RELU NEURAL NETWORK APPROXIMATION TO PIECEWISE CONSTANT FUNCTIONS *

ZHIQIANG CAI†, JUNPYO CHOI†, AND MIN LIU‡

Abstract. This paper studies the approximation property of ReLU neural networks (NNs) to piecewise constant functions with unknown interfaces in bounded regions in $\mathbb{R}^d$. Under the assumption that the discontinuity interface Γ may be approximated by a connected series of hyperplanes with a prescribed accuracy $\varepsilon > 0$, we show that a three-layer ReLU NN is sufficient to accurately approximate any piecewise constant function and establish its error bound. Moreover, if the discontinuity interface is convex, an analytical formula of the ReLU NN approximation with exact weights and biases is provided.

Key words. ReLU neural networks, Deep neural networks, Function approximation, Classification, Singularity of Function

MSC codes. 68T07, 41A25, 41A46

1. Introduction.

For simplicity, consider the d-dimensional unit cube $\Omega = (0,1)^d$ with $d \geq 2$. Let $\{\Omega_1, \Omega_2\}$ be a partition of the domain Ω; that is, Ω_1 and Ω_2 are open and connected subdomains of Ω such that

$$\Omega_1 \cap \Omega_2 = \emptyset \quad \text{and} \quad \bar{\Omega} = \bar{\Omega}_1 \cup \bar{\Omega}_2.$$

Let $\chi(\mathbf{x})$ be a piece-wise constant function defined on Ω given by

$$\chi(\mathbf{x}) = \begin{cases} 0, & \mathbf{x} \in \Omega_1, \\ 1, & \mathbf{x} \in \Omega_2. \end{cases} \tag{1.1}$$

Denote by $\Gamma = \partial\Omega_1 \cap \partial\Omega_2$ the discontinuity interface of $\chi(\mathbf{x})$, where $\partial\Omega_i$ is the boundary of the subdomain Ω_i. In this paper, we assume that the interface Γ is in C^0 and that its $(d-1)$-dimensional measure $|\Gamma|$ is finite.

Functions of the form in (1.1) are encountered in many applications such as classification tasks in data science and linear and nonlinear hyperbolic conservation laws with discontinuous solutions (see, e.g., [1, 13, 5, 7]). Generally, a piecewise constant function has the form

$$\chi(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i \chi_i(\mathbf{x}), \tag{1.2}$$

where α_i is a real number, $\chi_i(\mathbf{x}) = \mathbf{1}_{\Omega_i}(\mathbf{x})$ is the indicator function of a subdomain $\Omega_i \subset \Omega$, and $\{\Omega_i\}_{i=1}^m$ forms a partition of the domain Ω. The partition means that $\{\Omega_1, \ldots, \Omega_m\}$ are open, connected, and disjoint subdomains of Ω and that $\bar{\Omega} = \cup_{i=1}^m \bar{\Omega}_i$. Once we know how to approximate $\chi(\mathbf{x})$ in (1.1) by neural networks (NNs), then approximating (1.2) is a matter of concatenation or parallelization of the NNs (see, e.g., [10]).

A critical component of using NNs as a model is the use of a properly designed architecture (e.g., the number of layers), and carelessly chosen architectures could lead to poor performance regardless of the size of the network (see, e.g., [7, 8]). To efficiently approximate piecewise constant functions with unknown interface location, several practical guidelines on the architecture of NNs have been provided recently (see, e.g., [12, 15, 9, 7]). The first notable work was done by Petersen

*Submitted to the editors DATE.

Funding: This work was supported in part by the National Science Foundation under grant DMS-2110571.

†Department of Mathematics, Purdue University, 150 N. University Street, West Lafayette, IN 47907-2067 (caiz@purdue.edu, choi508@purdue.edu).

‡School of Mechanical Engineering, Purdue University, 585 Purdue Mall, West Lafayette, IN 47907-2088(liu66@purdue.edu).

and Voigtlaender in their 2018 paper [15]. For any prescribed accuracy $\varepsilon > 0$, if the discontinuity interface Γ is in C^β with $\beta > 0$, they showed that there exists a NN function $\mathcal{N}(\mathbf{x})$, generated by a ReLU NN with at most $(3 + \lceil \log_2 \beta \rceil)(11 + 2\beta/d)$ layers and at most $c\varepsilon^{-p(d-1)/\beta}$ nonzero weights for some constant $c > 0$, such that

$$\|\chi - \mathcal{N}\|_{L^p(\Omega)} \le \varepsilon. \tag{1.3}$$

In the case that Γ can locally be parametrized by functions of Barron-type, it was proved in [9] that for every $N \in \mathbb{N}$, there exists a NN function $\mathcal{N}(\mathbf{x})$, generated by a four-layer ReLU NN with a total of $\mathcal{O}(d + N)$ neurons, such that

$$\|\chi - \mathcal{N}\|_{L^p(\Omega)} \le C\, d^{\frac{3}{2p}} N^{-\frac{\alpha}{2p}},$$

where C and α are positive constants independent of N. Here, the magnitude of the weights and biases can be chosen to be $\mathcal{O}(d + N^{1/2})$.

Recently, we studied this problem in [7] through an explicit construction based on the *two-layer* ReLU approximation $p(\mathbf{x})$ in Lemma 3.2 of [6]. Under the assumption that the interface Γ may be approximated such that there exists a region of ε width containing the interface, we were able to construct a continuous piecewise linear (CPWL) function with a sharp transition layer of ε width whose approximation to the piecewise constant function $\chi(\mathbf{x})$ has the approximation accuracy ε. Combining with the main results in [2], this indicates that a ReLU NN with at most $\lceil \log_2(d+1) \rceil + 1$ layers is sufficient to achieve the prescribed accuracy ε. However, [7] does not provide an estimate of the minimum number of neurons in each layer.

The purpose of this paper is to address the following two questions:

(1) What is the minimum number of hidden-layers of a ReLU NN in order to approximate a piecewise constant function with the prescribed accuracy?
(2) How many neurons per each hidden-layer are needed?

Under the assumption that the interface Γ may be approximated by a connected series of hyperplanes with a prescribed accuracy $\varepsilon > 0$ (see Figure 1(b)), we show that a *three-layer* (two-hidden-layer) ReLU NN is sufficient and necessary to accurately approximate the piecewise constant function $\chi(\mathbf{x})$, in any dimensions, with an error bound of $\mathcal{O}(\varepsilon^{1/p})$ in the $L^p(\Omega)$ norm (see Theorem 3.2). Again, this is done through an explicit construction based on a novel *three-layer* ReLU NN approximation (see, e.g, $\mathcal{N}(\mathbf{x})$ in (4.2) when the interface is a hyperplane). Moreover, the number of neurons at the first hidden-layer and their locations depend on the hyperplanes used for approximating the interface and the number of neurons of the second hidden-layer depends on convexity of the interface.

For classification problems or partial differential equations with a discontinuous solution, our approximation results would provide a guideline on the choice of ReLU NN architectures and on initialization for any training algorithm. It is well-known that initialization is critical for success of any optimization/iterative/training scheme when the resulting discrete problem is a non-convex optimization.

The remainder of the paper is organized as follows. Three-layer ReLU NN functions with relevant concepts and terminology are described in Section 2. Then in Section 3, we describe how to approximate the interface Γ with necessary assumptions, and state the main result of the approximation theory by three-layer ReLU NN functions. The proof of a lemma for the theorem is provided in Section 4. Finally, multiple examples with $d \ge 2$ are given in Section 5 to confirm our theoretical findings.

2. Three-layer ReLU neural network functions.

In this paper, we will restrict our attention to three-layer (two-hidden-layer) neural network functions that are scalar-valued. A function $\mathcal{N} : \mathbb{R}^d \to \mathbb{R}$ is a three-layer neural network (NN) function if the function $\mathcal{N}$ has a representation as a composition of 3 functions $\mathbf{x}^{(l)} : \mathbb{R}^{n_{l-1}} \to \mathbb{R}^{n_l}$ $(n_0 = d,\, n_3 = 1)$ for $l = 1, 2, 3$:

$$\mathcal{N} = \mathbf{x}^{(3)} \circ \mathbf{x}^{(2)} \circ \mathbf{x}^{(1)}, \tag{2.1}$$

where $\mathbf{x}^{(3)}$ is affine linear, and $\mathbf{x}^{(2)}$ and $\mathbf{x}^{(1)}$ are affine linear with a function $\sigma : \mathbb{R} \to \mathbb{R}$, called an activation function, applied to each component of the functions. Such a function is called a d–n_1–n_2–1 NN function.

As the activation function, we use the rectified linear unit (ReLU):

$$\sigma(t) = \mathrm{ReLU}(t) := \max\{0, t\} = \begin{cases} 0, & \text{if } t \le 0, \\ t, & \text{if } t > 0, \end{cases}$$

and refer to such a three-layer NN function as a three-layer (two-hidden-layer) ReLU NN function. Therefore, the collection of all three-layer ReLU NN functions from $\mathbb{R}^d$ to $\mathbb{R}$ is the collection of all functions $\mathcal{N} : \mathbb{R}^d \to \mathbb{R}$ defined by

$$\mathcal{N}(\mathbf{x}) = \boldsymbol{\omega}^{(3)}\sigma\left(\boldsymbol{\omega}^{(2)}\sigma\left(\boldsymbol{\omega}^{(1)}\mathbf{x} - \mathbf{b}^{(1)}\right) - \mathbf{b}^{(2)}\right) - \mathbf{b}^{(3)},$$

where for each $l = 1, 2, 3$, $\boldsymbol{\omega}^{(l)} \in \mathbb{R}^{n_l \times n_{l-1}}$, $\mathbf{b}^{(l)} \in \mathbb{R}^{n_l}$ for $n_l, n_{l-1} \in \mathbb{N}$. We may also assume that each row of the matrix $\boldsymbol{\omega}^{(1)}$ has unit length by adjusting the entries of $\boldsymbol{\omega}^{(2)}$ and $\mathbf{b}^{(1)}$ (see, e.g., [10]).

We will follow the same terminology in [7]. Finally, in the numerical examples in this paper, as in [7], we will see the breaking hyperplanes of the first- and second-(hidden-) layers, which are defined as follows. For $l = 1, 2$, let

$$\boldsymbol{\omega}^{(l)} = (\mathbf{w}_1^{(l)}, \dots, \mathbf{w}_{n_l}^{(l)})^T \in \mathbb{R}^{n_l \times n_{l-1}}, \quad \text{and} \quad \mathbf{b}^{(l)} = (b_1^{(l)}, \dots, b_{n_l}^{(l)})^T.$$

Then the first- (hidden-) layer breaking hyperplanes are for $i = 1, \dots, n_1$,

$$P_i^{(1)} = \left\{\mathbf{x} \in \mathbb{R}^d : \mathbf{w}_i^{(1)}\mathbf{x} - b_i^{(1)} = 0\right\},$$

and the second- (hidden-) layer breaking (poly-) hyperplanes are for $i = 1, \dots, n_2$,

$$P_i^{(2)} = \left\{\mathbf{x} \in \mathbb{R}^d : \mathbf{w}_i^{(2)}\sigma\left(\boldsymbol{\omega}^{(1)}\mathbf{x} - \mathbf{b}^{(1)}\right) - b_i^{(2)} = 0\right\}.$$

ReLU NN functions are continuous piecewise linear with respect to the partition of $\Omega \subset \mathbb{R}^d$ determined by the breaking hyperplanes. The constructions of approximations to piecewise constant functions in this paper will be better understood with the help of breaking hyperplanes.

3. Main results.

Let $\Gamma = \partial\Omega_1 \cap \partial\Omega_2$ be an interface in C^0. For any given $\varepsilon > 0$, assume that there exists a connected series of hyperplanes

$$\mathbf{a}_i \cdot \mathbf{x} - b_i = 0 \quad \text{for } i = 1, \dots, n \tag{3.1}$$

approximating the interface Γ such that the hyperplanes divide the domain Ω by a partition $\left\{\hat{\Omega}_1, \hat{\Omega}_2\right\}$ (see Figures 1(a), 1(b), and 1(c)) and that

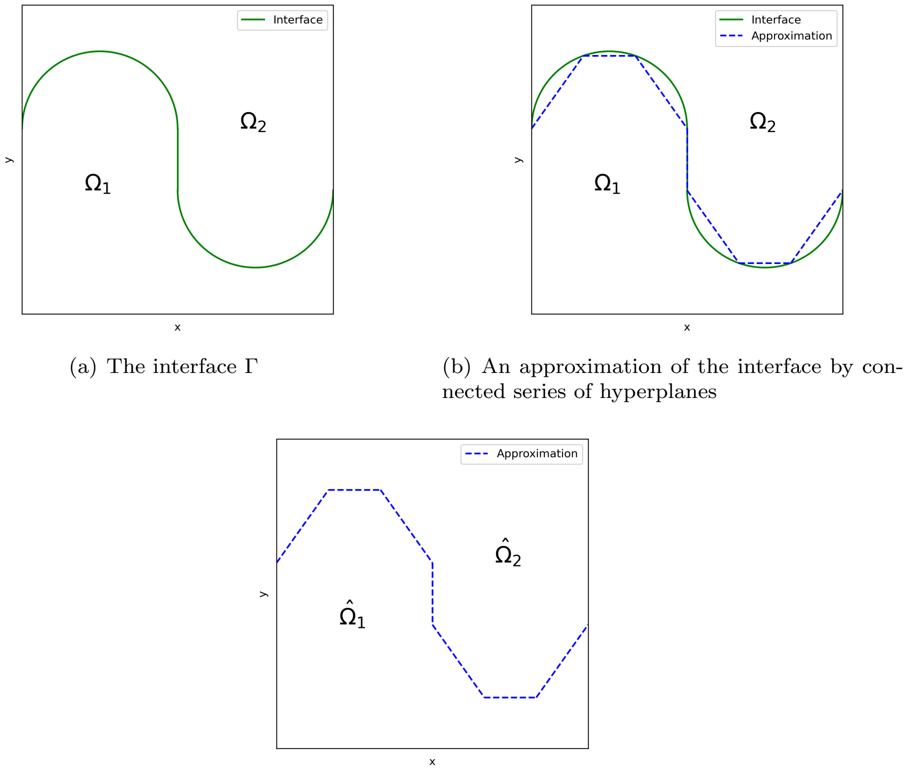
$$\left|\Omega_1 \setminus \hat{\Omega}_1\right| + \left|\Omega_2 \setminus \hat{\Omega}_2\right| \le \varepsilon, \tag{3.2}$$

where $\left|\Omega_i \setminus \hat{\Omega}_i\right|$ is the d-dimensional measure of $\Omega_i \setminus \hat{\Omega}_i$. Let $\hat{\chi}$ be the indicator function of the subdomain $\hat{\Omega}_2$, i.e.,

$$\hat{\chi}(\mathbf{x}) = \begin{cases} 0, & \mathbf{x} \in \hat{\Omega}_1, \\ 1, & \mathbf{x} \in \hat{\Omega}_2. \end{cases} \tag{3.3}$$

Then it is easy to see that (3.2) implies that

$$\|\chi - \hat{\chi}\|_{L^p(\Omega)} = \left(\left|\Omega_1 \setminus \hat{\Omega}_1\right| + \left|\Omega_2 \setminus \hat{\Omega}_2\right|\right)^{1/p} \le \varepsilon^{1/p}. \tag{3.4}$$



(a) The interface Γ

(b) An approximation of the interface by connected series of hyperplanes

(c) $\hat{\chi}$

FIG. 1. *An approximation of the interface* Γ

LEMMA 3.1. *Let $\hat{\Gamma} = \partial\hat{\Omega}_1 \cap \partial\hat{\Omega}_2$. There exists a d–n_1–n_2–1 ReLU NN function $\mathcal{N}$ such that*

$$\|\hat{\chi} - \mathcal{N}\|_{L^p(\Omega)} \le C(|\hat{\Gamma}|)\,\varepsilon^{1/p}, \tag{3.5}$$

where n_1 and n_2 are integers depending on, respectively, the number of the hyperplanes and convexity of $\hat{\Gamma}$, and $C(|\hat{\Gamma}|)$ is a positive constant depending on the $(d-1)$-dimensional measure of the interface $|\hat{\Gamma}|$.

Proof. The proof of the lemma is provided in Section 4. ∎

THEOREM 3.2. *Under the assumption on the interface Γ, there exists a d–n_1–n_2–1 ReLU NN function $\mathcal{N}$ such that*

$$\|\chi - \mathcal{N}\|_{L^p(\Omega)} \le C(|\hat{\Gamma}|)\,\varepsilon^{1/p}, \tag{3.6}$$

where n_1 and n_2 are integers depending on, respectively, the number of the hyperplanes and convexity of $\hat{\Gamma}$, and $C(|\hat{\Gamma}|)$ is a positive constant depending on the $(d-1)$-dimensional measure of the interface $|\hat{\Gamma}|$.

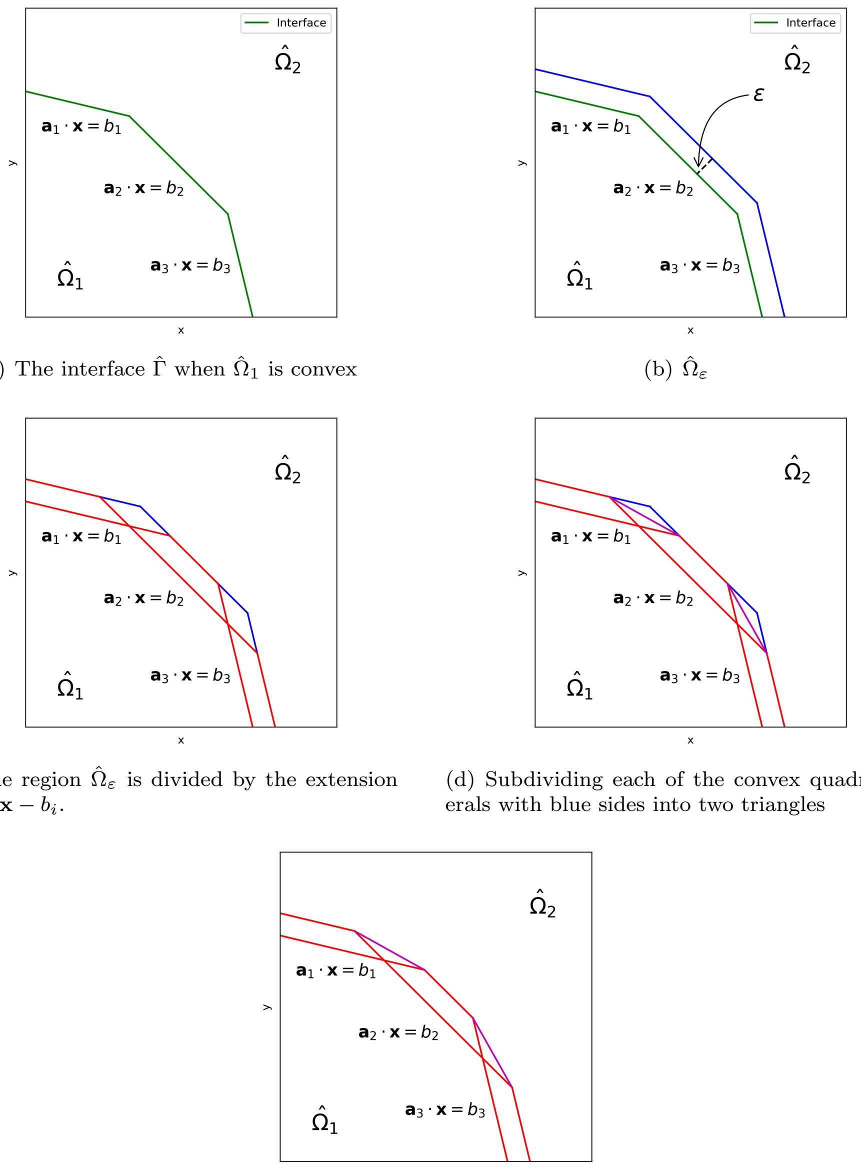
Proof. It follows from (3.4), Lemma 3.1 and the triangle inequality that there exists a d–n_1–n_2–1 ReLU NN function $\mathcal{N}$ for some $n_1, n_2 \in \mathbb{N}$ such that

$$\|\chi - \mathcal{N}\|_{L^p(\Omega)} = \|\chi - \hat{\chi} + \hat{\chi} - \mathcal{N}\|_{L^p(\Omega)} \le \|\chi - \hat{\chi}\|_{L^p(\Omega)} + \|\hat{\chi} - \mathcal{N}\|_{L^p(\Omega)} \le \Big(C(|\hat{\Gamma}|) + 1\Big)\,\varepsilon^{1/p}, \tag{3.7}$$

which completes the proof of the theorem. ∎

4. Proof of Lemma 3.1 . This section proves Lemma 3.1 in Subsection 4.1 and Subsection 4.2 when the subdomain $\hat{\Omega}_1$ is convex and non-convex, respectively.

4.1. Convex $\hat{\Omega}_1$. This section shows the validity of Lemma 3.1 in a special case that the subdomain $\hat{\Omega}_1$ is convex (see Figure 2(a)).



(a) The interface $\hat{\Gamma}$ when $\hat{\Omega}_1$ is convex

(b) $\hat{\Omega}_\varepsilon$

(c) The region $\hat{\Omega}_\varepsilon$ is divided by the extension of $\mathbf{a}_i \cdot \mathbf{x} - b_i$.

(d) Subdividing each of the convex quadrilaterals with blue sides into two triangles

(e) Removing the triangles not adjacent to the interface $\hat{\Gamma}$

FIG. 2. *The subdomain $\hat{\Omega}_1$ is convex.*

Without loss of generality, assume that the normal vectors $\mathbf{a}_i$ of the hyperplanes are the unit vectors and point toward $\hat{\Omega}_2$. Then we approximate the unit step function $\hat{\chi}(\mathbf{x})$ in (3.3) by the following ReLU NN function

$$\mathcal{N}(\mathbf{x}) = 1 - \sigma\left(1 - \frac{1}{\varepsilon}\sum_{i=1}^{n}\sigma(\mathbf{a}_i\cdot\mathbf{x} - b_i)\right). \tag{4.1}$$

The $\mathcal{N}(\mathbf{x})$ is a d–n–1–1 ReLU NN function.

When the interface $\hat{\Gamma}$ is a hyperplane $\mathbf{a}\cdot\mathbf{x} - b = 0$ in $\mathbb{R}^d$, i.e., $n = 1$, the $\mathcal{N}(\mathbf{x})$ has the form

$$\mathcal{N}(\mathbf{x}) = 1 - \sigma\left(1 - \frac{1}{\varepsilon}\sigma(\mathbf{a}\cdot\mathbf{x} - b)\right). \tag{4.2}$$

The second term of $\mathcal{N}(\mathbf{x})$, a three-layer ReLU NN function, is a ramp function that equals negative one in $\hat{\Omega}_1$ and vanishes in $\hat{\Omega}_2 \setminus Y_\varepsilon$, where $Y_\varepsilon = \{\mathbf{x}\in\Omega : 0 < \mathbf{a}\cdot\mathbf{x} - b < \varepsilon\}$ is a strip with ε-width. It is then easy to see that

$$\hat{\chi}(\mathbf{x}) - \mathcal{N}(\mathbf{x}) = \begin{cases} 0, & \mathbf{x}\in\hat{\Omega}_1\cup\left(\hat{\Omega}_2\setminus Y_\varepsilon\right),\\ \sigma\left(1 - \frac{1}{\varepsilon}\sigma(\mathbf{a}\cdot\mathbf{x} - b)\right), & \mathbf{x}\in Y_\varepsilon,\end{cases}$$

which, together with a simple calculation, implies the upper bound in (3.5).

Now, we consider the case $n \geq 2$. For simplicity of presentation, the proof of the error bound in (3.5) is carried out in two dimensions $d = 2$. Denote by $\hat{\Omega}_\varepsilon \subset \Omega$ the region produced by translating $\mathbf{a}_i\cdot\mathbf{x} - b_i = 0$ toward $\hat{\Omega}_2$ along $\mathbf{a}_i$ by ε (see Figure 2(b)). By extending the line segments $\mathbf{a}_i\cdot\mathbf{x} - b_i = 0$, we partition the region $\hat{\Omega}_\varepsilon$ into convex subregions (see Figure 2(c)).

The subregions of the first type are denoted by $\{\Upsilon_{1i}\}_{i=1}^n$, where Υ_{1i} is the subregion bounded by the line $\mathbf{a}_i\cdot\mathbf{x} - b_i = 0$, its translated line $\mathbf{a}_i\cdot\mathbf{x} - b_i = \varepsilon$, and two neighboring lines or one neighboring line and the boundary of Ω (the convex quadrilaterals with red sides in Figure 2(c)). More precisely, we have that

$$\Upsilon_{11} = \left\{\mathbf{x}\in\hat{\Omega}_\varepsilon : 0 < \mathbf{a}_1\cdot\mathbf{x} - b_1 < \varepsilon \text{ and } \mathbf{a}_2\cdot\mathbf{x} - b_2 < 0\right\},$$

$$\Upsilon_{1n} = \left\{\mathbf{x}\in\hat{\Omega}_\varepsilon : 0 < \mathbf{a}_n\cdot\mathbf{x} - b_n < \varepsilon \text{ and } \mathbf{a}_{n-1}\cdot\mathbf{x} - b_{n-1} < 0\right\},$$

and that for $i = 2, \ldots, n-1$

$$\Upsilon_{1i} = \left\{\mathbf{x}\in\hat{\Omega}_\varepsilon : 0 < \mathbf{a}_i\cdot\mathbf{x} - b_i < \varepsilon,\ \mathbf{a}_{i-1}\cdot\mathbf{x} - b_{i-1} < 0, \text{ and } \mathbf{a}_{i+1}\cdot\mathbf{x} - b_{i+1} < 0\right\}.$$

Notice that $\hat{\Omega}_\varepsilon \setminus (\cup_{i=1}^n \Upsilon_{1i})$ consists of $n-1$ convex quadrilaterals (with blue sides in Figure 2(c)). Subdividing each of these convex quadrilaterals into two triangles (see Figure 2(d)) and removing the triangles not adjacent to the interface $\hat{\Gamma}$ (see Figure 2(e)), the remaining triangles are denoted by $\{\Upsilon_{2i}\}_{i=1}^{n-1}$, where Υ_{2i} is given by

$$\Upsilon_{2i} = \left\{\mathbf{x}\in\hat{\Omega}_\varepsilon : \mathbf{a}_j\cdot\mathbf{x} - b_j > 0 \text{ for } j = i, i+1, \text{ and } (\mathbf{a}_i + \mathbf{a}_{i+1})\cdot\mathbf{x} - (b_i + b_{i+1}) < \varepsilon\right\}.$$

We then have the following lemma.

Lemma 4.1. *Let $\mathcal{N}(\mathbf{x})$ be the three-layer ReLU NN function defined in* (4.1)*, then we have*

$$\hat{\chi}(\mathbf{x}) - \mathcal{N}(\mathbf{x}) = \begin{cases} 0, & \mathbf{x}\in\Omega\setminus\left(\bigcup_{j=1}^{2}\bigcup_{i=1}^{n+1-j}\Upsilon_{ji}\right),\\ \hat{\chi}(\mathbf{x}) - \frac{1}{\varepsilon}(\mathbf{a}_i\cdot\mathbf{x} - b_i), & \mathbf{x}\in\Upsilon_{1i} \text{ for } i = 1,\ldots n,\\ \hat{\chi}(\mathbf{x}) - \frac{1}{\varepsilon}\left[(\mathbf{a}_i + \mathbf{a}_{i+1})\cdot\mathbf{x} - (b_i + b_{i+1})\right], & \mathbf{x}\in\Upsilon_{2i} \text{ for } i = 1,\ldots n-1.\end{cases} \tag{4.3}$$

Proof. Let

$$\hat{\Omega}_3 = \hat{\Omega}_2 \setminus \left(\bigcup_{j=1}^{2} \bigcup_{i=1}^{n+1-j} \Upsilon_{ji} \right).$$

Since $\mathbf{a}_i$ points toward $\hat{\Omega}_2$, clearly, we have $\sigma(\mathbf{a}_i \cdot \mathbf{x} - b_i) = 0$ for all $\mathbf{x} \in \hat{\Omega}_1$ and $i = 1, \dots, n$. This implies

$$\mathcal{N}(\mathbf{x}) = 1 - \sigma(1) = 0, \quad \forall\, \mathbf{x} \in \hat{\Omega}_1. \tag{4.4}$$

Clearly, we have

$$1 - \frac{1}{\varepsilon}\sum_{i=1}^{n} \sigma(\mathbf{a}_i \cdot \mathbf{x} - b_i) = \begin{cases} 1 - \dfrac{1}{\varepsilon}(\mathbf{a}_i \cdot \mathbf{x} - b_i), & \mathbf{x} \in \Upsilon_{1i}, \\[2ex] 1 - \dfrac{1}{\varepsilon}\left[(\mathbf{a}_i + \mathbf{a}_{i+1}) \cdot \mathbf{x} - (b_i + b_{i+1})\right], & \mathbf{x} \in \Upsilon_{2i}. \end{cases}$$

It is easy to see that

$$1 - \frac{1}{\varepsilon}(\mathbf{a}_i \cdot \mathbf{x} - b_i) \begin{cases} > 0, & 0 < \mathbf{a}_i \cdot \mathbf{x} - b_i < \varepsilon, \\ \le 0, & \varepsilon \le \mathbf{a}_i \cdot \mathbf{x} - b_i \end{cases}$$

and that similar inequalities hold for $1 - \frac{1}{\varepsilon}\left[(\mathbf{a}_i + \mathbf{a}_{i+1}) \cdot \mathbf{x} - (b_i + b_{i+1})\right]$; furthermore, by the definition of $\hat{\Omega}_3$, we have

$$1 - \frac{1}{\varepsilon}\sum_{i=1}^{n} \sigma(\mathbf{a}_i \cdot \mathbf{x} - b_i) < 0, \quad \forall\, \mathbf{x} \in \hat{\Omega}_3.$$

Now, applying the activation function σ, multiplying by -1, and adding 1 imply

$$\mathcal{N}(\mathbf{x}) = \begin{cases} (\mathbf{a}_i \cdot \mathbf{x} - b_i)/\varepsilon, & \mathbf{x} \in \Upsilon_{1i}, \\ \left[(\mathbf{a}_i + \mathbf{a}_{i+1}) \cdot \mathbf{x} - (b_i + b_{i+1})\right]/\varepsilon, & \mathbf{x} \in \Upsilon_{2i}, \\ 1 & \mathbf{x} \in \hat{\Omega}_3, \end{cases}$$

which, together with (4.4), leads to (4.3). This completes the proof of the lemma. ∎

Proof of Lemma 3.1 *for convex* $\hat{\Omega}_1$. When $\hat{\Omega}_1$ is convex, to show the validity of Lemma 3.1, notice that for all $p \in [1, \infty)$, we have by Lemma 4.1,

$$|\hat{\chi}(\mathbf{x}) - \mathcal{N}(\mathbf{x})|^p = \left|\hat{\chi}(\mathbf{x}) - \frac{1}{\varepsilon}(\mathbf{a}_i \cdot \mathbf{x} - b_i)\right|^p \le 1, \quad \forall\, \mathbf{x} \in \Upsilon_{1i},$$

$$\text{and} \quad |\hat{\chi}(\mathbf{x}) - \mathcal{N}(\mathbf{x})|^p = \left|\hat{\chi}(\mathbf{x}) - \frac{1}{\varepsilon}\left[(\mathbf{a}_i + \mathbf{a}_{i+1}) \cdot \mathbf{x} - (b_i + b_{i+1})\right]\right|^p \le 1, \quad \forall\, \mathbf{x} \in \Upsilon_{2i},$$

which implies

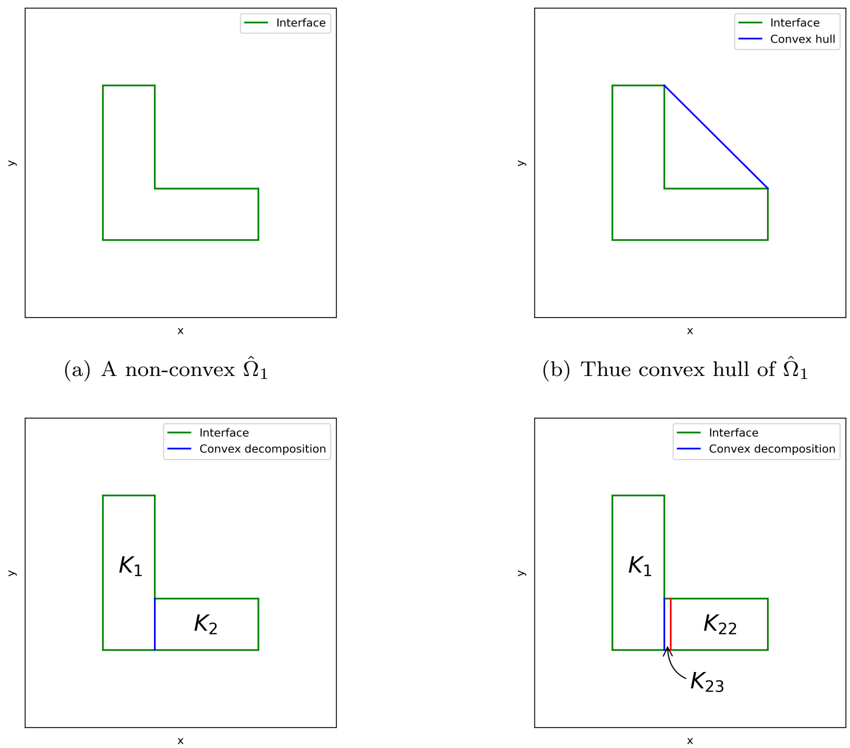
$$\|\hat{\chi} - \mathcal{N}\|^p_{L^p(\Upsilon_{1i})} \le |\Upsilon_{1i}| \quad \text{and} \quad \|\hat{\chi} - \mathcal{N}\|^p_{L^p(\Upsilon_{2i})} \le |\Upsilon_{2i}|, \tag{4.5}$$

where $|\Upsilon_{ji}|$ denotes the area of the quadrilateral Υ_{ji}. It follows from (4.3) and (4.5) that

$$\|\hat{\chi} - \mathcal{N}\|^p_{L^p(\Omega)} = \sum_{i=1}^{n} \|\hat{\chi} - \mathcal{N}\|^p_{L^p(\Upsilon_{1i})} + \sum_{i=1}^{n-1} \|\hat{\chi} - \mathcal{N}\|^p_{L^p(\Upsilon_{1i})} \le \sum_{i=1}^{n} |\Upsilon_{1i}| + \sum_{i=1}^{n-1} |\Upsilon_{2i}| \le |\hat{\Omega}_\varepsilon|, \tag{4.6}$$

which, together with the fact that $|\hat{\Omega}_\varepsilon| \le C\,|\hat{\Gamma}|\,\varepsilon$ for a positive constant C, implies the error bound in (3.5). This completes the proof of Lemma 3.1. ∎

4.2. Non-Convex $\hat{\Omega}_1$. This section shows the validity of Lemma 3.1 when $\hat{\Omega}_1$ is non-convex (see, e.g., Figure 3(a)). Our proof is again through explicit constructions. Specifically, we present two approaches: one is based on the convex hull of $\hat{\Omega}_1$ (see Subsubsection 4.2.1) and the other uses a convex decomposition of $\hat{\Omega}_1$ (see Subsubsection 4.2.2).



(a) A non-convex $\hat{\Omega}_1$ (b) Thue convex hull of $\hat{\Omega}_1$

(c) A convex decomposition of $\hat{\Omega}_1$ (d) A subset of K_2

FIG. 3. *The subdomain $\hat{\Omega}_1$ is non-convex.*

4.2.1. Convex hull. Let

$$\Omega_1^{(1)} = \hat{\Omega}_1 \cup \left(\cup_{i=1}^{k} K_i\right), \tag{4.7}$$

be the convex hull of $\hat{\Omega}_1$ (see Figure 3(b)) generated by a convex hull algorithm (see, e.g., [14, 4, 3]), where K_i are polytopes and pairwise disjoint. Without loss of generality, we assume that all K_i $(i = 1, \ldots, k)$ are convex. Otherwise, the procedure presented in this section may be applied to non-convex K_is for the indicator functions $\mathbf{1}_{\Omega \setminus K_i}(\mathbf{x})$ of the subdomains $\Omega \setminus K_i$. Note that the procedure may be needed for several times recursively.

Let $\hat{\chi}_0(\mathbf{x})$ be the unit step function defined on the convex hull $\Omega_1^{(1)}$ of the non-convex subdomain $\hat{\Omega}_1$:

$$\hat{\chi}_0(\mathbf{x}) = \begin{cases} 0, & \mathbf{x} \in \Omega_1^{(1)} \subset \Omega, \\ 1, & \mathbf{x} \in \Omega \setminus \Omega_1^{(1)}, \end{cases} \tag{4.8}$$

then its discontinuity interface is $\hat{\Gamma}_0 = \partial\Omega_1^{(1)} \cap \partial\left(\Omega \setminus \Omega_1^{(1)}\right)$ consisting of $n_{1,0}$ faces. As proved in Subsection 4.1 (see (4.6)), there exists a d–$n_{1,0}$–1–1 ReLU NN function approximation $\mathcal{N}_0(\mathbf{x})$ such that

$$\|\hat{\chi}_0 - \mathcal{N}_0\|_{L^p(\Omega)} = \|\hat{\chi}_0 - \mathcal{N}_0\|_{L^p(\hat{\Omega}_{\varepsilon,0})} \le \left|\hat{\Omega}_{\varepsilon,0}\right|^{1/p}, \tag{4.9}$$

where $\hat{\Omega}_{\varepsilon,0}$ is the region with ε-width containing the interface $\hat{\Gamma}_0$ by translating the faces of $\hat{\Gamma}_0$ towards the subdomain $\Omega \setminus \Omega_1^{(1)}$.

For each convex polytope K_i $(i = 1, \ldots, k)$ in (4.7) having $n_{1,i}$ faces, let $\hat{\chi}_i(\mathbf{x})$ be the unit step function defined on K_i:

$$\hat{\chi}_i(\mathbf{x}) = \begin{cases} 1, & \mathbf{x} \in K_i \subset \Omega, \\ 0, & \mathbf{x} \in \Omega \setminus K_i. \end{cases}$$

Define the following d–$n_{1,i}$–1–1 ReLU NN function

$$\mathcal{N}_i(\mathbf{x}) = \sigma\left(1 - \frac{1}{\varepsilon}\sum_{j=1}^{n_{1,i}} \sigma(\mathbf{a}_{i,j}\cdot\mathbf{x} - b_{i,j})\right), \tag{4.10}$$

where the hyperplanes $\mathbf{a}_{i,j}\cdot\mathbf{x} - b_{i,j} = 0$ $(j = 1, \ldots, n_{1,i})$ are the faces of ∂K_i with $\mathbf{a}_{i,j}$ pointing toward $\Omega \setminus K_i$. In a similar fashion as in Subsection 4.1, it is easy to check that

$$\|\hat{\chi}_i - \mathcal{N}_i\|_{L^p(\Omega)} = \|\hat{\chi}_i - \mathcal{N}_i\|_{L^p(\hat{\Omega}_{\varepsilon,i})} \le \left|\hat{\Omega}_{\varepsilon,i}\right|^{1/p}, \tag{4.11}$$

where $\hat{\Omega}_{\varepsilon,i}$ is a region having ε-width.

Now, we are ready to define the following d–n_1–n_2–1 ReLU NN function:

$$\mathcal{N}(\mathbf{x}) = \mathcal{N}_0(\mathbf{x}) + \sum_{i=1}^{k} \mathcal{N}_i(\mathbf{x}), \tag{4.12}$$

where $\mathcal{N}_0(\mathbf{x})$ is given in a similar fashion as in (4.1).

Proof of Lemma 3.1 *for non-convex* $\hat{\Omega}_1$. Note that

$$\hat{\chi} = \sum_{i=0}^{k} \hat{\chi}_i.$$

Then it follows from (4.12), the triangle inequality, (4.9), and (4.11) that

$$\|\hat{\chi} - \mathcal{N}\|_{L^p(\Omega)} \le \sum_{i=0}^{k} \|\hat{\chi}_i - \mathcal{N}_i\|_{L^p(\Omega)} \le \sum_{i=0}^{k} \left|\hat{\Omega}_{\varepsilon,i}\right|^{1/p}, \tag{4.13}$$

which, together with the fact that $\left|\hat{\Omega}_{\varepsilon,i}\right| \le C_i \left|\hat{\Gamma}_i\right| \varepsilon$ for a positive constant C_i, implies the error bound in (3.5). Here $\hat{\Gamma}_i = \partial K_i \cap \partial\left(\Omega \setminus K_i\right)$ for $i = 1, \ldots, k$ This completes the proof of Lemma 3.1. ∎

4.2.2. Convex decomposition. Assume that $\hat{\Omega}_1$ has a convex decomposition (see, e.g., [11]) given by

$$\hat{\Omega}_1 = \bigcup_{i=1}^{l} K_i,$$

where all K_i are convex polytopes. For simplicity of presentation, assume that $l = 2$, i.e., the decomposition has only two convex polytopes: $\hat{\Omega}_1 = K_1 \cup K_2$ (see Figure 3(c)).

Denote the indicator function of the subdomain K_1 by

$$\hat{\chi}_1(\mathbf{x}) = \begin{cases} 1, & \mathbf{x} \in K_1 \subset \Omega, \\ 0, & \mathbf{x} \in \Omega \setminus K_1. \end{cases}$$

Let $\mathbf{a} \cdot \mathbf{x} - b = 0$ be the hyperplane that divides $\hat{\Omega}_1$ into K_1 and K_2 (blue line in Figure 3(c)). Assume that $\mathbf{a}$ points toward K_2. Translate $\mathbf{a} \cdot \mathbf{x} - b = 0$ toward K_2 by ε to obtain the hyperplane $\mathbf{a} \cdot \mathbf{x} - b - \varepsilon = 0$ (red line in Figure 3(d)). Partition K_2 by $\{K_{22}, K_{23}\}$ (see Figure 3(d)), where

$$K_{22} = \{\mathbf{x} \in \hat{\Omega}_1 : \varepsilon < \mathbf{a} \cdot \mathbf{x} - b\} \quad \text{and} \quad K_{23} = \{\mathbf{x} \in \hat{\Omega}_1 : 0 < \mathbf{a} \cdot \mathbf{x} - b < \varepsilon\}.$$

Denote by $\hat{\chi}_{22}(\mathbf{x})$ and $\hat{\chi}_{23}(\mathbf{x})$ the respective indicator functions of K_{22} and K_{23}:

$$\hat{\chi}_{22}(\mathbf{x}) = \begin{cases} 1, & \mathbf{x} \in K_{22} \subset \Omega, \\ 0, & \mathbf{x} \in \Omega \setminus K_{22}. \end{cases} \quad \text{and} \quad \hat{\chi}_{23}(\mathbf{x}) = \begin{cases} 1, & \mathbf{x} \in K_{23} \subset \Omega, \\ 0, & \mathbf{x} \in \Omega \setminus K_{23}. \end{cases}$$

Assume that polygonal domains K_1 and K_{22} have n_1 and n_2 faces, respectively. In a similar fashion as in (4.10) and (4.11), there exist d–n_1–1–1 and d–n_2–1–1 ReLU NN functions $\mathcal{N}_1$ and $\mathcal{N}_{22}$ such that

$$(4.14) \qquad \begin{cases} \|\hat{\chi}_1 - \mathcal{N}_1\|_{L^p(\Omega)} = \|\hat{\chi}_1 - \mathcal{N}_1\|_{L^p(\hat{\Omega}_{\varepsilon,1})} \le \left|\hat{\Omega}_{\varepsilon,1}\right|^{1/p} \\ \text{and} \ \|\hat{\chi}_{22} - \mathcal{N}_{22}\|_{L^p(\Omega)} = \|\hat{\chi}_{22} - \mathcal{N}_{22}\|_{L^p(\hat{\Omega}_{\varepsilon,22})} \le \left|\hat{\Omega}_{\varepsilon,22}\right|^{1/p}, \end{cases}$$

where $\hat{\Omega}_{\varepsilon,1}$ and $\hat{\Omega}_{\varepsilon,22}$ are regions having ε-width. Clearly, there exist positive constants C_1, C_{22}, and C_{23} such that

$$(4.15) \qquad \left|\hat{\Omega}_{\varepsilon,1}\right| \le C_1 \left|\hat{\Gamma}_1\right| \varepsilon, \quad \left|\hat{\Omega}_{\varepsilon,22}\right| \le C_{22} \left|\hat{\Gamma}_{22}\right| \varepsilon, \quad \text{and} \quad \left|K_{23}\right| \le C_{23} \left|\hat{\Gamma}_{23}\right| \varepsilon$$

where $\hat{\Gamma}_1$, $\hat{\Gamma}_{22}$, and $\hat{\Gamma}_{23}$ are the boundaries of K_1, K_{22}, and K_{23}.

Proof of Lemma 3.1 *for non-convex* $\hat{\Omega}_1$. Let

$$\mathcal{N}(\mathbf{x}) = 1 - \mathcal{N}_1(\mathbf{x}) - \mathcal{N}_{22}(\mathbf{x}).$$

Note that

$$\hat{\chi}(\mathbf{x}) - \mathcal{N}(\mathbf{x}) = (\hat{\chi}_1 - \mathcal{N}_1) + (\hat{\chi}_{22} - \mathcal{N}_{22}) + \hat{\chi}_{23}.$$

It follows from the triangle inequality, (4.14), and (4.15) that

$$\begin{aligned} \|\hat{\chi} - \mathcal{N}\|_{L^p(\Omega)} &\le \|\hat{\chi}_1 - \mathcal{N}_1\|_{L^p(\Omega)} + \|\hat{\chi}_{22} - \mathcal{N}_{22}\|_{L^p(\Omega)} + \|\hat{\chi}_{23}\|_{L^p(\Omega)} \\ (4.16) \qquad &\le \left|\hat{\Omega}_{\varepsilon,1}\right|^{1/p} + \left|\hat{\Omega}_{\varepsilon,22}\right|^{1/p} + |K_{23}|^{1/p} \le C\,\varepsilon. \end{aligned}$$

This completes the approximation. ∎

Remark 4.2. The ε-width region K_{23} is a subdomain of $\hat{\Omega}_1$, and its boundary contains the hyperplanes $\mathbf{a} \cdot \mathbf{x} - b = 0$ (blue line in Figure 3(c)) and $\mathbf{a} \cdot \mathbf{x} - b - \varepsilon = 0$ (red line in Figure 3(d)) which are not part of the interface $\hat{\Gamma}$. For any $\mathbf{x} \in K_{23}$, it is easy to see that

$$\hat{\chi}(\mathbf{x}) - \mathcal{N}(\mathbf{x}) = \mathcal{N}_1(\mathbf{x}) + \mathcal{N}_{22}(\mathbf{x}) = \left(1 - \frac{1}{\varepsilon}(\mathbf{a} \cdot \mathbf{x} - b)\right) + \left(1 - \frac{1}{\varepsilon}(-\mathbf{a} \cdot \mathbf{x} + b + \varepsilon)\right) = 0,$$

which, together with (4.16), implies that the ReLU NN function $\mathcal{N}(\mathbf{x})$ approximates the discontinuous step function $\hat{\chi}(\mathbf{x})$ without overshooting. Moreover, it is clear from the construction that $\mathcal{N}(\mathbf{x})$ has no oscillation. No overshooting and no oscillation remain true for the ReLU NN approximation $\mathcal{N}(\mathbf{x})$ constructed in Subsubsection 4.2.1.

5. Examples. This section validates our theoretical findings with several examples in $d \geq 2$ dimensions. The first three examples demonstrate Theorem 3.2 for convex $\hat{\Omega}_1$, with the third example extending to the case of $d = 10000$. The final example illustrates a non-convex case using the twp decomposition procedures outlined in Subsection 4.2.

5.1. A two-dimensional circular interface. Let $\Omega = (0,1)^2$,

$$\Omega_1 = \{(x,y) \in \Omega : (x-0.5)^2 + (y-0.5)^2 < 0.25^2\}, \quad \text{and} \quad \Omega_2 = \Omega \setminus \Omega_1.$$

The piecewise constant function $\chi(\mathbf{x})$ is shown in Figure 4(a). The interface Γ is a circle centered at $(0.5, 0.5)$ with a radius of 0.25 (see Figure 4(b)):

$$\Gamma = \{(x,y) \in \Omega : (x-0.5)^2 + (y-0.5)^2 = 0.25^2\}.$$

Consider approximations of the interface Γ by $n = 6$ and 50 line segments (see Figures 4(c) and 4(d)), respectively. The 2–6–1–1 and 2–50–1–1 ReLU NN approximations given in (4.1) with $\varepsilon = 1/25$ and $1/2000$ are shown in Figures 4(e) and 4(f), respectively. Figures 4(g) and 4(h) illustrate the breaking lines of the first and second layers, with the distances between them equal to ε.

5.2. A three-dimensional spherical interface. Let $\Omega = (0,1)^3$,

$$\Omega_1 = \{(x,y,z) \in \Omega : z < \sqrt{0.7^2 - x^2 - y^2}\}, \quad \text{and} \quad \Omega_2 = \Omega \setminus \Omega_1.$$

The intersection between the piecewise constant function $\chi(x,y,z)$ and the hyperplane $z = 0.205$ is shown in Figure 5(a). The interface Γ is part of a sphere centered at $(0,0,0)$ with a radius of 0.7 (see Figure 5(b)):

$$\Gamma = \{(x,y,z) \in \Omega : x^2 + y^2 + z^2 = 0.7^2\},$$

which is approximated by $n = 9$ and 100 plane segments (see Figures 5(c) and 5(d)), respectively. The 3–9–1–1 and 3–100–1–1 ReLU NN approximations given in (4.1) with $\varepsilon = 1/15$ and $1/100$ are depicted in Figures 5(e) and 5(f), respectively. Figures 5(g) and 5(h) illustrate the first- and second-layer breaking hyperplanes on $z = 0.205$.

5.3. A 10000-dimensional hypercube interface. Let $d = 10000$, $\Omega = (0,1)^d$,

$$\Omega_1 = \{\mathbf{x} = (x_1, \ldots, x_d) \in \Omega : x_1 < 1/2, \ldots, x_d < 1/2\}, \quad \text{and} \quad \Omega_2 = \Omega \setminus \Omega_1.$$

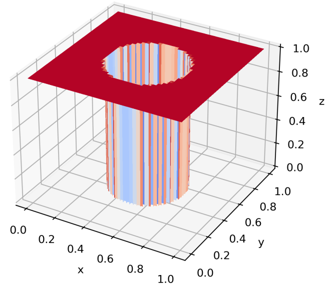
The intersection between the piecewise constant function $\chi(\mathbf{x})$ and the hyperplanes $x_i = 0.255$ for $i = 3, \ldots, 10000$ is shown in Figure 6(a). The interface Γ is the boundary of a hypercube in Ω (see Figure 6(b) for a three-dimensional section of it):

$$\Gamma = \bigcup_{i=1}^{d} \{\mathbf{x} = (x_1, \ldots, x_d) \in \Omega : x_i = 1/2, \text{ and } 0 \leq x_j \leq 1/2 \text{ for } j \neq i\}.$$

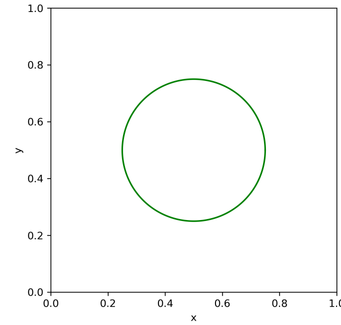
In this example, we can simply take $\hat{\chi} = \chi$. Noting that the hypercube consists of 10000 hyperplanes of the form $x_i - 1/2 = 0$ for $i = 1, \ldots, 10000$, the corresponding NN approximation is

$$\mathcal{N}(\mathbf{x}) = 1 - \sigma\left(1 - \frac{1}{\varepsilon}\sum_{i=1}^{d} \sigma(x_i - 1/2)\right). \tag{5.1}$$

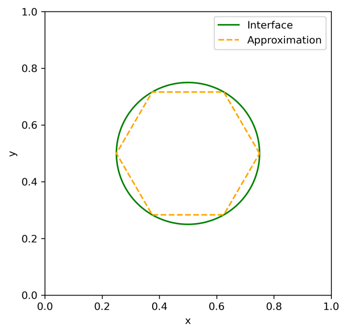
Two sectional views of $\mathcal{N}(\mathbf{x})$ are shown in Figures 6(c) and 6(d) with $\varepsilon = 1/20$ and $1/200$, respectively. Figures 6(e) and 6(f) plot the corresponding breaking hyperplanes.



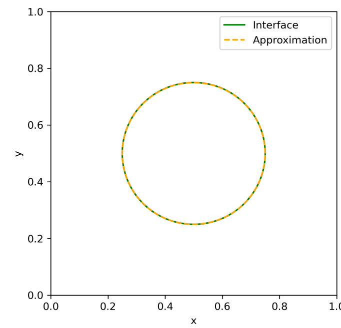
(a) The piecewise constant function $\chi(\mathbf{x})$



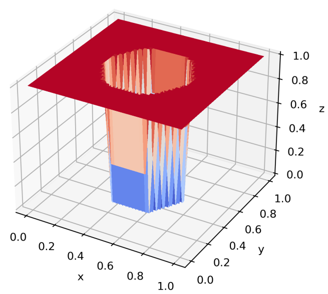
(b) The circular interface



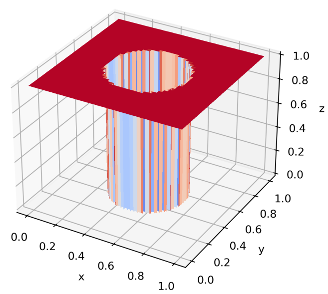
(c) An approximation of the interface by $n = 6$ line segments



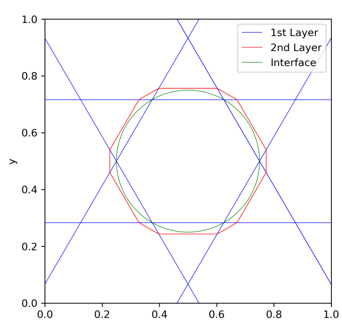
(d) An approximation of the interface by $n = 50$ line segments



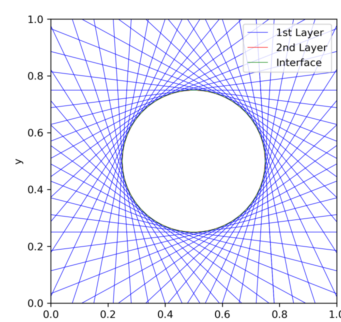
(e) An approximation of $\chi(\mathbf{x})$ by the 2–6–1–1 ReLU NN function in (4.1) with $\varepsilon = 1/25$



(f) An approximation of $\chi(\mathbf{x})$ by the 2–50–1–1 ReLU NN function in (4.1) with $\varepsilon = 1/2000$

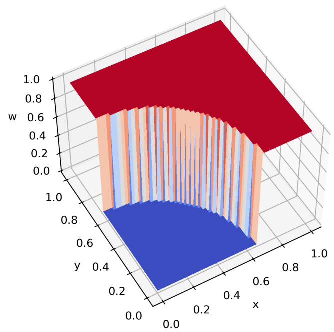


(g) The breaking hyperplanes of the approximation in Figure 4(e)

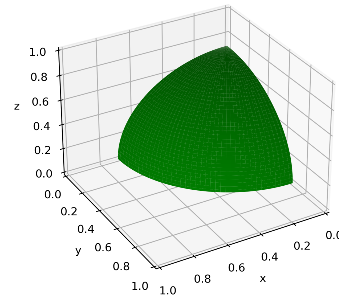


(h) The breaking hyperplanes of the approximation in Figure 4(f)

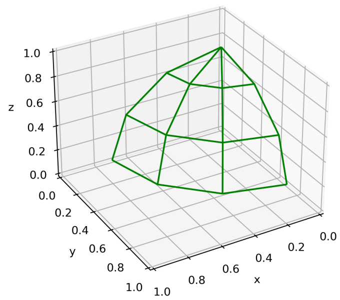
FIG. 4. *A convex example to illustrate Theorem 3.2 for the case* $d = 2$



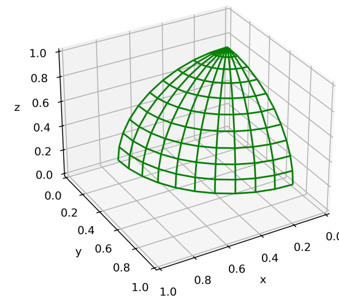
(a) The piecewise constant function $\chi(\mathbf{x})$ on $z = 0.205$



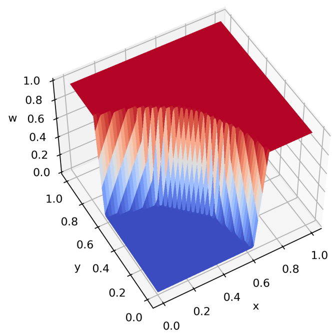
(b) The spherical interface



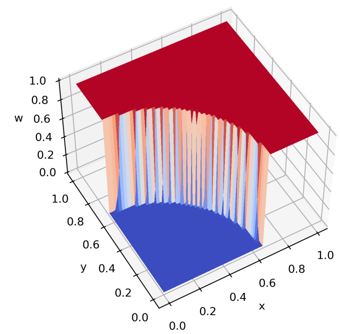
(c) An approximation of the interface by $n = 9$ plane segments



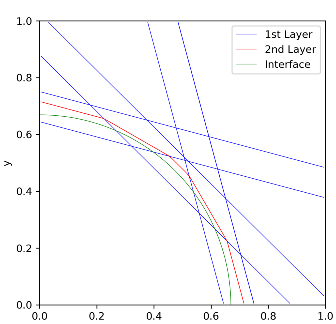
(d) An approximation of the interface by $n = 100$ plane segments



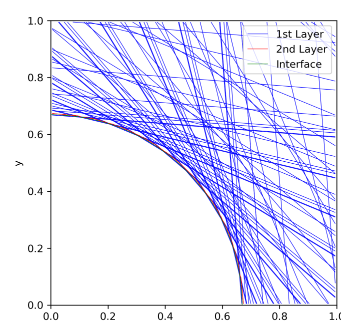
(e) An approximation of $\chi(\mathbf{x})$ by the 3–9–1–1 ReLU NN function in (4.1) with $\varepsilon = 1/15$ on $z = 0.205$



(f) An approximation of $\chi(\mathbf{x})$ by the 3–100–1–1 ReLU NN function in (4.1) with $\varepsilon = 1/100$ on $z = 0.205$

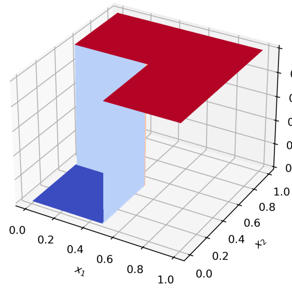


(g) The breaking hyperplanes of the approximation in Figure 5(e) on $z = 0.205$

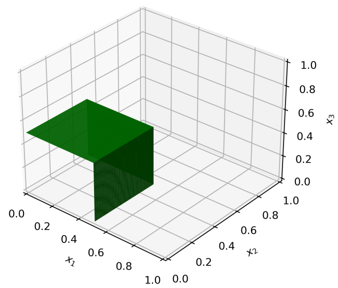


(h) The breaking hyperplanes of the approximation in Figure 5(f) on $z = 0.205$

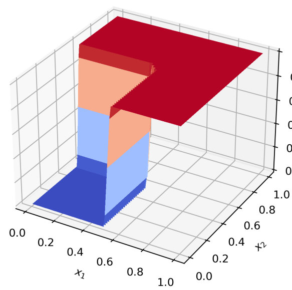
FIG. 5. *A convex example to illustrate Theorem 3.2 for the case* $d = 3$



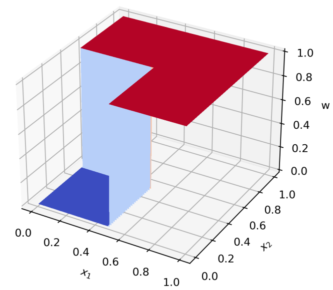
(a) The piecewise constant function $\chi(\mathbf{x})$ on $x_i = 0.255$ for $i = 3, \ldots, 10000$



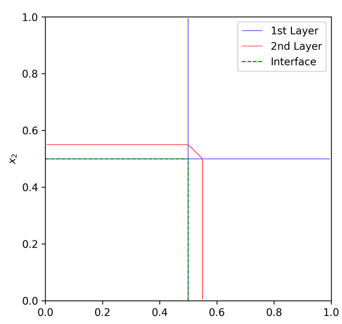
(b) The interface on $x_i = 0.255$ for $i = 4, \ldots, 10000$



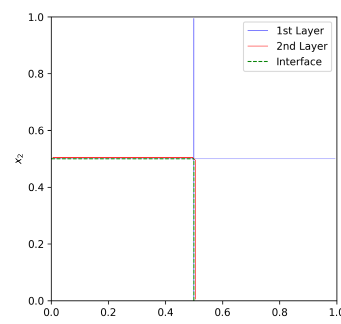
(c) An approximation of $\chi(\mathbf{x})$ by the 10000–10000–1–1 ReLU NN function in (4.1) with $\varepsilon = 1/20$ on $x_i = 0.255$ for $i = 3, \ldots, 10000$



(d) An approximation of $\chi(\mathbf{x})$ by the 10000–10000–1–1 ReLU NN function in (4.1) with $\varepsilon = 1/200$ on $x_i = 0.255$ for $i = 3, \ldots, 10000$



(e) The breaking hyperplanes of the approximation in Figure 6(c) on $x_i = 0.255$ for $i = 3, \ldots, 10000$



(f) The breaking hyperplanes of the approximation in Figure 6(d) on $x_i = 0.255$ for $i = 3, \ldots, 10000$

FIG. 6. *A convex example to illustrate Theorem 3.2 for the case* $d = 10000$

5.4. A two-dimensional non-convex example. Let $\Omega = (-2, 2)^2$ and Ω_1 be the H-shaped region depicted in Figure 7(b) whose boundary is the interface $\Gamma = \partial\Omega_1 \cap \partial\Omega_2 = \partial\Omega_1$. The unit step function $\chi(\mathbf{x})$ is depicted in Figure 7(a). Again, in this example, we can simply take $\hat{\chi} = \chi$. We construct 2–12–3–1 ReLU NN functions using 2–4–1–1 ReLU NN functions as discussed in Subsections 4.2.1 and 4.2.2 (see Figures 7(c) and 8(a)). The approximations with $\varepsilon = 1/12$ and 1/200 are depicted in Figures 7(d) and 7(e) for the convex hull and Figures 8(b) and 8(c) for the convex decomposition, respectively. Their corresponding breaking lines are plotted in Figures 7(f) and 7(g) for the convex hull and Figures 8(d) and 8(e) for the convex decomposition, respectively.

REFERENCES

[1] M. Anthony, P. L. Bartlett, P. L. Bartlett, et al., *Neural network learning: Theoretical foundations*, vol. 9, cambridge university press Cambridge, 1999.

[2] R. Arora, A. Basu, P. Mianjy, and A. Mukherjee, *Understanding deep neural networks with rectified linear units*, in International Conference on Learning Representations, 2018.

[3] D. Avis and D. Bremner, *How good are convex hull algorithms?*, in Proceedings of the eleventh annual symposium on Computational geometry, 1995, pp. 20–28.

[4] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa, *The quickhull algorithm for convex hulls*, ACM Transactions on Mathematical Software (TOMS), 22 (1996), pp. 469–483.

[5] Z. Cai, J. Chen, and M. Liu, *Least-squares neural network (LSNN) method for scalar nonlinear hyperbolic conservation laws: discrete divergence operator*, J. Comput. Appl. Math. 433 (2023) 115298, https://doi.org/10.1016/j.cam.2023.115298.

[6] Z. Cai, J. Chen, and M. Liu, *Least-squares ReLU neural network (LSNN) method for linear advection-reaction equation*, J. Comput. Phys. 443 (2021) 110514, https://doi.org/10.1016/j.jcp.2021.110514.

[7] Z. Cai, J. Choi, and M. Liu, *Least-squares neural network (LSNN) method for linear advection-reaction equation: Discontinuity interface*, SIAM J. Sci. Comput., 46 (2024), pp. C448–C478, https://doi.org/10.1137/23M1568107.

[8] Z. Cai, J. Choi, and M. Liu, *Least-squares neural network (LSNN) method for linear advection-reaction equation: non-constant jumps*, Int'l. J. Numer. Anal. Modeling, 21 (2024), p. 609–628.

[9] A. Caragea, P. Petersen, and F. Voigtlaender, *Neural network approximation and estimation of classifiers with classification boundary in a barron class*, The Annals of Applied Probability, 33 (2023), pp. 3039–3079.

[10] R. DeVore, B. Hanin, and G. Petrova, *Neural network approximation*, Acta Numer., 30 (2021), pp. 327–444, https://doi.org/10.1017/S0962492921000052.

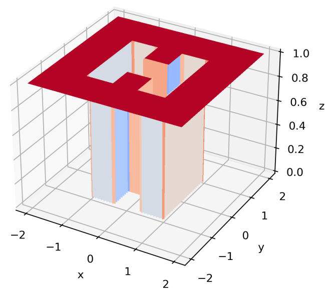
[11] S. Hertel and K. Mehlhorn, *Fast triangulation of simple polygons*, in Foundations of Computation Theory: Proceedings of the 1983 International FCT-Conference Borgholm, Sweden, August 21–27, 1983 4, Springer, 1983, pp. 207–218.

[12] M. Imaizumi and K. Fukumizu, *Deep neural networks learn non-smooth functions effectively*, in The 22nd international conference on artificial intelligence and statistics, PMLR, 2019, pp. 869–878.

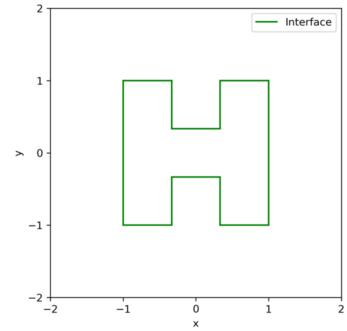
[13] Y. Kim, I. Ohn, and D. Kim, *Fast convergence rates of deep neural networks for classification*, Neural Networks, 138 (2021), pp. 179–197, https://doi.org/10.1016/j.neunet.2021.02.012.

[14] D. M. Mount, *Cmsc 754 computational geometry*, Lecture Notes, University of Maryland, (2002), pp. 1–122.

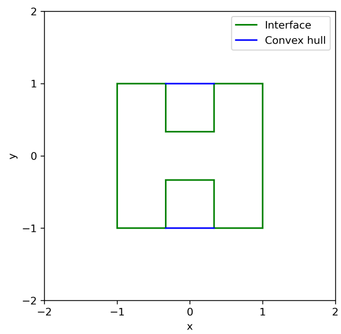
[15] P. Petersen and F. Voigtlaender, *Optimal approximation of piecewise smooth functions using deep ReLU neural networks*, Neural Networks, 108 (2018), pp. 296–330.



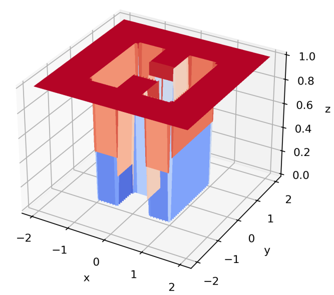
(a) The piecewise constant function $\chi(\mathbf{x})$



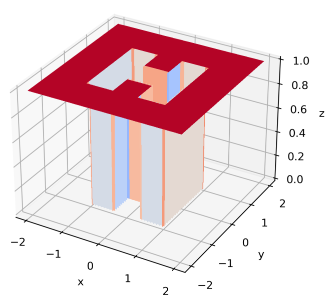
(b) The interface



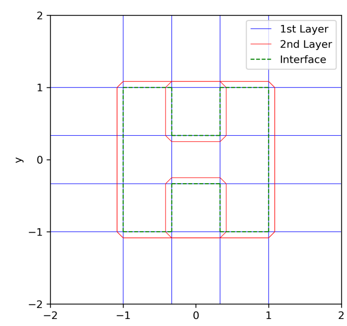
(c) The convex hull of $\hat{\Omega}_1$



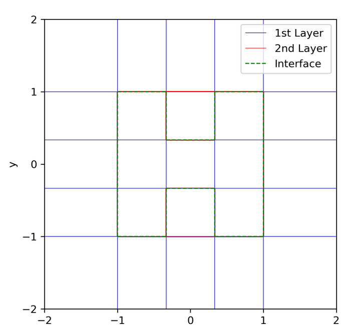
(d) An approximation of $\chi(\mathbf{x})$ by the 2–12–3–1 ReLU NN function with $\varepsilon = 1/12$



(e) An approximation of $\chi(\mathbf{x})$ by the 2–12–3–1 ReLU NN function with $\varepsilon = 1/200$

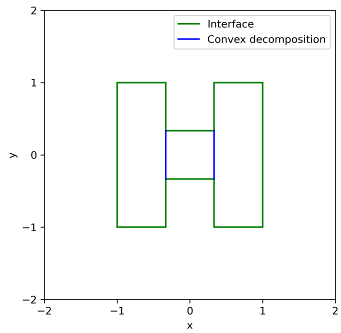


(f) The breaking hyperplanes of the approximation in Figure 7(d)

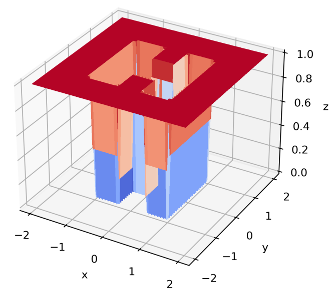


(g) The breaking hyperplanes of the approximation in Figure 7(e)

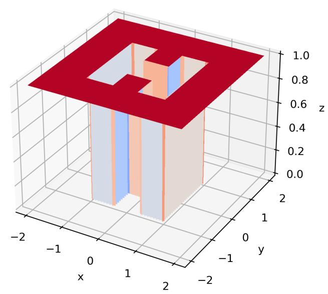
FIG. 7. *A non-convex example to illustrate Theorem 3.2 for the case* $d = 2$ *(convex hull)*



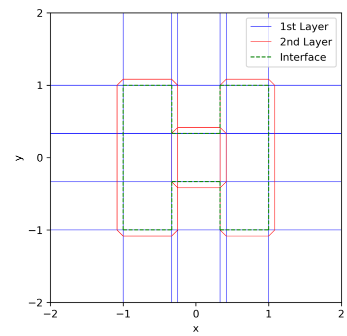
(a) A convex decomposition of $\hat{\Omega}_1$



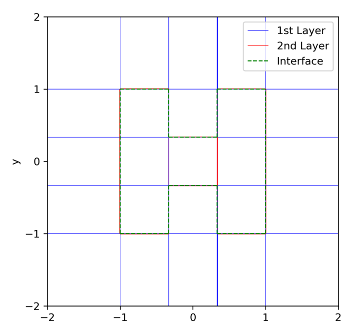
(b) An approximation of $\chi(\mathbf{x})$ by the 2–12–3–1 ReLU NN function with $\varepsilon = 1/12$



(c) An approximation of $\chi(\mathbf{x})$ by the 2–12–3–1 ReLU NN function with $\varepsilon = 1/200$



(d) The breaking hyperplanes of the approximation in Figure 8(b)



(e) The breaking hyperplanes of the approximation in Figure 8(c)

FIG. 8. *A non-convex example to illustrate Theorem 3.2 for the case* $d = 2$ *(convex decomposition)*