# Nearly Optimal VC-Dimension and Pseudo-Dimension Bounds for Deep Neural Network Derivatives

Yahong YANG

(joint work with Prof. Haizhao Yang (UMD) and Prof. Yang Xiang (HKUST))

Department of Mathematics
The Pennsylvania State University, University Park

NeurIPS 2023 conference

# Overview

1. Optimal bounds of VC-dimension and Pseudo-dimension of DNN derivatives

2. Error analysis in the Sobolev training

3. Application of the Bound of VC-Dimension and Pseudo-Dimension Bounds of DNN Derivatives

4. Recent Works

5. Proofs

# Definition of VC-dimension[1] of function sets

## Definition (VC-dimension)

*Let $H$ denote a class of functions from $\mathcal{X}$ to $\{0,1\}$. For any non-negative integer $m$, define the growth function of $H$ as*

$$\Pi_H(m) := \max_{x_1,x_2,\ldots,x_m \in \mathcal{X}} \left|\{(h(x_1), h(x_2), \ldots, h(x_m)) : h \in H\}\right| .$$

*The Vapnik–Chervonenkis dimension (VC-dimension) of $H$, denoted by $VCdim(H)$, is the largest $m$ such that $\Pi_H(m) = 2^m$. For a class $\mathcal{G}$ of real-valued functions, define $VCdim(\mathcal{G}) := VCdim(\mathrm{sgn}(\mathcal{G}))$, where $\mathrm{sgn}(\mathcal{G}) := \{\mathrm{sgn}(f) : f \in \mathcal{G}\}$ and $\mathrm{sgn}(x) = 1[x > 0]$.*

[1] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. Journal of the ACM (JACM), 36(4):929–965, 1989.

# Example of VC-dimension

The VC-dimension of linear function in $\mathbb{R}^2$ is 3 based on Radon's theorem.

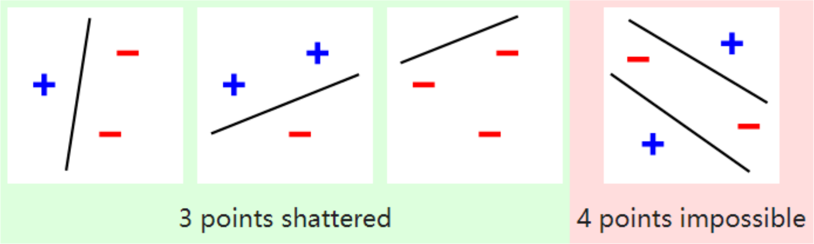

Figure: The VC-dimension of linear functions in two-dimension spaces is 3.

# Definition of pseudo-dimension[2] of function sets

### Definition (pseudo-dimension)

*Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathbb{R}$. The pseudo-dimension of $\mathcal{F}$, denoted by Pdim$(\mathcal{F})$, is the largest integer $m$ for which there exists $(x_1, x_2, \ldots, x_m, y_1, y_2, \ldots, y_m) \in \mathcal{X}^m \times \mathbb{R}^m$ such that for any $(b_1, \ldots, b_m) \in \{0,1\}^m$ there is $f \in \mathcal{F}$ such that $\forall i : f(x_i) > y_i \iff b_i = 1$.*

[2] D. Pollard. Empirical processes: theory and applications. Ims, 1990

# Applications about main results

In the work [3], they prove that

$$\mathrm{VCdim}(\Phi) \sim O(N^2 L^2 \log_2 L \log_2 N).$$

This result has found wide applications in the error analysis of DNN approximations.

[3] P. Bartlett, N. Harvey, C. Liaw, and A. Mehrabian. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. The Journal of Machine Learning Research, 20(1): 2285–2301, 2019

# Error analysis

Learn $f(\boldsymbol{x})$ defined on $(0,1)^d$ with $\|f\|_{\mathcal{W}^{n,\infty}((0,1)^d)} \leq 1$ from a finite set of data samples $\{(\boldsymbol{x}_i, f(\boldsymbol{x}_i))\}_{i=1}^{M}$. Denote

$$\boldsymbol{\theta}_D^{\text{class}} := \arg\inf_{\boldsymbol{\theta}} \mathcal{R}_D(\boldsymbol{\theta}) := \arg\inf_{\boldsymbol{\theta}} \int_{(0,1)^d} |f(\boldsymbol{x}) - \phi(\boldsymbol{x};\boldsymbol{\theta})|^2 \, \mathrm{d}\boldsymbol{x}, \tag{1}$$

$$\boldsymbol{\theta}_S^{\text{class}} := \arg\inf_{\boldsymbol{\theta}} \mathcal{R}_S(\boldsymbol{\theta}) := \arg\inf_{\boldsymbol{\theta}} \frac{1}{M} \sum_{i=1}^{M} |f(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_i;\boldsymbol{\theta})|^2. \tag{2}$$

# Error analysis

The overall inference error is $\mathbf{E}\mathcal{R}_D(\boldsymbol{\theta}_S^{\text{class}})$, which can be divided into two parts:

$$\begin{aligned}\mathbf{E}\mathcal{R}_D(\boldsymbol{\theta}_S^{\text{class}}) =& \mathcal{R}_D(\boldsymbol{\theta}_D^{\text{class}}) + \mathbf{E}\mathcal{R}_S(\boldsymbol{\theta}_D^{\text{class}}) - \mathcal{R}_D(\boldsymbol{\theta}_D^{\text{class}}) \\ &+ \mathbf{E}\mathcal{R}_S(\boldsymbol{\theta}_S^{\text{class}}) - \mathbf{E}\mathcal{R}_S(\boldsymbol{\theta}_D^{\text{class}}) + \mathbf{E}\mathcal{R}_D(\boldsymbol{\theta}_S^{\text{class}}) - \mathbf{E}\mathcal{R}_S(\boldsymbol{\theta}_S^{\text{class}}) \\ \leq & \underbrace{\mathcal{R}_D(\boldsymbol{\theta}_D^{\text{class}})}_{\text{approximation error}} + \underbrace{\mathbf{E}\mathcal{R}_S(\boldsymbol{\theta}_D^{\text{class}}) - \mathcal{R}_D(\boldsymbol{\theta}_D^{\text{class}}) + \mathbf{E}\mathcal{R}_D(\boldsymbol{\theta}_S^{\text{class}}) - \mathbf{E}\mathcal{R}_S(\boldsymbol{\theta}_S^{\text{class}}),}_{\text{generalization error}} \quad (3)\end{aligned}$$

where the last inequality is due to $\mathbf{E}\mathcal{R}_S(\boldsymbol{\theta}_S^{\text{class}}) \leq \mathbf{E}\mathcal{R}_S(\boldsymbol{\theta}_D^{\text{class}})$ by the definition of $\boldsymbol{\theta}_S^{\text{class}}$. For complex function sets used in approximation, the approximation error may be small, but the generalization error can become large.

# Lower bound of the approximation rate given by VC-dimension[4]

## Proposition

*Given any* $s, d \in \mathbb{N}^+$, *there exists a (small) positive constant* $C_{s,d}$ *determined by* $s$ *and* $d$ *such that the following holds: For any* $\varepsilon > 0$ *and a function set* $\mathscr{F}$ *with all elements defined on* $[0,1]^d$, *if* $\mathrm{VCDim}(\mathscr{F}) \geq 1$ *and*

$$\inf_{\phi \in \mathscr{F}} \|\phi - f\|_{L^\infty([0,1]^d)} \leq \varepsilon \quad \textit{for any } f \in C_u^s\left([0,1]^d\right)$$

*then* $\mathrm{VCDim}(\mathscr{F}) \geq C_{s,d}\varepsilon^{-d/s}$.

[4] J. Lu, Z. Shen, H. Yang, and S. Zhang. Deep network approximation for smooth functions. SIAM Journal on Mathematical Analysis, 53(5):5465–5506, 2021.

# Upper bound of the generalization error given by pseudo-dimension

According to Rademacher complexity and uniform covering number, we can derive that

$$\text{Generalization error} \leq C \left( \frac{\mathrm{Pdim}(\Phi)}{M} \right)^{\frac{1}{2}} \sqrt{\log \left( \frac{2eM}{\mathrm{Pdim}(\Phi)} \right)},$$

where $\Phi := \left\{ \phi : \phi \text{ is a } \sigma_1\text{-NN in } \mathbb{R}^d \text{ with width} \leq N \text{ and depth} \leq L \right\}$.

# Sobolev training

While the VC-dimension and pseudo-dimension have been extensively studied, they are not sufficient for analyzing the errors of DNNs in Sobolev training, as the loss functions used in such training contain derivatives of the DNNs. For example, For example, to solve the following PDEs

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ \frac{\partial u}{\partial \nu} = 0 & \text{on } \partial\Omega, \end{cases} \tag{4}$$

using the Deep Ritz method, the loss function can be written as

$$\mathcal{E}_D(\boldsymbol{\theta}) := \frac{1}{2}\int_\Omega |\nabla\phi(\boldsymbol{x};\boldsymbol{\theta})|^2 \mathrm{d}\boldsymbol{x} + \frac{1}{2}\left(\int_\Omega \phi(\boldsymbol{x};\boldsymbol{\theta})\mathrm{d}\boldsymbol{x}\right)^2 - \int_\Omega f\phi(\boldsymbol{x};\boldsymbol{\theta})\mathrm{d}\boldsymbol{x},$$

where $\boldsymbol{\theta}$ represents all the parameters in the neural network. Denote $\Omega$ as $(0,1)^d$. Proposition 1 in proves that the loss function $\mathcal{E}_D(\boldsymbol{\theta})$ is equivalent to $\|\phi(\boldsymbol{x};\boldsymbol{\theta}) - u^*(\boldsymbol{x})\|_{H^1((0,1)^d)}$, where $u^*(\boldsymbol{x})$ is the exact solution of the PDEs in equation (4). Thus, the Sobolev norm $H^1((0,1)^d)$ can measure the loss function.

# Generalization error of loss functions defined by Sobolev norms

Denote

$$\boldsymbol{\theta}_D := \arg\inf_{\boldsymbol{\theta}} \mathcal{R}_D(\boldsymbol{\theta}) := \arg\inf_{\boldsymbol{\theta}} \int_{(0,1)^d} |\nabla(f(\boldsymbol{x}) - \phi(\boldsymbol{x};\boldsymbol{\theta}))|^2 + |f(\boldsymbol{x}) - \phi(\boldsymbol{x};\boldsymbol{\theta})|^2 \,\mathrm{d}\boldsymbol{x}, \tag{5}$$

$$\boldsymbol{\theta}_S := \arg\inf_{\boldsymbol{\theta}} \mathcal{R}_S(\boldsymbol{\theta}) := \arg\inf_{\boldsymbol{\theta}} \frac{1}{M} \sum_{i=1}^{M} \left[|\nabla(f(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_i;\boldsymbol{\theta}))|^2 + |f(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_i;\boldsymbol{\theta})|^2\right]. \tag{6}$$

# Generalization error of loss functions defined by Sobolev norms

The overall inference error is $\mathbf{E}\mathcal{R}_D(\boldsymbol{\theta}_S)$, which can be divided into two parts:

$$\begin{aligned}\mathbf{E}\mathcal{R}_D(\boldsymbol{\theta}_S) =& \mathcal{R}_D(\boldsymbol{\theta}_D) + \mathbf{E}\mathcal{R}_S(\boldsymbol{\theta}_D) - \mathcal{R}_D(\boldsymbol{\theta}_D) + \mathbf{E}\mathcal{R}_S(\boldsymbol{\theta}_S) - \mathbf{E}\mathcal{R}_S(\boldsymbol{\theta}_D) + \mathbf{E}\mathcal{R}_D(\boldsymbol{\theta}_S) - \mathbf{E}\mathcal{R}_S(\boldsymbol{\theta}_S) \\ \leq & \quad \underbrace{\mathcal{R}_D(\boldsymbol{\theta}_D)}_{\text{approximation error}} \quad + \underbrace{\mathbf{E}\mathcal{R}_S(\boldsymbol{\theta}_D) - \mathcal{R}_D(\boldsymbol{\theta}_D) + \mathbf{E}\mathcal{R}_D(\boldsymbol{\theta}_S) - \mathbf{E}\mathcal{R}_S(\boldsymbol{\theta}_S),}_{\text{generalization error}} \end{aligned} \tag{7}$$

where the last inequality is due to $\mathbf{E}\mathcal{R}_S(\boldsymbol{\theta}_S) \leq \mathbf{E}\mathcal{R}_S(\boldsymbol{\theta}_D)$ by the definition of $\boldsymbol{\theta}_S$.

# Main results: Optimal bound of VC-dimension

## Theorem

**(i):** *For any $N, L, d \in \mathbb{N}_+$, there exists a constant $\bar{C}$ independent with $N, L$ such that*

$$VCdim(D\Phi) \leq \bar{C} N^2 L^2 \log_2 L \log_2 N, \tag{8}$$

*for $D\Phi := \{\psi = D_i\phi : \phi \in \Phi,\ \ i = 1, 2, \ldots, d\}$, where*
*$\Phi := \left\{\phi : \phi \text{ is a } \sigma_1\text{-NN in } \mathbb{R}^d \text{ with width} \leq N \text{ and depth} \leq L\right\}$, and $D_i$ is the weak derivative in the $i$-th variable.*
**(ii):** *For any $d \in \mathbb{N}_+$, $C, J_0, \varepsilon > 0$, there exists $N, L \in \mathbb{N}$ with $NL \geq J_0$ such that*

$$VCdim(D\Phi) > C N^{2-\varepsilon} L^{2-\varepsilon}. \tag{9}$$

# Main results: Optimal bound of and Pseudo-dimension

**Theorem**

**(i):** *For any* $N, L, d \in \mathbb{N}_+$, *there exists a constant* $\hat{C}$ *independent with* $N, L$ *such that*

$$Pdim(D\Phi) \leq \hat{C} N^2 L^2 \log_2 L \log_2 N. \tag{10}$$

**(ii):** *For any* $d \in \mathbb{N}_+$, $C, J_0, \varepsilon > 0$, *there exists* $N, L \in \mathbb{N}$ *with* $NL \geq J_0$ *such that*

$$Pdim(D\Phi) > CN^{2-\varepsilon} L^{2-\varepsilon}. \tag{11}$$

# Optimallity of DNN approximation in Sobolev Spaces

By utilizing Theorem 1, we prove that our DNN approximation rate for approximating functions in Sobolev spaces $\mathcal{W}^{n,\infty}((0,1)^d)$ using Sobolev norms in $\mathcal{W}^{1,\infty}((0,1)^d)$ is nearly optimal:

## Theorem

*For any $f \in \mathcal{W}^{n,\infty}((0,1)^d)$ with $n \geq 2$ and $\|f\|_{\mathcal{W}^{n,\infty}((0,1)^d)} \leq 1$, any $N, L \in \mathbb{N}_+$, there is a $\sigma_1$-NN $\phi$ with the width $(34+d)2^d n^{d+1}(N+1)\log_2(8N)$ and depth $56d^2n^2(L+1)\log_2(4L)$ such that*

$$\|f(\boldsymbol{x}) - \phi(\boldsymbol{x})\|_{\mathcal{W}^{1,\infty}((0,1)^d)} \leq C_9(n,d)N^{-2(n-1)/d}L^{-2(n-1)/d},$$

*where $C_9$ is the constant independent with $N, L$.*

# Optimallity of DNN approximation in Sobolev Spaces

## Theorem

*Given any $\rho, C_1, C_2, C_3, J_0 > 0$ and $n, d \in \mathbb{N}^+$, there exist $N, L \in \mathbb{N}$ with $NL \geq J_0$ and $f$ with $\|f\|_{\mathcal{W}^{n,\infty}((0,1)^d))} \leq 1$, satisfying for any $\sigma_1$-NN $\phi$ with the width smaller than $C_1 N \log N$ and depth smaller than $C_2 L \log L$, we have*

$$|\phi - f|_{\mathcal{W}^{1,\infty}((0,1)^d)} > C_3 L^{-2(n-1)/d-\rho} N^{-2(n-1)/d-\rho}. \tag{12}$$

In other words, the approximation rate of $O(N^{-2(n-1)/d-\rho} K^{-2(n-1)/d-\rho})$ cannot be achieved asymptotically when ReLU $\sigma_1$-NNs with width $O(N \log N)$ and depth $O(L \log L)$ to approximate functions in $\mathcal{F}_{n,d} := \left\{ f \in \mathcal{W}^{n,\infty}((0,1)^d) : \|f\|_{\mathcal{W}^{n,\infty}((0,1)^d)} \leq 1 \right\}$. The proof of Theorem 4 is based on the estimation of the VC-dimension of DNN derivatives, which is provided in Theorem 1.

# Generalization error of loss functions defined by Sobolev norms

## Theorem

*For any* $N, L, d, B, C_1, C_2$, *if* $\phi(\boldsymbol{x};\boldsymbol{\theta}_D), \phi(\boldsymbol{x};\boldsymbol{\theta}_S) \in \widetilde{\Phi}$, *we will have that there are constants* $C_5 = C_5(B, d, C_1, C_2)$ *and* $J = J(d, N, L, C_1, C_2)$ *such that for any* $M \geq J$, *we have*

$$\mathbf{E}\mathcal{R}_S(\boldsymbol{\theta}_D) - \mathcal{R}_D(\boldsymbol{\theta}_D) + \mathbf{E}\mathcal{R}_D(\boldsymbol{\theta}_S) - \mathbf{E}\mathcal{R}_S(\boldsymbol{\theta}_S) \leq C_5 \frac{NL(\log_2 L \log_2 N)^{\frac{1}{2}}}{\sqrt{M}} \log M. \tag{13}$$

*where* $\widetilde{\Phi} := \{\phi : \phi$ *with the width* $\leq C_1 N \log N$ *and depth* $\leq C_2 L \log L, \|\phi\|_{\mathcal{W}^{1,\infty}((0,1)^d)} \leq B\}$, *and* $\mathcal{R}_S, \mathcal{R}_D, \boldsymbol{\theta}_S, \boldsymbol{\theta}_D$ *are defined in Eqs. (5,6).*

# Recent Works

**Nearly Optimal Approximation Rate in Sobolev Spaces including Higher-order Derivatives**

$$\mathcal{N}_{C,L} := \{\psi(\boldsymbol{x}) = \psi_2 \circ \boldsymbol{\psi}_1(\boldsymbol{x}) : \psi_2 \text{ is a ReLU}^k\text{-NN with depth } L_2, \text{ each component of } \boldsymbol{\psi}_1 \text{ is a ReL} \tag{14}$$

We refer to the elements in $\mathcal{N}_{C,L}$ as deep super ReLU networks (DSRNs).

**Theorem**

*For any $f \in \mathcal{W}^{n,p}((0,1)^d)$ with $\|f\|_{\mathcal{W}^{n,p}((0,1)^d)} \le 1$ for $m \in \mathbb{N}$ with $m \ge 2$ and $1 \le p \le +\infty$, any $N, L \in \mathbb{N}_+$ with $N \log_2 L + 2^{\lfloor \log_2 N \rfloor} \ge \max\{d, n\}$ and $L \ge N, m < n$, there is a DSRN $\gamma(\boldsymbol{x})$ in $\mathcal{N}_{\eta_1, \eta_2 L \log_2 L}$ with the width $\eta_3 N \log_2 N$ such that*

$$\|f(\boldsymbol{x}) - \gamma(\boldsymbol{x})\|_{\mathcal{W}^{m,p}((0,1)^d)} \le 2^{d+7} C_{11}(n,d) N^{-2(n-m)/d} L^{-2(n-m)/d},$$

*where $\eta_i, C_{11}$ are the constants independent with $N, L$.*

# Proof of the approximation rate

The proof of Theorem 3 can be outlined in five parts:
**(i)**: First of all, define a sequence of subsets of $\Omega$:

## Definition

*Given* $K, d \in \mathbb{N}^+$, *and for any* $\boldsymbol{m} = (m_1, m_2, \ldots, m_d) \in \{1,2\}^d$, *we define* $\Omega_{\boldsymbol{m}} := \prod_{j=1}^{d} \Omega_{m_j}$, *where* $\Omega_1 := \bigcup_{i=0}^{K-1} \left[\frac{i}{K}, \frac{i}{K} + \frac{3}{4K}\right]$, $\Omega_2 := \bigcup_{i=0}^{K} \left[\frac{i}{K} - \frac{1}{2K}, \frac{i}{K} + \frac{1}{4K}\right] \cap [0,1]$.

Then we define a partition of unity $\{g_{\boldsymbol{m}}\}_{\boldsymbol{m} \in \{1,2\}^d}$ on $(0,1)^d$ with $\operatorname{supp} g_{\boldsymbol{m}} \cap (0,1)^d \subset \Omega_{\boldsymbol{m}}$ for each $\boldsymbol{m} \in \{1,2\}^d$:

# $g_1$ and $g_2$

## Definition

*Given $K, d \in \mathbb{N}_+$, we define*

$$g_1(x) := \begin{cases} 1, & x \in \left[\frac{i}{K} + \frac{1}{4K}, \frac{i}{K} + \frac{1}{2K}\right] \\ 0, & x \in \left[\frac{i}{K} + \frac{3}{4K}, \frac{i+1}{K}\right] \\ 4K\left(x - \frac{i}{K}\right), & x \in \left[\frac{i}{K}, \frac{i}{K} + \frac{1}{4K}\right] \\ -4K\left(x - \frac{i}{K} - \frac{3}{4K}\right), & x \in \left[\frac{i}{K} + \frac{1}{2K}, \frac{i}{K} + \frac{3}{4K}\right] \end{cases}, \quad g_2(x) := g_1\left(x + \frac{1}{2K}\right), \qquad (15)$$

*for $i \in \mathbb{Z}$. For any $\boldsymbol{m} = (m_1, m_2, \ldots, m_d) \in \{1, 2\}^d$, define $g_{\boldsymbol{m}}(\boldsymbol{x}) = \prod_{j=1}^{d} g_{m_j}(x_j)$, $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$.*

# $g_1$ and $g_2$

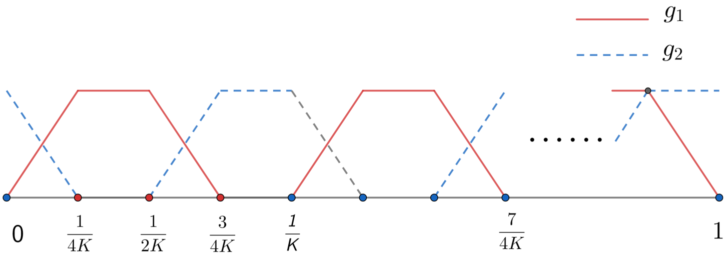

Figure: The schematic diagram of $g_i$ for $i = 1, 2$

# Proof of the approximation rate

**(ii)**: Then we use the following proposition to approximate $\{g_{\boldsymbol{m}}\}_{\boldsymbol{m}\in\{1,2\}^d}$ by $\sigma_1$-NNs and construct a sequence of $\sigma_1$-NNs $\{\phi_{\boldsymbol{m}}\}_{\boldsymbol{m}\in\{1,2\}^d}$:

## Proposition

*Given any $N, L, n \in \mathbb{N}_+$ for $K = \lfloor N^{1/d} \rfloor^2 \lfloor L^{2/d} \rfloor$, then for any $\boldsymbol{m} = (m_1, m_2, \ldots, m_d) \in \{1, 2\}^d$, there is a $\sigma_1$-NN with the width smaller than $(9 + d)(N + 1) + d - 1$ and depth smaller than $15d(d - 1)nL$ such as $\|\phi_{\boldsymbol{m}}(\boldsymbol{x}) - g_{\boldsymbol{m}}(\boldsymbol{x})\|_{\mathcal{W}^{1,\infty}((0,1)^d)} \leq 50d^{\frac{5}{2}}(N + 1)^{-4dnL}$.*

# Proof of the approximation rate

**(iii)**: For each $\Omega_{\boldsymbol{m}} \subset [0,1]^d$, where $\boldsymbol{m} \in \{1,2\}^d$, we find a function $f_{K,\boldsymbol{m}}$ satisfying

$$\begin{aligned} \|f - f_{K,\boldsymbol{m}}\|_{\mathcal{W}^{1,\infty}(\Omega_{\boldsymbol{m}})} &\leq C_1(n,d)K^{-(n-1)}, \\ \|f - f_{K,\boldsymbol{m}}\|_{L^\infty(\Omega_{\boldsymbol{m}})} &\leq C_1(n,d)K^{-n}, \end{aligned} \tag{16}$$

where $C_1$ is a constant independent of $K$. Moreover, each $f_{K,\boldsymbol{m}}$ can be expressed as $f_{K,\boldsymbol{m}} = \sum_{|\boldsymbol{\alpha}| \leq n-1} g_{f,\boldsymbol{\alpha},\boldsymbol{m}}(\boldsymbol{x})\boldsymbol{x}^{\boldsymbol{\alpha}}$, where $g_{f,\boldsymbol{\alpha},\boldsymbol{m}}(\boldsymbol{x})$ is a piecewise constant function on $\Omega_{\boldsymbol{m}}$. The proof of this result is based on the Bramble-Hilbert Lemma[5].

[5] S. Brenner, L. Scott, and L. Scott. The mathematical theory of finite element methods, volume 3. Springer, 2008

# Proof of the approximation rate

**(iv)**: Follow the work of Lu, etc. We obtain a neural network $\psi_{\boldsymbol{m}}$ with width $O(N \log N)$ and depth $O(L \log L)$ such that

$$\begin{aligned} \|f_{K,\boldsymbol{m}} - \psi_{\boldsymbol{m}}(\boldsymbol{x})\|_{\mathcal{W}^{1,\infty}(\Omega_{\boldsymbol{m}})} &\leq C_5(n,d) N^{-2(n-1)/d} L^{-2(n-1)/d} \\ \|f_{K,\boldsymbol{m}} - \psi_{\boldsymbol{m}}(\boldsymbol{x})\|_{L^{\infty}(\Omega_{\boldsymbol{m}})} &\leq C_5(n,d) N^{-2n/d} L^{-2n/d}, \end{aligned} \tag{17}$$

where $C_5$ is a constant independent of $N$ and $L$.
By combining (iii) and (iv) and setting $K = \lfloor N^{1/d} \rfloor^2 \lfloor L^{2/d} \rfloor$, we obtain that for each $\boldsymbol{m} \in \{1,2\}^d$, there exists a neural network $\psi_{\boldsymbol{m}}$ with width $O(N \log N)$ and depth $O(L \log L)$ such that

$$\begin{aligned} \|f(\boldsymbol{x}) - \psi_{\boldsymbol{m}}(\boldsymbol{x})\|_{\mathcal{W}^{1,\infty}(\Omega_{\boldsymbol{m}})} &\leq C_6(n,d) N^{-2(n-1)/d} L^{-2(n-1)/d} \\ \|f(\boldsymbol{x}) - \psi_{\boldsymbol{m}}(\boldsymbol{x})\|_{L^{\infty}(\Omega_{\boldsymbol{m}})} &\leq C_6(n,d) N^{-2n/d} L^{-2n/d}. \end{aligned} \tag{18}$$

# Proof of the approximation rate

**(v)**: The final step is to combine the sequences $\{\phi_{\boldsymbol{m}}\}_{\boldsymbol{m}\in\{1,2\}^d}$ and $\{\psi_{\boldsymbol{m}}\}_{\boldsymbol{m}\in\{1,2\}^d}$ to construct a network that can approximate $f$ over the entire space $[0,1]^d$. We define the sequence $\{\phi_{\boldsymbol{m}}\}_{\boldsymbol{m}\in\{1,2\}^d}$ because $\psi_{\boldsymbol{m}}$ may not accurately approximate $f$ on $[0,1]^d\backslash\Omega_{\boldsymbol{m}}$. The purpose of $\phi_{\boldsymbol{m}}$ is to remove this portion of the domain and allow other networks to approximate $f$ on $[0,1]^d\backslash\Omega_{\boldsymbol{m}}$.

# Proof of the bounds of VC-dimension of DNN derivatives

In the proof of Theorem 1, we use the following lemmas:

### Lemma

*Suppose $W \leq M$ and let $P_1, \ldots, P_M$ be polynomials of degree at most $D$ in $W$ variables. Define $K := \left|\{(\operatorname{sgn}(P_1(a)), \ldots, \operatorname{sgn}(P_M(a))) : a \in \mathbb{R}^W\}\right|$, then we have $K \leq 2(2eMD/W)^W$.*

### Lemma

*Suppose that $2^m \leq 2^t(mr/w)^w$ for some $r \geq 16$ and $m \geq w \geq t \geq 0$. Then, $m \leq t + w\log_2(2r\log_2 r)$.*

# Sketch of Proof

An element in $\Phi$ can be represented as $\phi = \boldsymbol{W}_{L+1}\sigma_1(\boldsymbol{W}_L\sigma_1(\dots\sigma_1(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1)\dots) + \boldsymbol{b}_L) + b_{L+1}$. Therefore, an element in $D\Phi$ can be represented as

$$\begin{aligned}\psi(\boldsymbol{x}) = D_i\phi(\boldsymbol{x}) =& \boldsymbol{W}_{L+1}\sigma_0(\boldsymbol{W}_L\sigma_1(\dots\sigma_1(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1)\dots) + \boldsymbol{b}_L) \\ &\cdot \boldsymbol{W}_L\sigma_0(\dots\sigma_1(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1)\dots)\dots\boldsymbol{W}_2\sigma_0(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1)(\boldsymbol{W}_1)_i, \qquad (19)\end{aligned}$$

where $(\boldsymbol{W})_i$ is $i$-th column of $\boldsymbol{W}$.

# Sketch of Proof

Let $\boldsymbol{x} \in \mathbb{R}^d$ be an input and $\boldsymbol{\theta} \in \mathbb{R}^W$ be a parameter vector in $\psi$. We denote the output of $\psi$ with input $\boldsymbol{x}$ and parameter vector $\boldsymbol{\theta}$ as $f(\boldsymbol{x}, \boldsymbol{\theta})$. For fixed $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$ in $\mathbb{R}^d$, we aim to bound

$$K := \left| \{ (\operatorname{sgn}(f(\boldsymbol{x}_1, \boldsymbol{\theta})), \ldots, \operatorname{sgn}(f(\boldsymbol{x}_m, \boldsymbol{\theta}))) : \boldsymbol{\theta} \in \mathbb{R}^W \} \right| . \tag{20}$$

For any partition $\mathcal{S} = \{P_1, P_2, \ldots, P_T\}$ of the parameter domain $\mathbb{R}^W$, we have $K \le \sum_{i=1}^{T} |\{ (\operatorname{sgn}(f(\boldsymbol{x}_1, \boldsymbol{\theta})), \ldots, \operatorname{sgn}(f(\boldsymbol{x}_m, \boldsymbol{\theta}))) : \boldsymbol{\theta} \in P_i \}|$.
We choose the partition such that within each region $P_i$, the functions $f(\boldsymbol{x}_j, \cdot)$ are all fixed polynomials of bounded degree. This allows us to bound each term in the sum using Lemma 1.

# Sketch of Proof

We define a sequence of sets of functions $\{\mathbb{F}_j\}_{j=0}^{L}$ with respect to parameters $\boldsymbol{\theta} \in \mathbb{R}^W$:

$$\begin{aligned}
\mathbb{F}_0 &:= \{(\boldsymbol{W}_1)_i, \boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1\} \\
\mathbb{F}_1 &:= \{(\boldsymbol{W}_1)_i, \boldsymbol{W}_2\sigma_0(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1), \boldsymbol{W}_2\sigma_1(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2\} \\
\mathbb{F}_2 &:= \{(\boldsymbol{W}_1)_i, \boldsymbol{W}_2\sigma_0(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1), \\
&\qquad \boldsymbol{W}_3\sigma_0(\boldsymbol{W}_2\sigma_1(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2), \boldsymbol{W}_3\sigma_1(\boldsymbol{W}_2\sigma_1(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2) + \boldsymbol{b}_3\} \\
&\ \vdots \\
\mathbb{F}_L &:= \{(\boldsymbol{W}_1)_i, \boldsymbol{W}_2\sigma_0(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1), \ldots, \boldsymbol{W}_{L+1}\sigma_0(\boldsymbol{W}_L\sigma_1(\ldots\sigma_1(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1)\ldots) + \boldsymbol{b}_L)\}. \qquad (21)
\end{aligned}$$

# Sketch of Proof

The partition of $\mathbb{R}^W$ is constructed layer by layer through successive refinements denoted by $\mathcal{S}_0, \mathcal{S}_1, \ldots, \mathcal{S}_L$. These refinements possess the following properties:

**1**. We have $|\mathcal{S}_0| = 1$, and for each $n = 1, \ldots, L$, we have $\frac{|\mathcal{S}_n|}{|\mathcal{S}_{n-1}|} \leq 2\left(\frac{2emnN_k}{\sum_{i=1}^n W_i}\right)^{\sum_{i=1}^n W_i}$.

**2**. For each $n = 0, \ldots, L-1$, each element $S$ of $\mathcal{S}_n$, when $\boldsymbol{\theta}$ varies in $S$, the output of each term in $\mathbb{F}_n$ is a fixed polynomial function in $\sum_{i=1}^n W_i$ variables of $\boldsymbol{\theta}$, with a total degree no more than $n+1$.

**3**. For each element $S$ of $\mathcal{S}_L$, when $\boldsymbol{\theta}$ varies in $S$, the $h$-th term in $\mathbb{F}_L$ for $h \in \{1, 2, \ldots, L+1\}$ is a fixed polynomial function in $W_h$ variables of $\boldsymbol{\theta}$, with a total degree no more than 1.

# Thanks for Listening!