Minimum Width of Leaky-ReLU Neural Networks for Uniform Universal Approximation

Li'ang Li [*1] **Yifei Duan** [*1] **Guanghua Ji** [1] **Yongqiang Cai** [1]

Abstract

The study of universal approximation properties (UAP) for neural networks (NN) has a long history. When the network width is unlimited, only a single hidden layer is sufficient for UAP. In contrast, when the depth is unlimited, the width for UAP needs to be not less than the critical width $w^*_{\min} = \max(d_x, d_y)$, where d_x and d_y are the dimensions of the input and output, respectively. Recently, (Cai, 2022) shows that a leaky-ReLU NN with this critical width can achieve UAP for L^p functions on a compact domain $\mathcal{K}$, *i.e.,* the UAP for $L^p(\mathcal{K}, \mathbb{R}^{d_y})$. This paper examines a uniform UAP for the function class $C(\mathcal{K}, \mathbb{R}^{d_y})$ and gives the exact minimum width of the leaky-ReLU NN as $w_{\min} = \max(d_x + 1, d_y) + 1_{d_y = d_x+1}$, which involves the effects of the output dimensions. To obtain this result, we propose a novel lift-flow-discretization approach that shows that the uniform UAP has a deep connection with topological theory.

1. Introduction

The universal approximation theorem is important for the development of artificial neural networks. Artificial neural networks can approximate functions with arbitrary precision, this fact reveals the great potential of neural networks, and provides important guarantees for its development. (Cybenko, 1989) produces the original universal approximation theorem, stating that an arbitrarily wide feedforward neural network with a single hidden layer and sigmoid activation function can arbitrarily approximate continuous function. (Hornik, 1991) later demonstrated that the key to the universal approximation property lies in the multilayer and neuron architecture rather than the choice of an activation function. Then, (Leshno et al., 1993) show that for a continuous activation function $f : X \to R^{d_y}$ defined on a compact set $X \subseteq R^{d_x}$ can be approximated by a single hidden layer neural network, if and only if, the activation function is a nonpolynomial function.

After solving the activation function's theoretical problem, the field of vision naturally shifted to a consideration of the width and depth of the neural network. With the gradual development of deep neural networks, researchers have begun to pay attention to how to theoretically analyze the expressiveness of networks. (Daniely, 2017) simplifies the proof that the expressive ability of the three-layer neural network is superior to that of the two-layer neural network. For any positive integer k, (Telgarsky, 2016) shows that there are neural networks with $\Theta(k^3)$ layers and fixed widths that cannot be approximated by networks with $\mathcal{O}(k)$ layers unless they have $\Omega(2^k)$ nodes [1]. The universal approximation theorem explains that deep–bounded neural networks with suitable activation functions are universal approximators. (Lu et al., 2017) explained that a neural network with a bounded width can also be a universal approximator, such as the width-$(d_x + 4)$ ReLU networks, where d_x is the input dimension. (Lu et al., 2017) also shows that a ReLU network of width d_x cannot be used for universal approximation.

Many studies, such as (Beise & Da Cruz, 2020; Hanin & Sellke, 2018; Park et al., 2021), have shown that for a narrow neural network (the width is not greater than the input dimension), it is difficult to attain the UAP. (Nguyen et al., 2018) noted that deep neural networks with a specific type of activation function generally need to have a width larger than the input dimension to guarantee that the network can produce disconnected decision regions. For ReLU networks, (Park et al., 2021) proved that the minimum width for L^p-UAP is $w_{\min} = \max(d_x + 1, d_y)$ and summarized the known upper/lower bounds on the minimum width for universal approximation. Furthermore, conclusions related to the UAP of continuous functions have yet to be studied.

[*]Equal contribution [1]School of Mathematical Sciences, Laboratory of Mathematics and Complex Systems, MOE, Beijing Normal University, Beijing, 100875, China. Correspondence to: Yongqiang Cai <caiyq.math@bnu.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

[1]$\Theta(k^3)$ means that it is bound both above and below by k^3 asymptotically; $\mathcal{O}(k)$ means that it is bounded above by k asymptotically; $\Omega(2^k)$ means that it is bounded below by 2^k asymptotically.

Table 1. A summary of the known minimum width of feed-forward neural networks for universal approximation. †

References	Functions	Activation	Minimum width
(Hanin & Sellke, 2018)	$\mathcal{C}(\mathcal{K}, \mathbb{R})$	ReLU	$w_{\min} = d_x + 1$
(Park et al., 2021)	$L^p(\mathbb{R}^{d_x}, \mathbb{R}^{d_y})$	ReLU	$w_{\min} = \max(d_x + 1, d_y)$
(Park et al., 2021)	$\mathcal{C}([0,1], \mathbb{R}^2)$	ReLU	$w_{min} = 3$
(Park et al., 2021)	$\mathcal{C}(\mathcal{K}, \mathbb{R}^{d_y})$	ReLU + STEP	$w_{\min} = \max(d_x + 1, d_y)$
(Cai, 2022)	$L^p(\mathcal{K}, \mathbb{R}^{d_y})$	Leaky-ReLU	$w_{\min} = \max(d_x, d_y, 2)$
(Cai, 2022)	$\mathcal{C}(\mathcal{K}, \mathbb{R}^{d_y})$	ReLU + FLOOR	$w_{\min} = \max(d_x, d_y, 2)$
Ours (Theorem 2.2)	$\mathcal{C}(\mathcal{K}, \mathbb{R}^{d_y})$	Leaky-ReLU	$w_{\min} = \max(d_x + 1, d_y) + 1_{d_y = d_x + 1}$
Ours (Lemma 2.4)	$\mathcal{C}(\mathcal{K}, \mathbb{R}^{d_x})$	Leaky-ReLU	$w_{\min} = d_x + 1$

† d_x and d_y are the input and output dimensions, respectively. $\mathcal{K} \subset \mathbb{R}^{d_x}$ is a compact domain and $p \in [1, \infty)$.

(Park et al., 2021) and (Cai, 2022) demonstrate the minimum width of some neural networks for C-UAP using noncontinuous activation functions. If only continuous monotonically increasing activation functions are used, the known minimum width is restricted to the ReLU NN for function class $\mathcal{C}([0,1], \mathbb{R}^2)$, where the critical width is $w_{\min} = 3$. Table 1 provides a summary of the known minimum width for UAP.

To determine the minimum width of uniform UAP on $\mathcal{C}(\mathcal{K}, \mathbb{R}^{d_y})$, we introduce a novel scheme called *lift-flow-discretization approach*. Based on the close relationship between uniform UAP and topology, the functions are embedded in high-dimensional diffeomorphisms, and feed-forward neural networks are used to approximate these flow maps. Finally, we determine the minimum width of leaky-ReLU neural networks for C-UAP on $\mathcal{C}(\mathcal{K}, \mathbb{R}^{d_y})$ to be $w_{\min} = \max(d_x + 1, d_y) + 1_{d_y = d_x + 1}$.

1.1. Contributions

1. Theorem 2.2 states that the minimum width of leaky-ReLU networks for $\mathcal{C}(\mathcal{K}, \mathbb{R}^{d_y})$ is exactly $\max(d_x + 1, d_y) + 1_{d_y = d_x + 1}$. This is the first time that the minimum width for the universal approximation of leaky-ReLU networks is fully provided. It is worth mentioning that the previous results for the minimum width for the uniform approximation are based on discontinuous activation functions. The conclusion of this paper is based on continuous activation functions such as the leaky-ReLU function.

2. Section 3 presents a novel approach for approximating continuous functions using a feedforwrd neural network from the perspective of topology. The lift-flow-discretization approach of combining topology and neural network approximation is the key to the proof in this paper. Our approach is generic for strictly monotone continuous activations, as they all correspond to diffeomorphisms.

1.2. Related work

Width and depth bounds. Theoretical analyses of the expressive power of neural networks have taken place over the years. (Cybenko, 1989) proposed a prototype of the early classic universal approximation theorem. Continuous univariate functions over bounded domains can be fitted with arbitrary precision using the sigmoid activation function. (Hornik et al., 1989; Leshno et al., 1993; Barron, 1994) obtained similar conclusions and extended them to a large class of activation functions, revealing the relationship between universal approximation and network structure.

The effect of neural network width on expressiveness is an enduring question. (Sutskever & Hinton, 2008; Le Roux & Bengio, 2008) and (Montufar, 2014) reveal the impact of depth and width, especially width, on the general approximation of belief networks, and networks with too narrow a width cannot complete the approximation task. The width has important research value for many emerging networks and different activation functions. Conventional conclusions tell us that networks with appropriate activation functions under bounded depths are universal approximators. Correspondingly, (Lu et al., 2017) proposed a general approximation theorem for ReLU networks with bounded widths. (Hanin & Sellke, 2018) also studied in the ReLU network, whose input dimension is d_x, hidden layer width is at most w and depth is not limited. To fit any continuous real-valued function, the minimum value of w is exactly $d_x + 1$.

For a deep neural network that satisfies the activation function $\sigma(\mathbb{R}) = \mathbb{R}$, to learn the disconnected regions, it is usually necessary to make the network width larger than the input dimension. If the network is narrow, the paths connecting the disconnected regions yield high-confidence predictions (Nguyen et al., 2018). (Chong, 2020) gives a direct algebraic proof of the universal approximation theorem, and (Beise et al., 2021) reveals the fundamental reason why the universal approximation of network functions with width $w \leq d_x$ from $\mathbb{R}^{d_x}$ to $\mathbb{R}$ is impossible.

(Park et al., 2021) gives the first definitive results for the critical width enabling the universal approximation of width-bounded networks. The minimum width for the L^p functions is $\max(d_x+1, d_y)$ using the ReLU activation functions. (Park et al., 2021) also shows that this conclusion is unsuitable for the uniform approximation of the ReLU network, but it still holds using the ReLU+STEP activation function. (Cai, 2022) shows that minimum widths for the C-UAP and L^p-UAP on compact domains have a universal lower bound $w_{min} = \max(d_x, d_y)$. (Cai, 2022) also shows the minimum width for the uniform approximation with some additional threshold activation functions.

Homeomorphism properties of networks. Residual networks (ResNets) are an advanced deep learning architecture for supervised learning problems. (Rousseau & Fablet, 2018) shows that a continuous flow of diffeomorphisms governed by ordinary differential equations can be numerically implemented using the mapping component of ResNets.

Neural ordinary differential equations (neural ODEs) turn the neural network training problem into a problem of solving differential equations and can make the discrete ResNet continuous. As a deep learning method, (Teshima et al., 2020b) shows the universality of discrete neural ODEs with the condition that the source vectors $f_i(z) \in \mathcal{H}$, where $\mathcal{H}$ is a universal approximator for the Lipschitz functions. (Ruiz-Balet & Zuazua, 2021) provide L^2-UAP for neural ODE $\dot{x} = W\sigma(Ax+b)$. (Zhang et al., 2019) shows that neural ODEs with extra dimensions are universal approximators for homeomorphisms.

Invertible neural networks have diffeomorphic properties, and many flow models can also be used as universal approximators. (Huang et al., 2018) shows that neural autoregressive flows are universal approximators for continuous probability distributions. (Teshima et al., 2020a) indicates that normalizing flow models based on affine coupling also have UAP. (Kong & Chaudhuri, 2021) shows that residual flows are universal approximators in maximum mean discrepancies.

1.3. Organization

We first define the necessary notation and the main results and give the proof ideas in Section 2. In Section 3, we present our *lift-flow-discretization approach* demonstrating the minimum width to achieve C-UAP. The detailed proof process is given in Section 4. Considering the influence of the output dimension, the final proof is divided into four parts. In Section 5, we give an outlook on the direction of our current work. All formal proofs are provided in the appendix.

2. Main results

We consider the standard feedforward neural network with the same number of neurons at each hidden layer. We say a σ-NN with depth L is a function with inputs $x \in \mathbb{R}^{d_x}$ and outputs $y \in \mathbb{R}^{d_y}$, which has the following form:

$$\begin{aligned} y &= f_{NN,L}(x) = y_L \\ &= W_{L+1}\sigma\left(W_L\left(\cdots\sigma\left(W_1 x + b_1\right) + \cdots\right) + b_L\right) + b_{L+1}, \end{aligned} \tag{1}$$

where b_i are vectors, W_i are matrices and $\sigma(\cdot)$ is the activation function. We mainly consider the number of neurons in all the layers to be the same N. In this case, $W_i \in \mathbb{R}^{N\times N}, b_i \in \mathbb{R}^N, i \in \{1, \cdots, L+1\}$, except $W_1 \in \mathbb{R}^{N\times d_x}$, $W_{L+1} \in \mathbb{R}^{d_y \times N}$ and $b_{L+1} \in \mathbb{R}^{d_y}$. We denote the set of all networks in Eq. (1) as $\mathcal{N}_{N,L}(\sigma)$, and $\mathcal{N}_N(\sigma) = \bigcup_L \mathcal{N}_{N,L}(\sigma)$. The activation function is crucial for the approximation power of the neural network. Our main results are for the following leaky-ReLU activations function with a fixed parameter $\alpha \in \mathbb{R}^+ \setminus \{1\}$,

$$\sigma(x) = \sigma_\alpha(x) = \begin{cases} x, & x > 0, \\ \alpha x, & x \le 0. \end{cases} \tag{2}$$

2.1. Main theorem

Our main theorem is the following Theorem 2.2, which provides the exact minimum width of the leaky-ReLU networks that process uniform universal approximations.

Definition 2.1. We say the leaky-ReLU networks with width N have C-UAP or L^p-UAP if the set $\mathcal{N}_N(\sigma)$ is dense in $C(\mathcal{K}, \mathbb{R}^{d_y})$ or $L^p(\mathcal{K}, \mathbb{R}^{d_y})$, respectively.

Theorem 2.2. *Let $\mathcal{K} \subset \mathbb{R}^{d_x}$ be a compact set; then, for the continuous function class $C(\mathcal{K}, \mathbb{R}^{d_y})$, the minimum width $w_{\min}$ of leaky-ReLU neural networks having C-UAP is exactly $w_{\min} = \max(d_x+1, d_y) + 1_{d_y = d_x+1}$. Thus, $\mathcal{N}_N(\sigma)$ is dense in $C(\mathcal{K}, \mathbb{R}^{d_y})$ if and only if $N \ge w_{\min}$.*

Before giving the proof, let's emphasize the points of Theorem 2.2. First, if the width N of the leaky-ReLU networks is smaller than $w_{\min}$, then there is a continuous function $f^* \in C(\mathcal{K}, \mathbb{R}^{d_y})$ that cannot be well approximated, *i.e.* there is a positive constant $\varepsilon > 0$ such that $\|f - f^*\| > \varepsilon$ for all $f \in \mathcal{N}_N(\sigma)$. For the case of $\mathcal{K} = [-1,1]^{d_x}, d_y = 1$, the function f^* can be chosen as $f^*(x) = \|x\|^2$. The reason will be given in the next section, which is based on the topological properties of the level sets (Johnson, 2019).

Second, if $N = w_{\min}$, then for any $f^* \in C(\mathcal{K}, \mathbb{R}^{d_y})$ and any $\varepsilon > 0$, we can construct a leaky-ReLU network f_L with width N and depth L such that $\|f - f^*\| < \varepsilon$. We will introduce the construction scheme later.

Lastly, the formula of $w_{\min}$ includes a characteristic func-

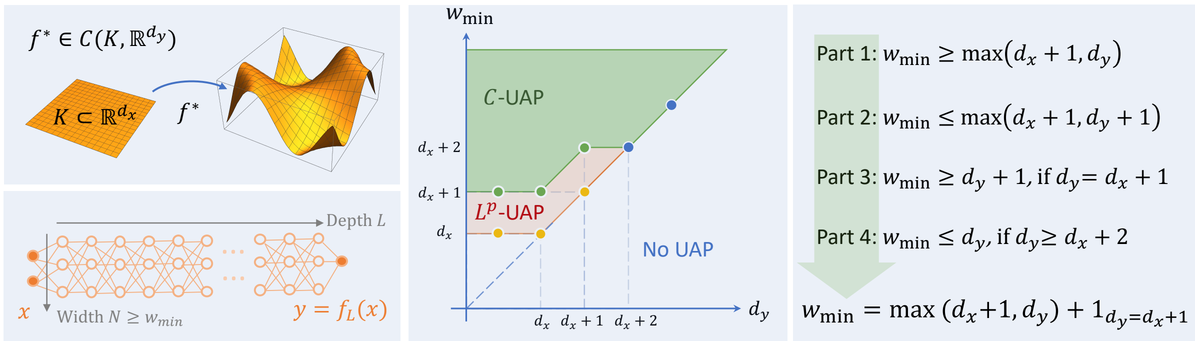


Figure 1. Minimum width of leaky-ReLU networks for universal approximation. (a) Example of function from $\mathcal{K} \subset \mathbb{R}^{d_x}$ to $\mathbb{R}^{d_y}$. (b) Feedforward neural networks with depth L and width N. (c) The minimum width of leaky-ReLU networks to reach UAP. (d) Proof parts of the main result.

tion $1_{d_y=d_x+1}$,

$$1_{d_y=d_x+1} = \begin{cases} 1, & d_y = d_x+1, \\ 0, & d_y \neq d_x+1, \end{cases}$$

which indicates that there is an obstacle of dimension. In fact, this is caused by the topology of the manifolds.

2.2. Proof ideas

Now, we provide the proof scheme, while the details will be given in the next section. As illustrated in Figure 1, the result of Theorem 2.2 is split into four parts: Part 1 and Part 2 give a lower bound and an upper bound for the general dimensions, Part 3 considers the exceptional case of $d_y = d_x + 1$, and Part 4 considers the case of $d_y \geq d_x + 2$.

Part 1 is based on the following lemma, which results in a lower bound of leaky-ReLU network to be $w_{\min} \geq \max(d_x + 1, d_y)$.

Lemma 2.3. *For any compact domain $\mathcal{K} \subset \mathbb{R}^{d_x}$, the leaky-ReLU networks with width $N < \max(d_x + 1, d_y)$ do not have UAP for $C(\mathcal{K}, \mathbb{R}^{d_y})$,* i.e. *$\mathcal{N}_N(\sigma)$ is not dense in $C(\mathcal{K}, \mathbb{R}^{d_y})$.*

The proof is based on the result of (Cai, 2022), which shows a universal lower bound $w_{\min} \geq \max(d_x, d_y)$ for arbitrary activations, and (Johnson, 2019), which shows that $w_{\min} \geq d_x + 1$ for monotone and continuous activations such as leaky-ReLU is sufficient.

Part 2 is based on the following lemma, which considers the case of $d_x = d_y = d$. If d_x and d_y are not the same, we can lift them to dimension $d = \max(d_x, d_y)$ by filling in zeros for the auxiliary dimensions.

Lemma 2.4. *For any continuous function $f^* \in C(\mathcal{K}, \mathbb{R}^d)$ on compact domain $\mathcal{K} \subset \mathbb{R}^d$, and $\varepsilon > 0$, there is a leaky-ReLU network $f_L(x)$ with depth L and width $d + 1$ such that $\|f_L(x) - f^*(x)\| \leq \varepsilon$ for all x in $\mathcal{K}$.*

Lemma 2.4 is our main result for Part 2, which shows that leaky-ReLU neural network with width $d + 1$ has enough expressive power to approximate continuous function f^* with $d_x = d_y = d$. The proof of Lemma 2.4 will be given in Section 3 as it is based on our lift-flow-discretization approach given in the next section.

The gap between the lower bound $w_{\min} \geq \max(d_x + 1, d_y)$ and the upper bound $w_{\min} \leq \max(d_x + 1, d_y + 1)$ is at most one. When $d_y \leq d_x$, it directly implies that $w_{\min} = \max(d_x + 1, d_y)$. Then, we consider the case of $d_y \geq d_x + 1$. In this case, the lower and upper bounds read $d_y \leq w_{\min} \leq d_y + 1$, and the question is whether width $N = d_y$ is sufficient for C-UAP.

Part 3 and Part 4 of our main result answer the question by showing that width d_y is enough for the case of $d_y \geq d_x + 2$ but not for the case of $d_y = d_x + 1$. The two cases are heavily related to topology theory. Here, we give a short example to show this phenomenon. Let $f^* \in C([0, 1], \mathbb{R}^2)$ be a parameterized curve shaped like '4', which has a self intersecting point. Then, f^* cannot be approximated by curve homeomorphic to a line segment. However, if $f^* \in C([0, 1], \mathbb{R}^3)$, the approximation is possible according to our lift-flow-discretization approach. We will show this example in detail in Section 4.

3. Lift-flow-discretization approach

Before presenting our proof of Part 3 and Part 4, we will first provide our key approach for proving Part 3 and Part 4 in Figure 1 which is called the lift-flow-discretization approach in this section. We reformulate network (1) as follows:

$$f_L(x) = W_{L+1} \Phi_L(W_1 x + b_1) + b_{L+1}, \tag{3}$$

where Φ_L is a map from $\mathbb{R}^N$ to $\mathbb{R}^N$ and $W_1 x + b_1$ and $W_{L+1}\Phi + b_{L+1}$ are linear maps. Since we use the leaky-ReLU activation and the weight matrix in (1), it can be assumed to be nonsingular, the map Φ_L is a homeomorphism. Motivated by the recent work of (Duan et al., 2022), which shows that leaky-ReLU networks can approximate flow maps, we propose an approach to approximate functions f^* in $C(\mathcal{K}, \mathbb{R}^{d_y})$ by lifting it as a diffeomorphism Φ and then we approximate Φ by flow maps and neural networks.

For any function f^* in $C(\mathcal{K}, \mathbb{R}^{d_y})$ and any $\varepsilon > 0$, our lift-flow-discretization approach includes three parts:

1) **(Lift)** A lift map $\Phi \in C(\mathbb{R}^N, \mathbb{R}^N)$, which is an orientation preserving (OP) diffeomorphism such that

$$\|f^*(x) - \beta \circ \Phi \circ \alpha(x)\| \le \varepsilon/3, \quad \forall x \in \mathcal{K}, \tag{4}$$

where α and β are two linear maps. Without loss of generality, we can assume that the Lipschitz constants of α and β are less than one. Within this notation, we say the map Φ is a lift of f^*.

2) **(Flow)** A flow map $\phi^\tau \in C(\mathbb{R}^N, \mathbb{R}^N)$ corresponding to a neural ODE

$$z'(t) = v(z(t), t), t \in (0, \tau), \quad z(0) = x, \tag{5}$$

which satisfies $\|\Phi(x) - \phi^\tau(x)\| \le \varepsilon/3$ for all x in $\alpha(\mathcal{K})$.

3) **(Discretization)** A discretization map $\psi \in C(\mathbb{R}^N, \mathbb{R}^N)$ is a leaky-ReLU network in $\mathcal{N}_N(\sigma)$ that approximates ϕ^τ such that $\|\psi(x) - \phi^\tau(x)\| \le \varepsilon/3$ for all x in $\alpha(\mathcal{K})$.

As a result, the composition $\beta \circ \psi \circ \alpha =: f_L$ is a leaky-ReLU network with width N, which approximates the target function f^* such that

$$\|f^*(x) - \beta \circ \psi \circ \alpha(x)\| \le \varepsilon, \quad \forall x \in \mathcal{K}. \tag{6}$$

3.1. Theory of the lift-flow-discretization approach

Note that the existence of ϕ^τ and ψ are guaranteed by the following lemmas based on the results of (Caponigro, 2011) and (Duan et al., 2022). We need to construct the lift map Φ, which will be constructed case by case.

Lemma 3.1. *Let Φ be an orientation preserving diffeomorphism of $\mathbb{R}^N$, Ω be a compact set in $\mathbb{R}^N$ and $\varepsilon > 0$. Then, there is an ODE with tanh neural fields, whose flow map is denoted by $\phi^\tau(x_0) = z(\tau)$,*

$$\begin{aligned} \dot{x}(t) &= v(x(t), t) \\ &\equiv \sum_{i=1}^{M} a_i(t) \tanh(w_i(t) \cdot x(t) + b_i(t)), t \in [0, \tau], \\ x(0) &= x_0 \in \mathbb{R}^N, \quad M \in \mathbb{Z}^+, \end{aligned} \tag{7}$$

where $a_i, w_i \in \mathbb{R}^N$ and $b_i \in \mathbb{R}$ are piecewise constant functions of t, such that $\|\phi^\tau(x_0) - \Phi(x_0)\| < \varepsilon$ for all x_0 in Ω.

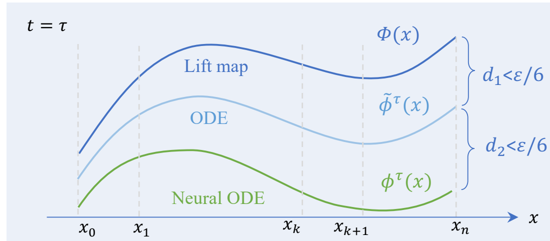


Figure 2. Sketch of the lift-flow-discretization approach. The target map $\Phi(x)$ is approximated by a flow map $\tilde{\phi}^\tau(x)$ of an ODE, which is further approximated by a flow map $\phi^\tau(x)$ of a neural ODE (7).

Lemma 3.1 ensures Step 2 ('flow') of our lift-flow-discretization approach, where we use the flow map of a neural ODE to approximate a given orientation preserving diffeomorphism.

The formal proof of the lemma can be seen in the appendix. Here, we provide the main idea of the proof. First, we refer to (Caponigro, 2011) to prove that for any $\varepsilon > 0$, there exists a flow map at the endpoint of time $\tilde{\phi}^\tau(x)$ of an ODE such that $\|\tilde{\phi}^\tau(x) - \Phi(x)\| < \varepsilon/2$ for all $x \in \alpha(\mathcal{K})$, then we use neural ODE (7) to approximate $\tilde{\phi}^\tau(x)$, there exist (a, w, b) such that the flow map (denoted as $\phi^\tau(x)$) of Eq. (7) satisfies $\|\phi^\tau(x) - \tilde{\phi}^\tau(x)\| < \varepsilon/2$, then $\|\phi^\tau(x) - \Phi(x)\| < \varepsilon$.

Lemma 3.2. *Let $\phi^\tau \in C(\mathbb{R}^N, \mathbb{R}^N)$ be the flow map in Lemma 3.1 and Ω be a compact set in $\mathbb{R}^N$ and $\varepsilon > 0$. Then, there is a leaky-ReLU network $\psi \in \mathcal{N}_N(\sigma)$ with width N and depth L such that $\|\phi^\tau(x_0) - \psi(x_0)\| < \varepsilon$ for all x_0 in Ω.*

This lemma ensures Step 3 (discretization) of our lift-flow-discretization approach, where we find a neural network to approximate ϕ^τ in Step 2.

The formal proof, motivated by (Duan et al., 2022), can be seen in the appendix. The main idea is to solve the ODE (7) by a splitting method and then approximate each split step by leaky-ReLU networks. Consider the following splitting for v in (7), $v(x,t) \equiv \sum_{i=1}^{N} \sum_{j=1}^{d} v_{ij}(x,t) e_j$ with $v_{ij}(x,t) = a_i^{(j)}(t) \tanh(w_i(t) \cdot x + b_i(t)) \in \mathbb{R}$. Then, the flow map can be approximated by an iteration with time step $\Delta t = \tau/n$, $n \in \mathbb{Z}^+$ large enough,

$$\begin{aligned} \phi^\tau(x_0) \approx x_n = T_n(x_{n-1}) &= T_n \circ \cdots \circ T_1(x_0), \\ &= T_n^{(N,d)} \circ \cdots \circ T_n^{(1,2)} \circ T_n^{(1,1)} \circ \cdots \circ \\ &\quad T_1^{(N,d)} \circ \cdots \circ T_1^{(1,2)} \circ T_1^{(1,1)}(x_0). \end{aligned}$$

where the k-th iteration is $x_{k+1} = T_k x_k = T_k^{(N,d)} \circ \cdots \circ$

$T_k^{(1,2)} \circ T_k^{(1,1)}(x_k)$, The map $T_k^{(i,j)} : x \to y$ in each split step is:

$$T_k^{(i,j)} : \begin{cases} y^{(l)} = x^{(l)}, l = 1, 2, .., j-1, j+1, ..., d, \\ y^{(j)} = x^{(j)} + \Delta t v_{ij}(x, k\Delta t). \end{cases}$$

Combining all the approximation networks, we have $\|\phi^\tau(x_0) - \psi(x_0)\| < \varepsilon$ for all $x_0 \in \Omega$.

Having reached the above conclusion, if the lift map Φ in Step 1 (lift) is constructed, we can obtain the following corollary.

Corollary 3.3. *Let $f^* \in C(\mathcal{K}, \mathbb{R}^d)(\mathcal{K} \subset \mathbb{R}^d)$ and $N \in \mathbb{Z}^+$. If for any $\varepsilon > 0$, there is an orientation preserving diffeomorphism Φ of $\mathbb{R}^N$ and two linear maps α and β such that $\|f^*(x) - \beta \circ \Phi \circ \alpha(x)\| < \varepsilon$ for all $x \in \mathcal{K}$, then there is a leaky-ReLU network $f_L \in \mathcal{N}_N(\sigma)$ with width N and depth L such that $\|f_L(x) - f^*(x)\| < \varepsilon$ for all $x \in \mathcal{K}$.*

This corollary shows the expressive power of the leaky-ReLU neural network, and it is our main result for Part 3 and Part 4, in Section 4, we will show that it holds for Part 4 ($d_y \geq d_x + 2$) while failing for Part 3 ($d_y = d_x + 1$).

3.2. Proof idea of Lemma 2.4

We can prove Lemma 2.4 by designing a proper lift map. According to the lift-flow-discretization approach, we only need to construct two linear maps and an orientation preserving diffeomorphism with $N = d + 1$, which satisfies the condition of Corollary 3.3. For any continuous function $f^* \in C(\mathcal{K}, \mathbb{R}^d)$ on compact domain $\mathcal{K} \subset \mathbb{R}^d$, and $\varepsilon > 0$, there is a locally Lipschitz smooth function p such that $\|f^*(x) - p(x)\| < \varepsilon$ for all $x \in \mathcal{K}$. The function p can be chosen as a polynomial according to the well-known Stone-Weierstass theorem. Then, the maps $\alpha : \mathbb{R}^d \to \mathbb{R}^{d+1}, \beta : \mathbb{R}^{d+1} \to \mathbb{R}^d$ and $\Phi : \mathbb{R}^{d+1} \to \mathbb{R}^{d+1}$ can be chosen as follows:

$$\alpha(x) = \begin{pmatrix} I_d \\ \mathbf{1}^T \end{pmatrix} x, \quad \Phi \begin{pmatrix} x \\ x_{d+1} \end{pmatrix} = \begin{pmatrix} p(x) + \kappa \mathbf{1}\mathbf{1}^T \\ x_{d+1} \end{pmatrix}, \quad (8)$$

$$\beta \begin{pmatrix} x \\ x_{d+1} \end{pmatrix} = (I, -\kappa \mathbf{1}) \begin{pmatrix} x \\ x_{d+1} \end{pmatrix} = x - \kappa \mathbf{1} x_{d+1}, \quad (9)$$

where $\mathbf{1} \in \mathbb{R}^d$ is a column vector with all being elements one, κ is a number larger than the Lipschitz constant of p on $\mathcal{K}$, and x_{d+1} is the coordinate of the auxiliary dimension. It is obvious that the maps α and β are linear and $p(x) = \beta \circ \Phi \circ \alpha(x)$. Our proof is constructive, and the formal proof is in appendix.

Our 'lift-flow-discretization' approach deeply connects the minimal width to topology theory, providing that the activation is a one-dimensional diffeomorphism.

4. Effect of the output dimension

Now we turn to Part 3 and Part 4 of the main results which consider the case of $d_y \geq d_x + 1$. We examine the approximation power of leaky-ReLU networks with width $N = d_y$.

We emphasize the homeomorphism properties. In fact, leaky-ReLU, the nonsingular linear transformer and their inverse are continuous and homeomorphic. Since compositions of homeomorphism are also homeomorphism, we have the input-output map as a homeomorphism. Note that a singular matrix can be approximated by nonsingular matrixes, therefore we can restrict the weight matrix in neural networks as nonsingular.

When $d_y > d_x$, we can reformulate the leaky-ReLU network with width $N = d_y$ as $f_L(x) = \psi(W_1 x + b_1), W_1 \in \mathbb{R}^{d_x \times d_y}, b \in \mathbb{R}^{d_y}$, where $\psi(\cdot)$ is a homeomorphism in dimension d_y.

4.1. The particular dimension $d_y = d_x + 1$

The following lemma shows that width $N = d_y$ is not enough for C-UAP which implies that $w_{\min} \geq d_y + 1$ when $d_y = d_x + 1$.

Lemma 4.1. *If $d_y = d_x + 1$, there exists a continuous function $f^*(x) \in C(\mathcal{K}, \mathbb{R}^{d_y})$ which can NOT be uniformly approximated by functions like $\psi(W_1 x + b_1)$ with homeomorphism maps $\psi : \mathbb{R}^{d_y} \to \mathbb{R}^{d_y}$.*

In order to prove this lemma, the counterexample (Figure 4(a)) we constructed seems very intuitive. In the counterexample, we need to prove that given target function $g(t) : \mathbb{R} \to \mathbb{R}^2$, for a sufficiently small $\varepsilon > 0$, h satisfying $\|g(t) - h(t)\| < \varepsilon$ for any $t \in [0, \tau]$ has a self-intersection point, we just need to prove that the curve starting from $(0, 1)$ to $(0, 1)$ and the curve starting from $(-1, 0)$ to $(1, 0)$ within $[0, 1] \times [0, 1]$ have at least one intersection within $[0, 1] \times [0, 1]$. This conclusion is so intuitive that even non-mathematics can also see at a glance that the two curves must have intersections. While the proof may seem complicated because we need knowledge of topology. Interested readers can refer to the proof in the appendix.

4.2. The case of $d_y \geq d_x + 2$

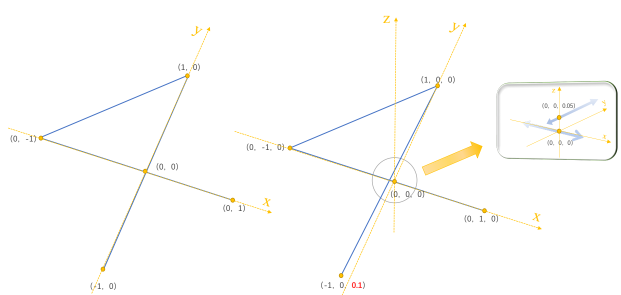
Note that we only need to consider the case of $d_y = d_x + 2$. The reason is that when $d_y > d_x + 2$, we can increase d_x by adding some auxiliary dimensions to the input. Then, employing the lift-flow-discretization approach, we only need to show that any $f^* \in C(\mathcal{K}, \mathbb{R}^{d_x+2})$ can be approximated by functions formulated as $\psi(Wx)$, where $\psi(\cdot)$ is an orientation preserving diffeomorphism in dimension $d_x + 2$.

Lemma 4.2. *For any $f^* \in C(\mathcal{K}, \mathbb{R}^{d_x+2})$ and $\varepsilon > 0$, there is a matrix $W \in \mathbb{R}^{(d_x+2) \times d_x}$ and an OP diffeomorphism map Φ such that $\|\Phi(Wx) - f^*(x)\| < \varepsilon$ for all x in $\mathcal{K}$.*

Here, we provide the main ideas. Let $f^* : I^d := [0,1]^d \to \mathbb{R}^{d+2}$ be a continuous map. Then, let f^* be the locally smooth homeomorphism and the transversal intersection at all the self-intersection points. Denote the set of self-intersection points of f as $\mathcal{A}$. We can prove that $\mathcal{A}$ is a compactly closed subset. For $\forall x \in \mathcal{A}$, there is an open neighborhood U of x; then, we can make perturbations in U to make the maps disjoint and not create new intersections. Thus, we obtain a smooth approximation f of f^* without intersections, which is the desired diffeomorphism approximation.

Because f is the embedding of $I^d \to \mathbb{R}^{d+2}$, it can be written as the composition of linear mapping and differential homeomorphism, which is $f(x) = \Phi(Wx), W \in \mathbb{R}^{(d_x+2)\times d_x}$. According to the lift and flow steps, there exists a flow map $\phi^\tau \in \mathcal{C}(\mathbb{R}^{d_x+2}, \mathbb{R}^{d_x+2})$ satisfying $\|\phi^\tau(x) - f(x)\| < \varepsilon/2$. According to the discretization step, there exists a leaky-ReLU network $\psi \in \mathcal{C}(\mathbb{R}^{d_x+2}, \mathbb{R}^{d_x+2})$ that satisfies $\|\psi(x) - \phi^\tau(x)\| < \varepsilon/2$.

By employing lift-flow-discretization approach, we can arrive at the desired result. To understand the result, we show an example of the case of $d_y \geq d_x + 2$. We have shown that a continuous function f^* from $[0,1]^{d_x} \to \mathbb{R}^{d_x+2}$ can be uniformly approximated by a diffeomorphism, we take a '4'-shape curve corresponding to a continuous function $f^*(t)$ from $[0,1] \to \mathbb{R}^3$ as an example.



(a) Target function (b) Approximate function

Figure 3. Example of $d_x = 1$. Approximate the '4'-shape curve (a) in $\mathbb{R}^2$ by lifting it to the three-dimensional curve (b) in $\mathbb{R}^3$.

From Figure 3, we lift the four vertices of the '4'-shape curve as $(-1,0,0), (1,0,0), (0,-1,0), (0,1,0)$, and connect them in turn to form a polyline, which is our target function $f^*(t), t \in [0,1]$. Then, we construct a curve without self-intersection points by changing one of the z-axis coordinates of the points to ε (such as $\varepsilon = 0.1$), the approximation function f is the curve connected by 4 vertices as $(-1,0,\varepsilon), (1,0,0), (0,-1,0), (0,1,0)$ in sequence, in Figure 3(b). Now the approximation function becomes a curve $\tilde{f}(t)$ without self-intersecting points, corresponding to a homeomorphic mapping Φ in $\mathbb{R}^3$ with $\tilde{f}(t) = \Phi(wt)$ for some $w \in \mathbb{R}^3$. Employing the flow and discretization steps, we can approximate Φ by leaky-ReLU networks. Consequently, we can conclude that the '4'-shape curve in Figure 3 can be approximated by leaky-ReLU networks.

5. Discussion

General Activation. It should be noted that our 'lift-flow-discretization' approach is generic for strictly monotone continuous activations. For example, our results are valid for strict monotone piecewise linear activations. We focus on leaky-ReLU networks mainly because 1) it is the simplest demo to prove our concept, and 2) the results of (Caponigro, 2011) and (Duan et al., 2022) allow us to finish the 'flow' and 'discretization' steps easily.

We also note that our result may not hold for ReLU networks as the ReLU function is not invertible. ReLU networks can be regarded as the limits of leaky-ReLU networks with parameter α tending to 0. However, in our construction, some weights of the network are $O(1/\alpha)$, which tend to ∞ as $\alpha \to 0$. This suggests that the narrow ReLU and leaky-ReLU networks are different. How to rediscuss these issues under the ReLU network, and show the differences between the two more clearly, maybe a very interesting topic.

L^p-UAP and C-UAP. Leaky-ReLU activation has been studied by (Duan et al., 2022) and (Cai, 2022) to connect neural ODEs, flow maps, and the minimum width of neural networks. However, the previous results are for the UAP under the L^p norm, which simplified the analysis because the diffeomorphisms are L^p approximations of maps (Brenier & Gangbo, 2003). Our 'lift-flow-discretization' approach can deeply connect the minimal width to the topology theory, properly lift the target function to higher dimensions and employ facts from topology theory to further obtain sharp bounds of width for the uniform/C-UAP.

Approximation rate. Determining the number of weights or layers to achieve ε approximation error is related to the approximation rate or the error-bound problems. Estimating the error bound of all three steps in our lift-flow-discretization approach is challenging since the error in the 'flow' step is hard to estimate. This may need to establish new construction tools. We leave it as future work.

Acknowledgments

We are grateful to Mr. Ziyi Lei for the helpful discussion on the proof of Lemma 4.2. We are also grateful to the anonymous reviewers for their useful comments and suggestions. This work was supported by the National Natural Science Foundation of China (Grant No. 12201053) and the National Natural Science Foundation of China (Grant No. 11871105).

References

Barron, A. R. Approximation and estimation bounds for artificial neural networks. *Machine learning*, 14(1):115–133, 1994.

Beise, H.-P. and Da Cruz, S. D. Expressiveness of Neural Networks Having Width Equal or Below the Input Dimension. *arxiv:2011.04923*, 2020.

Beise, H.-P., Da Cruz, S. D., and Schröder, U. On decision regions of narrow deep neural networks. *Neural Networks*, 2021.

Brenier, Y. and Gangbo, W. L^p Approximation of maps by diffeomorphisms. *Calculus of Variations and Partial Differential Equations*, 16(2):147–164, 2003.

Cai, Y. Achieve the minimum width of neural networks for universal approximation. *arXiv preprint arXiv:2209.11395*, 2022.

Caponigro, M. Orientation preserving diffeomorphisms and flows of control-affine systems. *IFAC Proceedings Volumes*, 44(1):8016–8021, 2011.

Chong, K. F. E. A closer look at the approximation capabilities of neural networks. In *International Conference on Learning Representations*, 2020.

Cybenko, G. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Daniely, A. Depth separation for neural networks. In *Conference on Learning Theory*, 2017.

Duan, Y., Li, L., Ji, G., and Cai, Y. Vanilla feedforward neural networks as a discretization of dynamic systems. *arXiv preprint arXiv:2209.10909*, 2022.

Hanin, B. and Sellke, M. Approximating Continuous Functions by ReLU Nets of Minimal Width. *arXiv preprint arXiv:1710.11278*, 2018.

Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.

Hornik, K., Stinchcombe, M., and White, H. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Huang, C.-W., Krueger, D., Lacoste, A., and Courville, A. Neural Autoregressive Flows. In *International Conference on Machine Learning*, 2018.

Johnson, J. Deep, Skinny Neural Networks are not Universal Approximators. In *International Conference on Learning Representations*, 2019.

Kong, Z. and Chaudhuri, K. Universal Approximation of Residual Flows in Maximum Mean Discrepancy. *International Conference on Machine Learning workshop*, 2021.

Le Roux, N. and Bengio, Y. Representational Power of Restricted Boltzmann Machines and Deep Belief Networks. *Neural Computation*, 20(6):1631–1649, 2008.

Leshno, M., Lin, V. Y., Pinkus, A., and Schocken, S. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993.

Lu, Z., Pu, H., Wang, F., Hu, Z., and Wang, L. The Expressive Power of Neural Networks: A View from the Width. In *Neural Information Processing Systems*, 2017.

Montufar, G. F. Universal approximation depth and errors of narrow belief networks with discrete units. *Neural Computation*, 26(7):1386–1407, 2014.

Nguyen, Q., Mukkamala, M. C., and Hein, M. Neural Networks Should Be Wide Enough to Learn Disconnected Decision Regions. In *International Conference on Machine Learning*, 2018.

Park, S., Yun, C., Lee, J., and Shin, J. Minimum Width for Universal Approximation. In *International Conference on Learning Representations*, 2021.

Rousseau, F. and Fablet, R. Residual networks as geodesic flows of diffeomorphisms. *arXiv preprint arXiv:1805.09585*, 2018.

Ruiz-Balet, D. and Zuazua, E. Neural ODE control for classification, approximation and transport. *arXiv: 2104.05278*, 2021.

Sutskever, I. and Hinton, G. E. Deep, Narrow Sigmoid Belief Networks Are Universal Approximators. *Neural Computation*, 20(11):2629–2636, 2008.

Telgarsky, M. Benefits of depth in neural networks. In *Conference on learning theory*, 2016.

Teshima, T., Ishikawa, I., Tojo, K., Oono, K., Ikeda, M., and Sugiyama, M. Coupling-based Invertible Neural Networks Are Universal Diffeomorphism Approximators. In *Neural Information Processing Systems*, 2020a.

Teshima, T., Tojo, K., Ikeda, M., Ishikawa, I., and Oono, K. Universal Approximation Property of Neural Ordinary Differential Equations. *Neural Information Processing Systems 2020 Workshop on Differential Geometry meets Deep Learning*, 2020b.

Zhang, H., Gao, X., Unterman, J., and Arodz, T. Approximation Capabilities of Neural ODEs and Invertible Residual Networks. In *International Conference on Machine Learning*, 2019.

A. Proofs

A.1. Notations

Flow map A flow on a set $\mathbb{X}$ is a group action of the additive group of real numbers on $\mathbb{X}$. More explicitly, a flow is a mapping $\phi : \mathbb{X} \times \mathbb{R} \to \mathbb{X}$ such that for all $x \in \mathbb{X}$ and $s, t \in \mathbb{R}$, $\phi(x, 0) = x, \phi(\phi(x, t), s) = \phi(x, s + t)$. If we fix $t = \tau$, we gain a flow map $\phi^\tau : \mathbb{X} \to \mathbb{X}$.

Diffeomorphism Given two manifolds M and N, a differentiable map $f : M \to N$ is called a diffeomorphism if it is a bijection and its inverse $f^{-1} : N \to M$ is differentiable as well.

A.2. Proof of Lemma 2.4

Proof. If we can find two linear maps and an orientation preserving diffeomorphism with $N = d + 1$ that satisfies the condition of Corollary 3.3. Then according to Lemma 3.1, for any $\varepsilon > 0$, there exists an ODE (7) whose flow map $\phi^\tau(x)$ satisfies $\|\phi^\tau(x) - \Phi(x)\| < \varepsilon$. Lemma 3.2 shows that for such ϕ^τ and any $\varepsilon > 0$, there exists a leaky-ReLU network $\psi(x)$ with width N and depth L such that $\|\phi^\tau(x) - \psi(x)\| < \varepsilon$ for all x in Ω. According to the lift-flow-discretization approach, we only need to construct two linear maps and an orientation preserving diffeomorphism with $N = d + 1$, which satisfies the condition of Corollary 3.3. In fact, for any continuous function $f^* \in C(\mathcal{K}, \mathbb{R}^d)$ on compact domain $\mathcal{K} \subset \mathbb{R}^d$ and $\varepsilon > 0$, there is a locally Lipschitz smooth function p such that $\|f^*(x) - p(x)\| < \varepsilon$ for all $x \in \mathcal{K}$. The function p can be chosen as a polynomial according to the Stone-Weierstass theorem. Then, the maps $\alpha : \mathbb{R}^d \to \mathbb{R}^{d+1}, \beta : \mathbb{R}^{d+1} \to \mathbb{R}^d$ and $\Phi : \mathbb{R}^{d+1} \to \mathbb{R}^{d+1}$ can be chosen as follows:

$$\alpha(x) = \begin{pmatrix} I_d \\ \mathbf{1}^T \end{pmatrix} x, \quad \Phi \begin{pmatrix} x \\ x_{d+1} \end{pmatrix} = \begin{pmatrix} p(x) + \kappa \mathbf{1}\mathbf{1}^T \\ x_{d+1} \end{pmatrix}, \quad \beta \begin{pmatrix} x \\ x_{d+1} \end{pmatrix} = (I, -\kappa \mathbf{1}) \begin{pmatrix} x \\ x_{d+1} \end{pmatrix} = x - \kappa \mathbf{1} x_{d+1}, \tag{10}$$

where $\mathbf{1} \in \mathbb{R}^d$ is a column vector with all elements equal to one, κ is a number larger than the Lipschitz constant of p on $\mathcal{K}$, and x_{d+1} is the coordinate of the auxiliary dimension. It is obvious that maps α and β are linear and $p(x) = \beta \circ \Phi \circ \alpha(x)$. In addition, Φ is an OP diffeomorphism as $\det(\nabla\Phi) > 0$ for all $x \in \mathcal{K}$ and $x_{d+1} \in \mathbb{R}$. Therefore, the Corollary 3.3 implies that there is a leaky-ReLU network $f_L(x)$ with depth L and width $d + 1$ such that $\|f_L(x) - f^*(x)\| \le \varepsilon$ for all x in $\mathcal{K}$.

□

A.3. Proof of Lemma 3.1

Proof. It is a corollary of Theorem 6 in (Caponigro, 2011). Here, we only provide the main ideas. Let $\mathcal{M} \subset \mathbb{R}^N$ be a compact connected manifold and $\{f_1, \cdots, f_n\}$ be a bracket-generating family of vector fields. (Caponigro, 2011) shows that for any OP diffeomorphism $P \in \text{Diff}_0(\mathcal{M})$ and $\varepsilon > 0$, there exist n time-varying feedback controls, $u_j(x, t)$, which are piecewise constant with respect to t, such that P can be represented by the flow map $\tilde{\phi}^\tau$ of the ODE $\dot{x}(t) = \sum_{j=1}^{n} u_j(x, t) f_j(x), t \in (0, \tau)$, which means $\|\tilde{\phi}^\tau(x) - \Phi(x)\| < \varepsilon$.

Then, we find a neural ODE to approximate $\tilde{\phi}^\tau$. For the approximation of $\tilde{\phi}^\tau$, we need to approximate each $u_j(x, t)$, which can be done by polynomial, trigonometric polynomials or neural networks. The neural ODE (7) is such an example that takes $\mathcal{M} \supset \Omega$, $n = N$, $f_j = e_j$ as the axis vectors, and according to the UAP of neural networks, we have $u_j(x, t) \approx \sum_{i=1}^{M} a_i^{(j)}(t) \tanh(w_i(t) \cdot x + b_i(t))$ where $a_i^{(j)}$ is the j-th coordinate of a_i. In this case, ϕ^τ, the flow map of (7) satisfies $\|\tilde{\phi}^\tau(x) - \phi^\tau(x)\| < \varepsilon, x \in \mathbb{R}^N$. □

A.4. Proof of Lemma 3.2

Proof. The main idea is to solve the ODE (7) by a splitting method and then approximate each split step by leaky-ReLU networks. Consider the following splitting for v in (7): $v(x, t) \equiv \sum_{i=1}^{N} \sum_{j=1}^{d} v_{ij}(x, t) e_j$ with $v_{ij}(x, t) = a_i^{(j)}(t) \tanh(w_i(t) \cdot x + b_i(t)) \in \mathbb{R}$. Then, the flow map can be approximated by an iteration with time step $\Delta t = \tau/n$,

$n \in \mathbb{Z}^+$,

$$\phi^\tau(x_0) \approx x_n = T_n(x_{n-1}) = T_n \circ \cdots \circ T_1(x_0),$$
$$= T_n^{(N,d)} \circ \cdots \circ T_n^{(1,2)} \circ T_n^{(1,1)} \circ \cdots \circ$$
$$T_1^{(N,d)} \circ \cdots \circ T_1^{(1,2)} \circ T_1^{(1,1)}(x_0).$$

where the k-th iteration is $x_{k+1} = T_k x_k = T_k^{(N,d)} \circ \cdots \circ T_k^{(1,2)} \circ T_k^{(1,1)}(x_k)$. The map $T_k^{(i,j)} : x \to y$ in each split step is:

$$T_k^{(i,j)} : \begin{cases} y^{(l)} = x^{(l)}, l = 1, 2, .., j-1, j+1, ..., d, \\ y^{(j)} = x^{(j)} + \Delta t v_{ij}(x, k\Delta t). \end{cases}$$

Here, the superscript in $x^{(l)}$ indicates the l-th coordinate of x. (Duan et al., 2022) constructed leaky-ReLU networks with width N to approximate each map $T_k^{(i,j)}$ and we finished the proof.

□

A.5. Proof of Lemma 4.1

Proof. Without loss of generality, we assume $\mathcal{K} = [0, 1]^{d_x}$. Here, we give a simple example in Figure 4. '4'-shape curve (a) corresponding to a continuous function $g(t)$ from $[0, 1] \subset \mathbb{R}$ to $\mathbb{R}^2$ cannot be uniformly approximated.

When $g(0) = (0, -1), g(1) = (1, 0)$, for some $t_1, t_2 \in [0, 1]$, $g(t_1) = g(t_2) = (0, 0)$, then $(0, 0)$ is a self-intersecting point.

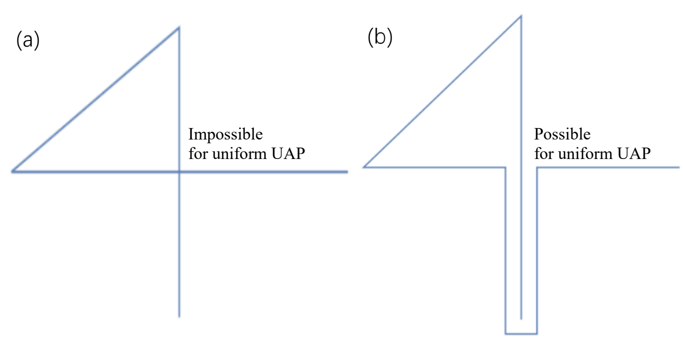


Figure 4. Illustration of the possibility of C-UAP when $d_x \leq d_y$. The curve in (b) is homeomorphic to the interval $[0, 1]$, while curve in (a) is not, and cannot be uniformly approximated by homeomorphisms. For comparison, the C-UAP is possible for (b).

In Figure 4, (a) is not homeomorphic because it has a self-intersecting point, if it can be approximated by some homeomorphism function, that function must avoid the self-intersecting points.

However, for some incredibly small $\varepsilon > 0$, if a function $h(t)$ satisfies $\|g(t) - h(t)\| < \varepsilon, \forall t \in [0, 1]$, $h(t)$ cannot avoid self-intersecting points, so it can not be corresponded to homeomorphism through simple dimension raising. Then Figure 4(a) is not possible to be approximated by a homeomorphism map for some incredibly small error, hence, any function $h(t)$ with $\|h(t) - g(t)\| < \varepsilon, \forall t \in [0, 1]$ has a self-intersecting point that prevents it from being a monomorphism.

We will show that for any $\varepsilon_0 < 1$, the function $h_0(t)$ (orange line in Figure 5(a)) satisfying $\|h_0(t) - g(t)\| < \varepsilon_0$ must have an intersecting point. See Figure 5(b), we construct a closed curve S^1 passing through $(0, 1)$ and $(0, -1)$ with $(-1, 0)$ inside it and $(1, 0)$ outside it, then according to the Jordan curve theorem, the curve passing through points $(1, 0)$ and $(-1, 0)$ must intersect with S^1 within $[0, 1] \times [0, 1]$. This means that any closed curve passing through $(0, 1)$ and$(0, -1)$, and the closed curve passing through $(-1, 0)$ and $(1, 0)$, must have intersection points in $[0, 1] \times [0, 1]$. Therefore, the approximation curve in Figure 5(a) cannot be homeomorphic. Thus, we finish the proof.

For function $\tilde{g}(t)$ (figure 4(b)), for any $\varepsilon > 0$, there exists $\tilde{h}(t)$ satisfying $\|\tilde{g}(t) - \tilde{h}(t)\| < \varepsilon$ which do not have intersection points, the proof is left to interested readers. In conclusion, Figure 4(a) is a counterexample we provide to finish the proof. □

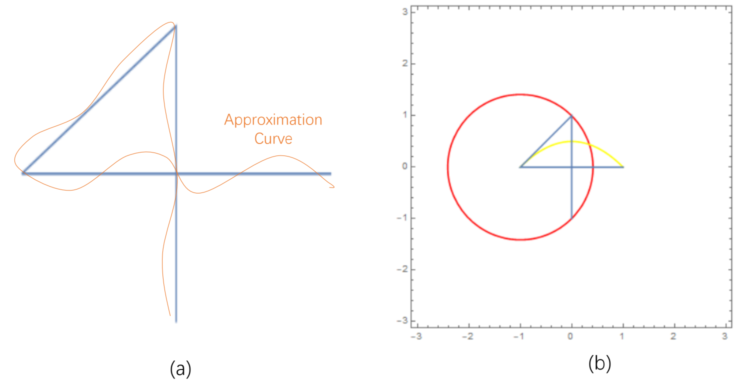


Figure 5. There is an intersection point in $[0, 1] \times [0, 1]$ between two closed curves that pass through $(0, 1), (0, -1)$ and $(1, 0), (-1, 0)$.

A.6. Proof of Lemma 4.2

Proof. Let $f^* : I^d := [0, 1]^d \to \mathbb{R}^{d+2}$ be a continuous map. Then, let f^* be a locally smooth homeomorphism and transversal intersection at all the self-intersection points. If f^* satisfy neither of the two conditions, it needs to be replaced by a smooth approximation or be slightly perturbed. Denote the set of self-intersection points of f as $\mathcal{A}$. According to the transversal intersection, $\partial\mathcal{A} = \mathcal{A}$, $\mathcal{A}$ is a closed set. Since f^* is continuous and I^d is a compact set, we know that $f^*(I^d)$ is also a compact set, so $\mathcal{A}$ as a closed subset of $f(I^d)$ is also compact. For any $x \in \mathcal{A}$, it is locally obtained by transversely intersecting at least two, at most $d + 1$ d-dimensional hyperplanes and set as k intersection points. That is, there is an open neighborhood U of x, $U \xrightarrow[\cong]{\varphi} \mathring{I}^{d+2}$ and $\varphi(f^*(I^d) \cap U) = \bigcup_{i=1}^{k} S_i^d$, where S_i^d denotes the restriction of the i-th d-dimensional hyperplane on $\mathbb{R}^{d+2}$ on $\mathring{I}^{d+2}$.

Let the corresponding equation be $\tilde{x}_i = 0, \tilde{x}_{d+2} = 0$, where $(\tilde{x}_1, \dots, \tilde{x}_{d+2})$ is the coordinate of $\mathbb{R}^{d+2}$. Denote the tubular neighborhood of $T := S_i^d \cap (\bigcup_{k \neq i} S_i^d)$ on S_i^d with a sufficiently small radius as N_i. Consider the mapping $f_i : \mathring{I}^d \to S_i^d, (x_1, \dots, x_d) \mapsto (x_1, \dots, x_{i-1}, 0, x_i, \dots, x_d, 0)$, and perturb it slightly:

$$\bar{f}_i = \begin{cases} \bar{f}_i|_{I^d \setminus \bar{f}_i^{-1}(N_i)} : f_i|_{I^d \setminus f_i{}^{-1}(N_i)}, \\ \bar{f}_i|_{\bar{f}_i^{-1}(N_i)} : (x_1, \dots, x_d) \mapsto \\ \qquad (x_1, \dots, x_{i-1}, 0, x_i, \dots, x_d, \varepsilon_i \tau(x_1, \dots, x_d)), \end{cases} \tag{11}$$

where $\tau(x_1, \dots, x_d)$ is a smooth function that is 1 when $x_j = 0, j \in \{1, \dots, d\}$. $\tau(x_1, \dots, x_d)$ and all its derivatives tends to 0 at the boundary of $\mathring{I}_d$.

For $\bar{f}_i$, make a smooth approximation $\tilde{f}_i$ that maintains zero on each component, and denote $F_i^d = \bar{f}_i(\mathring{I}^d)$. Therefore, $\varphi^{-1}(F_i^d)$ and $f^*(I_i^d)$ are smoothly connected on ∂U and for $\forall i \neq j \in \{1, \dots, k\}, F_i^d \cap F_j^d = \varnothing$. Therefore, all $\varphi^{-1}(S_i^d)$ in U can be replaced by $\varphi^{-1}(F_i^d)$, which will make f^* have no self-intersection point in U, and no new self-intersection point will be generated.

For all of the above x, U constitutes an open cover of $\mathcal{A}$, which has finite subcovers $\{U_1, \dots, U_m\}$. Let $V_1 = U_1, V_2 = U_2 \setminus U_1, \dots, V_m = U_m \setminus (U_1 \cup \dots \cup U_{m-1})$; then, all the V_i are disjoint and cover $\mathcal{A}$.

Execute the above substitution for V_i. Suppose that a similar operation can be performed for $V_1, \dots, V_l$, so that the image of f^* in $V_1 \cup \dots \cup V_l$ after substitution has no self-intersection point. We can prove that a similar operation can also be performed in V_{l+1}, so that f^* in $V_1 \cup \dots \cup V_{l+1}$ obtains the same conclusion after substitution. In fact, this only needs to change N_i at the definition of F_i in the above process to $N_i \setminus \varphi(V_1 \cup \dots \cup V_l)$ and then make a smooth approximation.

Thus far, we have obtained a smooth approximation f of f^*, which is the embedding of $I^d \to \mathbb{R}^{d+2}$. □