SHEN, H. YANG, AND S. ZHANG, *Deep network approximation character-*
1404 *ized by number of neurons*, Communications in Computational Physics, 28 (2020),
1405 pp. 1768–1811.
- 1406 [32] —, *Deep network with approximation error being reciprocal of width to power of*
1407 *square root of depth*, Neural Computation, 33 (2021), pp. 1005–1036.
- 1408 [33] —, *Neural network approximation: Three hidden layers are enough*, Neural Net-
1409 works, 141 (2021), pp. 160–173.
- 1410 [34] —, *Optimal approximation rate of ReLU networks in terms of width and depth*,
1411 Journal de Mathématiques Pures et Appliquées, (to appear).
- 1412 [35] J. W. SIEGEL AND J. XU, *Optimal approximation rates and metric entropy of*
1413 *ReLU^k and cosine networks*, arXiv e-prints, (2021), p. arXiv:2101.12365.

- 1414 [36] J. TURIAN, J. BERGSTRÄ, AND Y. BENGIO, *Quadratic features and deep archi-*
1415 *tectures for chunking*, in Proceedings of Human Language Technologies: The 2009
1416 Annual Conference of the North American Chapter of the Association for Compu-
1417 tational Linguistics, Companion Volume: Short Papers, NAACL-Short '09, USA,
1418 2009, Association for Computational Linguistics, pp. 245–248.
- 1419 [37] Y. YANG AND Y. WANG, *Approximation in shift-invariant spaces with deep ReLU*
1420 *neural networks*, arXiv e-prints, (2020), p. arXiv:2005.11949.
- 1421 [38] D. YAROTSKY, *Optimal approximation of continuous functions by very deep ReLU*
1422 *networks*, in Proceedings of the 31st Conference On Learning Theory, S. Bubeck,
1423 V. Perchet, and P. Rigollet, eds., vol. 75 of Proceedings of Machine Learning Re-
1424 search, PMLR, 06–09 Jul 2018, pp. 639–649.
- 1425 [39] —, *Elementary superexpressive activations*, arXiv e-prints, (2021),
1426 p. arXiv:2102.10911.
- 1427 [40] D. YAROTSKY AND A. ZHEVNERCHUK, *The phase diagram of approximation rates*
1428 *for deep neural networks*, in Advances in Neural Information Processing Systems,
1429 H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, eds., vol. 33,
1430 Curran Associates, Inc., 2020, pp. 13005–13015.
- 1431 [41] S. ZHANG, *Deep neural network approximation via function compositions*, PhD
1432 Thesis, National University of Singapore, (2020). URL: [https://scholarbank.](https://scholarbank.nus.edu.sg/handle/10635/186064)
1433 [nus.edu.sg/handle/10635/186064](https://scholarbank.nus.edu.sg/handle/10635/186064).