7 Supplementary Material

7.1 Proofs of Theorems 1 and 2.

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

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

where $\boldsymbol{W}_i \in \mathbb{R}^{N_i\times N_{i-1}}$ ($(\boldsymbol{W})_i$ is i-th column of $\boldsymbol{W}$) and $\boldsymbol{b}_i \in \mathbb{R}^{N_i}$ are the weight matrix and the bias vector in the i-th linear transform in ϕ, and $\sigma_0(x) = \operatorname{sgn}(x) = 1[x > 0]$, which is the derivative of the ReLU function and $\sigma_0(\boldsymbol{x}) = \operatorname{diag}(\sigma_0(x_i))$. Denote W_i as the number of parameters in $\boldsymbol{W}_i, \boldsymbol{b}_i$, i.e., $W_i = N_iN_{i-1} + N_i$.

Let $\boldsymbol{x} \in \mathbb{R}^d$ be an input and $\boldsymbol{\theta} \in \mathbb{R}^W$ be a parameter vector in ψ. We denote the output of ψ 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

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

The proof is inspired by [4, Theorem 7]. 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.

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

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

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 $\frac{|\mathcal{S}_n|}{|\mathcal{S}_{n-1}|} \le 2\left(\frac{2emnN_k}{\sum_{i=1}^n W_i}\right)^{\sum_{i=1}^n W_i}$.

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.

We define $\mathcal{S}_0 = \{\mathbb{R}^W\}$, which satisfies properties 1,2 above, since $\boldsymbol{W}_1\boldsymbol{x}_j + \boldsymbol{b}_1$ and $(\boldsymbol{W}_1)_i$ are affine functions of $\boldsymbol{W}_1, \boldsymbol{b}_1$.

To define $\mathcal{S}_n$, we use the last term of $\mathbb{F}_{n-1}$ as inputs for the last two terms in $\mathbb{F}_n$. Assuming that $\mathcal{S}_0, \mathcal{S}_1, \ldots, \mathcal{S}_{n-1}$ have already been defined, we observe that the last two terms are new additions to $\mathbb{F}_n$ when comparing it to $\mathbb{F}_{n-1}$. Therefore, all elements in $\mathbb{F}_n$ except the last two are fixed polynomial functions in W_n variables of $\boldsymbol{\theta}$, with a total degree no greater than n when $\boldsymbol{\theta}$ varies in $S \in \mathcal{S}_n$. This is because $\mathcal{S}_n$ is a finer partition than $\mathcal{S}_{n-1}$.

We denote $p_{\boldsymbol{x}_j,n-1,S,k}(\boldsymbol{\theta})$ as the output of the k-th node in the last term of $\mathbb{F}_{n-1}$ in response to $\boldsymbol{x}_j$ when $\boldsymbol{\theta}\in S$. The collection of polynomials

$$\{p_{\boldsymbol{x}_j,n-1,S,k}(\boldsymbol{\theta}) : j = 1, \ldots, m,\ k = 1, \ldots, N_n\}$$

can attain at most $2\left(\frac{2emnN_n}{\sum_{i=1}^n W_i}\right)^{\sum_{i=1}^n W_i}$ distinct sign patterns when $\boldsymbol{\theta}\in S$ due to Lemma 1 for sufficiently large m. Therefore, we can divide S into $2\left(\frac{2emnN_n}{\sum_{i=1}^n W_i}\right)^{\sum_{i=1}^n W_i}$ parts, each having the property that $p_{\boldsymbol{x}_j,n-1,S,k}(\boldsymbol{\theta})$ does not change sign within the subregion. By performing this for all $S \in \mathcal{S}_{n-1}$, we obtain the desired partition $\mathcal{S}_n$. This division ensures that the required property 1 is satisfied.

Additionally, since the input to the last two terms in $\mathbb{F}_n$ is $p_{\boldsymbol{x}_j,n-1,S,k}(\boldsymbol{\theta})$, and we have shown that the sign of this input will not change in each region of $\mathcal{S}_n$, it follows that the output of the last two terms in $\mathbb{F}_n$ is also a polynomial without breakpoints in each element of $\mathcal{S}_n$. Therefore, the required property 2 is satisfied.

In the context of DNNs, the last layer is characterized by all terms containing the activation function σ_0. Consequently, for any element S of the partition $\mathcal{S}_L$, when the vector of parameters $\boldsymbol{\theta}$ varies within S, the h-th term in $\mathbb{F}_L$ for $h \in \{1, 2, \ldots, L+1\}$ can be expressed as a polynomial function of at most degree 1, which depends on at most W_h variables of $\boldsymbol{\theta}$. Hence, the required property 3 is satisfied.

Due to property 3, we multiply all the terms in $\mathbb{F}_L$ and obtain a term in $D\Phi$. Hence, the output of each term in $D\Phi$ is a polynomial function in $\sum_{i=1}^{L+1} W_i$ variables of $\boldsymbol{\theta} \in S \in \mathcal{S}_L$, of total degree no more than $L+1$. Therefore, for each $S \in \mathcal{S}_L$ we have $|\{(\operatorname{sgn}(f(\boldsymbol{x}_1, \boldsymbol{\theta})), \ldots, \operatorname{sgn}(f(\boldsymbol{x}_m, \boldsymbol{\theta}))) : \boldsymbol{\theta} \in S\}| \leq 2\left(2em(L+1)/\sum_{i=1}^{L+1} W_i\right)^{\sum_{i=1}^{L+1} W_i}$. Then

$$\begin{aligned}
K \leq & 2\left(2em(L+1)/\sum_{i=1}^{L+1} W_i\right)^{\sum_{i=1}^{L+1} W_i} \cdot \prod_{n=1}^{L} 2\left(\frac{2emnN_n}{\sum_{i=1}^{n} W_i}\right)^{\sum_{i=1}^{n} W_i} \leq \prod_{n=1}^{L+1} 2\left(\frac{2emnN_n}{\sum_{i=1}^{n} W_i}\right)^{\sum_{i=1}^{n} W_i} \\
\leq & 2^{L+1}\left(\frac{2em(L+2)(L+1)N}{2U}\right)^{U}
\end{aligned} \tag{22}$$

where $U := \sum_{n=1}^{L+1}\sum_{i=1}^{n} W_i = O(N^2L^2)$, N is the width of the network, and the last inequality is due to weighted AM-GM. For the definition of the VC-dimension, we have

$$2^{\text{VCdim}(D\Phi)} \leq 2^{L+1}\left(\frac{e\text{VCdim}(D\Phi)(L+1)(L+2)N}{U}\right)^{U}. \tag{23}$$

Due to Lemma 2, we obtain that

$$\text{VCdim}(D\Phi) \leq L+1+U\log_2[2(L+1)(L+2)\log_2(L+1)(L+2)] = O(N^2L^2\log_2 L\log_2 N) \tag{24}$$

since $U = O(N^2L^2)$. □

Proof of Theorem 2. Denote $D\Phi_{\mathcal{N}} := \{\eta(\boldsymbol{x}, y) : \eta(\boldsymbol{x}, y) = \psi(\boldsymbol{x}) - y, \psi \in D\Phi, (\boldsymbol{x}, y) \in \mathbb{R}^{d+1}\}$. Based on the definition of VC-dimension and pseudo-dimension, we have that

$$\text{Pdim}(D\Phi) \leq \text{VCdim}(D\Phi_{\mathcal{N}}). \tag{25}$$

For the VCdim$(D\Phi_{\mathcal{N}})$, it can be bounded by $O(N^2L^2\log_2 L\log_2 N)$. The proof is similar to that for the estimate of VCdim$(D\Phi)$ as given in Theorem 1. □

We establish that Pdim$(D\Phi) \leq$ VCdim$(D\Phi_{\mathcal{N}})$, where $\Phi_{\mathcal{N}}$ represents DNNs with $N+1$ width and $L+1$ depth. This implies that Pdim$(D\Phi)$ is upper bounded by $\bar{C}(N+1)^2(L+1)^2\log_2(L+1)\log_2(N+1) \leq 64\bar{C}N^2L^2\log_2 L\log_2 N$. Therefore, we conclude that $64\bar{C} \geq \hat{C}$.

7.2 Proof of Theorem 3

7.2.1 Propositions of Sobolev spaces and ReLU neural networks

The following two lemmas estimate the Sobolev norms and Sobolev semi-norms for the composition and product, which will be used in later proof.

Lemma 3 ([18, Corollary B.5]). *Let $d, m \in \mathbb{N}_+$ and $\Omega_1 \subset \mathbb{R}^d$ and $\Omega_2 \subset \mathbb{R}^m$ both be open, bounded, and convex. Then for $\boldsymbol{f} \in W^{1,\infty}(\Omega_1, \mathbb{R}^m)$ and $g \in W^{1,\infty}(\Omega_2)$ with $\operatorname{ran}\boldsymbol{f} \subset \Omega_2$, we have*

$$\|g \circ \boldsymbol{f}\|_{W^{1,\infty}(\Omega_2)} \leq \sqrt{d}m\max\{\|g\|_{L^\infty(\Omega_2)}, |g|_{W^{1,\infty}(\Omega_2)}|\boldsymbol{f}|_{W^{1,\infty}(\Omega_1,\mathbb{R}^m)}\}.$$

Lemma 4 ([18, Corollary B.6]). *Let $d \in \mathbb{N}_+$ and $\Omega \subset \mathbb{R}^d$. Then for $f, g \in W^{1,\infty}(\Omega)$, we have*

$$\|gf\|_{W^{1,\infty}(\Omega)} \le \|g\|_{L^\infty(\Omega)}|f|_{W^{1,\infty}(\Omega)} + \|f\|_{L^\infty(\Omega)}|g|_{W^{1,\infty}(\Omega)}.$$

Then we collect and establish some propositions for ReLU neural networks.

Proposition 2 ([28, Proposition 4.3]). *Given any $N, L \in \mathbb{N}_+$ and $\delta \in \left(0, \frac{1}{3K}\right]$ for $K = \lfloor N^{1/d}\rfloor^2 \lfloor L^{2/d}\rfloor$, there exists a σ_1-NN ϕ with the width $4N+5$ and depth $4L+4$ such that*

$$\phi(x) = k, x \in \left[\frac{k}{K}, \frac{k+1}{K} - \delta \cdot 1_{k<K-1}\right],\ k = 0, 1, \ldots, K-1.$$

Proposition 3. *[28, Proposition 4.4] Given any $N, L, s \in \mathbb{N}_+$ and $\xi_i \in [0,1]$ for $i = 0, 1, \ldots N^2L^2 - 1$, there exists a σ_1-NN ϕ with the width $16s(N+1)\log_2(8N)$ and depth $(5L+2)\log_2(4L)$ such that*

1. $|\phi(i) - \xi_i| \le N^{-2s}L^{-2s}$ for $i = 0, 1, \ldots N^2L^2 - 1$.

2. $0 \le \phi(x) \le 1$, $x \in \mathbb{R}$.

Proposition 4. *For any $N, L \in \mathbb{N}_+$ and $a > 0$, there is a σ_1-NN ϕ with the width $15N$ and depth $2L$ such that $\|\phi\|_{W^{1,\infty}((-a,a)^2)} \le 12a^2$ and*

$$\|\phi(x,y) - xy\|_{W^{1,\infty}((-a,a)^2)} \le 6a^2N^{-L}. \tag{26}$$

Furthermore,

$$\phi(0,y) = \frac{\partial \phi(0,y)}{\partial y} = 0,\ y \in (-a,a). \tag{27}$$

Proof. We first need to construct a neural network to approximate x^2 on $(-1,1)$, and the idea is similar with [23, Lemma 3.2] and [28, Lemma 5.1]. The reason we do not use [23, Lemma 3.4] and [28, Lemma 4.2] directly is that constructing $\phi(x,y)$ by translating a neural network in $W^{1,\infty}[0,1]$ will lose the proposition of $\phi(0.y) = 0$. Here we need to define teeth functions T_i on $\widetilde{x} \in [-1,1]$:

$$T_1(\widetilde{x}) = \begin{cases} 2|\widetilde{x}|, & |\widetilde{x}| \le \frac{1}{2}, \\ 2(1-|\widetilde{x}|), & |\widetilde{x}| > \frac{1}{2}, \end{cases}$$

and

$$T_i = T_{i-1} \circ T_1, \quad \text{for } i = 2, 3, \cdots.$$

Define

$$\widetilde{\psi}(\widetilde{x}) = \widetilde{x} - \sum_{i=1}^{s} \frac{T_i(\widetilde{x})}{2^{2i}},$$

According to [23, Lemma 3.2] and [28, Lemma 5.1], we know ψ is a neural network with the width $5N$ and depth $2L$ such that $\|\widetilde{\psi}(\widetilde{x})\|_{W^{1,\infty}((-1,1))} \le 2$, $\|\widetilde{\psi}(\widetilde{x}) - \widetilde{x}^2\|_{W^{1,\infty}((-1,1))} \le N^{-L}$ and $\psi(0) = 0$.

By setting $x = a\widetilde{x} \in (-a,a)$ for $\widetilde{x} \in (-1,1)$, we define

$$\psi(x) = a^2 \widetilde{\psi}\left(\frac{x}{a}\right).$$

Note that $x^2 = a^2\left(\frac{x}{a}\right)^2$, we have

$$\begin{aligned} \|\psi(x) - x^2\|_{\mathcal{W}^{1,\infty}(-a,a)} &= a^2 \left\|\widetilde{\psi}\left(\frac{x}{a}\right) - \left(\frac{x}{a}\right)^2\right\|_{\mathcal{W}^{1,\infty}((-a,a))} \\ &\le a^2 N^{-L}, \end{aligned}$$

and $\psi(0) = 0$, which will be used to prove Eq. (27).

Then we can construct $\phi(x,y)$ as

$$\phi(x,y) = 2\left[\psi\left(\frac{|x+y|}{2}\right) - \psi\left(\frac{|x|}{2}\right) - \psi\left(\frac{|y|}{2}\right)\right] \tag{28}$$

where $\phi(x)$ is a neural network with the width $15N$ and depth $2L$ such that $\|\phi\|_{W^{1,\infty}((-a,a)^2)} \leq 12a^2$ and

$$\|\phi(x,y) - xy\|_{W^{1,\infty}((-a,a)^2)} \leq 6a^2N^{-L}. \tag{29}$$

For the last equation Eq. (27) is due to $\phi(x,y)$ in the proof can be read as Eq. (28) with $\psi(0) = 0$. $\square$

Proposition 5. *For any $N, L, s \in \mathbb{N}_+$ with $s \geq 2$, there exists a σ_1-NN ϕ with the width $9(N+1) + s - 1$ and depth $14s(s-1)L$ such that $\|\phi\|_{\mathcal{W}^{1,\infty}((0,1)^s)} \leq 18$ and*

$$\|\phi(\boldsymbol{x}) - x_1x_2\cdots x_s\|_{\mathcal{W}^{1,\infty}((0,1)^s)} \leq 10(s-1)(N+1)^{-7sL}. \tag{30}$$

Furthermore, for any $i = 1, 2, \ldots, s$, if $x_i = 0$, we will have

$$\phi(x_1, x_2, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_s) = \frac{\partial \phi(x_1, x_2, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_s)}{\partial x_j} = 0, \ i \neq j. \tag{31}$$

Proof. The proof of the first inequality Eq. (30) can be found in [23, Lemma 3.5]. The proof of Eq. (31) can be obtained via induction. For $s = 2$, based on Proposition 4, we know there is a neural network ϕ_2 satisfied Eq. (31).

Now assume that for any $i \leq n-1$, there is a neural network ϕ_i satisfied Eq. (31). ϕ_n in [23] is constructed as

$$\phi_n(x_1, x_2, \ldots, x_n) = \phi_2(\phi_{n-1}(x_1, x_2, \ldots, x_{n-1}), \sigma(x_n)), \tag{32}$$

which satisfies Eq. (30). Then $\phi_n(x_1, x_2, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n) = 0$ for any $i = 1, 2, \ldots, n$. For $i = n$, we have

$$\frac{\phi(x_1, x_2, \ldots, 0)}{\partial x_j} = \underbrace{\frac{\partial \phi_2(\phi_{n-1}(x_1, x_2, \ldots, x_{n-1}), 0)}{\partial \phi_{n-1}(x_1, x_2, \ldots, x_{n-1})}}_{=0, \text{ by the property of } \phi_2.} \cdot \frac{\partial \phi_{n-1}(x_1, x_2, \ldots, x_{n-1})}{\partial x_j} = 0. \tag{33}$$

For $i < n$ and $j < n$, we have

$$\begin{aligned} &\frac{\phi(x_1, x_2, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n)}{\partial x_j} \\ =&\frac{\partial \phi_2(\phi_{n-1}(x_1, x_2, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_{n-1}), \sigma(x_n))}{\partial \phi_{n-1}(x_1, \ldots, 0, x_{i+1}, \ldots, x_{n-1})} \cdot \underbrace{\frac{\partial \phi_{n-1}(x_1, \ldots, 0, x_{i+1}, \ldots, x_{n-1})}{\partial x_j}}_{=0, \text{ via induction.}} = 0. \end{aligned} \tag{34}$$

For $i < n$ and $j = n$, we have

$$\begin{aligned} &\frac{\phi(x_1, x_2, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n)}{\partial x_n} \\ =&\underbrace{\frac{\partial \phi_2(\phi_{n-1}(x_1, x_2, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_{n-1}), \sigma(x_n))}{\partial \sigma(x_n)}}_{=0, \text{ by the property of } \phi_2.} \cdot \frac{\mathrm{d}\sigma(x_n)}{\mathrm{d}x_n} = 0. \end{aligned} \tag{35}$$

Therefore, Eq. (31) is valid. $\square$

Proposition 6 ([23, Propositiion 3.6])**.** *For any $N, L, s \in \mathbb{N}_+$ and $|\boldsymbol{\alpha}| \leq s$, there is a σ_1-NN ϕ with the width $9(N+1) + s - 1$ and depth $14s^2L$ such that $\|\phi\|_{W^{1,\infty}((0,1)^d)} \leq 18$ and*

$$\|\phi(\boldsymbol{x}) - \boldsymbol{x}^{\boldsymbol{\alpha}}\|_{W^{1,\infty}((0,1)^d)} \leq 10s(N+1)^{-7sL}. \tag{36}$$

Proposition 7 ([39, Proposition 1])**.** *Given a sequence of the neural network $\{p_i\}_{i=1}^M$, and each p_i is a σ_1-NN from $\mathbb{R} \to \mathbb{R}$ with the width N and depth L_i, then $\sum_{i=1}^M p_i$ is a σ_1-NN with the width $N+4$ and depth $\sum_{i=1}^M L_i$.*

We present the proof of Proposition 1 below.

Proof of Proposition 1. First, we construct g_1 and g_2 by neural networks in $[0,1]$. Note that $\lfloor L^{2/d}\rfloor \leq L^{2/d} \leq \left(\lfloor L^{1/d}\rfloor+1\right)^2$. We first construct a σ_1-NN in the small set $\left[0, \lfloor N^{1/d}\rfloor\lfloor L^{2/d}\rfloor\right]$. It is easy to check there is a neural network $\hat{\psi}$ with the width 4 and one layer such as

$$\hat{\psi}(x) := \begin{cases} 1, & x \in \left[\frac{1}{8K}, \frac{3}{8K}\right] \\ 4K\left(x-\frac{1}{8K}\right), & x \in \left[\frac{1}{8K}, \frac{3}{8K}\right] \\ -4K\left(x-\frac{7}{8K}\right), & x \in \left[\frac{5}{8K}, \frac{7}{8K}\right] \\ 0, & \text{Otherwise.} \end{cases} \tag{37}$$

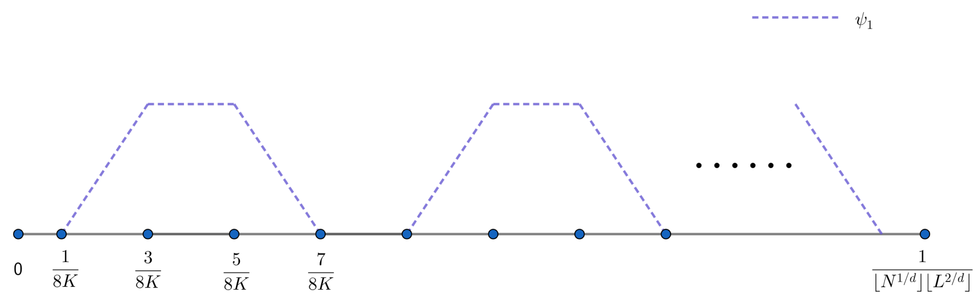

Figure 2: ψ_1

Hence, we have a network ψ_1 with the width $4\lfloor N^{1/d}\rfloor$ and one layer such as

$$\psi_1(x) := \sum_{i=0}^{\lfloor N^{1/d}\rfloor-1} \hat{\psi}\left(x-\frac{i}{K}\right).$$

Next, we construct ψ_i for $i=2,3,4$ based on the symmetry and periodicity of g_i. ψ_2 is the function with period $\frac{2}{\lfloor N^{1/d}\rfloor\lfloor L^{2/d}\rfloor}$ in $\left[0, \frac{1}{\lfloor L^{2/d}\rfloor}\right]$, and each period is a hat function with gradient 1. ψ_3 is the function with period $\frac{2}{\lfloor L^{2/d}\rfloor}$ in $\left[0, \frac{\lfloor L^{1/d}\rfloor+1}{\lfloor L^{2/d}\rfloor}\right]$, and each period is a hat function with gradient 1. ψ_4 is the function with period $\frac{2\left(\lfloor L^{1/d}\rfloor+1\right)}{\lfloor L^{2/d}\rfloor}$ in $\left[0, \frac{\left(\lfloor L^{1/d}\rfloor+1\right)^2}{\lfloor L^{2/d}\rfloor}\right]$, and each period is a hat function with gradient 1. The schematic diagram is in Fig. 3 (The diagram is shown the case for $\lfloor N^{1/d}\rfloor$ and $\lfloor L^{1/d}\rfloor+1$ is a even integer.).

Note that $\psi_2\circ\psi_3\circ\psi_4(x)$ is the function with period $\frac{2}{\lfloor N^{1/d}\rfloor\lfloor L^{2/d}\rfloor}$ in $[0,1] \subset \left[0, \frac{\left(\lfloor L^{1/d}\rfloor+1\right)^2}{\lfloor L^{2/d}\rfloor}\right]$, and each period is a hat function with gradient 1. Then function $\psi_1\circ\psi_2\circ\psi_3\circ\psi_4(x)$ is obtained by repeating reflection ψ_1 in $\left[0, \frac{\left(\lfloor L^{1/d}\rfloor+1\right)^2}{\lfloor L^{2/d}\rfloor}\right]$, which is the function we want.

Similar with ψ_1, ψ_2 is a network with $4\lfloor N^{1/d}\rfloor$ width and one layer. Due to Proposition 7, we know that ψ_3 and ψ_4 is a network with 7 width and $\lfloor L^{1/d}\rfloor+1$ depth. Hence

$$\psi(x) := \psi_1\circ\psi_2\circ\psi_3\circ\psi_4(x) \tag{38}$$

is a network with $4\lfloor N^{1/d}\rfloor$ width and $2\lfloor L^{1/d}\rfloor+4$ depth and $g_1 = \psi\left(x+\frac{1}{8K}\right)$ and $g_1 = \psi\left(x+\frac{5}{8K}\right)$.

Now we can construct $g_{\boldsymbol{m}}$ for $m \in \{1,2\}^d$ based on Proposition 5: There is a neural network ϕ_{prod} with the width $9(N+1)+d-1$ and depth $14d(d-1)nL$ such that $\|\phi_{\text{prod}}\|_{\mathcal{W}^{1,\infty}((0,1)^d)} \leq 18$ and

$$\|\phi_{\text{prod}}(\boldsymbol{x}) - x_1x_2\cdots x_d\|_{\mathcal{W}^{1,\infty}((0,1)^d)} \leq 10(d-1)(N+1)^{-7dnL}.$$

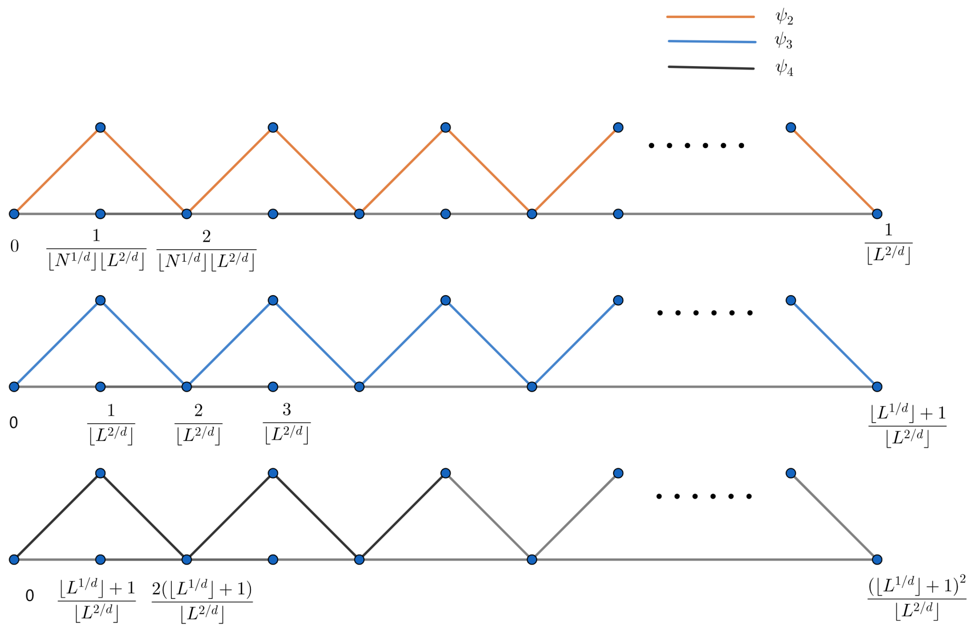

Figure 3: ψ_i for $i = 2, 3, 4$

Then denote $\phi_{\boldsymbol{m}}(\boldsymbol{x}) := \phi_{\text{prod}}(g_{m_1}, g_{m_2}, \ldots, g_{m_d})$ which is a neural network with the width smaller than $(9+d)(N+1)+d-1$ and depth smaller than $15d(d-1)nL$. Furthermore, due to Lemma 3, we have

$$\begin{aligned}
\|\phi_{\boldsymbol{m}}(\boldsymbol{x}) - g_{\boldsymbol{m}}(\boldsymbol{x})\|_{\mathcal{W}^{1,\infty}((0,1)^d)} \leq & d^{\frac{3}{2}} \|\phi_{\text{prod}}(\boldsymbol{x}) - x_1x_2\cdots x_d\|_{L^\infty((0,1)^d)} \\
& + d^{\frac{3}{2}} \|\phi_{\text{prod}}(\boldsymbol{x}) - x_1x_2\cdots x_d\|_{\mathcal{W}^{1,\infty}((0,1)^d)} |\psi|_{W^{1,\infty}(0,1)} \\
\leq & d^{\frac{3}{2}} 10(d-1)(N+1)^{-7ndL} \left(1 + 4\lfloor N^{1/d}\rfloor^2 \lfloor L^{2/d}\rfloor\right) \\
\leq & 50 d^{\frac{5}{2}} (N+1)^{-4dnL}, \qquad (39)
\end{aligned}$$

where the last inequality is due to

$$\frac{\lfloor N^{1/d}\rfloor^2 \lfloor L^{2/d}\rfloor}{(N+1)^{3dnL}} \leq \frac{N^2L^2}{(N+1)^{3dnL}} \leq \frac{L^2}{(N+1)^{3dnL-2}} \leq \frac{L^2}{2^{dnL}} \leq 1.$$

□

In the final of this subsection, we establish three lemmas for $\{\Omega_{\boldsymbol{m}}\}_{\boldsymbol{m}\in\{1,2\}^d}$, $\{g_{\boldsymbol{m}}\}_{\boldsymbol{m}\in\{1,2\}^d}$ and $\{\phi_{\boldsymbol{m}}\}_{\boldsymbol{m}\in\{1,2\}^d}$ defined in Subsection 3.1.

Lemma 5. *For $\{\Omega_{\boldsymbol{m}}\}_{\boldsymbol{m}\in\{1,2\}^d}$ defined in Definition 5, we have*

$$\bigcup_{\boldsymbol{m}\in\{1,2\}^d} \Omega_{\boldsymbol{m}} = [0,1]^d.$$

Proof. We prove this lemma via induction. $d = 1$ is valid due to $\Omega_1 \cup \Omega_2 = [0,1]$. Assume that the lemma is true for $d-1$, then

$$\begin{aligned}
\bigcup_{\boldsymbol{m}\in\{1,2\}^d} \Omega_{\boldsymbol{m}} = & [0,1]^d = \bigcup_{\boldsymbol{m}\in\{1,2\}^{d-1}} \Omega_{\boldsymbol{m}} \times \Omega_1 + \bigcup_{\boldsymbol{m}\in\{1,2\}^{d-1}} \Omega_{\boldsymbol{m}} \times \Omega_2 \\
= & \left([0,1]^{d-1} \times \Omega_1\right) \bigcup \left([0,1]^{d-1} \times \Omega_2\right) = [0,1]^d, \qquad (40)
\end{aligned}$$

hence the case of d is valid, and we finish the proof of the lemma. □

Lemma 6. *$\{g_{\boldsymbol{m}}\}_{\boldsymbol{m}\in\{1,2\}^d}$ defined in Definition 6 satisfies:*

(i): $\sum_{\boldsymbol{m}\in\{1,2\}^d} g_{\boldsymbol{m}}(\boldsymbol{x}) = 1$ for every $x \in [0,1]^d$.

(ii): $\operatorname{supp} g_{\boldsymbol{m}} \cap [0,1]^d \subset \Omega_{\boldsymbol{m}}$, where $\Omega_{\boldsymbol{m}}$ is defined in Definition 5.

(ii): For any $\boldsymbol{m} = (m_1, m_2, \ldots, m_d) \in \{1,2\}^d$ and $\boldsymbol{x} = (x_1, x_2, \ldots, x_d) \in [0,1]^d \backslash \Omega_{\boldsymbol{m}}$, there exists j such as $g_{m_j}(x_j) = 0$ and $\frac{\mathrm{d}g_{m_j}(x_j)}{\mathrm{d}x_j} = 0$.

Proof. (i) can be proved via induction as Lemma 5, and we leave it to readers.

As for (ii) and (iii), without loss of generality, we show the proof for $\boldsymbol{m}_* := (1, 1, \ldots, 1)$. For any $\boldsymbol{x} \in [0,1]^d \backslash \Omega_{\boldsymbol{m}_*}$, there is $x_j \in [0,1] \backslash \Omega_1$. Then $g_1(x_j) = 0$ and $g_{\boldsymbol{m}_*}(\boldsymbol{x}) = \prod_{j=1}^d g_1(x_j) = 0$, therefore $\operatorname{supp} g_{\boldsymbol{m}_*} \cap [0,1]^d \subset \Omega_{\boldsymbol{m}_*}$. Furthermore, $\frac{\mathrm{d}g_{m_j}(x_j)}{\mathrm{d}x_j} = 0$ for $x_j \in [0,1] \in \Omega_1$ due to the definition of g_1 (Definition 6), then we finish this proof. $\square$

The following lemma demonstrates that $\phi_{\boldsymbol{m}}$, as defined in Proposition 1, can restrict the Sobolev norm of the entire space to $\Omega_{\boldsymbol{m}}$.

Lemma 7. *For any $\chi(\boldsymbol{x}) \in W^{1,\infty}((0,1)^d)$, denote*

$$M = \max\{\|\chi\|_{W^{1,\infty}((0,1)^d)}, \|\phi_{\boldsymbol{m}}\|_{W^{1,\infty}((0,1)^d)}\},$$

then we have

$$\begin{aligned} \|\phi_{\boldsymbol{m}}(\boldsymbol{x}) \cdot \chi(\boldsymbol{x})\|_{W^{1,\infty}((0,1)^d)} =& \|\phi_{\boldsymbol{m}}(\boldsymbol{x}) \cdot \chi(\boldsymbol{x})\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}})} \\ \|\phi_{\boldsymbol{m}}(\boldsymbol{x}) \cdot \chi(\boldsymbol{x}) - \phi_M(\phi_{\boldsymbol{m}}(\boldsymbol{x}), \chi(\boldsymbol{x}))\|_{W^{1,\infty}((0,1)^d)} =& \|\phi_{\boldsymbol{m}}(\boldsymbol{x}) \cdot \chi(\boldsymbol{x}) - \phi_M(\phi_{\boldsymbol{m}}(\boldsymbol{x}), \chi(\boldsymbol{x}))\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}})} \end{aligned} \tag{41}$$

for any $\boldsymbol{m} \in \{1,2\}^d$, where $\phi_{\boldsymbol{m}}(\boldsymbol{x})$ and $\Omega_{\boldsymbol{m}}$ is defined in Proposition 1 and Definition. 5, and ϕ_M is from Proposition 4 (choosing $a = M$ in the proposition).

Proof. For the first equality, we only need to show that

$$\|\phi_{\boldsymbol{m}}(\boldsymbol{x}) \cdot \chi(\boldsymbol{x})\|_{W^{1,\infty}((0,1)^d \backslash \Omega_{\boldsymbol{m}})} = 0. \tag{42}$$

According to the Proposition 1, we have $\phi_{\boldsymbol{m}}(\boldsymbol{x}) = \phi_{\text{prod}}(g_{m_1}, g_{m_2}, \ldots, g_{m_d})$, and for any $\boldsymbol{x} = (x_1, x_2, \ldots, x_d) \in (0,1)^d \backslash \Omega_{\boldsymbol{m}}$, there is m_j such as $g_{m_j}(x_j) = 0$ and $\frac{\mathrm{d}g_{m_j}(x_j)}{\mathrm{d}x_j} = 0$ due to Lemma 6. Based on Eq. (31) in Proposition 5, we have

$$\phi_{\boldsymbol{m}}(\boldsymbol{x}) = \frac{\partial \phi_{\boldsymbol{m}}(\boldsymbol{x})}{\partial x_s} = 0, \ x \in (0,1)^d \backslash \Omega_{\boldsymbol{m}}, s \neq j.$$

Furthermore,

$$\frac{\partial \phi_{\boldsymbol{m}}(\boldsymbol{x})}{\partial x_j} = \frac{\partial \phi_{\text{prod}}(g_{m_1}, g_{m_2}, \ldots, g_{m_d})}{\partial g_{m_j}} \frac{\mathrm{d}g_{m_j}(x_j)}{\mathrm{d}x_j} = 0. \tag{43}$$

Hence we have

$$|\phi_{\boldsymbol{m}}(\boldsymbol{x}) \cdot \chi(\boldsymbol{x})| + \sum_{q=1}^d \left| \frac{\partial\left[\phi_{\boldsymbol{m}}(\boldsymbol{x}) \cdot \chi(\boldsymbol{x})\right]}{\partial x_q} \right| = 0 \tag{44}$$

for all $\boldsymbol{x} \in (0,1)^d \backslash \Omega_{\boldsymbol{m}}$.

Similarly, for the second equality in this lemma, we have

$$\begin{aligned} &|\phi_M(\phi_{\boldsymbol{m}}(\boldsymbol{x}), \chi(\boldsymbol{x}))| + \sum_{q=1}^d \left| \frac{\partial\left[\phi_M(\phi_{\boldsymbol{m}}(\boldsymbol{x}), \chi(\boldsymbol{x}))\right]}{\partial x_q} \right| \\ =& |\phi_M(0, \chi(\boldsymbol{x}))| + \sum_{q=1}^d \left[\left| \frac{\partial\left[\phi_M(0, \chi(\boldsymbol{x}))\right]}{\partial \chi(\boldsymbol{x})} \cdot \frac{\partial \chi(\boldsymbol{x})}{\partial x_q} \right| + \left| \frac{\partial\left[\phi_M(\phi_{\boldsymbol{m}}(\boldsymbol{x}), \chi(\boldsymbol{x}))\right]}{\partial \phi_{\boldsymbol{m}}(\boldsymbol{x})} \cdot \frac{\partial \phi_{\boldsymbol{m}}(\boldsymbol{x})}{\partial x_q} \right| \right] \\ =& 0, \end{aligned} \tag{45}$$

for all $\boldsymbol{x} \in (0,1)^d \backslash \Omega_{\boldsymbol{m}}$ based on

$$\phi_M(0,y) = \frac{\partial \phi_M(0,y)}{\partial y} = 0,\ y \in (-M, M),$$

and $\frac{\partial \phi_{\boldsymbol{m}}(\boldsymbol{x})}{\partial x_q} = 0$. Hence we finish our proof. □

7.2.2 An approximation of functions in Sobolev spaces based on the Bramble–Hilbert Lemma [7, Lemma 4.3.8]

In this subsection, we establish $\{f_{K,\boldsymbol{m}}\}_{\boldsymbol{m}\in\{1,2\}^d}$ as mentioned in Subsection 3.1, which is presented in Theorem 6. To prove this result, we build upon the work of [18], which leverages the average Taylor polynomials and the Bramble-Hilbert Lemma to approximate functions in Sobolev spaces.

Before we show Theorem 6, we define subsets of $\Omega_{\boldsymbol{m}}$ for simplicity notations.

For any $\boldsymbol{m} \in \{1,2\}^d$, we define

$$\Omega_{\boldsymbol{m},\boldsymbol{i}} := [0,1]^d \cap \prod_{j=1}^{d} \left[\frac{2i_j - 1_{m_j \le 2}}{2K}, \frac{3 + 4i_j - 2 \cdot 1_{m_j \le 2}}{4K}\right] \tag{46}$$

$\boldsymbol{i} = (i_1, i_2, \ldots, i_d) \in \{0, 1 \ldots, K\}^d$, and it is easy to check $\bigcup_{\boldsymbol{i}\in\{0,1\ldots,K\}^d} \Omega_{\boldsymbol{m},\boldsymbol{i}} = \Omega_{\boldsymbol{m}}$.

Theorem 6. *Let $K \in \mathbb{N}_+$ and $n \ge 2$. Then for any $f \in W^{n,\infty}((0,1)^d)$ with $\|f\|_{W^{n,\infty}((0,1)^d)} \le 1$ and $\boldsymbol{m} \in \{1,2\}^d$, there exist piece-wise polynomials function $f_{K,\boldsymbol{m}} = \sum_{|\boldsymbol{\alpha}|\le n-1} g_{f,\boldsymbol{\alpha},\boldsymbol{m}}(\boldsymbol{x})\boldsymbol{x}^{\boldsymbol{\alpha}}$ on $\Omega_{\boldsymbol{m}}$ (Definition 5) with the following properties:*

$$\begin{aligned} \|f - f_{K,\boldsymbol{m}}\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}})} &\le C_1(n,d)K^{-(n-1)}, \\ \|f - f_{K,\boldsymbol{m}}\|_{L^{\infty}(\Omega_{\boldsymbol{m}})} &\le C_1(n,d)K^{-n}. \end{aligned} \tag{47}$$

Furthermore, $g_{f,\boldsymbol{\alpha},\boldsymbol{m}}(\boldsymbol{x}) : \Omega_{\boldsymbol{m}} \to \mathbb{R}$ is a constant function with on each $\Omega_{\boldsymbol{m},\boldsymbol{i}}$ for $\boldsymbol{i} \in \{0, 1 \ldots, K\}^d$. And

$$|g_{f,\boldsymbol{\alpha},\boldsymbol{m}}(\boldsymbol{x})| \le C_2(n,d) \tag{48}$$

for all $\boldsymbol{x} \in \Omega_{\boldsymbol{m}}$, where C_1 and C_2 are constants independent with K.

This proof is similar to that of [18, Lemma C.4.], but we provide detailed proof as follows for readability. Before the proof, we must introduce the partition of unity, average Taylor polynomials, and a lemma.

Definition 7 (The partition of unity)**.** *Let $d, K \in \mathbb{N}_+$, then*

$$\Psi = \left\{h_{\boldsymbol{i}} : \boldsymbol{i} \in \{0,1,\ldots,K\}^d\right\}$$

with $h_{\boldsymbol{i}} : \mathbb{R}^d \to \mathbb{R}$ for all $\boldsymbol{i} \in \{0,1,\ldots,K\}^d$ is called the partition of unity $[0,1]^d$ if it satisfies

(i): $0 \le h_{\boldsymbol{i}}(\boldsymbol{x}) \le 1$ for every $h_{\boldsymbol{i}} \in \Psi$.

(ii): $\sum_{h_{\boldsymbol{i}}\in\Psi} h_{\boldsymbol{i}} = 1$ for every $x \in [0,1]^d$.

Definition 8. *Let $n \ge 1$ and $f \in W^{n,\infty}((0,1)^d)$, $\boldsymbol{x}_0 \in ((0,1)^d)$ and $r > 0$ such that for the ball $B(\boldsymbol{x}_0) := B(\boldsymbol{x}_0)_{r,|\cdot|}$ which is a compact subset of $((0,1)^d)$. The corresponding Taylor polynomial of order n of f averaged over B is defined for*

$$Q^n f(x) := \int_B T^n_{\boldsymbol{y}} f(\boldsymbol{x}) b_r(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y} \tag{49}$$

where

$$\begin{aligned} T^n_{\boldsymbol{y}} f(\boldsymbol{x}) &:= \sum_{|\boldsymbol{\alpha}|\le n-1} \frac{1}{\boldsymbol{\alpha}!} D^{\boldsymbol{\alpha}} f(\boldsymbol{y})(\boldsymbol{x}-\boldsymbol{y})^{\boldsymbol{\alpha}}, \\ b_r(\boldsymbol{x}) &:= \begin{cases} \frac{1}{c_r} e^{-\left(1-(|\boldsymbol{x}-\boldsymbol{x}_0|/r)^2\right)^{-1}}, & |\boldsymbol{x}-\boldsymbol{x}_0| < r, \\ 0, & |\boldsymbol{x}-\boldsymbol{x}_0| \le r, \end{cases} \\ c_r &= \int_{\mathbb{R}^d} e^{-\left(1-(|\boldsymbol{x}-\boldsymbol{x}_0|/r)^2\right)^{-1}} \mathrm{d}x. \end{aligned} \tag{50}$$

Lemma 8. *Let $n \geq 1$ and $f \in W^{n,\infty}((0,1)^d)$, $\boldsymbol{x}_0 \in \Omega$ and $r > 0$ such that for the ball $B(\boldsymbol{x}_0) := B_{r,|\cdot|}(\boldsymbol{x}_0)$ which is a compact subset of $((0,1)^d)$. The corresponding Taylor polynomial of order n of f averaged over B can be read as*

$$Q^n f(\boldsymbol{x}) = \sum_{|\boldsymbol{\alpha}| \leq n-1} c_{f,\boldsymbol{\alpha}} \boldsymbol{x}^{\boldsymbol{\alpha}}.$$

Furthermore,

$$|c_{f,\boldsymbol{\alpha}}| \leq C_2(n,d) \|f\|_{W^{n-1,\infty}(B)}. \tag{51}$$

where $C_2(n,d) = \sum_{|\boldsymbol{\alpha}+\boldsymbol{\beta}| \leq n-1} \frac{1}{\boldsymbol{\alpha}!\boldsymbol{\beta}!}$.

Proof. Based on [18, Lemma B.9.], $Q^n f(x)$ can be read as

$$Q^n f(\boldsymbol{x}) = \sum_{|\boldsymbol{\alpha}| \leq n-1} c_{f,\boldsymbol{\alpha}} \boldsymbol{x}^{\boldsymbol{\alpha}} \tag{52}$$

where

$$c_{f,\boldsymbol{\alpha}} = \sum_{|\boldsymbol{\alpha}+\boldsymbol{\beta}| \leq n-1} \frac{1}{(\boldsymbol{\beta}+\boldsymbol{\alpha})!} a_{\boldsymbol{\beta}+\boldsymbol{\alpha}} \int_B D^{\boldsymbol{\alpha}+\boldsymbol{\beta}} f(\boldsymbol{x}) \boldsymbol{y}^{\boldsymbol{\beta}} b_r(\boldsymbol{y}) \,\mathrm{d}\boldsymbol{y} \tag{53}$$

for $a_{\boldsymbol{\beta}+\boldsymbol{\alpha}} \leq \frac{(\boldsymbol{\alpha}+\boldsymbol{\beta})!}{\boldsymbol{\alpha}!\boldsymbol{\beta}!}$. Note that

$$\left| \int_B D^{\boldsymbol{\alpha}+\boldsymbol{\beta}} f(\boldsymbol{x}) \boldsymbol{y}^{\boldsymbol{\beta}} b_r(\boldsymbol{y}) \,\mathrm{d}\boldsymbol{y} \right| \leq \|f\|_{W^{n-1,\infty}(B)} \|b_r(x)\|_{L^1(B)} = \|f\|_{W^{n-1,\infty}(B)}. \tag{54}$$

Then

$$|c_{f,\boldsymbol{\alpha}}| \leq C_2(n,d) \|f\|_{W^{n-1,\infty}(B_{\boldsymbol{m},N})}. \tag{55}$$

where $C_2(n,d) = \sum_{|\boldsymbol{\alpha}+\boldsymbol{\beta}| \leq n-1} \frac{1}{\boldsymbol{\alpha}!\boldsymbol{\beta}!}$. ∎

The proof of Theorem 6 is based on average Taylor polynomials and the Bramble–Hilbert Lemma [7, Lemma 4.3.8].

Definition 9. *Let $\Omega,\ B \in \mathbb{R}^d$. Then Ω is called stared-shaped with respect to B if*

$$\overline{\mathit{conv}}\left(\{\boldsymbol{x}\} \cup B \subset \Omega\right), \textit{ for all } \boldsymbol{x} \in \Omega.$$

Definition 10. *Let $\Omega \in \mathbb{R}^d$ be bounded, and define*

$$\mathcal{R} := \left\{ r > 0 : \begin{array}{l} \textit{there exists } \boldsymbol{x}_0 \in \Omega \textit{ such that } \Omega \textit{ is} \\ \textit{star-shaped with respect to } B_{r,|\cdot|}(\boldsymbol{x}_0) \end{array} \right\}.$$

Then we define

$$r^\star_{\max} := \sup \mathcal{R} \quad \textit{and call} \quad \gamma := \frac{\mathrm{diam}(\Omega)}{r^\star_{\max}}$$

the chunkiness parameter of Ω if $\mathcal{R} \neq \emptyset$.

Lemma 9 (Bramble–Hilbert Lemma [7, Lemma 4.3.8])**.** *Let $\Omega \in \mathbb{R}^d$ be open and bounded, $\boldsymbol{x}_0 \in \Omega$ and $r > 0$ such that Ω is the stared-shaped with respect to $B := B_{r,|\cdot|}(\boldsymbol{x}_0)$, and $r \geq \frac{1}{2} r^\star_{\max}$. Moreover, let $n \in \mathbb{N}_+$, $1 \leq p \leq \infty$ and denote by γ by the chunkiness parameter of Ω. Then there is a constant $C(n,d,\gamma) > 0$ such that for all $f \in W^{n,p}(\Omega)$*

$$|f - Q^n f|_{W^{k,p}(\Omega)} \leq C(n,d,\gamma) h^{n-k} |f|_{W^{n,p}(\Omega)} \quad \textit{for } k = 0,1,\ldots,n$$

where $Q^n f$ denotes the Taylor polynomial of order n of f averaged over B and $h = \mathrm{diam}(\Omega)$.

Proof of Theorem 6. Without loss of generalization, we prove the case for $\boldsymbol{m} = (1,1,\ldots,1) =: \boldsymbol{m}_*$.

Denote $E : W^{n,\infty}((0,1)^d) \to W^{n,\infty}(\mathbb{R}^d)$ be an extension operator [43] and set $\tilde{f} := Ef$ and C_E is the norm of the extension operator.

Define $p_{f,\boldsymbol{i}}$ as the average Taylor polynomial Definition 8 in $B_{\boldsymbol{i},K} := B_{\frac{1}{4K},|\cdot|}\left(\frac{8\boldsymbol{i}+3}{8K}\right)$ i.e.

$$p_{f,\boldsymbol{i}} := \int_{B_{\boldsymbol{i},K}} T^n_{\boldsymbol{y}} \tilde{f}(\boldsymbol{x}) b_{\frac{1}{4K}}(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y}. \tag{56}$$

Based on Lemma 8, $p_{f,\boldsymbol{i}}$ can be read as

$$p_{f,\boldsymbol{i}} = \sum_{|\boldsymbol{\alpha}|\le n-1} c_{f,\boldsymbol{i},\boldsymbol{\alpha}} \boldsymbol{x}^{\boldsymbol{\alpha}} \tag{57}$$

where

$$|c_{f,\boldsymbol{i},\boldsymbol{\alpha}}| \le C_2(n,d). \tag{58}$$

The reason to define average Taylor polynomial on $B_{\boldsymbol{i},K}$ is to use the Bramble–Hilbert Lemma 9 on

$$\Omega_{\boldsymbol{m}_*,\boldsymbol{i}} = B_{\frac{3}{8K},\|\cdot\|_{\ell_\infty}}\left(\frac{8\boldsymbol{i}+3}{8K}\right) = \prod_{j=1}^d \left[\frac{i_j}{K}, \frac{3+4i_j}{4K}\right].$$

Note that

$$\frac{1}{4K} \ge \frac{1}{2}\cdot\frac{3}{8K} = \frac{1}{2} r^\star_{\max}(\Omega_{\boldsymbol{m}_*,\boldsymbol{i}}),\ \gamma(\Omega_{\boldsymbol{m}_*,\boldsymbol{i}}) = \frac{\operatorname{diam}(\Omega_{\boldsymbol{m}_*,\boldsymbol{i}})}{r^\star_{\max}(\Omega_{\boldsymbol{m}_*,\boldsymbol{i}})} = 2\sqrt{d}.$$

Therefore we can apply the Bramble–Hilbert Lemma 9 and have

$$\begin{aligned} \|\tilde{f} - p_{f,\boldsymbol{i}}\|_{L^\infty(\Omega_{\boldsymbol{m}_*,\boldsymbol{i}})} &\le C_{BH}(n,d)K^{-n} \\ |\tilde{f} - p_{f,\boldsymbol{i}}|_{W^{1,\infty}(\Omega_{\boldsymbol{m}_*,\boldsymbol{i}})} &\le C_{BH}(n,d)K^{-(n-1)} \end{aligned} \tag{59}$$

where $C_{BH}(n,d) = |\{|\boldsymbol{\alpha}| = n\}| \frac{1}{d\int_0^1 x^{d-1}e^{-(1-x^2)^{-1}}\,\mathrm{d}x}\left(2+4\sqrt{d}\right)^d C_E$ by following the proof of Lemma [7, Lemma 4.3.8]. Therefore,

$$\|\tilde{f} - p_{f,\boldsymbol{i}}\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}_*,\boldsymbol{i}})} \le C_1(n,d)K^{-(n-1)}$$

where $C_1(n,d) = 2C_{BH}(n,d)$.

Now we construct a partition of unity that we use in this theorem. First of all, given any integer K, define $\{h_i\}_{i=0}^K$ from $\mathbb{R} \to \mathbb{R}$:

$$h_i(x) := h\left(4K\left(x - \frac{8i+3}{8K}\right)\right),\ h(x) := \begin{cases} 1, & |x| < \frac{3}{2} \\ 0, & |x| > 2 \\ 4-2|x|, & \frac{3}{2} \le |x| \le 2. \end{cases} \tag{60}$$

It is easy to check that $\{h_i\}_{i=0}^K$ is a partition of unity of $[0,1]$ and $h_i(x) = 1$ for $x \in \left[\frac{i}{K}, \frac{3+4i}{4K}\right]$. Hence we can define $h_{\boldsymbol{i}}(\boldsymbol{x})$ for $\boldsymbol{i} = (i_1, i_2, \ldots, i_d) \in \{0,1,\ldots,K\}^d$ and $\boldsymbol{x} = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$:

$$h_{\boldsymbol{i}}(\boldsymbol{x}) = \prod_{j=1}^d h_{i_j}(x_j), \tag{61}$$

and $\left\{h_{\boldsymbol{i}} : \boldsymbol{i} \in \{0,1,\ldots,K\}^d\right\}$ is a partition of unity of $[0,1]^d$ and $h_{\boldsymbol{i}}(\boldsymbol{x}) = 1$ for $\boldsymbol{x} \in \prod_{j=1}^d \left[\frac{i_j}{K}, \frac{3+4i_j}{4K}\right] = \Omega_{\boldsymbol{m}_*,\boldsymbol{i}}$ and $\boldsymbol{i} = (i_1, i_2, \ldots, i_d) \in \{0,1,\ldots,K\}^d$.

Furthermore,

$$\|h_{\boldsymbol{i}}(\tilde{f} - p_{f,\boldsymbol{i}})\|_{L^\infty(\Omega_{\boldsymbol{m}_*,\boldsymbol{i}})} \le \|\tilde{f} - p_{f,\boldsymbol{i}}\|_{L^\infty(\Omega_{\boldsymbol{m}_*,\boldsymbol{i}})} \le C_{BH}(n,d)K^{-n} \tag{62}$$

and

$$|h_{\boldsymbol{i}}(\tilde{f} - p_{f,\boldsymbol{i}})|_{W^{1,\infty}(\Omega_{\boldsymbol{m}_*,\boldsymbol{i}})} \le |\tilde{f} - p_{f,\boldsymbol{i}}|_{W^{1,\infty}(\Omega_{\boldsymbol{m}_*,\boldsymbol{i}})} \le C_{BH}(n,d)K^{-(n-1)} \tag{63}$$

which is due to $h_{\boldsymbol{i}} = 1$ on $\Omega_{\boldsymbol{m}_*,\boldsymbol{i}}$.

Then

$$\|h_{\boldsymbol{i}}(\tilde{f} - p_{f,\boldsymbol{i}})\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}_*,\boldsymbol{i}})} \le C_1(n,d)K^{-(n-1)}.$$

Finally,

$$\left\|f-\sum_{\boldsymbol{i}\in\{0,1,\ldots,K\}^d} h_{\boldsymbol{i}}p_{f,\boldsymbol{i}}\right\|_{W^{1,\infty}(\Omega_{m_*})} \le \max_{\boldsymbol{i}\in\{0,1,\ldots,K\}^d}\|h_{\boldsymbol{i}}(\tilde f-p_{f,\boldsymbol{i}})\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}_*,\boldsymbol{i}})}$$
$$\le C_1(n,d)K^{-(n-1)}, \tag{64}$$

which is due to $\cup_{\boldsymbol{i}\in\{0,1,\ldots,K\}^d}\Omega_{\boldsymbol{m}_*,\boldsymbol{i}}=\Omega_{m_*}$ and supp $h_{\boldsymbol{i}}\cap\Omega_{m_*}=\Omega_{\boldsymbol{m}_*,\boldsymbol{i}}$.

Similarly,

$$\left\|f-\sum_{\boldsymbol{i}\in\{0,1,\ldots,K\}^d} h_{\boldsymbol{i}}p_{f,\boldsymbol{i}}\right\|_{L^{\infty}(\Omega_{1,d})} \le C_1(n,d)K^{-n}. \tag{65}$$

Last of all,

$$\begin{aligned} f_{k,\boldsymbol{m}_*}(\boldsymbol{x}) &:= \sum_{\boldsymbol{i}\in\{0,1,\ldots,K\}^d} h_{\boldsymbol{i}}p_{f,\boldsymbol{i}} = \sum_{\boldsymbol{i}\in\{0,1,\ldots,K\}^d}\sum_{|\boldsymbol{\alpha}|\le n-1} h_{\boldsymbol{i}}c_{f,\boldsymbol{i},\boldsymbol{\alpha}}\boldsymbol{x}^{\boldsymbol{\alpha}} \\ &= \sum_{|\boldsymbol{\alpha}|\le n-1}\sum_{\boldsymbol{i}\in\{0,1,\ldots,K\}^d} h_{\boldsymbol{i}}c_{f,\boldsymbol{i},\boldsymbol{\alpha}}\boldsymbol{x}^{\boldsymbol{\alpha}} \\ &=: \sum_{|\boldsymbol{\alpha}|\le n-1} g_{f,\boldsymbol{\alpha},\boldsymbol{m}_*}(\boldsymbol{x})\boldsymbol{x}^{\boldsymbol{\alpha}} \end{aligned} \tag{66}$$

with $|g_{f,\boldsymbol{\alpha},\boldsymbol{m}_*}(\boldsymbol{x})|\le C_2(n,d)$ for $x\in\Omega_{\boldsymbol{m}_*}$. Note that $g_{f,\boldsymbol{\alpha},\boldsymbol{m}_*}(\boldsymbol{x})$ is a step function from $\Omega_{\boldsymbol{m}_*}\to\mathbb{R}$:

$$g_{f,\boldsymbol{\alpha},\boldsymbol{m}_*}(\boldsymbol{x}) = c_{f,\boldsymbol{i},\boldsymbol{\alpha}} \tag{67}$$

for $\boldsymbol{x}\in\prod_{j=1}^d\left[\frac{i_j}{K},\frac{3+4i_j}{4K}\right]$ and $\boldsymbol{i}=(i_1,i_2,\ldots,i_d)$ since $h_{\boldsymbol{i}}(\boldsymbol{x})=0$ for $\boldsymbol{x}\in\Omega_{\boldsymbol{m}_*}\setminus\prod_{j=1}^d\left[\frac{i_j}{K},\frac{3+4i_j}{4K}\right]$ and $h_{\boldsymbol{i}}(\boldsymbol{x})=1$ for $\boldsymbol{x}\in\prod_{j=1}^d\left[\frac{i_j}{K},\frac{3+4i_j}{4K}\right]$. □

7.2.3 Approximation of functions in $W^{n,\infty}$ with $W^{1,\infty}$ norm by ReLU neural networks in the whole space except a small set

Theorem 7. *For any $f\in W^{n,\infty}((0,1)^d)$ with $\|f\|_{W^{n,\infty}((0,1)^d)}\le 1$, any $N,L\in\mathbb{N}_+$, and $\boldsymbol{m}=(m_1,m_2,\ldots,m_d)\in\{1,2\}^d$, there is a neural network $\psi_{\boldsymbol{m}}$ with the width $25n^{d+1}(N+1)\log_2(8N)$ and depth $27n^2(L+2)\log_2(4L)$ such that*

$$\begin{aligned} \|f(\boldsymbol{x})-\psi_{\boldsymbol{m}}(\boldsymbol{x})\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}})} &\le C_6(n,d)N^{-2(n-1)/d}L^{-2(n-1)/d} \\ \|f(\boldsymbol{x})-\psi_{\boldsymbol{m}}(\boldsymbol{x})\|_{L^{\infty}(\Omega_{\boldsymbol{m}})} &\le C_6(n,d)N^{-2n/d}L^{-2n/d}, \end{aligned} \tag{68}$$

where C_6 is the constant independent with N,L.

Proof. Without loss of the generalization, we consider the case for $\boldsymbol{m}_*=(1,1,\ldots,1)$. Due to Theorem 6 and setting $K=\lfloor N^{1/d}\rfloor^2\lfloor L^{2/d}\rfloor$, we have

$$\begin{aligned} \|f-f_{K,\boldsymbol{m}_*}\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}_*})} &\le C_1(n,d)K^{-(n-1)} \le C_1(n,d)N^{-2(n-1)/d}L^{-2(n-1)/d} \\ \|f-f_{K,\boldsymbol{m}_*}\|_{L^{\infty}(\Omega_{\boldsymbol{m}_*})} &\le C_1(n,d)K^{-n} \le C_1(n,d)N^{-2n/d}L^{-2n/d}, \end{aligned} \tag{69}$$

where $f_{K,\boldsymbol{m}_*}=\sum_{|\boldsymbol{\alpha}|\le n-1}g_{f,\boldsymbol{\alpha},\boldsymbol{m}_*}(\boldsymbol{x})\boldsymbol{x}^{\boldsymbol{\alpha}}$ for $x\in\Omega_{\boldsymbol{m}_*}$. Note that $g_{f,\boldsymbol{\alpha},\boldsymbol{m}_*}(\boldsymbol{x})$ is a constant function for $\boldsymbol{x}\in\prod_{j=1}^d\left[\frac{i_j}{K},\frac{3+4i_j}{4K}\right]$ and $\boldsymbol{i}=(i_1,i_2,\ldots,i_d)\in\{0,1,\ldots,K-1\}^d$. The remaining part is to approximate $f_{K,\boldsymbol{m}_*}$ by neural networks.

The way to approximate $g_{f,\boldsymbol{\alpha},\boldsymbol{m}_*}(\boldsymbol{x})$ is similar with [23, Theorem 3.1]. First of all, due to Proposition 2, there is a neural network $\phi_1(x)$ with the width $4N+5$ and depth $4L+4$ such that

$$\phi(x)=k, x\in\left[\frac{k}{K},\frac{k+1}{K}-\frac{1}{4K}\right],\ k=0,1,\ldots,K-1. \tag{70}$$

Note that we choose $\delta = \frac{1}{4K} \le \frac{1}{3K}$ in Proposition 2. Then define

$$\boldsymbol{\phi}_2(\boldsymbol{x}) = \left[\frac{\phi_1(x_1)}{K}, \frac{\phi_1(x_2)}{K}, \ldots, \frac{\phi_1(x_d)}{K}\right]^{\intercal}.$$

For each $p = 0, 1, \ldots, K^d - 1$, there is a bijection

$$\boldsymbol{\eta}(p) = [\eta_1, \eta_2, \ldots, \eta_d] \in \{0, 1, \ldots, K-1\}^d$$

such that $\sum_{j=1}^d \eta_j K^{j-1} = p$. Then define

$$\xi_{\boldsymbol{\alpha},p} = \frac{g_{f,\boldsymbol{\alpha},\boldsymbol{m}_*}\left(\frac{\boldsymbol{\eta}(p)}{K}\right) + C_2(n,d)}{2C_2(n,d)} \in [0,1],$$

where $C_2(n,d)$ is the bounded of $g_{f,\boldsymbol{\alpha},\boldsymbol{m}_*}$ defined in Theorem 6. Therefore, based on Proposition 3, there is a neural network $\tilde{\phi}_{\boldsymbol{\alpha}}(x)$ with the width $16n(N+1)\log_2(8N)$ and depth $(5L+2)\log_2(4L)$ such that $|\tilde{\phi}_{\boldsymbol{\alpha}}(p) - \xi_{\boldsymbol{\alpha},p}| \le N^{-2n}L^{-2n}$ for $p = 0, 1, \ldots K^d - 1$. Denote

$$\phi_{\boldsymbol{\alpha}}(\boldsymbol{x}) = 2C_2(n,d)\tilde{\phi}_{\boldsymbol{\alpha}}\left(\sum_{j=1}^d x_j K^j\right) - C_2(n,d)$$

and obtain that

$$\left|\phi_{\boldsymbol{\alpha}}\left(\frac{\boldsymbol{\eta}(p)}{K}\right) - g_{f,\boldsymbol{\alpha},\boldsymbol{m}_*}\left(\frac{\boldsymbol{\eta}(p)}{K}\right)\right| = 2C_2(n,d)|\tilde{\phi}_{\boldsymbol{\alpha}}(p) - \xi_{\boldsymbol{\alpha},p}| \le 2C_2(n,d)N^{-2n}L^{-2n}.$$

Then we obtain that

$$\begin{aligned}\|\phi_{\boldsymbol{\alpha}}(\boldsymbol{\phi}_2(\boldsymbol{x})) - g_{f,\boldsymbol{\alpha},\boldsymbol{m}_*}(\boldsymbol{x})\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}_*})} =& \|\phi_{\boldsymbol{\alpha}}(\boldsymbol{\phi}_2(\boldsymbol{x})) - g_{f,\boldsymbol{\alpha},\boldsymbol{m}_*}(\boldsymbol{x})\|_{L^\infty(\Omega_{\boldsymbol{m}_*})} \\ \le& 2C_2(n,d)N^{-2n}L^{-2n} \end{aligned} \tag{71}$$

which is due to $\phi_{\boldsymbol{\alpha}}(\boldsymbol{\phi}_2(\boldsymbol{x})) - g_{f,\boldsymbol{\alpha},\boldsymbol{m}_*}(\boldsymbol{x})$ is a step function, and the first order weak derivative is 0 in $\Omega_{\boldsymbol{m}_*}$.

Due to Proposition 6, there is a neural network $\phi_{3,\boldsymbol{\alpha}}$ with the width $9(N+1)+n-1$ and depth $14n^2L$ such that $\|\phi_{3,\boldsymbol{\alpha}}\|_{W^{1,\infty}((0,1)^d)} \le 18$ and

$$\|\phi_{3,\boldsymbol{\alpha}}(\boldsymbol{x}) - \boldsymbol{x}^{\boldsymbol{\alpha}}\|_{W^{1,\infty}((0,1)^d)} \le 10n(N+1)^{-7nL}. \tag{72}$$

Due to Proposition 4, there is a neural network ϕ_4 with the width $15(N+1)$ and depth $4n(L+1)$ such that $\|\phi_4\|_{W^{1,\infty}(-C_3,C_3)^2} \le 12(C_2(n,d))^2$ and

$$\|\phi_4(x,y) - xy\|_{W^{1,\infty}((-C_3,C_3)^2)} \le 6(C_2(n,d))^2(N+1)^{-2n(L+1)}. \tag{73}$$

where $C_3(n,d) = \max\{3C_2(n,d), 18\}$.

Now we define the neural network $\phi_{\boldsymbol{m}_*}(\boldsymbol{x})$ to approximate $f_{K,\boldsymbol{m}_*}(\boldsymbol{x})$ in $\Omega_{\boldsymbol{m}_*}$:

$$\psi_{\boldsymbol{m}_*}(\boldsymbol{x}) = \sum_{|\boldsymbol{\alpha}| \le n-1} \phi_4\left[\phi_{\boldsymbol{\alpha}}(\boldsymbol{\phi}_2(\boldsymbol{x})), \phi_{3,\boldsymbol{\alpha}}(\boldsymbol{x})\right]. \tag{74}$$

The remaining question is to find the error $\mathcal{E}$:

$$
\begin{aligned}
\mathcal{E} :=& \left\| \sum_{|\boldsymbol{\alpha}|\le n-1} \phi_4\left[\phi_{\boldsymbol{\alpha}}(\boldsymbol{\phi}_2(\boldsymbol{x})), \phi_{3,\boldsymbol{\alpha}}(\boldsymbol{x})\right] - f_{K,\boldsymbol{m}_*}(\boldsymbol{x}) \right\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}_*})} \\
\le& \sum_{|\boldsymbol{\alpha}|\le n-1} \left\| \phi_4\left[\phi_{\boldsymbol{\alpha}}(\boldsymbol{\phi}_2(\boldsymbol{x})), \phi_{3,\boldsymbol{\alpha}}(\boldsymbol{x})\right] - g_{f,\boldsymbol{\alpha},\boldsymbol{m}_*}(\boldsymbol{x})\boldsymbol{x}^{\boldsymbol{\alpha}} \right\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}_*})} \\
\le& \underbrace{\sum_{|\boldsymbol{\alpha}|\le n-1} \left\| \phi_4\left[\phi_{\boldsymbol{\alpha}}(\boldsymbol{\phi}_2(\boldsymbol{x})), \phi_{3,\boldsymbol{\alpha}}(\boldsymbol{x})\right] - \phi_{\boldsymbol{\alpha}}(\boldsymbol{\phi}_2(\boldsymbol{x}))\phi_{3,\boldsymbol{\alpha}}(\boldsymbol{x}) \right\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}_*})}}_{=:\mathcal{E}_1} \\
&+ \underbrace{\sum_{|\boldsymbol{\alpha}|\le n-1} \left\| \phi_{\boldsymbol{\alpha}}(\boldsymbol{\phi}_2(\boldsymbol{x}))\phi_{3,\boldsymbol{\alpha}}(\boldsymbol{x}) - g_{f,\boldsymbol{\alpha},\boldsymbol{m}_*}(\boldsymbol{x})\phi_{3,\boldsymbol{\alpha}}(\boldsymbol{x}) \right\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}_*})}}_{=:\mathcal{E}_2} \\
&+ \underbrace{\sum_{|\boldsymbol{\alpha}|\le n-1} \left\| g_{f,\boldsymbol{\alpha},\boldsymbol{m}_*}(\boldsymbol{x})\phi_{3,\boldsymbol{\alpha}}(\boldsymbol{x}) - g_{f,\boldsymbol{\alpha},\boldsymbol{m}_*}(\boldsymbol{x})\boldsymbol{x}^{\boldsymbol{\alpha}} \right\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}_*})}}_{=:\mathcal{E}_3}.
\end{aligned} \tag{75}
$$

As for $\mathcal{E}_1$, due to Lemma 3, we have

$$
\begin{aligned}
\mathcal{E}_1 \le& \sum_{|\boldsymbol{\alpha}|\le n-1} 2\sqrt{d}\max\Big\{ \|\phi_4(x,y)-xy\|_{L^\infty((-C_3,C_3)^2)}, \|\phi_4(x,y)-xy\|_{W^{1,\infty}((-C_3,C_3)^2)} \\
&\cdot \max\{\|\phi_{\boldsymbol{\alpha}}(\boldsymbol{\phi}_2(\boldsymbol{x}))\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}_*})}, \|\phi_{3,\boldsymbol{\alpha}}(\boldsymbol{x})\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}_*})}\}\Big\} \\
\le& \sum_{|\boldsymbol{\alpha}|\le n-1} 2\sqrt{d}\max\left\{ \|\phi_4(x,y)-xy\|_{L^\infty((-C_3,C_3)^2)}, C_3(n,d)\|\phi_4(x,y)-xy\|_{W^{1,\infty}((-C_3,C_3)^2)}\right\} \\
\le& \sum_{|\boldsymbol{\alpha}|\le n-1} 12\sqrt{d}\left[C_3(n,d)+1\right](C_2(n,d))^2(N+1)^{-2n(L+1)} \\
\le& C_4(n,d)(N+1)^{-2n(L+1)}
\end{aligned} \tag{76}
$$

where $C_4(n,d) = 12\sqrt{d}n^d\left[C_3(n,d)+1\right](C_2(n,d))^2$.

As for $\mathcal{E}_2$, due to Lemma 4, we have

$$
\begin{aligned}
\mathcal{E}_2 \le& \sum_{|\boldsymbol{\alpha}|\le n-1} 2\left\|\phi_{\boldsymbol{\alpha}}(\boldsymbol{\phi}_2(\boldsymbol{x})) - g_{f,\boldsymbol{\alpha},\boldsymbol{m}_*}(\boldsymbol{x})\right\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}_*})} \cdot \|\phi_{3,\boldsymbol{\alpha}}(\boldsymbol{x})\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}_*})} \\
\le& 72n^d C_2(n,d)N^{-2n}L^{-2n}.
\end{aligned} \tag{77}
$$

The estimation of $\mathcal{E}_3$ is similar with that of $\mathcal{E}_2$ which is

$$
\begin{aligned}
\mathcal{E}_3 \le& \sum_{|\boldsymbol{\alpha}|\le n-1} \|g_{f,\boldsymbol{\alpha},\boldsymbol{m}_*}\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}_*})} \cdot \|\phi_{3,\boldsymbol{\alpha}}(\boldsymbol{x}) - \boldsymbol{x}^{\boldsymbol{\alpha}}\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}_*})} \\
\le& 10n^d C_2(n,d)n(N+1)^{-7nL}.
\end{aligned} \tag{78}
$$

Therefore, using

$$
(N+1)^{-7nL} \le (N+1)^{-2n(L+1)} \le N^{-2n}L^{-2n}
$$

the total error is

$$
\mathcal{E} \le \mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3 \le C_5(n,d)K^{-2n}L^{-2n}, \tag{79}
$$

where $C_5(n,d) = C_4(n,d) + 72n^d C_2(n,d) + 10n^d C_2(n,d)n$.

At last, we finish the proof by estimating the network's width and depth, implementing $\psi_{\boldsymbol{m}_*}(\boldsymbol{x})$. From Eq. (74), we know that $\psi_{\boldsymbol{m}_*}(\boldsymbol{x})$ consists of the following subnetworks:

1. $\phi_{3,\boldsymbol{\alpha}}(\boldsymbol{x})$ with the width $9(N+1)+n-1$ and depth $14n^2L$.

2. $\phi_2(\boldsymbol{x})$ with the width $4N+5$ and depth $4L+4$.

3. $\phi_{\boldsymbol{\alpha}}$ with the width $16n(N+1)\log_2(8N)$ and depth $(5L+2)\log_2(4L)$.

4. $\phi_4(x,y)$ with the width $15(N+1)$ and depth $4n(L+1)$.

Therefore $\phi(\boldsymbol{x})$ is a neural network with the width $25n^{d+1}(N+1)\log_2(8N)$ and depth $27n^2(L+2)\log_2(4L)$.

Combining Eqs. (69) and (79), we have that there is a neural network $\psi_{\boldsymbol{m}_*}$ with the width $25n^{d+1}(N+1)\log_2(8N)$ and depth $27n^2(L+2)\log_2(4L)$ such that

$$\begin{aligned}\|f(\boldsymbol{x})-\psi_{\boldsymbol{m}_*}(\boldsymbol{x})\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}_*})} &\leq C_6(n,d)N^{-2(n-1)/d}L^{-2(n-1)/d}\\ \|f(\boldsymbol{x})-\psi_{\boldsymbol{m}_*}(\boldsymbol{x})\|_{L^{\infty}(\Omega_{\boldsymbol{m}_*})} &\leq C_6(n,d)N^{-2n/d}L^{-2n/d},\end{aligned}\tag{80}$$

where $C_6=C_1+C_5$ is the constant independent with N,L.

Similarly, we can construct a neural network $\psi_{\boldsymbol{m}}$ with the width $25n^{d+1}(N+1)\log_2(8N)$ and depth $27n^2(L+2)\log_2(4L)$ which can approximate f on $\Omega_{\boldsymbol{m}}$ with same order of Eq. (80). $\square$

7.2.4 Proof of Theorem 3

Now we can prove Theorem 3 based on Theorem 7 and Proposition 1.

Proof of Theorem 3. Based on Theorem 7, there is a sequence of the neural network $\{\psi_{\boldsymbol{m}}(\boldsymbol{x})\}_{\boldsymbol{m}\in\{1,2\}^d}$ such that

$$\begin{aligned}\|f(\boldsymbol{x})-\psi_{\boldsymbol{m}}(\boldsymbol{x})\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}})} &\leq C_6(n,d)N^{-2(n-1)/d}L^{-2(n-1)/d}\\ \|f(\boldsymbol{x})-\psi_{\boldsymbol{m}}(\boldsymbol{x})\|_{L^{\infty}(\Omega_{\boldsymbol{m}})} &\leq C_6(n,d)N^{-2n/d}L^{-2n/d},\end{aligned}\tag{81}$$

where $C_6=C_1+C_5$ is the constant independent with N,L, and each $\psi_{\boldsymbol{m}}$ is a neural network with the width $25n^{d+1}(N+1)\log_2(8N)$ and depth $27n^2(L+2)\log_2(4L)$. According to Proposition 1, there is a sequence of the neural network $\{\phi_{\boldsymbol{m}}(\boldsymbol{x})\}_{\boldsymbol{m}\in\{1,2\}^d}$ such that

$$\|\phi_{\boldsymbol{m}}(\boldsymbol{x})-g_{\boldsymbol{m}}(\boldsymbol{x})\|_{W^{1,\infty}((0,1)^d)}\leq 50d^{\frac{5}{2}}(N+1)^{-4dnL},$$

where $\{g_{\boldsymbol{m}}\}_{\boldsymbol{m}\in\{1,2\}^d}$ is defined in Definition 6 with $\sum_{\boldsymbol{m}\in\{1,2\}^d}g_{\boldsymbol{m}}(\boldsymbol{x})=1$ and $\operatorname{supp} g_{\boldsymbol{m}}\cap[0,1]^d=\Omega_{\boldsymbol{m}}$. For each $\phi_{\boldsymbol{m}}$, it is a neural network with the width smaller than $(9+d)(N+1)+d-1$ and depth smaller than $15d(d-1)nL$.

Due to Proposition 4, there is a neural network $\widehat{\phi}$ with the width $15(N+1)$ and depth $14n^2L$ such that $\|\widehat{\phi}\|_{W^{1,\infty}(-C_7,C_7)^2}\leq 12(C_7(n,d))^2$ and

$$\left\|\widehat{\phi}(x,y)-xy\right\|_{W^{1,\infty}(-C_7,C_7)^2}\leq 6(C_7)^2(N+1)^{-7n(L+1)},\tag{82}$$

where $C_7=C_6+50d^{\frac{5}{2}}+1$.

Now we define

$$\phi(\boldsymbol{x})=\sum_{\boldsymbol{m}\in\{1,2\}^d}\widehat{\phi}(\phi_{\boldsymbol{m}}(\boldsymbol{x}),\psi_{\boldsymbol{m}}(\boldsymbol{x})).\tag{83}$$

Note that

$$\begin{aligned}\mathcal{R}:=&\|f(\boldsymbol{x})-\phi(\boldsymbol{x})\|_{W^{1,\infty}((0,1)^d)}=\left\|\sum_{\boldsymbol{m}\in\{1,2\}^d}g_{\boldsymbol{m}}\cdot f(\boldsymbol{x})-\phi(\boldsymbol{x})\right\|_{W^{1,\infty}((0,1)^d)}\\ \leq&\left\|\sum_{\boldsymbol{m}\in\{1,2\}^d}[g_{\boldsymbol{m}}\cdot f(\boldsymbol{x})-\phi_{\boldsymbol{m}}(\boldsymbol{x})\cdot\psi_{\boldsymbol{m}}(\boldsymbol{x})]\right\|_{W^{1,\infty}((0,1)^d)}\\ &+\left\|\sum_{\boldsymbol{m}\in\{1,2\}^d}\left[\phi_{\boldsymbol{m}}(\boldsymbol{x})\cdot\psi_{\boldsymbol{m}}(\boldsymbol{x})-\widehat{\phi}(\phi_{\boldsymbol{m}}(\boldsymbol{x}),\psi_{\boldsymbol{m}}(\boldsymbol{x}))\right]\right\|_{W^{1,\infty}((0,1)^d)}.\end{aligned}\tag{84}$$

As for the first part,

$$
\begin{aligned}
&\left\|\sum_{\boldsymbol{m}\in\{1,2\}^d}[g_{\boldsymbol{m}}\cdot f(\boldsymbol{x})-\phi_{\boldsymbol{m}}(\boldsymbol{x})\cdot\psi_{\boldsymbol{m}}(\boldsymbol{x})]\right\|_{W^{1,\infty}((0,1)^d)}\\
\le&\sum_{\boldsymbol{m}\in\{1,2\}^d}\|g_{\boldsymbol{m}}\cdot f(\boldsymbol{x})-\phi_{\boldsymbol{m}}(\boldsymbol{x})\cdot\psi_{\boldsymbol{m}}(\boldsymbol{x})\|_{W^{1,\infty}((0,1)^d)}\\
\le&\sum_{\boldsymbol{m}\in\{1,2\}^d}\left[\|(g_{\boldsymbol{m}}-\phi_{\boldsymbol{m}}(\boldsymbol{x}))\cdot f(\boldsymbol{x})\|_{W^{1,\infty}((0,1)^d)}+\|(f_{\boldsymbol{m}}-\psi_{\boldsymbol{m}}(\boldsymbol{x}))\cdot\phi_{\boldsymbol{m}}(\boldsymbol{x})\|_{W^{1,\infty}((0,1)^d)}\right]\\
=&\sum_{\boldsymbol{m}\in\{1,2\}^d}\left[\|(g_{\boldsymbol{m}}-\phi_{\boldsymbol{m}}(\boldsymbol{x}))\cdot f(\boldsymbol{x})\|_{W^{1,\infty}((0,1)^d)}+\|(f_{\boldsymbol{m}}-\psi_{\boldsymbol{m}}(\boldsymbol{x}))\cdot\phi_{\boldsymbol{m}}(\boldsymbol{x})\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}})}\right],
\end{aligned}
\tag{85}
$$

where the last equality is due to Lemma 7. Based on Lemma 4 and $\|f\|_{W^{1,\infty}((0,1)^d)}\le 1$, we have

$$
\|(g_{\boldsymbol{m}}-\phi_{\boldsymbol{m}}(\boldsymbol{x}))\cdot f(\boldsymbol{x})\|_{W^{1,\infty}((0,1)^d)}\le\|(g_{\boldsymbol{m}}-\phi_{\boldsymbol{m}}(\boldsymbol{x}))\|_{W^{1,\infty}((0,1)^d)}\le 50d^{\frac{5}{2}}(N+1)^{-4dnL}.
\tag{86}
$$

And

$$
\begin{aligned}
&\|(f_{\boldsymbol{m}}-\psi_{\boldsymbol{m}}(\boldsymbol{x}))\cdot\phi_{\boldsymbol{m}}(\boldsymbol{x})\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}})}\\
\le&\|(f_{\boldsymbol{m}}-\psi_{\boldsymbol{m}}(\boldsymbol{x}))\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}})}\cdot\|\phi_{\boldsymbol{m}}\|_{L^{\infty}(\Omega_{\boldsymbol{m}})}+\|(f_{\boldsymbol{m}}-\psi_{\boldsymbol{m}}(\boldsymbol{x}))\|_{L^{\infty}(\Omega_{\boldsymbol{m}})}\cdot\|\phi_{\boldsymbol{m}}\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}})}\\
\le&C_6(n,d)N^{-2(n-1)/d}L^{-2(n-1)/d}\cdot\left(1+50d^{\frac{5}{2}}\right)+C_6(n,d)N^{-2n/d}L^{-2n/d}\cdot 54d^{\frac{5}{2}}\lfloor N^{1/d}\rfloor^2\lfloor L^{2/d}\rfloor\\
\le&C_7(n,d)N^{-2(n-1)/d}L^{-2(n-1)/d},
\end{aligned}
\tag{87}
$$

where the second inequality is due to

$$
\begin{aligned}
&\|\phi_{\boldsymbol{m}}\|_{L^{\infty}(\Omega_{\boldsymbol{m}})}\le\|\phi_{\boldsymbol{m}}\|_{L^{\infty}([0,1]^d)}\le\|g_{\boldsymbol{m}}\|_{L^{\infty}([0,1]^d)}+\|\phi_{\boldsymbol{m}}-g_{\boldsymbol{m}}\|_{L^{\infty}([0,1]^d)}\le 1+50d^{\frac{5}{2}}\\
&\|\phi_{\boldsymbol{m}}\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}})}\le\|\phi_{\boldsymbol{m}}\|_{W^{1,\infty}([0,1]^d)}\le\|g_{\boldsymbol{m}}\|_{W^{1,\infty}([0,1]^d)}+\|\phi_{\boldsymbol{m}}-g_{\boldsymbol{m}}\|_{W^{1,\infty}([0,1]^d)}\\
&\le 4\lfloor N^{1/d}\rfloor^2\lfloor L^{2/d}\rfloor+50d^{\frac{5}{2}}.
\end{aligned}
\tag{88}
$$

Therefore

$$
\left\|\sum_{\boldsymbol{m}\in\{1,2\}^d}[g_{\boldsymbol{m}}\cdot f(\boldsymbol{x})-\phi_{\boldsymbol{m}}(\boldsymbol{x})\cdot\psi_{\boldsymbol{m}}(\boldsymbol{x})]\right\|_{W^{1,\infty}((0,1)^d)}\le 2^d(C_7(n,d)+50d^{\frac{5}{2}})N^{-2(n-1)/d}L^{-2(n-1)/d}
\tag{89}
$$

due to $(N+1)^{-4dnL}\le N^{-2n}L^{-2n}$.

For the second part, due to Lemma 7, we have

$$
\begin{aligned}
&\left\|\sum_{\boldsymbol{m}\in\{1,2\}^d}\left[\phi_{\boldsymbol{m}}(\boldsymbol{x})\cdot\psi_{\boldsymbol{m}}(\boldsymbol{x})-\widehat{\phi}(\phi_{\boldsymbol{m}}(\boldsymbol{x}),\psi_{\boldsymbol{m}}(\boldsymbol{x}))\right]\right\|_{W^{1,\infty}((0,1)^d)}\\
\le&\sum_{\boldsymbol{m}\in\{1,2\}^d}\left\|\phi_{\boldsymbol{m}}(\boldsymbol{x})\cdot\psi_{\boldsymbol{m}}(\boldsymbol{x})-\widehat{\phi}(\phi_{\boldsymbol{m}}(\boldsymbol{x}),\psi_{\boldsymbol{m}}(\boldsymbol{x}))\right\|_{W^{1,\infty}((0,1)^d)}\\
=&\sum_{\boldsymbol{m}\in\{1,2\}^d}\left\|\phi_{\boldsymbol{m}}(\boldsymbol{x})\cdot\psi_{\boldsymbol{m}}(\boldsymbol{x})-\widehat{\phi}(\phi_{\boldsymbol{m}}(\boldsymbol{x}),\psi_{\boldsymbol{m}}(\boldsymbol{x}))\right\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}})}.
\end{aligned}
\tag{90}
$$

Similarly with the estimation of $\mathcal{E}_1$ (76), we have that

$$
\begin{aligned}
&\left\|\phi_{\boldsymbol{m}}(\boldsymbol{x})\cdot\psi_{\boldsymbol{m}}(\boldsymbol{x})-\widehat{\phi}(\phi_{\boldsymbol{m}}(\boldsymbol{x}),\psi_{\boldsymbol{m}}(\boldsymbol{x}))\right\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}})}\\
\le&C_8(n,d)(N+1)^{-7n(L+1)}\le C_8(n,d)N^{-2(n-1)/d}L^{-2(n-1)/d}.
\end{aligned}
\tag{91}
$$

Combining (89) and (91), we have that there is a σ_1-NN ϕ 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

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

where C_9 is the constant independent with N, L.

□

The method proposed in [28, 23, 39, 38, 37] may not be applied to prove Theorems 3. These works approximate the target function f using a deep neural network ϕ in the unit cube except for an arbitrarily small region Ω_δ, as per [36, Lemma 2.2]. Since $\|\phi\|_{L^\infty(\Omega)}$ can be bounded and is independent of the size of $\Omega\delta$, $\|f-\phi\|_{L^p(\Omega)}$ can be well estimated across the entire space for $p\in[1,+\infty)$. For approximations measured in the $L^\infty(\Omega)$ norm, [28] translates the deep neural network ϕ, while [39] constructs different neural networks in the unit cube away from various negligible regions. Both methods aim to find neural networks $\{\phi_i(\boldsymbol{x})\}i=1^N$ that approximate the target function f well in different regions. They then observe that the middle value of $\{\phi_i(\boldsymbol{x})\}_{i=1}^N$ is close to $f(\boldsymbol{x})$ for all $\boldsymbol{x}^*\in\Omega$, and construct the middle-value function using a ReLU neural network. However, these methods may not be generalized to prove the theorems presented in this paper.

Neither of the methods previously proposed can be applied to the approximation measured in Sobolev space. In the first method, $\|\phi\|_{W^{1,\infty}(\Omega)}$ depends on the length of $\Omega\delta$, and the derivative is substantial in the negligible region, as shown in [36, Lemma 2.2]. Thus, $\|f-\phi\|_{W^{1,p}(\Omega)}$ will be excessively large. In the second method, median value functions can only identify the median values, not the median values of functions and their derivatives simultaneously. In this paper, we overcome this difficulty using a partition of unity. We construct a partition of unity of Ω and approximate them using ReLU DNNs denoted as $\{\phi_{\boldsymbol{m}}\}_{\boldsymbol{m}\in\{1,2\}^d}$. For each $\phi\boldsymbol{m}$, its support set is the unit cube away from a small region, and we can construct a deep neural network $\psi_{\boldsymbol{m}}$ that approximates the target function f well on supp $\phi_{\boldsymbol{m}}$. We then combine $\{\phi_{\boldsymbol{m}}\}_{\boldsymbol{m}\in\{1,2\}^d}$ and $\{\psi_{\boldsymbol{m}}\}_{\boldsymbol{m}\in\{1,2\}^d}$ to obtain a deep neural network that can approximate the target function f well across the entire space. This approach resolves the issue of simultaneous approximation of both functions and their derivatives in Sobolev spaces.

7.3 Proofs of Corollaries 1 and 2

7.3.1 Preliminaries

First, we list a few basic lemmas of σ_2 neural networks repeatedly applied in our main analysis.

Lemma 10 ([23, Lemma 3.7])**.** *The following basic lemmas of σ_2 neural networks hold:*

(i) σ_1 neural networks are σ_2 neural networks.

(ii) Any identity map in $\mathbb{R}^d$ can be realized exactly by a σ_2 neural network with one hidden layer and $2d$ neurons.

(iii) $f(x)=x^2$ can be realized exactly by a σ_2 neural network with one hidden layer and two neurons.

(iv) $f(x,y)=xy=\frac{(x+y)^2-(x-y)^2}{4}$ can be realized exactly by a σ_2 neural network with one hidden layer and four neurons.

(v) Assume $\boldsymbol{x}^{\boldsymbol{\alpha}}=x_1^{\alpha_1}x_2^{\alpha_2}\cdots x_d^{\alpha_d}$ for $\boldsymbol{\alpha}\in\mathbb{N}^d$. For any $N,L\in\mathbb{N}^+$ such that $NL+2^{\lfloor\log_2 N\rfloor}\ge|\boldsymbol{\alpha}|$, there exists a σ_2 neural network $\phi(\boldsymbol{x})$ with the width $4N+2d$ and depth $L+\lceil\log_2 N\rceil$ such that

$$\phi(\boldsymbol{x})=\boldsymbol{x}^{\boldsymbol{\alpha}}$$

for any $\boldsymbol{x}\in\mathbb{R}^d$.

(vi) Assume $P(\boldsymbol{x})=\sum_{j=1}^J c_j\boldsymbol{x}^{\boldsymbol{\alpha}_j}$ for $\boldsymbol{\alpha}_j\in\mathbb{N}^d$. For any $N,L,a,b\in\mathbb{N}^+$such that $ab\ge J$ and $(L-2b-b\log_2 N)N\ge b\max_j|\boldsymbol{\alpha}_j|$, there exists a σ_2 neural network $\phi(\boldsymbol{x})$ with the width $4Na+2d+2$ and depth L such that

$$\phi(\boldsymbol{x})=P(\boldsymbol{x})\text{ for any }\boldsymbol{x}\in\mathbb{R}^d.$$

Next, we define a function which will be repeatly used in the proof of Corollary 1 in this section.

Definition 11. *Define $s(x)$ from $\mathbb{R} \to [0,1]$ as*

$$s(x) := \begin{cases} 2x^2, & x \in \left[0, \frac{1}{2}\right] \\ -2(x-1)^2+1, & x \in \left[\frac{1}{2}, 1\right] \\ 1, & x \in [1,2] \\ -2(x-2)^2+1, & x \in \left[2, \frac{5}{2}\right] \\ 2(x-3)^2, & x \in \left[\frac{5}{2}, 3\right] \\ 0, & \textit{Otherwise}. \end{cases} \tag{92}$$

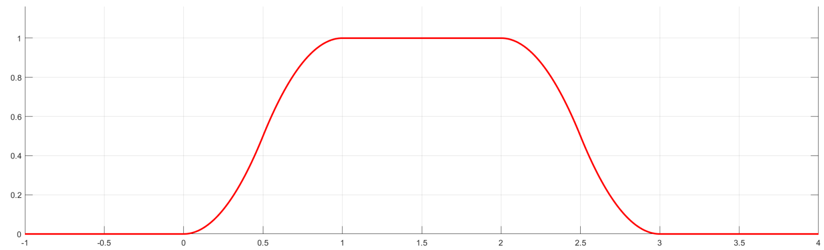

Figure 4: $s(x)$ in $\mathbb{R}$.

Definition 12. *Given $K \in \mathbb{N}_+$, then we define two functions in $\mathbb{R}$:*

$$s_1(x) = \sum_{i=0}^{K} s\left(4Kx+1-4i\right), \ s_2(x) = s_1\left(x + \frac{1}{2K}\right). \tag{93}$$

Then for any $\boldsymbol{m} = (m_1, m_2, \dots, m_d) \in \{1,2\}^d$, we define

$$s_{\boldsymbol{m}}(\boldsymbol{x}) := \prod_{j=1}^{d} s_{m_j}(x_j) \tag{94}$$

for any $\boldsymbol{x} = (x_1, x_2, \dots, x_d) \in \mathbb{R}^d$.

Proposition 8. *Given $N, L, d \in \mathbb{N}_+$ with $NL + 2^{\lfloor \log_2 N \rfloor} \geq d$ and $L \geq \lceil \log_2 N \rceil$, and setting $K = \lfloor N^{1/d} \rfloor \lfloor L^{2/d} \rfloor$, $\{s_{\boldsymbol{m}}(\boldsymbol{x})\}_{\boldsymbol{m} \in \{1,2\}^d}$ defined in Definition 12 satisfies:*

(i): $\|s_{\boldsymbol{m}}(\boldsymbol{x})\|_{L^\infty((0,1)^d)} \leq 1$, $\|s_{\boldsymbol{m}}(\boldsymbol{x})\|_{W^{1,\infty}((0,1)^d)} \leq 8K$ and $\|s_{\boldsymbol{m}}(\boldsymbol{x})\|_{W^{1,\infty}((0,1)^d)} \leq 64K^2$ for any $\boldsymbol{m} \in \{1,2\}^d$.

(ii): $\{s_{\boldsymbol{m}}(\boldsymbol{x})\}_{\boldsymbol{m} \in \{1,2\}^d}$ is a partition of the unity $[0,1]^d$ with $\operatorname{supp} s_{\boldsymbol{m}}(\boldsymbol{x}) \cap [0,1]^d = \Omega_{\boldsymbol{m}}$ defined in Definition 5.

(iii):For any $\boldsymbol{m} \in \{1,2\}^d$, there is a σ_2 neural network $\lambda_{\boldsymbol{m}}(\boldsymbol{x})$ with the width $16N + 2d$ and depth $4L + 5$ such as

$$\lambda_{\boldsymbol{m}}(\boldsymbol{x}) = \prod_{j=1}^{d} s_{m_j}(x_j) = s_{\boldsymbol{m}}(\boldsymbol{x}), \boldsymbol{x} \in [0,1]^d.$$

Proof. (i) and (ii) are proved by direct calculation. The proof of (iii) follows:

First, we architect $s(x)$ by a σ_2 neural network. The is a σ_1 neural network $g(x)$ with 3 the width and one layer such that:

$$g(x) := \begin{cases} x, & x \in \left[0, \frac{1}{2}\right] \\ \frac{1}{2}, & x \in \left[\frac{1}{2}, +\infty\right) \\ 0, & \text{Otherwise}. \end{cases} \tag{95}$$

Based on (iii) in Lemma 10, $g^2(x)$ is a σ_2 neural network with 3 the width and two layers. Then by direct calculation, we notice that

$$s(x) = 2g^2(x) - 2g^2(-x+1) + 2g^2\left(3-x\right) - 2g^2\left(2+x\right) + \frac{1}{2}, \tag{96}$$

which is a σ_2 neural network with 12 the width and two layers. The $\widetilde{g}(x)$ defined as

$$\widetilde{g}(x)=\sum_{i=0}^{\lfloor N^{1/d}\rfloor-1} s\left(4Kx-4i-\frac{1}{2}\right) \tag{97}$$

is a σ_2 neural network with $12(\lfloor N^{1/d}\rfloor)$ the width and two layers.

Similar with Lemma 1, we know that

$$\hat{g}=\widetilde{g}\circ\psi_2\circ\psi_3\circ\psi_4(x)$$

is a σ_2 neural network with $12(\lfloor N^{1/d}\rfloor)$ the width and $5+2\lfloor L^{1/d}\rfloor$, and

$$s_1(x)=\hat{g}\left(x+\frac{1}{8K}\right),\ s_2(x)=s_1\left(x+\frac{1}{2K}\right),\ x\in[0,1]. \tag{98}$$

Based on (v) in Lemma 10, we have there is a σ_2 neural network $\lambda_{\boldsymbol{m}}(\boldsymbol{x})$ with the width $16N+2d$ and depth $4L+5$ such as

$$\lambda_{\boldsymbol{m}}(\boldsymbol{x})=\prod_{j=1}^{d} s_{m_j}(x_j)=s_{\boldsymbol{m}}(\boldsymbol{x}),\boldsymbol{x}\in[0,1]^d.$$

$\square$

7.3.2 Proof of Corollaries 1 and 2

The proof is comprised of three parts, which include Theorem 8 and 9, followed by the combination of these results. Theorem 8 is to apply the Bramble–Hilbert Lemma 9 measured in the norm of $W^{2,\infty}$:

Theorem 8. *Let $K\in\mathbb{N}_+$ and $n\geq 2$. Then for any $f\in W^{n,\infty}((0,1)^d)$ with $\|f\|_{W^{n,\infty}((0,1)^d)}\leq 1$ and $\boldsymbol{m}\in\{1,2\}^d$, there exist piece-wise polynomials function $f_{K,\boldsymbol{m}}=\sum_{|\boldsymbol{\alpha}|\leq n-1} g_{f,\boldsymbol{\alpha},\boldsymbol{m}}(\boldsymbol{x})\boldsymbol{x}^{\boldsymbol{\alpha}}$ on $\Omega_{\boldsymbol{m}}$ (Definition 5) with the following properties:*

$$\begin{aligned}\|f-f_{K,\boldsymbol{m}}\|_{W^{2,\infty}(\Omega_{\boldsymbol{m}})}&\leq C_1(n,d)K^{-(n-2)},\\ \|f-f_{K,\boldsymbol{m}}\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}})}&\leq C_1(n,d)K^{-(n-1)},\\ \|f-f_{K,\boldsymbol{m}}\|_{L^{\infty}(\Omega_{\boldsymbol{m}})}&\leq C_1(n,d)K^{-n}.\end{aligned} \tag{99}$$

Furthermore, $g_{f,\boldsymbol{\alpha},\boldsymbol{m}}(\boldsymbol{x}):\Omega_{\boldsymbol{m}}\to\mathbb{R}$ is a constant function with on each $\Omega_{\boldsymbol{m},\boldsymbol{i}}$ for $\boldsymbol{i}\in\{0,1\ldots,K\}^d$. And

$$|g_{f,\boldsymbol{\alpha},\boldsymbol{m}}(\boldsymbol{x})|\leq C_2(n,d) \tag{100}$$

for all $\boldsymbol{x}\in\Omega_{\boldsymbol{m}}$, where C_1 and C_2 are constants independent with K.

The proof is the same as that of Theorem 6. Note that $\{f_{K,\boldsymbol{m}}\}_{\boldsymbol{m}\in\{1,2\}^d}$ will be same in two theorems if $f\in W^{n,\infty}((0,1)^d)$ in two theorem are same.

Theorem 9 is to establish σ_2 neural networks $\{\gamma_{\boldsymbol{m}}\}_{\{1,2\}^d}$, and each $\gamma_{\boldsymbol{m}}$ can approximate f well on $\Omega_{\boldsymbol{m}}$.

Theorem 9. *For any $f\in W^{n,\infty}((0,1)^d)$ with $\|f\|_{W^{n,\infty}((0,1)^d)}\leq 1$, any $N,L\in\mathbb{N}_+$ with $NL+2^{\lfloor\log_2 N\rfloor}\geq n$ and $L\geq\lceil\log_2 N\rceil$, and $\boldsymbol{m}=(m_1,m_2,\ldots,m_d)\in\{1,2\}^d$, there is a σ_2 neural network $\gamma_{\boldsymbol{m}}$ with the width $28n^{d+1}(N+d)\log_2(8N)$ and depth $11n^2(L+2)\log_2(4L)$ such that*

$$\begin{aligned}\|f(\boldsymbol{x})-\gamma_{\boldsymbol{m}}(\boldsymbol{x})\|_{W^{2,\infty}(\Omega_{\boldsymbol{m}})}&\leq C_{10}(n,d)N^{-2(n-2)/d}L^{-2(n-2)/d}\\ \|f(\boldsymbol{x})-\gamma_{\boldsymbol{m}}(\boldsymbol{x})\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}})}&\leq C_{10}(n,d)N^{-2(n-1)/d}L^{-2(n-1)/d}\\ \|f(\boldsymbol{x})-\gamma_{\boldsymbol{m}}(\boldsymbol{x})\|_{L^{\infty}(\Omega_{\boldsymbol{m}})}&\leq C_{10}(n,d)N^{-2n/d}L^{-2n/d},\end{aligned} \tag{101}$$

where C_{10} is the constant independent with N,L.

Proof. The proof is similar to that of Theorem 7; the difference is that xy and $\boldsymbol{x}^{\boldsymbol{\alpha}}$ can be architected precisely by σ_2 neural networks.

Without loss of the generalization, we consider the case for $\boldsymbol{m}_* = (1, 1, \dots, 1)$. Due to Theorem 8 and setting $K = \lfloor N^{1/d} \rfloor^2 \lfloor L^{2/d} \rfloor$, we have

$$\begin{aligned}
&\|f - f_{K,\boldsymbol{m}_*}\|_{W^{2,\infty}(\Omega_{\boldsymbol{m}_*})} \leq C_1(n,d) K^{-(n-2)} \leq C_1(n,d) N^{-2(n-2)/d} L^{-2(n-2)/d}\\
&\|f - f_{K,\boldsymbol{m}_*}\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}_*})} \leq C_1(n,d) K^{-(n-1)} \leq C_1(n,d) N^{-2(n-1)/d} L^{-2(n-1)/d}\\
&\|f - f_{K,\boldsymbol{m}_*}\|_{L^{\infty}(\Omega_{\boldsymbol{m}_*})} \leq C_1(n,d) K^{-n} \leq C_1(n,d) N^{-2n/d} L^{-2n/d},
\end{aligned} \tag{102}$$

where $f_{K,\boldsymbol{m}_*} = \sum_{|\boldsymbol{\alpha}| \leq n-1} g_{f,\boldsymbol{\alpha},\boldsymbol{m}_*}(\boldsymbol{x}) \boldsymbol{x}^{\boldsymbol{\alpha}}$ for $x \in \Omega_{\boldsymbol{m}_*}$. Note that $g_{f,\boldsymbol{\alpha},\boldsymbol{m}_*}(\boldsymbol{x})$ is a constant function for $\boldsymbol{x} \in \prod_{j=1}^{d} \left[\frac{i_j}{K}, \frac{3+4i_j}{4K}\right]$ and $\boldsymbol{i} = (i_1, i_2, \dots, i_d) \in \{0, 1, \dots, K-1\}^d$. The remaining part is to approximate $f_{K,\boldsymbol{m}_*}$ by neural networks.

The way to approximate $g_{f,\boldsymbol{\alpha},\boldsymbol{m}_*}(\boldsymbol{x})$ is same with Theorem 7, and we have that

$$\begin{aligned}
\|\phi_{\boldsymbol{\alpha}}(\boldsymbol{\phi}_2(\boldsymbol{x})) - g_{f,\boldsymbol{\alpha},\boldsymbol{m}_*}(\boldsymbol{x})\|_{W^{2,\infty}(\Omega_{\boldsymbol{m}_*})} =& \|\phi_{\boldsymbol{\alpha}}(\boldsymbol{\phi}_2(\boldsymbol{x})) - g_{f,\boldsymbol{\alpha},\boldsymbol{m}_*}(\boldsymbol{x})\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}_*})}\\
=& \|\phi_{\boldsymbol{\alpha}}(\boldsymbol{\phi}_2(\boldsymbol{x})) - g_{f,\boldsymbol{\alpha},\boldsymbol{m}_*}(\boldsymbol{x})\|_{L^{\infty}(\Omega_{\boldsymbol{m}_*})}\\
\leq& 2C_2(n,d) N^{-2n} L^{-2n}
\end{aligned} \tag{103}$$

which is due to $\phi_{\boldsymbol{\alpha}}(\boldsymbol{\phi}_2(\boldsymbol{x})) - g_{f,\boldsymbol{\alpha},\boldsymbol{m}_*}(\boldsymbol{x})$ is a step function, and the first order weak derivative is 0 in $\Omega_{\boldsymbol{m}_*}$.

Due to (v) in Lemma 10, there is a σ_2 neural network $\phi_{5,\boldsymbol{\alpha}}(\boldsymbol{x})$ with the width $4N + 2d$ and depth $L + \lceil \log_2 N \rceil$ such that

$$\phi_{5,\boldsymbol{\alpha}}(\boldsymbol{x}) = \boldsymbol{x}^{\boldsymbol{\alpha}},\ \boldsymbol{x} \in \mathbb{R}^d. \tag{104}$$

Due to (iv) in Lemma 10, there is a σ_2 neural network $\phi_6(\boldsymbol{x})$ with the width 4 and depth 1 such that

$$\phi_6(x, y) = xy,\ x, y \in \mathbb{R}. \tag{105}$$

Now we define the neural network $\gamma_{\boldsymbol{m}_*}(\boldsymbol{x})$ to approximate $f_{K,\boldsymbol{m}_*}(\boldsymbol{x})$ in $\Omega_{\boldsymbol{m}_*}$:

$$\gamma_{\boldsymbol{m}_*}(\boldsymbol{x}) = \sum_{|\boldsymbol{\alpha}| \leq n-1} \phi_6\left[\phi_{\boldsymbol{\alpha}}(\boldsymbol{\phi}_2(\boldsymbol{x})), \phi_{5,\boldsymbol{\alpha}}(\boldsymbol{x})\right]. \tag{106}$$

The remaining question is to find the error $\mathcal{E}$:

$$\begin{aligned}
\widetilde{\mathcal{E}} :=& \left\| \sum_{|\boldsymbol{\alpha}| \leq n-1} \phi_6\left[\phi_{\boldsymbol{\alpha}}(\boldsymbol{\phi}_2(\boldsymbol{x})), \phi_{5,\boldsymbol{\alpha}}(\boldsymbol{x})\right] - f_{K,\boldsymbol{m}_*}(\boldsymbol{x}) \right\|_{W^{2,\infty}(\Omega_{\boldsymbol{m}_*})}\\
\leq& \sum_{|\boldsymbol{\alpha}| \leq n-1} \|\phi_6\left[\phi_{\boldsymbol{\alpha}}(\boldsymbol{\phi}_2(\boldsymbol{x})), \phi_{5,\boldsymbol{\alpha}}(\boldsymbol{x})\right] - g_{f,\boldsymbol{\alpha},\boldsymbol{m}_*}(\boldsymbol{x}) \boldsymbol{x}^{\boldsymbol{\alpha}}\|_{W^{2,\infty}(\Omega_{\boldsymbol{m}_*})}\\
=& \sum_{|\boldsymbol{\alpha}| \leq n-1} \|\phi_{\boldsymbol{\alpha}}(\boldsymbol{\phi}_2(\boldsymbol{x})) \boldsymbol{x}^{\boldsymbol{\alpha}} - g_{f,\boldsymbol{\alpha},\boldsymbol{m}_*}(\boldsymbol{x}) \boldsymbol{x}^{\boldsymbol{\alpha}}\|_{W^{2,\infty}(\Omega_{\boldsymbol{m}_*})}\\
\leq& n^2 \sum_{|\boldsymbol{\alpha}| \leq n-1} \|\phi_{\boldsymbol{\alpha}}(\boldsymbol{\phi}_2(\boldsymbol{x})) - g_{f,\boldsymbol{\alpha},\boldsymbol{m}_*}(\boldsymbol{x})\|_{W^{2,\infty}(\Omega_{\boldsymbol{m}_*})}\\
\leq& 2n^{d+2} C_2(n,d) N^{-2n} L^{-2n}.
\end{aligned} \tag{107}$$

At last, we finish the proof by estimating the network's the width and depth, implementing $\gamma_{\boldsymbol{m}_*}(\boldsymbol{x})$. From Eq. (106), we know that $\gamma_{\boldsymbol{m}_*}(\boldsymbol{x})$ consists of the following subnetworks:

1. $\phi_{5,\boldsymbol{\alpha}}(\boldsymbol{x})$ with the width $4N + 2d$ and depth $L + \lceil \log_2 N \rceil$.
2. $\phi_2(\boldsymbol{x})$ with the width $4N + 5$ and depth $4L + 4$.
3. $\phi_{\boldsymbol{\alpha}}$ with the width $16n(N+1)\log_2(8N)$ and depth $(5L+2)\log_2(4L)$.
4. $\phi_6(x, y)$ with the width 4 and depth 1.

Therefore $\phi(\boldsymbol{x})$ is a neural network with the width $28n^{d+1}(N+d)\log_2(8N)$ and depth $11n^2(L+2)\log_2(4L)$.

Combining Eqs. (102) and (107), we have that there is a neural network $\gamma_{\boldsymbol{m}_*}$ with the width $28n^{d+1}(N+d)\log_2(8N)$ and depth $11n^2(L+2)\log_2(4L)$ such that

$$\begin{aligned}
\|f(\boldsymbol{x})-\psi_{\boldsymbol{m}_*}(\boldsymbol{x})\|_{W^{2,\infty}(\Omega_{\boldsymbol{m}_*})} &\leq C_{10}(n,d)N^{-2(n-2)/d}L^{-2(n-2)/d} \\
\|f(\boldsymbol{x})-\psi_{\boldsymbol{m}_*}(\boldsymbol{x})\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}_*})} &\leq C_{10}(n,d)N^{-2(n-1)/d}L^{-2(n-1)/d} \\
\|f(\boldsymbol{x})-\psi_{\boldsymbol{m}_*}(\boldsymbol{x})\|_{L^{\infty}(\Omega_{\boldsymbol{m}_*})} &\leq C_{10}(n,d)N^{-2n/d}L^{-2n/d},
\end{aligned} \tag{108}$$

where $C_{10}=C_1+2n^{d+2}C_2$ is the constant independent with N,L.

Similarly, we can construct a neural network $\gamma_{\boldsymbol{m}}$ with the width $28n^{d+1}(N+d)\log_2(8N)$ and depth $11n^2(L+2)\log_2(4L)$ which can approximate f on $\Omega_{\boldsymbol{m}}$ with same order of Eq. (108). $\square$

The last part is to combine $\{\lambda_{\boldsymbol{m}}\}_{\boldsymbol{m}\in\{1,2\}^d}$ and $\{\gamma_{\boldsymbol{m}}\}_{\boldsymbol{m}\in\{1,2\}^d}$ in $[0,1]^d$ and obtain a σ_2 neural network to approximate f measured in the norm of W^2.

Proof of Corollary 1. Based on Theorem 9, there is a sequence of the neural network $\{\gamma_{\boldsymbol{m}}(\boldsymbol{x})\}_{\boldsymbol{m}\in\{1,2\}^d}$ such that

$$\begin{aligned}
\|f(\boldsymbol{x})-\gamma_{\boldsymbol{m}}(\boldsymbol{x})\|_{W^{2,\infty}(\Omega_{\boldsymbol{m}})} &\leq C_{10}(n,d)N^{-2(n-2)/d}L^{-2(n-2)/d} \\
\|f(\boldsymbol{x})-\gamma_{\boldsymbol{m}}(\boldsymbol{x})\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}})} &\leq C_{10}(n,d)N^{-2(n-1)/d}L^{-2(n-1)/d} \\
\|f(\boldsymbol{x})-\gamma_{\boldsymbol{m}}(\boldsymbol{x})\|_{L^{\infty}(\Omega_{\boldsymbol{m}})} &\leq C_{10}(n,d)N^{-2n/d}L^{-2n/d},
\end{aligned} \tag{109}$$

where C_{10} is the constant independent with N,L, and each $\gamma_{\boldsymbol{m}}$ is a neural network with the width $28n^{d+1}(N+d)\log_2(8N)$ and depth $11n^2(L+2)\log_2(4L)$. According to Proposition 8, there is a sequence of the neural network $\{s_{\boldsymbol{m}}(\boldsymbol{x})\}_{\boldsymbol{m}\in\{1,2\}^d}$ satisfies:

(i): $\|s_{\boldsymbol{m}}(\boldsymbol{x})\|_{L^\infty((0,1)^d)}\leq 1$, $\|s_{\boldsymbol{m}}(\boldsymbol{x})\|_{W^{1,\infty}((0,1)^d)}\leq 8K$ and $\|s_{\boldsymbol{m}}(\boldsymbol{x})\|_{W^{1,\infty}((0,1)^d)}\leq 64K^2$ for any $\boldsymbol{m}\in\{1,2\}^d$.

(ii): $\{s_{\boldsymbol{m}}(\boldsymbol{x})\}_{\boldsymbol{m}\in\{1,2\}^d}$ is a partition of the unity $[0,1]^d$ with supp $s_{\boldsymbol{m}}(\boldsymbol{x})\cap[0,1]^d=\Omega_{\boldsymbol{m}}$ defined in Definition 5.

For each $s_{\boldsymbol{m}}$, it is a σ_2 neural network with the width $16N+2d$ and depth $4L+5$.

Due to (iv) in Lemma 10, there is a σ_2 neural network $\phi_6(\boldsymbol{x})$ with the width 4 and depth 1 such that

$$\phi_6(x,y)=xy,\ x,y\in\mathbb{R}. \tag{110}$$

Now we define

$$\gamma(\boldsymbol{x})=\sum_{\boldsymbol{m}\in\{1,2\}^d}\phi_6(s_{\boldsymbol{m}}(\boldsymbol{x}),\gamma_{\boldsymbol{m}}(\boldsymbol{x})). \tag{111}$$

Note that

$$\begin{aligned}
\widetilde{\mathcal{R}}:=&\|f(\boldsymbol{x})-\gamma(\boldsymbol{x})\|_{W^{2,\infty}((0,1)^d)}\leq\sum_{\boldsymbol{m}\in\{1,2\}^d}\|s_{\boldsymbol{m}}(x)\cdot f(\boldsymbol{x})-s_{\boldsymbol{m}}(\boldsymbol{x})\gamma_{\boldsymbol{m}}(\boldsymbol{x})\|_{W^{2,\infty}((0,1)^d)} \\
=&\sum_{\boldsymbol{m}\in\{1,2\}^d}\|s_{\boldsymbol{m}}(x)\cdot f(\boldsymbol{x})-s_{\boldsymbol{m}}(\boldsymbol{x})\gamma_{\boldsymbol{m}}(\boldsymbol{x})\|_{W^{2,\infty}(\Omega_{\boldsymbol{m}})}.
\end{aligned} \tag{112}$$

where the last equality is due to supp $s_{\boldsymbol{m}}(\boldsymbol{x})\cap[0,1]^d=\Omega_{\boldsymbol{m}}$.

Then due to chain rule, for each $\boldsymbol{m}\in\{1,2\}^d$, we have

$$\begin{aligned}
&\|s_{\boldsymbol{m}}(x)\cdot f(\boldsymbol{x})-s_{\boldsymbol{m}}(\boldsymbol{x})\gamma_{\boldsymbol{m}}(\boldsymbol{x})\|_{W^{2,\infty}(\Omega_{\boldsymbol{m}})} \\
\leq&\|s_{\boldsymbol{m}}(x)\|_{W^{2,\infty}(\Omega_{\boldsymbol{m}})}\|f(\boldsymbol{x})-\gamma_{\boldsymbol{m}}(\boldsymbol{x})\|_{L^{\infty}(\Omega_{\boldsymbol{m}})}+2\|s_{\boldsymbol{m}}(x)\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}})}\|f(\boldsymbol{x})-\gamma_{\boldsymbol{m}}(\boldsymbol{x})\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}})} \\
&+\|s_{\boldsymbol{m}}(x)\|_{L^{\infty}(\Omega_{\boldsymbol{m}})}\|f(\boldsymbol{x})-\gamma_{\boldsymbol{m}}(\boldsymbol{x})\|_{W^{2,\infty}(\Omega_{\boldsymbol{m}})}+\|s_{\boldsymbol{m}}(x)\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}})}\|f(\boldsymbol{x})-\gamma_{\boldsymbol{m}}(\boldsymbol{x})\|_{L^{\infty}(\Omega_{\boldsymbol{m}})} \\
&+\|s_{\boldsymbol{m}}(x)\|_{L^{\infty}(\Omega_{\boldsymbol{m}})}\|f(\boldsymbol{x})-\gamma_{\boldsymbol{m}}(\boldsymbol{x})\|_{W^{1,\infty}(\Omega_{\boldsymbol{m}})}+\|s_{\boldsymbol{m}}(x)\|_{L^{\infty}(\Omega_{\boldsymbol{m}})}\|f(\boldsymbol{x})-\gamma_{\boldsymbol{m}}(\boldsymbol{x})\|_{L^{\infty}(\Omega_{\boldsymbol{m}})} \\
\leq&91C_{10}(n,d)N^{-2(n-2)/d}L^{-2(n-2)/d}.
\end{aligned} \tag{113}$$

Hence

$$\widetilde{\mathcal{R}} \le 2^{d+7} C_{10}(n,d) N^{-2(n-2)/d} L^{-2(n-2)/d}.$$

At last, we finish the proof by estimating the network's width and depth, implementing $\gamma(\boldsymbol{x})$. From Eq. (111), we know that $\gamma(\boldsymbol{x})$ consists of the following subnetworks:

1. $\gamma_{\boldsymbol{m}}(\boldsymbol{x})$ with the width $28n^{d+1}(N+d)\log_2(8N)$ and depth $11n^2(L+2)\log_2(4L)$.
2. $s_{\boldsymbol{m}}(\boldsymbol{x})$ with the width $16N+2d$ and depth $4L+5$.
3. $\phi_6(x,y)$ with the width 4 and depth 1.

Therefore $\gamma(\boldsymbol{x})$ is a neural network with the width $2^{d+6}n^{d+1}(N+d)\log_2(8N)$ and depth $15n^2(L+2)\log_2(4L)$.

□

Our method can easily extend to approximations measured by the norm of $W^{m,\infty}$. The primary difference in the proof lies in the need to establish a differential $\{s_{\boldsymbol{m}}(\boldsymbol{x})\}_{\{1,2\}^d}$, which can be achieved by constructing architected $s_{\boldsymbol{m}}(\boldsymbol{x})$ as piece-wise m-degree polynomial functions. By extanding this approach, we can obtain Corollary 2 using our method.

7.4 Proof of Theorem 4

Proof. The Theorem 4 will be proved by contradiction. The idea of the proof is inspired by Ref. [28].

Claim 1. *There exist $\rho, C_1, C_2, C_3, J_0 > 0$ and $s, d \in \mathbb{N}^+$ such that, for any $f \in \mathcal{F}_{n,d}$, we have*

$$\inf_{\phi\in\widehat{\Phi}} |\phi - f|_{W^{1,\infty}((0,1)^d)} \le C_3 L^{-2(n-1)/d-\rho} N^{-2(n-1)/d-\rho}. \tag{114}$$

for all $NL \ge J_0$, where

$$\widehat{\Phi} := \{\phi : \textit{ReLU FNNs } \phi \textit{ with the width} \le C_1 N\log N \textit{ and depth} \le C_2 L \log L\}.$$

The remaining question is to show Claim 1 is invalid.

Denote

$$D\widehat{\Phi} := \{\psi : \psi = D_i\phi, \phi \in \widehat{\Phi}, i = 1,\ldots,d\},$$

Due to Theorem 1, we obtain

$$\mathrm{VCDim}(D\widehat{\Phi}) \le C_4 N^2 L^2 \log_2 L \log_2 N =: b_u. \tag{115}$$

Now we will use Claim 1 to estimate a lower bound

$$b_l := \lfloor (NL)^{\frac{2}{d}+\frac{\rho}{2(n-1)}} \rfloor^d$$

of $\mathrm{VCDim}(D\widehat{\Phi})$. In other words, we will construct $\{\psi_\beta(\boldsymbol{x}) : \psi_\beta(\boldsymbol{x}) \in D\widehat{\Phi}, \beta \in \mathcal{B}\}$ to scatter b_l points. $\mathcal{B}$ will be defined later.

First, fix $i = 1,\ldots,d$, and there exists $\widetilde{g} \in C^\infty(0,1)^d$ such that $\frac{\partial \widetilde{g}(\boldsymbol{0})}{\partial x_i} = 1$ and $\widetilde{g}(\boldsymbol{x}) = 0$ for $\|\boldsymbol{x}\|_2 \ge 1/3$. And we can find a constant $C_5 > 0$ such that $g := \widetilde{g}/C_5 \in \mathcal{F}_{n,d}$.

Denote $M = \lfloor (NL)^{\frac{2}{d}+\frac{\rho}{2(n-1)}} \rfloor$. Divide $[0,1]^d$ into M^d non-overlapping sub-cubes $\{Q_{\boldsymbol{\theta}}\}_\theta$ as follows:

$$Q_{\boldsymbol{\theta}} := \left\{ \boldsymbol{x} = [x_1, x_2, \cdots, x_d]^T \in [0,1]^d : x_i \in \left[\frac{\theta_i - 1}{M}, \frac{\theta_i}{M}\right], i = 1,2,\cdots,d \right\},$$

for any index vector $\boldsymbol{\theta} = [\theta_1, \theta_2, \cdots, \theta_d]^T \in \{1,2,\cdots,M\}^d$. Denote the center of $Q_{\boldsymbol{\theta}}$ by $\boldsymbol{x}_{\boldsymbol{\theta}}$ for all $\boldsymbol{\theta} \in \{1,2,\cdots,M\}^d$. Define

$$\mathcal{B} := \left\{ \beta : \beta \text{ is a map from } \{1,2,\cdots,M\}^d \text{ to } \{-1,1\} \right\}.$$

For each $\beta \in \mathcal{B}$, we define, for any $\boldsymbol{x} \in \mathbb{R}^d$,

$$h_\beta(\boldsymbol{x}) := \sum_{\boldsymbol{\theta} \in \{1,2,\cdots,M\}^d} M^{-n}\beta(\boldsymbol{\theta}) g_{\boldsymbol{\theta}}(\boldsymbol{x}), \quad \text{where } g_{\boldsymbol{\theta}}(\boldsymbol{x}) = g\left(M \cdot (\boldsymbol{x} - \boldsymbol{x}_{\boldsymbol{\theta}})\right).$$

Due to $|\mathrm{supp}\widetilde{g}(\boldsymbol{x})| \leq \frac{2}{3}$ and $|D^{\boldsymbol{\alpha}} h_\beta(\boldsymbol{x})| \leq M^{-n+|\boldsymbol{\alpha}|}\|g\|_{W^{n,\infty}} \leq 1$, we obtain that

$$|D^{\boldsymbol{\alpha}} f_\beta(\boldsymbol{x})| \leq 1$$

for any $|\boldsymbol{\alpha}| \leq n$ Therefore, $f_\beta \in \mathcal{F}_{n,d}$. And it is easy to check $\{D_i h_\beta = h_\beta : \beta \in \mathcal{B}\}$ can scatters b_l points since $\frac{\partial \widetilde{g}(\boldsymbol{0})}{\partial x_i} = 1$ and $\widetilde{g}(\boldsymbol{x}) = 0$ for $\|\boldsymbol{x}\|_2 \geq 1/3$.

Note that for any $h_\beta \in \mathcal{F}_{n,d}$, there is a $\phi_\beta \in \widehat{\Phi}$ such that $C_3(NL)^{\frac{-2(n-1)}{d}-\frac{\rho}{2}} \geq |D_i h_\beta(\boldsymbol{x}_{\boldsymbol{\theta}}) - D_i\phi_\beta(\boldsymbol{x}_{\boldsymbol{\theta}})|$ for any $J_\beta \leq NL$ due to Claim 1. Denote $J_1 = \max_{\beta\in\mathcal{B}}\{J_\beta\}$. There is a constant J_2 such that $\frac{M^{-n+1}}{C_5} \geq C_3(NL)^{\frac{-2(n-1)}{d}-\rho}$ for $J_2 \leq NL$. Define $J := \max\{J_1, J_2\}$, then for any $J \leq NL$, we have

$$|D_i h_\beta(\boldsymbol{x}_{\boldsymbol{\theta}})| = \left|M^{-n+1}\frac{\partial g(\boldsymbol{x}_{\boldsymbol{\theta}})}{\partial x_i}\right| = \frac{M^{-n+1}}{C_5} \geq C_3(NL)^{\frac{-2(n-1)}{d}-\rho} \geq |D_i h_\beta(\boldsymbol{x}_{\boldsymbol{\theta}}) - D_i\phi_\beta(\boldsymbol{x}_{\boldsymbol{\theta}})|. \tag{116}$$

In other words, for any $\beta \in \mathcal{B}$ and $\boldsymbol{\theta} \in \{1,2,\cdots,M\}^d$, $D_i f_\beta(\boldsymbol{x}_\theta)$ and $D_i\phi_\beta(\boldsymbol{x}_\theta)$ have the same sign. Then $\{D_i\phi_\beta : \beta \in \mathcal{B}\}$ shatters $\left\{\boldsymbol{x}_\theta : \boldsymbol{\theta} \in \{1,2,\cdots,M\}^d\right\}$ since $\{D_i h_\beta : \beta \in \mathcal{B}\}$ shatters $\left\{\boldsymbol{x}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \{1,2,\cdots,M\}^d\right\}$ as discussed above. Hence,

$$\mathrm{VCDim}\left(\{\phi_\beta : \beta \in \mathcal{B}\}\right) \geq M^d = b_l, \tag{117}$$

for $N, L \in \mathbb{N}$ with $NL \geq J$.

By Eqs. (115,117), for any $N, L \in \mathbb{N}$ with $NL \geq J$, we have $b_l \leq \mathrm{VCDim}\left(\{\phi_\beta : \beta \in \mathcal{B}\}\right) \leq \mathrm{VCDim}(D\widehat{\Phi}) \leq b_u$, implying that

$$\lfloor (NL)^{\frac{2}{d}+\frac{\rho}{2(n-1)}} \rfloor^d \leq C_4 N^2 L^2 \log_2 L \log_2 N \tag{118}$$

which is a contradiction for sufficiently large $N, L \in \mathbb{N}$. So we finish the proof of Theorem 4. $\square$

Based on the proof of Theorem 4, we can easily check that the estimation of VC-dimension of DNN derivatives (Theorem 1) is nearly optimal and prove Corollary 3. Assume $\mathrm{VCDim}(D\widehat{\Phi}) \leq b_u = O(N^{2-\varepsilon}L^{2-\varepsilon})$ in Eq. (118) for $\varepsilon > 0$, and b_l must be larger than $\lfloor (NL)^{\frac{2}{d}} \rfloor^d$ according the construction in the proof of Theorem 4 and Theorem 3. Hence we still obtain a contradiction in Eq. (118), and the estimation in Theorem 1 is nearly optimal.

7.5 Proof of Theorem 5

7.5.1 Bounding generalization error by Rademacher complexity

Definition 13 (Rademacher complexity [3]). *Given a sample set $S = \{z_1, z_2, \ldots, z_M\}$ on a domain $\mathcal{Z}$, and a class $\mathcal{F}$ of real-valued functions defined on $\mathcal{Z}$, the empirical Rademacher complexity of $\mathcal{F}$ in S is defined as*

$$\mathbf{R}_S(\mathcal{F}) := \frac{1}{M}\mathbf{E}_{\Sigma_M}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^M \sigma_i f(z_i)\right],$$

where $\Sigma_M := \{\sigma_1, \sigma_2, \ldots, \sigma_M\}$ are independent random variables drawn from the Rademacher distribution, i.e., $\mathbf{P}(\sigma_i = +1) = \mathbf{P}(\sigma_i = -1) = \frac{1}{2}$ for $i = 1, 2, \ldots, M$. For simplicity, if $S = \{z_1, z_2, \ldots, z_M\}$ is an independent random variable set with the uniform distribution, denote

$$\mathbf{R}_M(\mathcal{F}) := \mathbf{E}_S \mathbf{R}_S(\mathcal{F}).$$

The following lemma will be used to bounded generalization error by Rademacher complexities:

Lemma 11 ([47], Proposition 4.11)**.** *Let $\mathcal{F}$ be a set of functions. Then*

$$\mathbf{E}_X \sup_{u\in\mathcal{F}} \left| \frac{1}{M}\sum_{i=1}^{M} u(x_j) - \mathbf{E}_{x\sim\mathcal{P}_\Omega} u(x) \right| \leq 2\mathbf{R}_M(\mathcal{F}),$$

where $X := \{x_1, \ldots, x_M\}$ is an independent random variable set with the uniform distribution.

Now we can show that generalization error can be bounded by Rademacher complexities of two function sets.

Lemma 12. *Let $d, N, L, M \in \mathbb{N}_+$, $B, C_1, C_2 \in \mathbb{R}_+$. For any $f \in W^{1,\infty}((0,1)^d)$ with $\|f\|_{W^{1,\infty}((0,1)^d)} \leq 1$, set*

$$\begin{aligned} \widetilde{\Phi} &:= \{\phi : \phi \text{ with the width} \leq C_1 N \log N \text{ and depth} \leq C_2 L \log L, \|\phi\|_{W^{1,\infty}((0,1)^d)} \leq B\} \\ D\widetilde{\Phi} &:= \{\psi : \psi = D_i\phi, i = 1, \ldots, d\}. \end{aligned} \tag{119}$$

We have

$$2 \sup_{\boldsymbol{\theta}, \phi(\boldsymbol{x};\boldsymbol{\theta})\in\widetilde{\Phi}} |\mathbf{E}(\mathcal{R}_S(\boldsymbol{\theta})) - \mathcal{R}_D(\boldsymbol{\theta})| \leq 4(B+1)(d\mathbf{R}_M(D\widetilde{\Phi}) + \mathbf{R}_M(\widetilde{\Phi})),$$

where $\mathbf{E}$ is expected responding to X, and $X := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ is an independent random variables set uniformly distributed on $(0,1)^d$.

Proof. For any $\phi(\boldsymbol{x};\boldsymbol{\theta}) \in \widetilde{\Phi}$, we have

$$\begin{aligned} &|\mathbf{E}(\mathcal{R}_S(\boldsymbol{\theta})) - \mathcal{R}_D(\boldsymbol{\theta})| \\ =&\sum_{j=1}^{d}\left(\mathbf{E}\frac{1}{M}\sum_{i=1}^{M}\left|\frac{\partial(f(\boldsymbol{x}_i)-\phi(\boldsymbol{x}_i;\boldsymbol{\theta}))}{\partial x_j}\right|^2 - \int_{(0,1)^d}\left|\frac{\partial(f(\boldsymbol{x})-\phi(\boldsymbol{x};\boldsymbol{\theta}))}{\partial x_j}\right|^2 \mathrm{d}\boldsymbol{x}\right) \\ &+\mathbf{E}\frac{1}{M}\sum_{i=1}^{M}|(f(\boldsymbol{x}_i)-\phi(\boldsymbol{x}_i;\boldsymbol{\theta}))|^2 - \int_{(0,1)^d}|(f(\boldsymbol{x})-\phi(\boldsymbol{x};\boldsymbol{\theta}))|^2\,\mathrm{d}\boldsymbol{x} \\ \leq&(B+1)\sum_{j=1}^{d}\left(\mathbf{E}\left|\frac{1}{M}\sum_{i=1}^{M}\frac{\partial(f(\boldsymbol{x}_i)-\phi(\boldsymbol{x}_i;\boldsymbol{\theta}))}{\partial x_j} - \int_{(0,1)^d}\frac{\partial(f(\boldsymbol{x})-\phi(\boldsymbol{x};\boldsymbol{\theta}))}{\partial x_j}\mathrm{d}\boldsymbol{x}\right|\right) \\ &+(B+1)\mathbf{E}\left|\frac{1}{M}\sum_{i=1}^{M}(f(\boldsymbol{x}_i)-\phi(\boldsymbol{x}_i;\boldsymbol{\theta})) - \int_{(0,1)^d}(f(\boldsymbol{x})-\phi(\boldsymbol{x};\boldsymbol{\theta}))\,\mathrm{d}\boldsymbol{x}\right| \\ \leq&2(B+1)(d\mathbf{R}_M(D\widetilde{\Phi}) + \mathbf{R}_M(\widetilde{\Phi})) \end{aligned} \tag{120}$$

where the last inequality is due to Lemma 12. □

7.5.2 Bounding the Rademacher complexity and the proof of Theorem 5

In this subsection, we aim to estimate the Rademacher complexity using the covering number. We then estimate the covering number using the pseudo-dimension.

Definition 14 (covering number [3])**.** *Let $(V, \|\cdot\|)$ be a normed space, and $\Theta \in V$. $\{V_1, V_2, \ldots, V_n\}$ is an ε-covering of Θ if $\Theta \subset \cup_{i=1}^n B_{\varepsilon,\|\cdot\|}(V_i)$. The covering number $\mathcal{N}(\varepsilon, \Theta, \|\cdot\|)$ is defined as*

$$\mathcal{N}(\varepsilon, \Theta, \|\cdot\|) := \min\{n : \exists\varepsilon\text{-covering over } \Theta \text{ of size } n\}.$$

Definition 15 (Uniform covering number [3])**.** *Suppose the $\mathcal{F}$ is a class of functions from $\mathcal{F}$ to $\mathbb{R}$. Given n samples $\boldsymbol{Z}_n = (z_1, \ldots, z_n) \in \mathcal{X}^n$, define*

$$\mathcal{F}|_{\boldsymbol{Z}_n} = \{(u(z_1), \ldots, u(z_n)) : u \in \mathcal{F}\}.$$

The uniform covering number $\mathcal{N}(\varepsilon, \mathcal{F}, n)$ is defined as

$$\mathcal{N}(\varepsilon, \mathcal{F}, n) = \max_{\boldsymbol{Z}_n\in\mathcal{X}^n} \mathcal{N}\left(\varepsilon, \mathcal{F}|_{\boldsymbol{Z}_n}, \|\cdot\|_\infty\right),$$

where $\mathcal{N}\left(\varepsilon, \mathcal{F}|_{\boldsymbol{Z}_n}, \|\cdot\|_\infty\right)$ denotes the ε-covering number of $\mathcal{F}|_{\boldsymbol{Z}_n}$ w.r.t the L_∞-norm.

Then we use a lemma to estimate the Rademacher complexity using the covering number.

Lemma 13 (Dudley's theorem [3]). *Let $\mathcal{F}$ be a function class such that $\sup_{f\in\mathcal{F}}\|f\|_\infty \le B$. Then the Rademacher complexity $\mathbf{R}_n(\mathcal{F})$ satisfies that*

$$\mathbf{R}_n(\mathcal{F}) \le \inf_{0\le\delta\le B}\left\{4\delta + \frac{12}{\sqrt{n}}\int_\delta^B \sqrt{\log 2\mathcal{N}(\varepsilon,\mathcal{F},n)}\,\mathrm{d}\varepsilon\right\}$$

To bound the Rademacher complexity, we employ Lemma 13, which bounds it by the uniform covering number. We estimate the uniform covering number by the pseudo-dimension based on the following lemma.

Lemma 14 ([3]). *Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $[-B,B]$. For any $\varepsilon>0$, we have*

$$\mathcal{N}(\varepsilon,\mathcal{F},n) \le \left(\frac{2enB}{\varepsilon Pdim(\mathcal{F})}\right)^{Pdim(\mathcal{F})}$$

for $n \ge Pdim(\mathcal{F})$.

The remaining problem is to bound $\mathrm{Pdim}(\widetilde{\Phi})$ and $\mathrm{Pdim}(D\widetilde{\Phi})$. Based on [4], $\mathrm{Pdim}(\widetilde{\Phi}) = O(L^2N^2\log_2 L\log_2 N)$. For the $\mathrm{Pdim}(D\widetilde{\Phi})$, we can estimate it by Theorem 2.

Now we can estimate generalization error based on Lemma 12.

Proof of Theorem 5. Let $J=\max\{\mathrm{Pdim}(D\widetilde{\Phi}),\mathrm{Pdim}(\widetilde{\Phi})\}$. Due to Lemma 13, 14 and Theorem 2, for any $M\ge J$, we have

$$\begin{aligned}\mathbf{R}_M(D\widetilde{\Phi}) \le& 4\delta + \frac{12}{\sqrt{M}}\int_\delta^B \sqrt{\log 2\mathcal{N}(\varepsilon,D\widetilde{\Phi},M)}\,\mathrm{d}\varepsilon\\ \le& 4\delta + \frac{12}{\sqrt{M}}\int_\delta^B \sqrt{\log 2\left(\frac{2eMB}{\varepsilon\mathrm{Pdim}(D\widetilde{\Phi})}\right)^{\mathrm{Pdim}(D\widetilde{\Phi})}}\,\mathrm{d}\varepsilon\\ \le& 4\delta + \frac{12B}{\sqrt{M}} + 12\left(\frac{\mathrm{Pdim}(D\widetilde{\Phi})}{M}\right)^{\frac{1}{2}}\int_\delta^B\sqrt{\log\left(\frac{2eMB}{\varepsilon\mathrm{Pdim}(D\widetilde{\Phi})}\right)}\,\mathrm{d}\varepsilon. \end{aligned} \tag{121}$$

By the direct calculation for the integral, we have

$$\int_\delta^B\sqrt{\log\left(\frac{2eMB}{\varepsilon\mathrm{Pdim}(D\widetilde{\Phi})}\right)}\,\mathrm{d}\varepsilon \le B\sqrt{\log\left(\frac{2eMB}{\delta\mathrm{Pdim}(D\widetilde{\Phi})}\right)}.$$

Then choosing $\delta = B\left(\frac{\mathrm{Pdim}(D\widetilde{\Phi})}{M}\right)^{\frac{1}{2}} \le B$, we have

$$\mathbf{R}_M(D\widetilde{\Phi}) \le 28B\left(\frac{\mathrm{Pdim}(D\widetilde{\Phi})}{M}\right)^{\frac{1}{2}}\sqrt{\log\left(\frac{2eM}{\mathrm{Pdim}(D\widetilde{\Phi})}\right)}. \tag{122}$$

Therefore, due to Theorem 2, there is a constant C_4 independent with L,N,M such as

$$\mathbf{R}_M(D\widetilde{\Phi}) \le C_4\frac{NL(\log_2 L\log_2 N)^{\frac{1}{2}}}{\sqrt{M}}\log M. \tag{123}$$

$\mathbf{R}_M(\widetilde{\Phi})$ can be estimate in the similar way. Due to Lemma 12, we have that there is a constant $C_5=C_5(B,d,C_1,C_2)$ such that

$$\mathbf{E}\mathcal{R}_S(\boldsymbol{\theta}_D) - \mathcal{R}_D(\boldsymbol{\theta}_D) + \mathbf{E}\mathcal{R}_D(\boldsymbol{\theta}_S) - \mathbf{E}\mathcal{R}_S(\boldsymbol{\theta}_S) \le C_5\frac{NL(\log_2 L\log_2 N)^{\frac{1}{2}}}{\sqrt{M}}\log M. \tag{124}$$

□