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

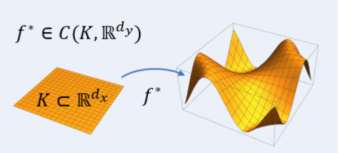
Li'ang Li

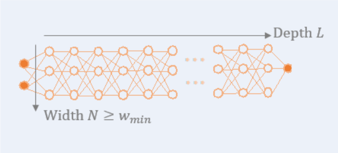
Beijing Normal University

2023.07.23 - 29 ICML

Outline

Universal approximation property (UAP)





Theorem (Universal Approximation)

Let $\sigma \in L^\infty(\mathbb{R})$ and σ is not an algebraic polynomial (a.e.). Then finite sums of the form

$$f(x) = \sum_{i=1}^{N} a_i \sigma\left(w_i^{\mathrm{T}} x + b_i\right), \quad N \in \mathbb{N},$$

are dense in $C(\mathcal{K})$, i.e., for any $f^* \in C(\mathcal{K})$ and $\varepsilon > 0$, there is a N and $f(x)$, such that $\|f(x) - f^*(x)\| < \varepsilon \quad$ for all $x \in \mathcal{K}$.

Refs: Cybenko (1989), Hornik (1991), Leshno et al. (1993), ...

The minimum width of FNN for universal approximation

Table: 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	$w_{\min} = d_x + 1$
Johnson(2019)	$\mathcal{C}(\mathcal{K}, \mathbb{R})$	uniformly conti.‡	$w_{\min} \geqslant d_x + 1$
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	$w_{min} = 3$
Cai(2022)	$L^p(\mathcal{K}, \mathbb{R}^{d_y})$	Leaky-ReLU	$w_{\min} = \max(d_x, d_y, 2)$
Cai(2022)	$\mathcal{C}(\mathcal{K}, \mathbb{R}^{d_y})$	Arbitrary	$w_{\min} \geqslant \max(d_x, d_y)$

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

‡ requires that ρ is uniformly approximated by a sequence of one-to-one functions.

We focus on the minimum width of leaky-ReLU networks for the UAP in $\mathcal{C}(\mathcal{K}, \mathbb{R}^{d_y})$.

Outline

1. Introduction
2. Main Results
3. Summary

lift-flow-discretization approach

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 orienation preserving diffeomorphism such that

$$\|f^*(x) - \beta \circ \Phi \circ \alpha(x)\| \le \varepsilon/3, \forall x \in \mathcal{K}, \tag{1}$$

where α and β are two linear maps. Without loss of generality, we can assume the Lipschitz constant 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

$$z'(t) = v(z(t), t), t \in (0, \tau), \quad z(0) = x, \tag{2}$$

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

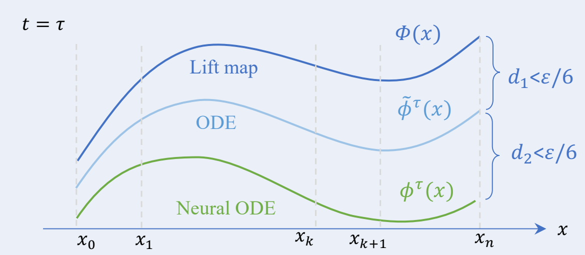


Figure: The target map $\Phi(x)$ is approximated by a flow map $\tilde{\phi}^\tau(x)$ of an ODE, which is further approximated by a flow map $\phi^\tau(x)$ of a neural ODE.

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 $\|\psi(x) - \phi^\tau(x)\| \leq \varepsilon/3$ 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

$$\|f^*(x) - \beta \circ \psi \circ \alpha(x)\| \leq \varepsilon, \forall x \in \mathcal{K}. \tag{3}$$

$\max(d_x+1, d_y) \leqslant w_{min} \leqslant \max(d_x+1, d_y+1)$

References	Functions	Activation	Minimum width
Johnson(2019)	$\mathcal{C}(\mathcal{K}, \mathbb{R})$	uniformly conti.	$w_{\min} \geqslant d_x + 1$
Cai(2022)	$\mathcal{C}(\mathcal{K}, \mathbb{R}^{d_y})$	Arbitrary	$w_{\min} \geqslant \max(d_x, d_y)$

Lemma 1

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 $\mathcal{K}$.

$d_y = d_x + 1, w_{min} \geqslant d_y + 1$

Lemma 2

If $d_y = d_x + 1$, there exists a continuous function $f^*(x) \in C([0,1]^{d_x}, \mathbb{R}^{d_y})$ which can NOT be uniformly approximated by functions like $\psi(W_1 x + b_1)$ with homeomorphism maps $\psi : \mathbb{R}^{d_y} \to \mathbb{R}^{d_y}$.

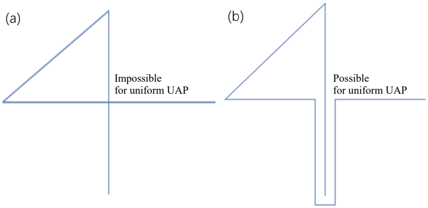


Figure: Illustration of the possibility of C-UAP when $d_x \leq d_y$. The curve in (b) is homeomorphic to the interval $[0, 1]$, while curve in (a) is not, and cannot be uniformly approximated by homeomorphisms. For comparison, the C-UAP is possible for (b).

$d_y \geqslant d_x + 2, w_{min} \leqslant d_y$

Lemma 3

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 $\mathcal{K}$.

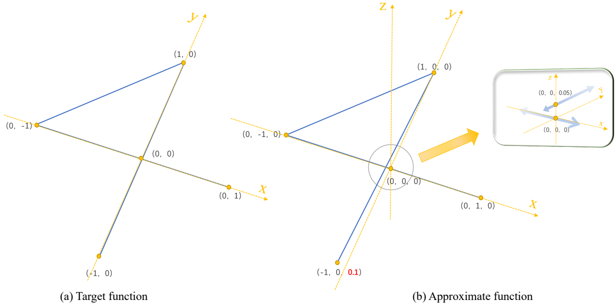


Figure: Example of $d_x = 1$. Approximate the '4'-shape curve (a) in $\mathbb{R}^2$ by lifting it to the three-dimensional curve (b) in $\mathbb{R}^3$.

The main result

Theorem 4

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 $N \geq w_{\min}$.

Part 1: $w_{\min} \geq \max(d_x + 1, d_y)$

Part 2: $w_{\min} \leq \max(d_x + 1, d_y + 1)$

Part 3: $w_{\min} \geq d_y + 1$, if $d_y = d_x + 1$

Part 4: $w_{\min} \leq d_y$, if $d_y \geq d_x + 2$

$$w_{\min} = \max(d_x + 1, d_y) + 1_{d_y = d_x + 1}$$

Figure: Proof parts of the main result.

Outline

Summary

References	Functions	Activation	Minimum width
Johnson(2019)	$\mathcal{C}(\mathcal{K}, \mathbb{R})$	uniformly conti.	$w_{\min} \geqslant d_x + 1$
Park et al.(2021)	$L^p(\mathbb{R}^{d_x}, \mathbb{R}^{d_y})$	ReLU	$w_{\min} = \max(d_x + 1, d_y)$
Cai(2022)	$L^p(\mathcal{K}, \mathbb{R}^{d_y})$	Leaky-ReLU	$w_{\min} = \max(d_x, d_y, 2)$
Cai(2022)	$\mathcal{C}(\mathcal{K}, \mathbb{R}^{d_y})$	Arbitrary	$w_{\min} \geqslant \max(d_x, d_y)$
Ours (Theorem 4)	$\mathcal{C}(\mathcal{K}, \mathbb{R}^{d_y})$	Leaky-ReLU	$w_{\min} = \max(d_x + 1, d_y) + 1_{d_y = d_x + 1}$
Ours (Lemma 1)	$\mathcal{C}(\mathcal{K}, \mathbb{R}^{d_x})$	Leaky-ReLU	$w_{\min} = d_x + 1$

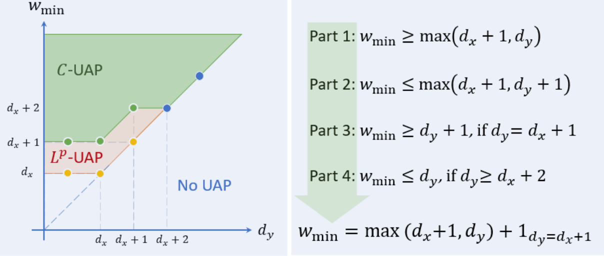


Figure: The minimum width of leaky-ReLU networks to reach UAP.

Contributions

- Theorem 4 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.
- 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.

Li'ang Li et al. Minimum Width of Leaky-ReLU Neural Networks for Uniform Universal Approximation. ICML ,2023.

Thanks for your attention!