Achieve the Minimum Width of Neural Networks for Universal Approximation

Yongqiang Cai

Beijing Normal University

caiyq.math@bnu.edu.cn

2023.05.01-05 ICLR

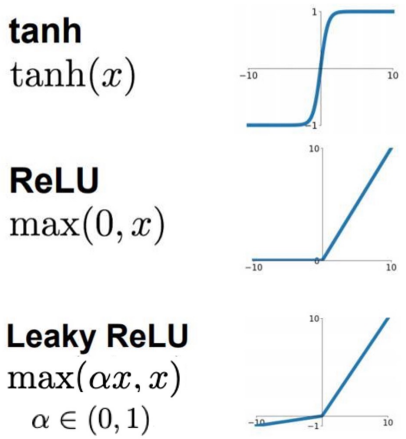
Outline

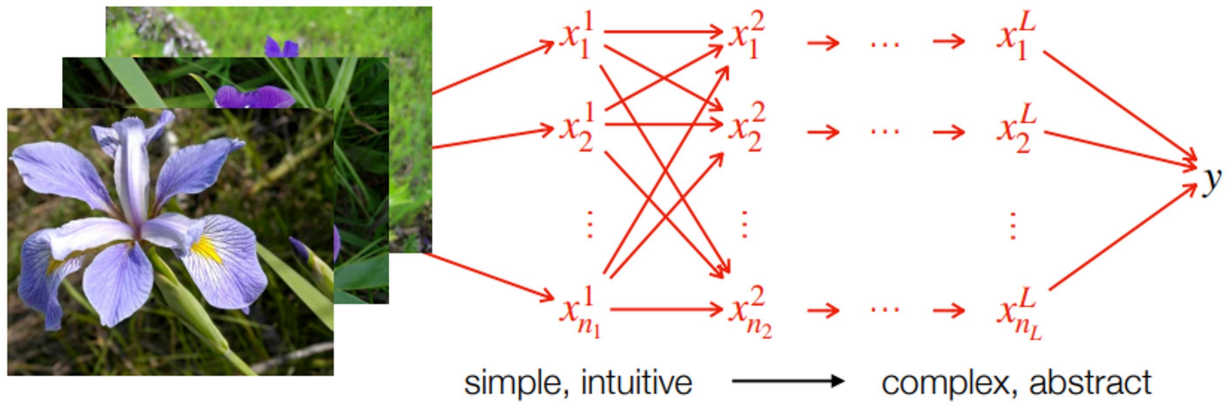
Feedforward neural network

A map from input $x \in \mathbb{R}^{d_x}$ to output $y = f_L(x) \in \mathbb{R}^{d_y}$, (depth L)

$$f_0(x) = W^{[0]}x + b^{[0]},$$
$$f_k(x) = W^{[k]}\sigma\left(f_{k-1}(x)\right) + b^{[k]} \in \mathbb{R}^{n_k}, k = 1, 2, \cdots, L,$$

where $\sigma(\cdot)$ is the activation such as tanh, ReLU, leaky-ReLU, etc.





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 an N and $f(x)$, such that

$$\|f(x) - f^*(x)\| < \varepsilon \quad \text{for all } x \in \mathcal{K}.$$

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

What is the minimum width of FNN that have the UAP?

Previous known minimum width (from Park et al. (2021))

Reference	Function class	Activation ρ	Upper/lower bounds
Lu et al. (2017)	$L^1(\mathbb{R}^{d_x}, \mathbb{R})$	RELU	$d_x + 1 \le w_{\min} \le d_x + 4$
	$L^1(\mathcal{K}, \mathbb{R})$	RELU	$w_{\min} \ge d_x$
Hanin and Sellke (2017)	$C(\mathcal{K}, \mathbb{R}^{d_y})$	RELU	$d_x + 1 \le w_{\min} \le d_x + d_y$
Johnson (2019)	$C(\mathcal{K}, \mathbb{R})$	uniformly conti.†	$w_{\min} \ge d_x + 1$
Kidger and Lyons (2020)	$C(\mathcal{K}, \mathbb{R}^{d_y})$	conti. nonpoly‡	$w_{\min} \le d_x + d_y + 1$
	$C(\mathcal{K}, \mathbb{R}^{d_y})$	nonaffine poly	$w_{\min} \le d_x + d_y + 2$
	$L^p(\mathbb{R}^{d_x}, \mathbb{R}^{d_y})$	RELU	$w_{\min} \le d_x + d_y + 1$
Park et al. (2021)	$L^p(\mathbb{R}^{d_x}, \mathbb{R}^{d_y})$	RELU	$w_{\min} = \max\{d_x + 1, d_y\}$
	$C([0,1], \mathbb{R}^2)$	RELU	$w_{\min} = 3 > \max\{d_x + 1, d_y\}$
	$C(\mathcal{K}, \mathbb{R}^{d_y})$	RELU+STEP	$w_{\min} = \max\{d_x + 1, d_y\}$
	$L^p(\mathcal{K}, \mathbb{R}^{d_y})$	conti. nonpoly‡	$w_{\min} \le \max\{d_x + 2, d_y + 1\}$

† requires that ρ is uniformly approximated by a sequence of one-to-one functions.
‡ requires that ρ is continuously differentiable at some z with $\rho'(z) \neq 0$.

Q: Are the bounds optimal?

Outline

Universal lower bound $w^*_{\min}$ for C-UAP and L^p-UAP

Lemma 1 (Universal lower bound is $w^*_{\min} \equiv \max(d_x, d_y)$)

For any compact domain $\mathcal{K} \subset \mathbb{R}^{d_x}$ and any finite set of activation functions $\{\sigma_i\}$, the $\{\sigma_i\}$ networks with width $w < w^*_{\min} \equiv \max(d_x, d_y)$ do not have the UAP for both $L^p(\mathcal{K}, \mathbb{R}^{d_y})$ and $C(\mathcal{K}, \mathbb{R}^{d_y})$, i.e., L^p-UAP and C-UAP, respectively.

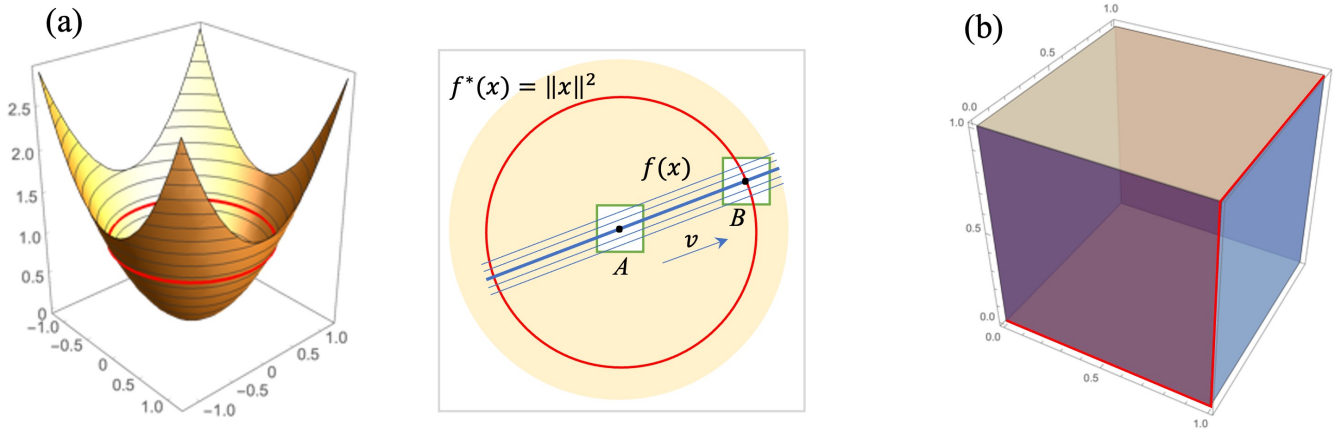


Figure: UAP failed when (a) $w \leq d_x - 1$ or (b) $w \leq d_y - 1$. (a) Points A and B on a level set of networks $f(x)$; $f(A) = f(B)$ but $f^*(A) - f^*(B)$ is not small. (b) The curve from $\mathbf{0}$ to $\mathbf{1}$ along the edge of the cubic has a positive distance to any hyperplane.

Achieve $w^*_{\min}$ for L^p-UAP with continuous activations

Theorem 2 (L^p-UAP of leaky-ReLU NN)

For the function class $L^p(\mathcal{K}, \mathbb{R}^{d_y})$, the minimum width of leaky-ReLU networks having UAP is exactly $w_{\min} = \max(d_x, d_y, 2)$.

Remark: leaky-ReLU+ABS networks achieve the critical width.

$f^* \approx$ OP diffeomorphism $\approx$ flow map of NODE $\approx$ leaky-ReLU NN.

Fact 1: Leaky-ReLU networks can approximate the flow map of neural ODEs. Duan et al. (2022).

Fact 2: Approximation power of neural ODEs. Li et al. (2022), Tabuada & Gharesifard (2020),Ruiz-Balet & Zuazua (2021).

Fact 3: Orientation preserving diffeomorphisms can approximate continuous functions if $d \geq 2$. Brenier & Gangbo (2003).

Achieve $w^*_{\min}$ for C-UAP with discontinuous activations

Lemma 4 (C-UAP of ReLU+Floor NN)

For the function class $C(\mathcal{K}, \mathbb{R}^{d_y})$, the minimum width of ReLU + Floor networks having UAP is exactly $w_{\min} = \max(d_x, d_y, 2)$.

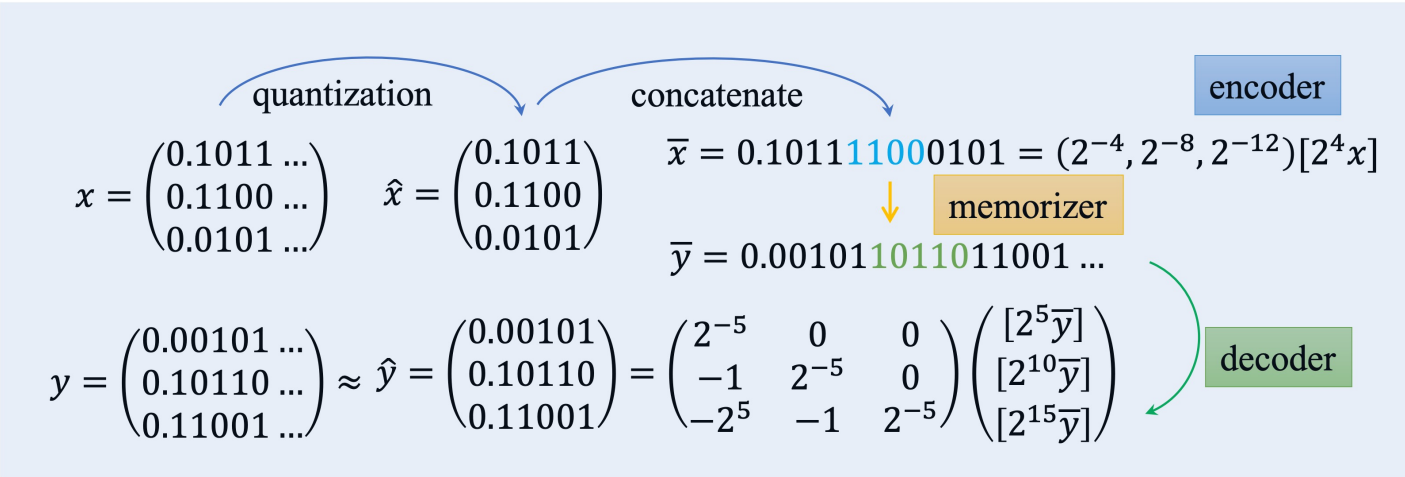


Figure: The encoder-memorizer-decoder scheme for C-UAP by an example where $d_x = d_y = 3$, 4 bits for the input and 5 bits for the output.

Corner case of dimension one, $d = 1$

Theorem 5 (UOE networks)

The UOE networks with width d_y have C-UAP for functions in $C([0,1], \mathbb{R}^{d_y})$.

Universal ordering of extrema (UOE) functions: any possible ordering(s) of values at the extrema can be found in the extrema of the function.

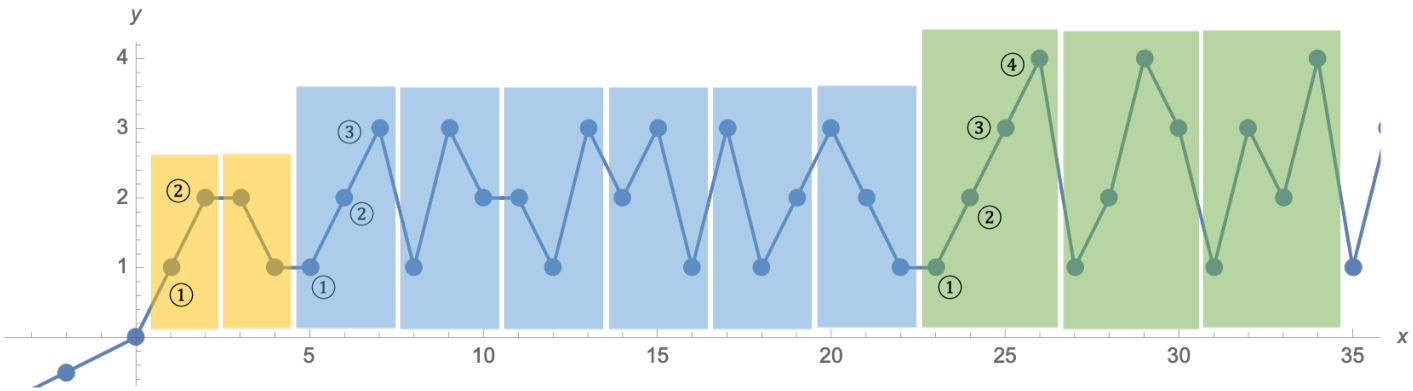


Figure: An example of the UOE function $\rho(x)$, which has an infinite number of pieces.

Corner case of dimension one, $d = 1$

Theorem 5 (UOE networks)

The UOE networks with width d_y have C-UAP for functions in $C([0,1], \mathbb{R}^{d_y})$.

Key: $f^*(x) \approx g \circ u(x) = v \circ \rho \circ u(x)$ with monotonic $v(x)$ and $u(x)$.

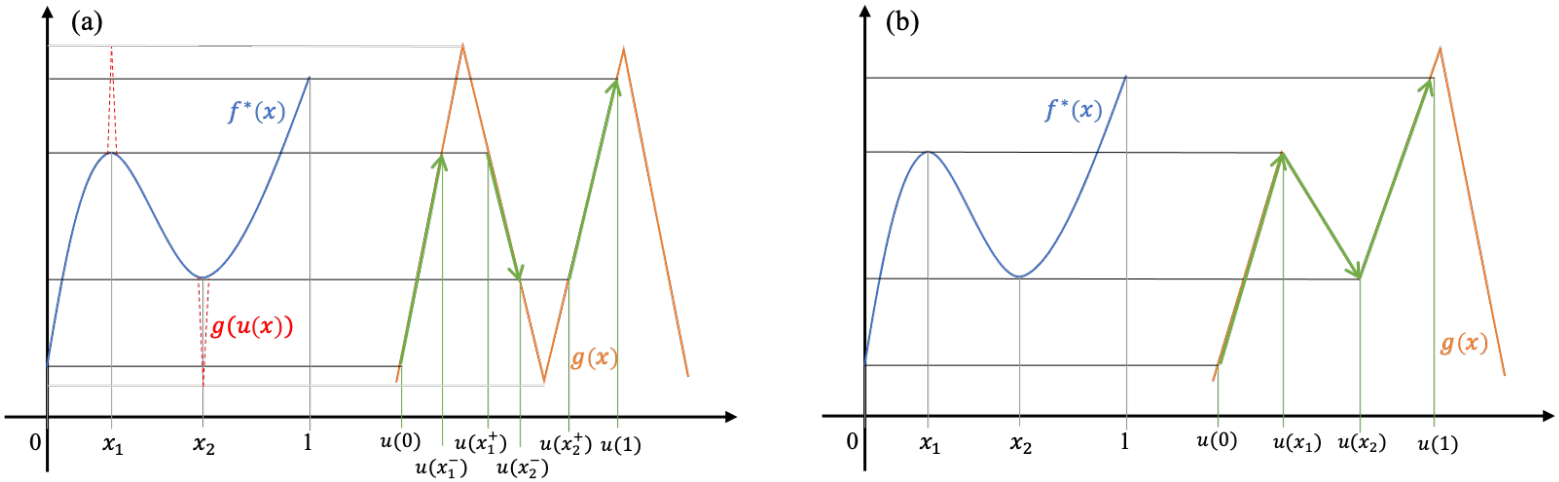


Figure: Approximate $f^*(x)$ by the composition of a monotonically increasing function $u(x)$ and a nonmonotone function $g(x)$. (a) only matching the ordering of extrema values, (b) matching the values as well.

Outline

Summary

Functions	Activation	Minimum width	References
$C(\mathcal{K}, \mathbb{R})$	ReLU	$w_{\min} = d_x + 1$	Hanin & Sellke (2017)
$L^p(\mathbb{R}^{d_x}, \mathbb{R}^{d_y})$	ReLU	$w_{\min} = \max(d_x + 1, d_y)$	Park et al. (2021)
$C([0,1], \mathbb{R}^2)$	ReLU	$w_{\min} = 3 = \max(d_x, d_y) + 1$	Park et al. (2021)
$C(\mathcal{K}, \mathbb{R}^{d_y})$	ReLU+STEP	$w_{\min} = \max(d_x + 1, d_y)$	Park et al. (2021)
$L^p(\mathcal{K}, \mathbb{R}^{d_y})$	Conti. nonpoly‡	$w_{\min} \leq \max(d_x + 2, d_y + 1)$	Park et al. (2021)
$L^p(\mathcal{K}, \mathbb{R}^{d_y})$	Arbitrary	$w_{\min} \geq \max(d_x, d_y) =: w^*_{\min}$	**Ours** (Lemma 1)
	Leaky-ReLU	$w_{\min} = \max(d_x, d_y, 2)$	**Ours** (Theorem 2)
	Leaky-ReLU+ABS	$w_{\min} = \max(d_x, d_y)$	**Ours** (Theorem 3)
$C(\mathcal{K}, \mathbb{R}^{d_y})$	Arbitrary	$w_{\min} \geq \max(d_x, d_y) =: w^*_{\min}$	**Ours** (Lemma 1)
	ReLU+FLOOR	$w_{\min} = \max(d_x, d_y, 2)$	**Ours** (Lemma 4)
	UOE†+FLOOR	$w_{\min} = \max(d_x, d_y)$	**Ours** (Corollary 6)
$C([0,1], \mathbb{R}^{d_y})$	UOE†	$w_{\min} = d_y$	**Ours** (Theorem 5)

‡ Continuous nonpolynomial ρ that is continuously differentiable at some z with $\rho'(z) \neq 0$.

† UOE means the function having *universal ordering of extrema*, see Definition 7.

[1] B. Hanin and M. Sellke. Approximating continuous functions by relu nets of minimal width. *arXiv:1710.11278*.
[2] Sejun Park, et al. Minimum Width for Universal Approximation. In *ICLR*, 2021.
[3] Cai, Y. Achieve the Minimum Width of Neural Networks for Universal Approximation. ICLR, 2023.

Contributions

- Obtained the universal lower bound of width w^*_{min} for feedforward neural networks (FNNs) that have universal approximation properties.
- Achieved the critical width w^*_{min} by leaky-ReLU+ABS networks and UOE+FLOOR networks.
- Proposed a novel construction scheme from a differential geometry perspective that could deepen our understanding of UAP through topology theory.

Cai, Yongqiang. Achieve the Minimum Width of Neural Networks for Universal Approximation. ICLR, 2023.

Thanks for your attention!