Improved Bounds on Neural Complexity for Representing Piecewise Linear Functions

Kuan-Lin Chen
Department of Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA 92093, USA
kuc029@ucsd.edu

Harinath Garudadri
Qualcomm Institute
University of California, San Diego
La Jolla, CA 92093, USA
hgarudadri@ucsd.edu

Bhaskar D. Rao
Department of Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA 92093, USA
brao@ucsd.edu

Abstract

A deep neural network using rectified linear units represents a continuous piecewise linear (CPWL) function and vice versa. Recent results in the literature estimated that the number of neurons needed to exactly represent any CPWL function grows exponentially with the number of pieces or exponentially in terms of the factorial of the number of distinct linear components. Moreover, such growth is amplified linearly with the input dimension. These existing results seem to indicate that the cost of representing a CPWL function is expensive. In this paper, we propose much tighter bounds and establish a polynomial time algorithm to find a network satisfying these bounds for any given CPWL function. We prove that the number of hidden neurons required to exactly represent any CPWL function is at most a quadratic function of the number of pieces. In contrast to all previous results, this upper bound is invariant to the input dimension. Besides the number of pieces, we also study the number of distinct linear components in CPWL functions. When such a number is also given, we prove that the quadratic complexity turns into bilinear, which implies a lower neural complexity because the number of distinct linear components is always not greater than the minimum number of pieces in a CPWL function. When the number of pieces is unknown, we prove that, in terms of the number of distinct linear components, the neural complexity of any CPWL function is at most polynomial growth for low-dimensional inputs and a factorial growth for the worst-case scenario, which are significantly better than existing results in the literature.

1 Introduction

The rectified linear unit (ReLU) [Fukushima, 1980, Nair and Hinton, 2010] activation has been by far the most widely used nonlinearity and successful building block in deep neural networks (DNNs). Numerous architectures based on ReLU DNNs have achieved remarkable performance or state-of-the-art accuracy in speech processing [Zeiler et al., 2013, Maas et al., 2013], computer vision [Krizhevsky et al., 2012, Simonyan and Zisserman, 2015, He et al., 2016], medical image segmentation [Ronneberger et al., 2015], game playing [Mnih et al., 2015, Silver et al., 2016], and natural language processing [Vaswani et al., 2017], just to name a few. Besides such unprecedented

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

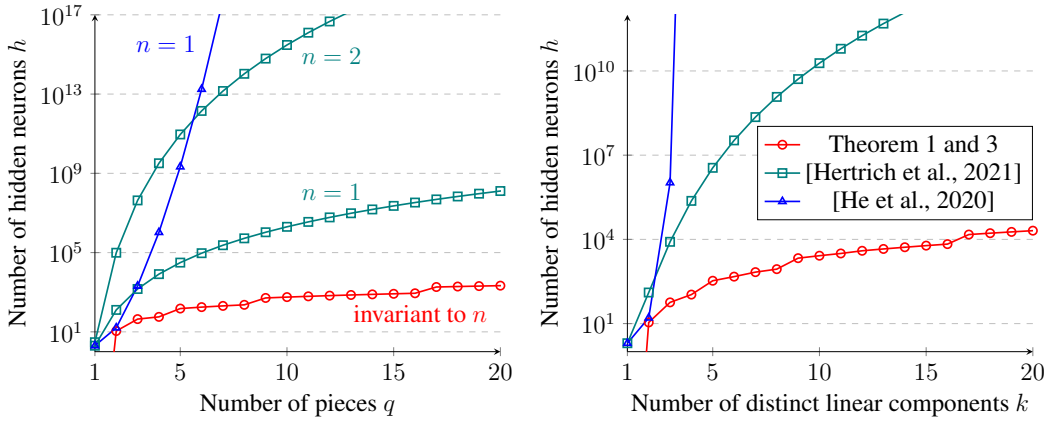

Figure 1: Any CPWL function $\mathbb{R}^n \to \mathbb{R}$ with q pieces or k distinct linear components can be exactly represented by a ReLU network with at most h hidden neurons. In Theorem 1 and 3, $h = 0$ when $q = 1$ or $k = 1$. The figure on the right-hand side uses $n = 1$ (see Figure 2 for different n). (5) is used to infer the bounds based on the depth and width given by Hertrich et al. [2021]. The upper bounds given by Theorem 1 and 3 are substantially lower than existing bounds in the literature, implying that any CPWL function can be exactly realized by a ReLU network at a much lower cost.

empirical success, ReLU DNNs are also probably the most understandable nonlinear deep learning models due to their ability to be "un-rectified" [Hwang and Heinecke, 2019].

The ability to demystify ReLU DNNs via "un-rectifying ReLUs" dates back to a seminal work by Pascanu et al. in 2014. Because each of ReLUs in a hidden layer divides the space of the preceding layer's output into two half spaces whose ReLU response is affine in one half space and exactly zero in the other, the layer of ReLUs can be replaced by an input-dependent diagonal matrix whose diagonal elements are ones for firing ReLUs and zeros for non-firing ReLUs. Based on this rationale, Pascanu et al. [2014] proved that a neural network using ReLUs divides the input space into many linear regions such that the network itself is an affine function within every region. Two excellent visualizations are shown in Figure 2 in [Hanin and Rolnick, 2019a] and [Hanin and Rolnick, 2019b]. At this point, it is quite evident that any ReLU network exactly represents a CPWL function. Pascanu et al. also proved that the maximum number of linear regions for any ReLU network with a single hidden layer is equivalent to the number of connected components induced by arrangements of hyperplanes in general position where each hyperplane corresponds to a ReLU in the hidden layer. Such a number can be computed in a closed form by Zaslavsky's Theorem [Zaslavsky, 1975]. Furthermore, they showed that the maximum number of linear regions can be bounded from below by exponential growth in terms of the number of hidden layers, leading to a conclusion that ReLU DNNs can generate more linear regions than their shallow counterparts. In the same year, Montúfar et al. improved such a lower bound and gave the first upper bound for the maximum number of linear regions. These bounds and their assumptions were later improved by [Montúfar, 2017, Raghu et al., 2017, Arora et al., 2018, Serra et al., 2018, Hinz and van de Geer, 2019], just to name a few. We refer readers to Hinz's doctoral thesis for a thorough discussion on the upper bound of the number of linear regions. Because a CPWL function with more pieces can better approximate any given continuous function and a ReLU DNN exactly represents a CPWL function [Arora et al., 2018], a ReLU DNN with more linear regions in general exhibits stronger expressivity. In summary, this "un-rectifying" perspective provides us a new angle to understand ReLU DNNs, and the results in some ways align with the advance in approximation theory demonstrating the expressivity. [1]

Despite these advancements in linear regions, the complexity of a ReLU DNN that exactly represents a given CPWL function remains largely unexplored. One can find that this question is the opposite direction of the above-mentioned line of research. Although Arora et al. [2018] proved that any CPWL function can be exactly represented by a ReLU DNN with a bounded depth, any estimates

[1] The approximation viewpoint is not the focus of this paper. The literature on approximation is vast and we refer readers to [Vardi et al., 2021, Lu et al., 2017, Eldan and Shamir, 2016, Telgarsky, 2016, Hornik et al., 1989, Cybenko, 1989], just to name a few.

regarding the width or number of neurons of such a network were not given. The resources required for a ReLU neural network to exactly represent a CPWL function remained unknown until He et al. [2020] first gave an estimate to the complexity of a ReLU network that realizes any given CPWL function. They proved that the number of neurons is bounded from above by exponential growth in terms of the product between the number of pieces and the number of distinct linear components of a given CPWL function. Such an exponential bound also grows linearly with the input dimension. Because the number of pieces is an upper bound of the number of distinct linear components for any CPWL function [Tarela and Martinez, 1999, He et al., 2020], the bound grows exponentially with the quadratic number of pieces, which seems to imply that the cost for representing a CPWL function by a ReLU DNN is exceedingly high.

The most recent upper bound can be inferred from a recent work by Hertrich et al. [2021] although the number of hidden neurons was not directly given. Hertrich et al. [2021] proved the first width bound in terms of the number of pieces under the same depth used by Arora et al. [2018] and He et al. [2020]. In particular, they proved that the width of a ReLU network that represents any given CPWL function can be polynomially bounded from above in terms of the number of pieces. However, the order of such a polynomial is a quadratic function of the input dimension, which can be immensely large for a small number of pieces if the input dimension is large. This bound grows larger with the input dimension even though the underlying CPWL function is just a one-hidden-layer ReLU network using only one ReLU.

In this paper, we provide improved bounds showing that any CPWL function can be represented by a ReLU DNN whose neural complexity is bounded from above by functions with much slower growth (see Figure 1). Our results imply that one can exactly realize any given CPWL function by a ReLU network at a much lower cost. Moreover, in addition to guarantee the existence, we also give a *polynomial time* algorithm to exactly find a network satisfying our bounds. To the best of our knowledge, our results regarding the computational resource for a ReLU network, i.e., the number of hidden neurons, are the lowest upper bounds in the existing literature and the algorithm is the first tailored procedure to find a network representation from any given CPWL function. Key results and main contributions of this paper are highlighted below.

1.1 Key results and contributions

Quadratic bounds. We prove that any CPWL function with q pieces can be represented by a ReLU network whose total number of hidden neurons is bounded from above by a quadratic function of q. We also give the corresponding upper bounds for the maximum width, i.e., the maximum number of neurons per hidden layer, and the number of layers for such a network. The maximum width is bounded from above by $\mathcal{O}(q^2)$ and the number of layers is bounded from above by a logarithmic function of q, i.e., $\mathcal{O}(\log_2 q)$. *These bounds are invariant to the input dimension.* For any affine function, the upper bounds for the maximum width and the number of hidden neurons are zero.

Further improvements on neural complexity. When the number of distinct linear components k of any CPWL function is given along with the number of pieces q, the quadratic bounds $\mathcal{O}(q^2)$ for the number of hidden neurons and the maximum width turn into *bilinear* bounds of k and q, i.e., $\mathcal{O}(kq)$. Such a change reduces the neural complexity because $k \leq q$ and q is usually much larger than k. Still, these bounds are independent of the input dimension.

Finding a network satisfying bilinear bounds. We establish a polynomial time algorithm that finds a ReLU network representing any given CPWL function. The network found by the algorithm satisfies the bilinear bounds on the number of hidden neurons and the maximum width, and the logarithmic bound on the number of layers. Note that such an algorithm also guarantees that one can always reverse-engineer at least one ReLU network from the function it computes. Compared to the general-purpose reverse-engineering algorithm proposed by Rolnick and Kording [2020], our algorithm specializes in the situation when pieces of a CPWL function are given.

Improved bounds from a perspective of linear components. When the number of pieces of a CPWL function is unknown and only the number of linear components k is available, we prove that the number of hidden neurons and maximum width are bounded from above by factorial growth. More precisely, $\mathcal{O}\left(k \cdot k!\right)$. The number of layers is bounded from above by linearithmic growth,

or $\mathcal{O}(k \log_2 k)$. However, if the input dimension n grows sufficiently slower than k, e.g., $\mathcal{O}\left(\sqrt{k}\right)$, then bounds for the number of hidden neurons and maximum width reduce to *polynomial* growth functions of order $2n+1$; and the linearithmic growth reduces to $\mathcal{O}\left(n \log_2 k\right)$ for the depth.

A new approach to choosing depth. Instead of scaling the depth of a ReLU network with the input dimension [Arora et al., 2018, He et al., 2020, Hertrich et al., 2021], we reveal that constructing a ReLU network whose depth is scaled with the number of pieces of the given CPWL function is more advantageous. Such a scaling turns out to be the key to derive better upper bounds. This insight is provided by the max-min representation of CPWL functions [Tarela et al., 1990]. The importance of this scaling on the depth in ReLU networks has not been well recognized by existing bounds in the literature. We discuss implications of different representations in Section 4.

2 Preliminaries

Notation and definitions used in this paper are set up and clarified in this section. The set $\{1, 2, \cdots, m\}$ is denoted by $[m]$. $\mathbb{I}\left[condition\right]$ is an indicator function that gives 1 if the *condition* is true, and 0 otherwise. The CPWL function is defined by Definition 1 below.

Definition 1. *A function $p\colon \mathbb{R}^n \to \mathbb{R}$ is said to be CPWL if there exists a finite number of closed subsets of $\mathbb{R}^n$, say $\{\mathcal{X}_i\}_{i\in[m]}$, such that (a) $\mathbb{R}^n = \bigcup_{i\in[m]} \mathcal{X}_i$; (b) p is affine on $\mathcal{X}_i, \forall i \in [m]$.*

A family of closed *convex* subsets of $\mathbb{R}^n$ satisfying Definition 1 are also referred to as convex regions, affine pieces or simply *pieces* for a CPWL function in this paper. Definition 1 follows the definition of CPWL functions by Ovchinnikov [2002]. Notice that there are different definitions in the literature. For example, Chua and Deng [1988] and Arora et al. [2018] defined a CPWL function on a finite number of polyhedral regions. However, their definitions are essentially the same as Definition 1 because any family of closed subsets satisfying Definition 1 can be decomposed into polyhedral regions. It is possible that some of the closed subsets satisfying Definition 1 are non-convex even though the number of them reaches the minimum (see Figure 2 in [Wang and Sun, 2005]). The continuity is implied by Definition 1 due to the subsets being closed.

Because the goal of this paper is to bound the complexity of a ReLU DNN that exactly represents any given CPWL function, it is necessary to be able to measure the complexity of a CPWL function. The complexity of a CPWL function can be described using two quite different perspectives. One is the number of pieces q, which is the number of closed convex subsets satisfying Definition 1. Because this number has a minimum and any finite number above the minimum can be a valid m in Definition 1, the bounds become obviously loose when the number of pieces is not the minimum. Without loss of generality, we are interested in the number q when it is the minimum. The other is the number of distinct linear components k. A linear component of a CPWL function is defined in Definition 2.

Definition 2. *An affine function f is said to be a linear component of a CPWL function p if there exists a nonempty subset $\mathcal{M} \subseteq [m]$ such that $f(\mathbf{x}) = p(\mathbf{x}), \forall \mathbf{x} \in \bigcup_{i\in\mathcal{M}} \mathcal{X}_i$ where $\{\mathcal{X}_i\}_{i\in[m]}$ is a family of minimum number of closed subsets satisfying Definition 1.*

A greater q or k gives a CPWL function more degrees of freedom because a CPWL function allowed to use $q+1$ pieces or $k+1$ arbitrary linear components can represent any CPWL function with q pieces or k distinct linear components and still have a flexibility to modify existing affine maps or increase the number of distinct affine maps of the CPWL function. Although increasing them both leads to a CPWL function with a higher flexibility, the speed of upgrading degrees of freedom is different from each other. Note that a CPWL function with q pieces can never have more than q distinct linear components and a CPWL function with k distinct linear components can easily have more than k minimum number of pieces. Such a difference in a 1-dimensional case can be clearly observed from Figure 1 in [Tarela and Martinez, 1999]. Note that it is possible for two disjoint subsets from a minimum number of closed subsets satisfying Definition 1 to share the same linear component. In other words, a linear component can be reused by multiple pieces. Hence, increasing k gives much faster growth than increasing q for the complexity and expressivity of CPWL functions.

We define the ReLU activation function in Definition 3. The ReLU network defined in Definition 4 is a simple architecture which is usually referred to as a *ReLU multi-layer perceptron*. Definition 5 defines the corresponding number of hidden neurons, depth, and the maximum width.

Definition 3. *The rectified linear unit (ReLU) activation function σ is defined as $\sigma(x) = \max(0, x)$. The ReLU layer or vector-valued rectified linear activation function σ_k is defined as*

$$\sigma_k(\mathbf{x}) = \begin{bmatrix}\sigma(x_1) & \sigma(x_2) & \cdots & \sigma(x_k)\end{bmatrix}^\mathsf{T} \tag{1}$$

where $\mathbf{x} = \begin{bmatrix}x_1 & x_2 & \cdots & x_k\end{bmatrix}^\mathsf{T}$.

Definition 4. *Let l be any positive integer. A function $g\colon \mathbb{R}^{k_0} \to \mathbb{R}^{k_l}$ is said to be an l-layer ReLU network if there exist weights $\mathbf{W}_i \in \mathbb{R}^{k_i \times k_{i-1}}$ and $\mathbf{b}_i \in \mathbb{R}^{k_i}$ for $i \in [l]$ such that the input-output relationship of the network satisfies $g(\mathbf{x}) = h_l(\mathbf{x})$ where $h_1(\mathbf{x}) = \mathbf{W}_1\mathbf{x} + \mathbf{b}_1$ and $h_i(\mathbf{x}) = \mathbf{W}_i\sigma_{k_{i-1}}\left(h_{i-1}(\mathbf{x})\right) + \mathbf{b}_i$ for every $i \in [l] \setminus [1]$.*

Definition 5. *The sum $\sum_{l=1}^{L-1} k_l$ and the maximum $\max_{l\in[L-1]} k_l$ for $L > 1$ are referred to as the number of hidden neurons and the maximum width of an L-layer ReLU network, respectively. Any 1-layer ReLU network is said to have 0 hidden neurons and a maximum width of 0. An l-layer ReLU network is said to have depth l and $l - 1$ hidden layers.*

3 Upper bounds on neural complexity for representing CPWL functions

The correspondence between CPWL functions and ReLU networks was first clearly confirmed by Theorem 2.1 in [Arora et al., 2018] although a weaker version of the correspondence can be inferred from Proposition 4.1 in [Goodfellow et al., 2013]. Arora et al. [2018] proved that every ReLU network $\mathbb{R}^n \to \mathbb{R}$ exactly represents a CPWL function, and the converse is also true, i.e., every CPWL function can be exactly represented by a ReLU network. One of the key steps used by Arora et al. [2018] to construct a ReLU network from any given CPWL function relies on an important representation result by Wang and Sun [2005], stating that any CPWL function can be represented by a sum of a finite number of *max-η-affine* functions [Magnani and Boyd, 2009] whose signs may be flipped and η is bounded from above by $n + 1$. The implication of using this representation is later discussed in Section 4.1. The bound $\eta \leq n + 1$ in the representation allowed Arora et al. [2018] to further prove that there exists a ReLU DNN with at most

$$\lceil \log_2(n + 1) \rceil \tag{2}$$

hidden layers to exactly realize any given CPWL function. However, the computational resource required for a ReLU network to exactly represent any CPWL function had not been available in the literature until the work by He et al. [2020].

3.1 Upper bounds in prior work

He et al. [2020] proved that a CPWL function $\mathbb{R}^n \to \mathbb{R}$ with q pieces and k linear components can be represented by a ReLU network whose number of neurons is given by

$$\mathcal{O}\left(n2^{kq+(n+1)(k-n-1)}\right) \tag{3}$$

for $k \geq n + 1$ and $\mathcal{O}\left(n2^{kq}\right)$ for $k < n + 1$. The number of hidden layers in such a ReLU DNN is also bounded from above by $\lceil \log_2(n + 1) \rceil$, the same as the bound derived by Arora et al. [2018]. One of their significant contributions in our view is that they utilize the number of pieces and linear components of a CPWL function to bound the complexity of the equivalent ReLU network. In general, the number of pieces q is much larger than the number of linear components k for a CPWL function. He et al. [2020] also proved the relationship

$$k \leq q \leq k! \tag{4}$$

for any CPWL function. Note that the bounds in (4) on the number of pieces q and linear components k were first mentioned by Tarela and Martinez [1999] who developed the lattice representation of CPWL function. Asymptotically, the bounds proved by He et al. [2020] are amplified linearly with the input dimension n for any fixed k. Due to (4), they can be further bounded from above by $\mathcal{O}\left(n2^{q^2+(n+1)(q-n-1)}\right)$ and $\mathcal{O}\left(n2^{q^2}\right)$ in terms of q and n. On the other hand, in terms of k and n, they can be further bounded from above by $\mathcal{O}\left(n2^{k\cdot k!+(n+1)(k-n-1)}\right)$ and $\mathcal{O}\left(n2^{k\cdot k!}\right)$. Because these

bounds grow much faster than exponential growth, they seem to suggest that the cost of computing a CPWL function via a ReLU network is exceptionally high.

Hertrich et al. [2021] proved that any CPWL function $\mathbb{R}^n \to \mathbb{R}$ with q pieces can be represented by a ReLU network whose maximum width is $\mathcal{O}\left(q^{2n^2+3n+1}\right)$ under the same number of hidden layers $\lceil \log_2(n+1) \rceil$. Hence, the number of hidden neurons must be bounded from above by

$$\mathcal{O}\left(q^{2n^2+3n+1} \log_2(n+1)\right). \tag{5}$$

Note that we infer this bound by taking the product of the depth and the maximum width. The bound given by (5) can grow slower than $\mathcal{O}\left(n2^{q^2}\right)$, but it grows faster than $\mathcal{O}\left(n2^{q^2}\right)$ if the input dimension n grows sufficiently faster than the number of pieces q. Bounds expressed in terms of the number of distinct linear components were not given in [Hertrich et al., 2021].

3.2 Improved upper bounds

We show that any CPWL function can be represented by a ReLU network whose number of hidden neurons is bounded by much slower growth functions. We state our main results in Theorem 1, 2 and 3, and focus on their impact in this subsection. Each one of them is tailored to a specific complexity measure of the CPWL function. Their proof sketches are deferred to Section 4.2. We first focus on the case when the number of linear components is unknown and the complexity of the CPWL is only measured by the number of pieces, a scenario used in [Hertrich et al., 2021].

Theorem 1. *Any CPWL function $p\colon \mathbb{R}^n \to \mathbb{R}$ with q pieces can be represented by a ReLU network whose number of layers l, maximum width w, and number of hidden neurons h satisfy*

$$l \leq 2\lceil \log_2 q \rceil + 1, \tag{6}$$

$$w \leq \mathbb{I}[q > 1] \left\lceil \frac{3q}{2} \right\rceil q, \tag{7}$$

and

$$h \leq \left(3 \cdot 2^{\lceil \log_2 q \rceil} + 2\lceil \log_2 q \rceil - 3\right) q + 3 \cdot 2^{\lceil \log_2 q \rceil} - 2\lceil \log_2 q \rceil - 3. \tag{8}$$

Furthermore, Algorithm 1 finds such a network in $\mathrm{poly}(n, q)$ *time.*

Algorithm 1 Find a ReLU network that computes a given continuous piecewise linear function

Input: A continuous piecewise linear function p with pieces $\{\mathcal{X}_i\}_{i\in[q]}$ of $\mathbb{R}^n$ satisfying Definition 1.
Output: A ReLU network g computing $g(\mathbf{x}) = p(\mathbf{x}), \forall \mathbf{x} \in \mathbb{R}^n$.

```
f_1, f_2, ··· , f_k ← run Algorithm 6 to find all distinct linear components of p
for i = 1, 2, ··· , q do
    A_i ← ∅
    for j = 1, 2 ··· , k do
        if f_j(x) ≥ p(x), ∀x ∈ X_i then
            A_i ← A_i ⋃ {j}
        end if
    end for
    v_i ← run Algorithm 2 with {f_m}_{m∈A_i} using the minimum type
end for
v ← run Algorithm 3 with v_1, v_2, ··· , v_q                ▷ Combine q ReLU networks in parallel
u ← run Algorithm 2 with {[s_1  s_2  ···  s_q]^T ↦ s_m}_{m∈[q]} using the maximum type
g ← run Algorithm 4 with v and u                ▷ Find a ReLU network for the composition u ∘ v
```

The proof of Theorem 1 is deferred to Appendix B.4 in the supplementary material. Algorithm 6, 2, 3, and 4 used by Algorithm 1 are deferred to Appendix C in the supplementary material and will be discussed later (before presenting Theorem 2). Because $2^{\lceil \log_2 q \rceil} < 2q$, the upper bound in (8) can be

further bounded from above by $6q^2 + 2\lceil \log_2 q \rceil q + 3q - 2\lceil \log_2 q \rceil - 3$, leading to the asymptotic bounds $l = \mathcal{O}(\log_2 q)$, $w = \mathcal{O}(q^2)$, and $h = \mathcal{O}(q^2)$. Since the bound given by Theorem 5.2 in He et al. [2020] can be lower bounded by $\mathcal{O}\left(n2^{q^2}\right)$, it grows exponentially faster than our bound of h given in Theorem 1. On the other hand, the upper bound given by (5) is at least polynomially larger than our bound of h and the order of this polynomial grows quadratically with the input dimension n. Note that such a polynomial becomes an exponential when the growth in n is not slower than q. Such differences are illustrated by the figure on the left-hand side of Figure 1. The bounds in Theorem 1 are independent of the input dimension n. Hence, one can realize any given CPWL function using a relatively small ReLU network even though the input dimension n is huge.

In terms of the maximum width, the upper bound given by (7) is at least polynomially smaller than the one given by Hertrich et al. [2021]. In contrast to the bound for the number of layers in [Arora et al., 2018, He et al., 2020, Hertrich et al., 2021] that grows logarithmically with the input dimension n, our bound in Theorem 1 grows logarithmically with the number of pieces q. Therefore, the ReLU network found by Algorithm 1 in general becomes deeper when the CPWL becomes more complex for a fixed input dimension. On the other hand, the network remains the same depth even for arbitrarily larger n as long as q is fixed. Taking an affine function for example, a 1-layer ReLU network with 0 hidden neurons is the solution given by Theorem 1. However, the bound given by [Arora et al., 2018, He et al., 2020, Hertrich et al., 2021] keeps increasing the depth for a larger n.

Algorithm 1 is a polynomial time algorithm, following from the proof of Theorem 2. Table 1 in Appendix C in the supplementary material gives a complexity analysis for Algorithm 1. Algorithm 2 finds a ReLU network that computes a *max-affine* or *min-affine* function [Magnani and Boyd, 2009]. Algorithm 3 concatenates two given ReLU networks in parallel and returns another ReLU network computing the concatenation of two outputs. Algorithm 4 finds a ReLU network that represents a composition of two given ReLU networks. These algorithms are basic manipulations of ReLU networks required in Algorithm 1.

Notice that Algorithm 1 does not need to be given any linear components or completely know the CPWL function because every distinct linear component can be found by Algorithm 6, which only needs to be given a closed ϵ-ball in the interior of every piece of a CPWL function p and observe the output of p when feeding an input. Algorithm 6 solves a system of linear equations for every piece of p to find the corresponding linear component. Every system of linear equations here has a unique solution because the interior of each of pieces is nonempty. The nonemptyness is guaranteed by Lemma 12(a) in Appendix A in the supplementary material.

The 5th step of Algorithm 1 can be executed by checking the optimization result of the following linear programming problem

$$\begin{aligned} &\text{minimize} \quad f_j(\mathbf{x}) - p(\mathbf{x}), \\ &\text{subject to} \quad \mathbf{x} \in \mathcal{X}_i. \end{aligned} \tag{9}$$

The condition in the 5th step can only be true when the problem is feasible and the optimal value is nonnegative. Because every piece of p is given to Algorithm 1, the piece $\mathcal{X}_i$ is available for the linear program as a system of linear inequalities. The objective function is also available since p is affine on $\mathcal{X}_i$ and all distinct linear components are available from Algorithm 6. The corresponding linear component of p on $\mathcal{X}_i$ can be found by first feeding at most $n+1$ affinely independent points from the closed ϵ-ball to p and every candidate linear component, and then matching their output values.

Theorem 2 discusses the case when the number of linear components and pieces are both known.

Theorem 2. *Any CPWL function $p\colon \mathbb{R}^n \to \mathbb{R}$ with k linear components and q pieces can be represented by a ReLU network whose number of layers l, maximum width w, and number of hidden neurons h satisfy*

$$l \leq \lceil \log_2 q \rceil + \lceil \log_2 k \rceil + 1, \tag{10}$$

$$w \leq \mathbb{I}[k > 1] \left\lceil \frac{3k}{2} \right\rceil q, \tag{11}$$

and

$$h \leq \left(3 \cdot 2^{\lceil \log_2 k \rceil} + 2\lceil \log_2 k \rceil - 3\right) q + 3 \cdot 2^{\lceil \log_2 q \rceil} - 2\lceil \log_2 k \rceil - 3. \tag{12}$$

Furthermore, Algorithm 1 finds such a network in $\mathrm{poly}(n, k, q)$ *time.*

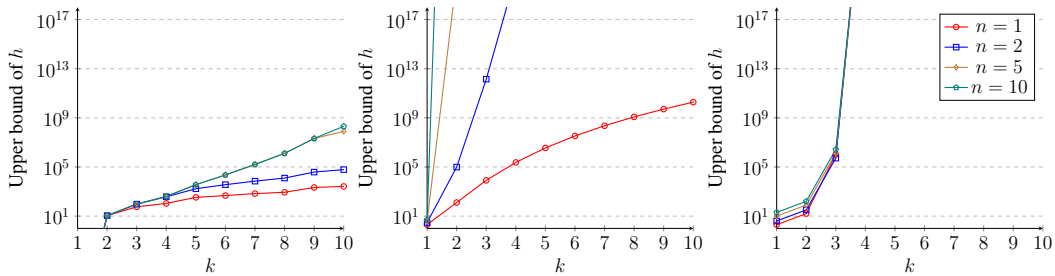

Figure 2: Left: The upper bound of h in Theorem 3 grows much slower when n grows sufficiently slower than k, leading to a much better upper bound compared to the worst-case asymptotic bound $\mathcal{O}(k \cdot k!)$ in Theorem 3. Middle: the bound in (5) inferred from [Hertrich et al., 2021] using the bound $\phi(n, k)$ for q. Right: Theorem 5.2 in [He et al., 2020].

The proof of Theorem 2 is deferred to Appendix B.3 in the supplementary material. The bounds in Theorem 2 are in general tighter and always no worse than those in Theorem 1 because q is never less than k but can be much larger than k. Asymptotically, $l = \mathcal{O}(\log_2 q)$, $w = \mathcal{O}(kq)$, and $h = \mathcal{O}(kq)$. The bound given by Theorem 5.2 in He et al. [2020] increases exponentially faster than the bound of h in Theorem 2.

When the number of linear components is the only complexity measure of the CPWL function, we resort to Theorem 3 below.

Theorem 3. *Any CPWL function $p\colon \mathbb{R}^n \to \mathbb{R}$ with k linear components can be represented by a ReLU network whose number of layers l, maximum width w, and number of hidden neurons h satisfy*

$$l \leq \lceil \log_2 \phi(n, k) \rceil + \lceil \log_2 k \rceil + 1, \tag{13}$$

$$w \leq \mathbb{I}[k > 1] \left\lceil \frac{3k}{2} \right\rceil \phi(n, k), \tag{14}$$

and

$$h \leq \left(3 \cdot 2^{\lceil \log_2 k \rceil} + 2\lceil \log_2 k \rceil - 3\right) \phi(n, k) + 3 \cdot 2^{\lceil \log_2 \phi(n,k) \rceil} - 2\lceil \log_2 k \rceil - 3 \tag{15}$$

where

$$\phi(n, k) = \min\left(\sum_{i=0}^{n} \binom{\frac{k^2-k}{2}}{i}, k!\right). \tag{16}$$

The proof of Theorem 3 is deferred to Appendix B.5 in the supplementary material. Because $\phi(n, k) \leq k!$, the worst-case asymptotic bounds for l, w and h are $l = \mathcal{O}(k \log_2 k)$, $w = \mathcal{O}(k \cdot k!)$, and $h = \mathcal{O}(k \cdot k!)$, respectively. However, it holds that $\sum_{i=0}^{n} \binom{\frac{k^2-k}{2}}{i} \leq k^{2n}$, so the asymptotic bounds are $l = \mathcal{O}(n \log_2 k)$, $w = \mathcal{O}\left(k^{2n+1}\right)$, and $h = \mathcal{O}\left(k^{2n+1}\right)$ when n grows sufficiently slower than k. For example, $n = \mathcal{O}\left(\sqrt{k}\right)$. In this case, w and h are bounded from above by a polynomial of order $2\sqrt{k} + 1$, which grows slower than factorial growth. Such an advantage for small n is illustrated by the figure on the left-hand side of Figure 2.

Since the bound given by Theorem 5.2 in [He et al., 2020] can be bounded from below by $\mathcal{O}\left(n2^{k \cdot k!}\right)$, it at the minimum grows exponentially larger than the upper bound of h in Theorem 3. For sufficiently small n, the relative order of growth is even larger. The figure on the right-hand side of Figure 2 illustrates such a large difference. For $k = 5$, $n2^{k \cdot k!} \approx 7.92 \times 10^{28}$ if $n = 1$, while our bound of h is less than 3.7×10^3 for any n. The difference is extremely large even though k is small under $n = 1$. The middle plot in Figure 2 shows that (5) increases much faster when n becomes larger. The upper bound of h in Theorem 3 is better than (5) for any n, which can be clearly seen from Figure 2.

Lemma 1. *Let $\mathcal{P}_{n,k}$ be the set of all CPWL functions with exactly k distinct linear components such that $p\colon \mathbb{R}^n \to \mathbb{R}, \forall p \in \mathcal{P}_{n,k}$. Let $\mathcal{C}_{n,k}(p)$ be the collection of all families of closed convex subsets satisfying Definition 1 for any $p \in \mathcal{P}_{n,k}$. Then,*

$$k \leq \min_{\mathcal{Q} \in \mathcal{C}_{n,k}(p)} |\mathcal{Q}| \leq \phi(n, k). \tag{17}$$

The proof of Lemma 1 is deferred to Appendix B.1 in the supplementary material. Clearly, $\phi(n, k)$ is a better upper bound of q compared to the bound $q \leq k!$ given by He et al. [2020]. When n grows sufficiently slower than k, the bound $\phi(n, k)$ can be exponentially smaller than $k!$.

3.3 Limitations

Although the upper bounds in our main results are much better, it is still possible to find a ReLU network whose hidden neurons are fewer than the bounds in Theorem 2 to exactly represent a given CPWL function. Tight bounds for the case when $n = 1$ was first given by Theorem 2.2 in [Arora et al., 2018]. However, it seems more difficult to bound the size of a network from below for a high dimensional input. To the best of our knowledge, we are not aware of any tight bounds in the literature for the exact ReLU network representation of general CPWL functions using an arbitrary n.

4 Representations of CPWL functions have different implications on depth

We reveal implications of using different representations of CPWL functions and their impact on constructing ReLU networks. We first discuss the popular representation used by prior work and the implicit constraint imposed by such a representation.

4.1 Constrained depth

Arora et al. [2018], He et al. [2020], and Hertrich et al. [2021] proved the same bound for the number of layers, relying on the following representation of a CPWL function

$$p(\mathbf{x}) = \sum_{j=1}^{J} \sigma_j \max_{i \in \eta(j)} f_i(\mathbf{x}) \tag{18}$$

where $\sigma_j \in \{+1, -1\}$ and $\eta(j)$ is a subset of $[J]$ such that $\left|\eta(j)\right| \leq n + 1$ for all $j \in [J]$. That is, a sum of a finite number of *max-η-affine* functions whose signs may be flipped. (18) was established by Theorem 1 in Wang and Sun [2005] which is essentially the same as Theorem 1 in Wang [2004] that emphasizes on the difference of two convex piecewise linear functions. This result was also used by Goodfellow et al. [2013] for the maxout network (Proposition 4.1).

The depth given by (2) does not scale with the complexity of the CPWL function. This feature directly comes from using a ReLU network to realize each of *max-η-affine* functions in (18) and concatenating all of them together. Because the size of $\eta(j)$ is bounded from above by $n + 1$, the depth can be made to depend solely on n. Such a treatment seems to be the only way if one considers a CPWL function represented by (18). As a result, the ReLU network is forced to use a depth constrained by the input dimension to represent the given CPWL function, which in turn requires more hidden neurons. Because we do not use (18), our networks are not limited by such an implication.

4.2 Proof sketch for unconstrained depth

We give a proof sketch in this subsection for our main results. By using a different representation, the depth of a ReLU network is able to be scaled with the complexity measure, i.e., number of pieces, of any given CPWL function to accommodate the high expressivity.

By Theorem 4.2 in [Tarela and Martinez, 1999], any CPWL function p can be represented as

$$p(\mathbf{x}) = \max_{\mathcal{X} \in \mathcal{Q}} \min_{i \in \mathcal{A}(\mathcal{X})} f_i(\mathbf{x}) \tag{19}$$

for all $\mathbf{x} \in \mathbb{R}^n$ where

$$\mathcal{A}(\mathcal{X}) = \left\{ i \in [k] \mid f_i(\mathbf{x}) \geq p(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X} \right\} \tag{20}$$

is the set of indices of linear components that have values greater than or equal to $p(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$, and $\mathcal{Q}$ is any family of closed convex subsets of $\mathbb{R}^n$ satisfying Definition 1. We have used $f_1, f_2, \cdots, f_k$ to denote the k distinct linear components of p. Notice that Theorem 4.2 in [Tarela and Martinez, 1999] was first stated by Theorem 7 in [Tarela et al., 1990]. Both are essentially the same, but Theorem 4.2 in [Tarela and Martinez, 1999] emphasizes the convexity of each of regions in the domain. Both theorems are also fundamentally equivalent to Theorem 2.1 in [Ovchinnikov, 2002].

Notice that one of the concluding remarks in [Ovchinnikov, 2002] pointed out that the convexity of the input space is an essential assumption. The entire space $\mathbb{R}^n$ satisfies such an assumption. In addition, Ovchinnikov pointed out that the max-min representation also holds for vector-valued CPWL functions. Hence, it is possible to generalize our bounds to vector-valued CPWL functions.

Using the representation in (19) and Lemma 2 below, we are able to prove Theorem 2 by bounding the size of $\mathcal{Q}$ and $\mathcal{A}(\mathcal{X})$. Theorem 1 and 3 can be proved by applying Lemma 1 to Theorem 2. Note that the size $|\mathcal{Q}|$ in (19) is the key for the depth of a ReLU network to be able to scale with q.

Lemma 2. *Let m be any positive integer. Define*

$$l(m) = \lceil \log_2 m \rceil + 1, \tag{21}$$

$$w(m) = \mathbb{I}[m > 1] \left\lceil \frac{3m}{2} \right\rceil, \tag{22}$$

and the following sequence for any positive integer k,

$$r(k) = \begin{cases} 0, & \text{if } k = 1 \\ \frac{3k}{2} + r\left(\frac{k}{2}\right), & \text{if } k \text{ is even} \\ 2 + \frac{3(k-1)}{2} + r\left(\frac{k+1}{2}\right), & \text{if } k \neq 1 \text{ and } k \text{ is odd} \end{cases}. \tag{23}$$

Then, there exists an $l(m)$-layer ReLU network $g\colon \mathbb{R}^n \to \mathbb{R}$ with $r(m)$ hidden neurons and a maximum width of $w(m)$ such that g computes the extremum of $f_1(\mathbf{x}), f_2(\mathbf{x}), \cdots, f_m(\mathbf{x})$, i.e., $g(\mathbf{x}) = \max_{i\in[m]} f_i(\mathbf{x})$ or $g(\mathbf{x}) = \min_{i\in[m]} f_i(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$ under any m scalar-valued affine functions $f_1, f_2, \cdots, f_m$. Furthermore, Algorithm 2 finds such a network in $\text{poly}(m, n)$ *time.*

The proof of Lemma 2 is deferred to Appendix B.2 in the supplementary material. One can also view $l(m)$, $w(m)$, and $r(m)$ as upper bounds for the number of layers, maximum width, and the number of hidden neurons. Because $r(m) < 6m - 3$ by Lemma 6, the bound for the number of hidden neurons $r(m)$ is tighter than the bound $8m - 4$ given by Lemma D.3 in [Arora et al., 2018] or Lemma 5.4 in [He et al., 2020] (these two lemmas are essentially the same). The bound for the number of layers remains the same as the one given by Lemma D.3 in [Arora et al., 2018]. By combining Lemma 2 with Lemma 3, Lemma 4, and Lemma 8, we can easily perform the same job on computing the extremum of multiple scalar-valued ReLU networks as Lemma D.3 does in [Arora et al., 2018]. Lemma 6, 3, 4, and 8 are given in Appendix A in the supplementary material.

5 Broader impact

Our results guarantee that any CPWL function can be exactly computed by a ReLU neural network at a more manageable cost. This assurance is crucial because CPWL functions are important tools in many applications. Such an assurance also relates DNNs closer to CPWL functions and allows researchers and engineers to understand the expressivity of DNNs from a different perspective. We focus on simple ReLU networks (ReLU multi-layer perceptrons) in this paper, but it may be possible to derive bounds for other activation functions and advanced neural network architectures such as maxout networks [Goodfellow et al., 2013], residual networks [He et al., 2016], densely connected networks [Huang et al., 2017], and other nonlinear networks [Chen et al., 2021], by making some (possibly mild) assumptions. Our contributions advance the fundamental understanding of the link between ReLU networks and CPWL functions.

Acknowledgments and disclosure of funding

We would like to thank the anonymous reviewers for their constructive comments and Tai-Hsuan Chung for answering our mathematical questions. This work was supported in part by NSF under Grant CCF-2225617, Grant CCF-2124929, and Grant IIS-1838897, in part by NIH/NIDCD under Grant R01DC015436, and in part by KIBM Innovative Research Grant Award.

References

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

K.-L. Chen, C.-H. Lee, H. Garudadri, and B. D. Rao. ResNEsts and DenseNEsts: Block-based DNN models with improved representation guarantees. In *Advances in Neural Information Processing Systems*, pages 3413–3424, 2021.

L. O. Chua and A.-C. Deng. Canonical piecewise-linear representation. *IEEE Transactions on Circuits and Systems*, 35(1):101–111, 1988.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

R. Eldan and O. Shamir. The power of depth for feedforward neural networks. In *Conference on Learning Theory*, pages 907–940. PMLR, 2016.

K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, 1980.

I. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio. Maxout networks. In *International Conference on Machine Learning*, pages 1319–1327. PMLR, 2013.

B. Hanin and D. Rolnick. Complexity of linear regions in deep networks. In *International Conference on Machine Learning*, pages 2596–2604. PMLR, 2019a.

B. Hanin and D. Rolnick. Deep ReLU networks have surprisingly few activation patterns. In *Advances in Neural Information Processing Systems*, 2019b.

J. He, L. Li, J. Xu, and C. Zheng. ReLU deep neural networks and linear finite elements. *Journal of Computational Mathematics*, 38(3):502–527, 2020.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, pages 770–778. IEEE, 2016.

C. Hertrich, A. Basu, M. Di Summa, and M. Skutella. Towards lower bounds on the depth of ReLU neural networks. In *Advances in Neural Information Processing Systems*, pages 3336–3348, 2021.

P. Hinz. *An analysis of the piece-wise affine structure of ReLU feed-forward neural networks*. PhD thesis, ETH Zurich, 2021.

P. Hinz and S. van de Geer. A framework for the construction of upper bounds on the number of affine linear regions of relu feed-forward neural networks. *IEEE Transactions on Information Theory*, 65(11):7304–7324, 2019.

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Conference on Computer Vision and Pattern Recognition*, pages 4700–4708. IEEE, 2017.

W.-L. Hwang and A. Heinecke. Un-rectifying non-linear networks for signal representation. *IEEE Transactions on Signal Processing*, 68:196–210, 2019.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang. The expressive power of neural networks: A view from the width. In *Advances in Neural Information Processing Systems*, 2017.

A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *International Conference on Machine Learning*, 2013.

A. Magnani and S. P. Boyd. Convex piecewise-linear fitting. *Optimization and Engineering*, 10(1): 1–17, 2009.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

G. Montúfar. Notes on the number of linear regions of deep neural networks. In *International Conference on Sampling Theory and Applications*, 2017.

G. Montúfar, R. Pascanu, K. Cho, and Y. Bengio. On the number of linear regions of deep neural networks. In *Advances in Neural Information Processing Systems*, pages 2924–2932, 2014.

V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *International Conference on Machine Learning*, pages 807–814, 2010.

S. Ovchinnikov. Max-min representation of piecewise linear functions. *Contributions to Algebra and Geometry*, 43(1):297–302, 2002.

R. Pascanu, G. Montúfar, and Y. Bengio. On the number of response regions of deep feed forward networks with piece-wise linear activations. *International Conference on Learning Representations*, 2014.

M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. Sohl-Dickstein. On the expressive power of deep neural networks. In *International Conference on Machine Learning*, pages 2847–2854. PMLR, 2017.

D. Rolnick and K. Kording. Reverse-engineering deep ReLU networks. In *International Conference on Machine Learning*, pages 8178–8187. PMLR, 2020.

O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 234–241. Springer, 2015.

T. Serra, C. Tjandraatmadja, and S. Ramalingam. Bounding and counting linear regions of deep neural networks. In *International Conference on Machine Learning*, pages 4558–4566. PMLR, 2018.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

J. Tarela and M. Martinez. Region configurations for realizability of lattice piecewise-linear models. *Mathematical and Computer Modelling*, 30(11-12):17–27, 1999.

J. Tarela, E. Alonso, and M. Martinez. A representation method for PWL functions oriented to parallel processing. *Mathematical and Computer Modelling*, 13(10):75–83, 1990.

M. Telgarsky. Benefits of depth in neural networks. In *Conference on Learning Theory*, pages 1517–1539. PMLR, 2016.

G. Vardi, D. Reichman, T. Pitassi, and O. Shamir. Size and depth separation in approximating benign functions with neural networks. In *Conference on Learning Theory*, pages 4195–4223. PMLR, 2021.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

S. Wang. General constructive representations for continuous piecewise-linear functions. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 51(9):1889–1896, 2004.

S. Wang and X. Sun. Generalization of hinging hyperplanes. *IEEE Transactions on Information Theory*, 51(12):4425–4431, 2005.

T. Zaslavsky. *Facing up to arrangements: Face-count formulas for partitions of space by hyperplanes*, volume 154. American Mathematical Society, 1975.

M. D. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q. V. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean, et al. On rectified linear units for speech processing. In *International Conference on Acoustics, Speech and Signal Processing*, pages 3517–3521. IEEE, 2013.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] The main claims are summarized in Section 1.1.
   (b) Did you describe the limitations of your work? [Yes] See Section 3.3.
   (c) Did you discuss any potential negative societal impacts of your work? [No] This is a theoretical paper and we are not aware of any negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] All definitions and assumptions are stated or referenced before the results.
   (b) Did you include complete proofs of all theoretical results? [Yes] Appendix B in the supplementary material.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Lemmas

Lemma 3. *Let l be any positive integer. Then, there exists an l-layer ReLU network g with $2n(l-1)$ hidden neurons and a maximum width of $2n$ such that $g(\mathbf{x}) = \mathbf{x}$ for all $\mathbf{x} \in \mathbb{R}^n$. Furthermore, Algorithm 5 finds such a network in* $\text{poly}(n, l)$ *time.*

Proof. Appendix B.6. □

Definition 6. *Let $g_{(l,n,w)}$ denote an l-layer ReLU network g with n hidden neurons and a maximum width bounded from above by w.*

Lemma 4. *There exists $g_{\left(l_1+l_2-1, n_1+n_2, \max(w_1,w_2)\right)}$ that represents any composition of $g_{(l_1,n_1,w_1)}$ and $g_{(l_2,n_2,w_2)}$. Algorithm 4 finds such a network computing the composition in* $\text{poly}\left(\max(w_1, w_2), \max(l_1, l_2)\right)$ *time.*

Proof. Appendix B.7. □

Lemma 5. *The sequence $r(k)$ defined by (23) is a strictly increasing sequence.*

Proof. Appendix B.8. □

Lemma 6. *For any positive integer k, the sequence $r(k)$ defined by (23) satisfies*

$$r(k) \leq 3\left(2^{\lceil \log_2 k \rceil} - 1\right) < 6k - 3. \tag{24}$$

Proof. Appendix B.9 □

Lemma 7. *Let m_1 and m_2 be the output dimensions of $g_{(l_1,n_1,w_1)}$ and $g_{(l_2,n_2,w_2)}$, respectively. Define*

$$l = \max(l_1, l_2), \tag{25}$$

$$w = w_j + \max(w_i, 2m_i), \tag{26}$$

and

$$n = n_1 + n_2 + 2m_i|l_1 - l_2|, \tag{27}$$

where $i = \arg\min_{k\in[2]} l_k$ and $j = [2] \setminus \{i\}$. Then, there exists $g_{(l,n,w)}$ such that

$$g_{(l,n,w)}(\mathbf{x}) = \begin{bmatrix} g_{(l_1,n_1,w_1)}(\mathbf{x}) \\ g_{(l_2,n_2,w_2)}(\mathbf{x}) \end{bmatrix} \tag{28}$$

for all $\mathbf{x} \in \mathbb{R}^n$.

Proof. Appendix B.10. □

Lemma 8. *Let $m_1, m_2, \cdots, m_k$ be the output dimensions of $g_{(l_1,n_1,w_1)}, g_{(l_2,n_2,w_2)}, \cdots, g_{(l_k,n_k,w_k)}$, respectively. Define*

$$l = \max_{i\in[k]} l_i, \tag{29}$$

$$w = \sum_{i\in[k]} \max(w_i, 2m_i), \tag{30}$$

and

$$n = \sum_{i\in[k]} n_i + 2m_i(l - l_i). \tag{31}$$

Then, there exists $g_{(l,n,w)}$ such that

$$g_{(l,n,w)}(\mathbf{x}) = \begin{bmatrix} g_{(l_1,n_1,w_1)}(\mathbf{x}) \\ g_{(l_2,n_2,w_2)}(\mathbf{x}) \\ \vdots \\ g_{(l_k,n_k,w_k)}(\mathbf{x}) \end{bmatrix} \tag{32}$$

for all $\mathbf{x} \in \mathbb{R}^n$. Furthermore, Algorithm 3 finds such a network in $\text{poly}\left(\max_{i\in[k]} w_i, k, l\right)$ *time.*

Proof. Appendix B.11. □

Lemma 9. *Let $f_1, f_2, \cdots, f_k$ be any affine functions such that $f_i \colon \mathbb{R}^n \to \mathbb{R}$ for all $i \in [k]$. Define the set of feasible ascending orders as*

$$\mathcal{S}^n_{f_1,f_2,\cdots,f_k} = \left\{(s_1, s_2, \cdots, s_k) \in \mathfrak{S}(k) \mid f_{s_1}(\mathbf{x}) \leq f_{s_2}(\mathbf{x}) \leq \cdots \leq f_{s_k}(\mathbf{x}), \mathbf{x} \in \mathbb{R}^n\right\} \tag{33}$$

where $\mathfrak{S}(k)$ is the collection of all permutations of the set $[k]$. It holds true that

$$\left|\mathcal{S}^n_{f_1,f_2,\cdots,f_k}\right| \leq \min\left(\sum_{i=0}^{n}\binom{\frac{k^2-k}{2}}{i}, k!\right). \tag{34}$$

Proof. Appendix B.12. □

Lemma 10. *If Definition 1 is satisfied for a non-affine function, then every nonempty subset has a nonempty intersection with at least one of the other subset(s).*

Proof. Appendix B.13. □

Assumption 1. *The number of closed connected subsets satisfying Definition 1 is a minimum.*

The interior and frontier (boundary) of a set $\mathcal{X}$ are denoted as Int$\mathcal{X}$ and Fr$\mathcal{X}$, respectively.

Lemma 11. *Let f_i denote the affine function associated with $\mathcal{X}_i$ for $i \in [I]$ where $\{\mathcal{X}_i\}_{i\in[I]}$ is a family of closed connected subsets satisfying Assumption 1. Then, for any $i \in [I], j \in [I]$ such that $i \neq j$ and $\mathcal{X}_i \bigcap \mathcal{X}_j \neq \emptyset$,*

(a) f_i and f_j are different, and $\{\mathbf{x} \in \mathbb{R}^n \mid f_i(\mathbf{x}) = f_j(\mathbf{x})\} \neq \emptyset$.

(b) $\{\mathbf{x} \in \mathbb{R}^n \mid f_i(\mathbf{x}) = f_j(\mathbf{x})\}$ is an affine subspace of $\mathbb{R}^n$ with dimension $n-1$.

(c) $\mathcal{X}_i \bigcap \mathcal{X}_j \subseteq \{\mathbf{x} \in \mathbb{R}^n \mid f_i(\mathbf{x}) = f_j(\mathbf{x})\}$.

(d) $\mathbf{x} \notin$ Int$\mathcal{X}_i$ and $\mathbf{x} \notin$ Int$\mathcal{X}_j$ for all $\mathbf{x} \in \mathcal{X}_i \bigcap \mathcal{X}_j$.

Proof. Appendix B.14, B.15, B.16, and B.17. □

Lemma 12. *If a family of closed connected subsets $\{\mathcal{X}_i\}_{i\in[I]}$ satisfies Assumption 1, then, for all $i \in [I]$,*

(a) Int$\mathcal{X}_i \neq \emptyset$.

(b) Fr$\mathcal{X}_i = \bigcup_{k\in[I]\setminus i} \mathcal{X}_k \bigcap \mathcal{X}_i$.

(c) Int$\mathcal{X}_i \bigcap$ Int$\mathcal{X}_j = \emptyset$ for all $j \in [I]$ such that $j \neq i$.

Proof. Appendix B.18, B.19, and B.20. □

Lemma 13. *Let $\{\mathcal{X}_i\}_{i\in[m]}$ be any finite family of subsets satisfying Assumption 1. Let $\{\mathcal{H}_j\}_{j\in[k]}$ be any finite family of affine subspaces of $\mathbb{R}^n$ with dimension $n-1$. Then, for $i \in [m]$,*

$$\mathcal{X}_i \bigcap \left(\mathbb{R}^n \setminus \bigcup_{j\in[k]} \mathcal{H}_j\right) \neq \emptyset. \tag{35}$$

Proof. Appendix B.21. □

Proposition 1. *For any family of closed connected subsets satisfying Definintion 1, all subsets are the largest closed connected subsets if and only if Assumption 1 is satisfied.*

Proof. Appendix B.22. □

B Proofs

B.1 Proof of Lemma 1

Proof. Let the family of closed connected subsets $\bar{\mathcal{Q}} = \{\mathcal{X}_i\}_{i\in[I]}$ satisfy Assumption 1 for any $p \in \mathcal{P}_{n,k}$. Let the k distinct linear components of p be $f_1, f_2, \cdots, f_k$ and $\mathcal{H}_{lm}$ be the intersection between f_l and f_m for $l \in [k], m \in [k], l \neq m$. Note that every $\mathcal{H}_{lm}$ is an affine subspace of $\mathbb{R}^n$ with dimension $n-1$ (a hyperplane) or an empty set. Because the linear components are distinct, it must be true that $k \leq I$ by Definition 2. If p is an affine function, then it follows that $k = \min_{\bar{\mathcal{Q}}\in\mathcal{C}_{n,k}(p)} \left|\bar{\mathcal{Q}}\right| = \phi(n,k) = 1$, the claim holds. For the non-affine case, we must have $k > 1$.

Let $\mathcal{R} = \mathbb{R}^n \setminus \mathcal{H}$ where

$$\mathcal{H} = \bigcup_{k\in[m], l\in[m], k\neq l} \mathcal{H}_{kl}. \tag{36}$$

Note that $\mathcal{H} \neq \emptyset$ according to Lemma 10 and 11(c). By Lemma 12(b), the boundary or frontier of $\mathcal{X}_i$ for $i \in [I]$ is given by

$$\text{Fr}\mathcal{X}_i = \bigcup_{j\in[I]\setminus i} \left(\mathcal{X}_i \bigcap \mathcal{X}_j\right). \tag{37}$$

Because every $\mathcal{X}_i \bigcap \mathcal{X}_j$ for $i \in [I], j \in [I], i \neq j$ is a subset of some $\mathcal{H}_{lm}$ for $l \in [k], m \in [k], l \neq m$ by Lemma 11(c), it follows that the boundary of $\mathcal{X}_i$, $\text{Fr}\mathcal{X}_i$, satisfies

$$\text{Fr}\mathcal{X}_i \subseteq \mathcal{H} \tag{38}$$

for $i \in [I]$. The interior of $\mathcal{X}_i$, $\text{Int}\mathcal{X}_i$, is a nonempty subset of $\mathbb{R}^n$ according to Lemma 12(a). Furthermore, by Lemma 13,

$$\mathcal{X}_i \bigcap \mathcal{R} \neq \emptyset. \tag{39}$$

Now, define

$$\mathcal{Z}_i = (\text{Int}\mathcal{X}_i) \bigcap \mathcal{R} \tag{40}$$

for $i \in [I]$. Note that $\mathcal{Z}_i = \mathcal{X}_i \bigcap \mathcal{R} \neq \emptyset$ due to (38) and (39). Let $\mathcal{A}$ be any subset of $\mathbb{R}^n$ and $\lambda(\mathcal{A})$ be the number of connected components of $\mathcal{A}$ in $\mathbb{R}^n$. It must be true that

$$1 = \lambda(\mathcal{X}_i) \leq \lambda\left(\text{Int}\mathcal{X}_i\right) \leq \lambda(\mathcal{Z}_i). \tag{41}$$

By Lemma 12(c), $\text{Int}\mathcal{X}_i \bigcap \text{Int}\mathcal{X}_j = \emptyset$ for $i \in [I], j \in [I], i \neq j$. We have

$$I \leq \lambda\left(\bigcup_{i\in[I]} \text{Int}\mathcal{X}_i\right) = \sum_{i\in[I]} \lambda(\text{Int}\mathcal{X}_i) \leq \sum_{i\in[I]} \lambda(\mathcal{Z}_i) = \lambda\left(\bigcup_{i\in[I]} \mathcal{Z}_i\right) = \lambda\left(\bigcup_{i\in[I]} \mathcal{X}_i \bigcap \mathcal{R}\right). \tag{42}$$

Notice that

$$\bigcup_{i\in[I]} \mathcal{X}_i \bigcap \mathcal{R} = \mathbb{R}^n \bigcap \mathcal{R} = \mathcal{R} \tag{43}$$

by the property $\bigcup_{i\in[I]} \mathcal{X}_i = \mathbb{R}^n$ in Definition 1. Plugging (43) into (42) leads to

$$I \leq \lambda(\mathcal{R}) \tag{44}$$

which states that I is bounded from above by the number of connected components of $\mathcal{R}$ in $\mathbb{R}^n$. Notice that every component is an open convex set because every component is the intersection of a finite number of open half spaces. Therefore,

$$I = \left|\bar{\mathcal{Q}}\right| = \min_{\mathcal{Q}'\in\mathcal{C}'_{n,k}(p)} \left|\mathcal{Q}'\right| \leq \min_{\mathcal{Q}\in\mathcal{C}_{n,k}(p)} |\mathcal{Q}| \leq \lambda(\mathcal{R}) \tag{45}$$

where $\mathcal{C}'_{n,k}(p)$ denotes the collection of all families of closed connected subsets satisfying Definition 1 for any $p \in \mathcal{P}_{n,k}$. Because the ascending order of these k linear components does not change within a connected component of $\mathcal{R}$, $\lambda(\mathcal{R})$ can be bounded from above by the number of feasible ascending orders. Let $\mathfrak{S}(k)$ be the collection of all permutations of the set $[k]$. It follows that

$$\lambda(\mathcal{R}) \leq \left|\left\{(s_1, s_2, \cdots, s_k) \in \mathfrak{S}(k) \mid f_{s_1}(\mathbf{x}) \leq f_{s_2}(\mathbf{x}) \leq \cdots \leq f_{s_k}(\mathbf{x}), \mathbf{x} \in \mathbb{R}^n\right\}\right|. \tag{46}$$

Finally, Lemma 9 proves the statement by bounding number of feasible ascending orders. □

B.2 Proof of Lemma 2

Proof. It suffices to show that

$$g(\mathbf{x}) = \max_{i\in[k]} x_i. \tag{47}$$

for all $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_k \end{bmatrix}^\mathsf{T} \in \mathbb{R}^k$ since the composition of affine functions is still affine. The affine functions can be absorbed into the first layer of the ReLU network g. We prove the case for taking the maximum of m real numbers since the same procedure below can be applied to prove the case of taking the minimum due to the following identity

$$\min_{i\in[k]} f_i(\mathbf{x}) = -\max_{i\in[k]} -f_i(\mathbf{x}). \tag{48}$$

Because $\max(x_1, x_2) = \max(0, x_2 - x_1) + \max(0, x_1) - \max(0, -x_1)$ for any $x_1 \in \mathbb{R}$ and $x_2 \in \mathbb{R}$, it holds true that

$$\max_{j\in[k]} x_j = \begin{cases} \max_{j\in\left[\frac{k}{2}\right]} \max_{i\in\{2j-1,2j\}} x_i, & \text{if } k \text{ is even} \\ \max_{j\in\left[\frac{k+1}{2}\right]} \alpha(j; x_1, x_2, \cdots, x_k), & \text{if } k \text{ is odd} \end{cases} \tag{49}$$

for $x_j \in \mathbb{R}, j \in [k]$ where

$$\alpha(j; x_1, x_2, \cdots, x_k) = \begin{cases} \max_{i\in\{2j-1,2j\}} x_i, & \text{if } j \in \left[\frac{k-1}{2}\right] \\ \max(0, x_k) - \max(0, -x_k), & \text{if } j = \frac{k+1}{2} \end{cases}. \tag{50}$$

Let $r(k)$ be the number of operations of taking the maximum between a zero and a real number, i.e., $\max(0, x), x \in \mathbb{R}$ for computing the maximum of k real numbers using (49). One can find $r(2) = 3$ and $r(3) = 8$ by expanding all operations in (49). Because we do not need any maximum operations to compute the maximum over a singleton, we define $r(1) = 0$. For any positive integer k such that $k \geq 2$, we have the recursion

$$r(k) = \begin{cases} \frac{3k}{2} + r\left(\frac{k}{2}\right), & \text{if } k \text{ is even} \\ 2 + \frac{3(k-1)}{2} + r\left(\frac{k+1}{2}\right), & \text{if } k \text{ is odd} \end{cases} \tag{51}$$

according to (49). Note that $r(n)$ is the number of ReLUs in a ReLU network g that computes the maximum of n real numbers or a *max-affine* function. The number of ReLUs here is equivalent to the number of hidden neurons according to Definition 4. We shall note that the number of ReLU layers is equivalent to the number of hidden layers.

Obviously, we only need a 1-layer ReLU network with no ReLUs to compute the maximum of a singleton. Suppose that we aim to compute the maximum of $m = 2^n$ real numbers for any positive integer n. Then, every time the recursion goes to the next level in (51), the number of variables considered for computing the maximum is halved. Hence, the number of ReLU layers is n. When m is not a power of two, i.e., $2^n < m < 2^{n+1}$, then we can always construct a ReLU network with $n + 1$ ReLU layers and 2^{n+1} input neurons, and set weights connected to the $2^{n+1} - m$ "phantom input neurons" to zeros. Because $\lceil \log_2 m \rceil = n + 1$ for $2^n < m < 2^{n+1}$, the number of ReLU layers is $\lceil \log_2 m \rceil$ for any positive integer m. By Definition 5, we have $l(m) = \lceil \log_2 m \rceil + 1$.

By Lemma 5, $r(k)$ is a strictly increasing sequence. Therefore, the maximum width of the network is given by the width of the first hidden layer. When $L = 1$ or $m = 1$, the width is 0 due to Definition 5. When $L > 1$ or $m > 1$,

$$\begin{aligned} \max_{l\in[L-1]} k_l &= \begin{cases} \frac{3m}{2}, & \text{if } m \text{ is even} \\ 2 + \frac{3(m-1)}{2}, & \text{if } m \text{ is odd} \end{cases} \\ &= \left\lceil \frac{3m}{2} \right\rceil. \end{aligned} \tag{52}$$

Algorithm 2 directly follows from the above construction. Its complexity analysis is deferred to Table 2 in Appendix C. □

B.3 Proof of Theorem 2

Proof. Let $f_1, f_2, \cdots, f_k$ be k distinct linear components of p and $\mathcal{Q}$ be any family of closed convex subsets of $\mathbb{R}^n$ satisfying Definition 1. By Theorem 4.2 in [Tarela and Martinez, 1999], p can be represented as

$$p(\mathbf{x}) = \max_{\mathcal{X} \in \mathcal{Q}} \min_{i \in \mathcal{A}(\mathcal{X})} f_i(\mathbf{x}) \tag{53}$$

for all $\mathbf{x} \in \mathbb{R}^n$ where

$$\mathcal{A}(\mathcal{X}) = \left\{ i \in [k] \mid f_i(\mathbf{x}) \geq p(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X} \right\} \tag{54}$$

is the set of indices of linear components that have values greater than or equal to $p(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. A thorough discussion of the representation (53) is given in Section 4.2.

According to (53), there are $|\mathcal{Q}|$ minima required to be computed where each of them is a minimum of $\left|\mathcal{A}(\mathcal{X})\right|$ real numbers. Then, the value of p can be computed by taking the maximum of the resulting $|\mathcal{Q}|$ minima. We will show that these operations are realizable by a ReLU network. By Lemma 2, an $l(m)$-layer ReLU network with $r(m)$ hidden neurons and a maximum width of $w(m)$ can compute the extremum of m real numbers given by m affine functions.

We realize (53) in three steps. First, we create $|\mathcal{Q}|$ ReLU networks where each of them is an $l\left(\left|\mathcal{A}(\mathcal{X})\right|\right)$-layer ReLU network with $r\left(\left|\mathcal{A}(\mathcal{X})\right|\right)$ hidden neurons and a maximum width of $w\left(\left|\mathcal{A}(\mathcal{X})\right|\right)$ that computes $\min_{i \in \mathcal{A}(\mathcal{X})} f_i(\mathbf{x})$ for $\mathcal{X} \in \mathcal{Q}$. Second, we parallelly concatenate these $|\mathcal{Q}|$ networks, i.e., put them in parallel and let them share the same input to obtain a ReLU network that takes $\mathbf{x}$ and outputs $|\mathcal{Q}|$ real numbers. Finally, we create an $l\left(|\mathcal{Q}|\right)$-layer ReLU network with $r\left(|\mathcal{Q}|\right)$ hidden neurons and a maximum width of $w\left(|\mathcal{Q}|\right)$ that takes the maximum of $|\mathcal{Q}|$ real numbers.

The parallel combination of $|\mathcal{Q}|$ networks in the second step can be realized by Lemma 8. The third step can be fulfilled by Lemma 4. With the above construction, we can now count the number of layers, the upper bound for the maximum width, and the number of hidden neurons for a ReLU network that realizes p. The number of layers is given by

$$l\left(|\mathcal{Q}|\right) + \max_{\mathcal{X} \in \mathcal{Q}} l\left(\left|\mathcal{A}(\mathcal{X})\right|\right) - 1. \tag{55}$$

The maximum width is bounded from above by

$$\max\left(\sum_{\mathcal{X} \in \mathcal{Q}} \max\left(w\left(\left|\mathcal{A}(\mathcal{X})\right|\right), 2\right), w\left(|\mathcal{Q}|\right)\right). \tag{56}$$

The number of hidden neurons is give by

$$r\left(|\mathcal{Q}|\right) + \sum_{\mathcal{X} \in \mathcal{Q}} r\left(\left|\mathcal{A}(\mathcal{X})\right|\right) + 2\left(\max_{\mathcal{Y} \in \mathcal{Q}} l\left(\left|\mathcal{A}(\mathcal{Y})\right|\right) - l\left(\left|\mathcal{A}(\mathcal{X})\right|\right)\right). \tag{57}$$

Because $\mathcal{A}(\mathcal{X})$ for every $\mathcal{X} \in \mathcal{Q}$ is a subset of $[k]$, it holds that

$$1 \leq \left|\mathcal{A}(\mathcal{X})\right| \leq k \tag{58}$$

for all $\mathcal{X} \in \mathcal{Q}$. Therefore, the number of layers in (55) can be bounded from above by

$$l\left(|\mathcal{Q}|\right) + l(k) - 1 = \lceil \log_2 |\mathcal{Q}| \rceil + \lceil \log_2 k \rceil + 1 \tag{59}$$

where we have used the definition of the function l in Lemma 2. Again, using (58), the upper bound for the maximum width in (56) can be further bounded from above by

$$\begin{aligned}
\max\left(\sum_{\mathcal{X} \in \mathcal{Q}} \max\left(w(k), 2\right), w\left(|\mathcal{Q}|\right)\right) &= \max\left(|\mathcal{Q}| \max\left(w(k), 2\right), w\left(|\mathcal{Q}|\right)\right) \\
&\leq \max\left(|\mathcal{Q}| \max\left(\left\lceil \frac{3k}{2} \right\rceil, 2\right), \left\lceil \frac{3|\mathcal{Q}|}{2} \right\rceil\right) \\
&= \left\lceil \frac{3k}{2} \right\rceil |\mathcal{Q}|
\end{aligned} \tag{60}$$

where we have used the definition of the function w in Lemma 2. Note that the maximum width is zero when the number of layers is one. Finally, again, using (58), the number of neurons in (57) can be bounded from above by

$$\begin{aligned}
& r\left(|\mathcal{Q}|\right) - 2l(k) + 2l(1) + \sum_{\mathcal{X}\in\mathcal{Q}} \left(r\left(k\right) + 2l\left(k\right) - 2l(1)\right) \\
&= r\left(|\mathcal{Q}|\right) - 2l(k) + 2l(1) + |\mathcal{Q}| \left(r\left(k\right) + 2l\left(k\right) - 2l(1)\right) \\
&= r\left(|\mathcal{Q}|\right) - 2\left\lceil \log_2 k \right\rceil + |\mathcal{Q}| \left(r\left(k\right) + 2\left\lceil \log_2 k \right\rceil\right) \\
&\leq 3\left(2^{\left\lceil \log_2 |\mathcal{Q}| \right\rceil} - 1\right) - 2\left\lceil \log_2 k \right\rceil + |\mathcal{Q}| \left(3\left(2^{\left\lceil \log_2 k \right\rceil} - 1\right) + 2\left\lceil \log_2 k \right\rceil\right) \\
&= 3\left(2^{\left\lceil \log_2 |\mathcal{Q}| \right\rceil} - 1\right) + 3|\mathcal{Q}| \left(2^{\left\lceil \log_2 k \right\rceil} - 1\right) + 2\left(|\mathcal{Q}| - 1\right)\left\lceil \log_2 k \right\rceil
\end{aligned} \tag{61}$$

where we have used Lemma 6 for the upper bound in the fourth line of (61). Expanding and rearranging terms in (61) lead to (12).

Algorithm 1 directly follows from the above construction. Its complexity analysis is deferred to Table 1 in Appendix C. □

B.4 Proof of Theorem 1

Proof. By Lemma 1, the number of distinct linear components k is bounded from above by the number of pieces, i.e., $k \leq q$, implying that the bounds in Theorem 2 can be written in terms of q. Substituting k with q in Theorem 2 proves the claim.

According to Theorem 2, the time complexity of Algorithm 1 is poly(n, k, q). Using the bound $k \leq q$ proves the claim for the time complexity. □

B.5 Proof of Theorem 3

Proof. By Lemma 1, the minimum number of closed convex subsets q of a CPWL function $p\colon \mathbb{R}^n \to \mathbb{R}$ can be bounded from above by $\phi(n, k)$, i.e.,

$$q \leq \phi(n, k) = \min\left(\sum_{i=0}^{n} \binom{\frac{k^2-k}{2}}{i}, k!\right). \tag{62}$$

Substituting q with $\phi(n, k)$ in Theorem 2 proves the claim. □

B.6 Proof of Lemma 3

Proof. Obviously, a one-layer ReLU network is an affine function whose weights can be set to fulfill the identify mapping in $\mathbb{R}^n$. We prove the case when the number of layers is more than one in the next paragraph. We start with a scalar case, and then work on the vector case.

For any $x \in \mathbb{R}$, it holds that $\max(0, x) - \max(0, -x) = x$. In other words, a hidden layer of two ReLUs with $+1$ and -1 weights can represent an identity mapping for any scalar. For any vector input in $\mathbb{R}^n$, we can concatenate such structures of two ReLUs in parallel because the identity mapping can be decomposed into n individual identity mappings from n coordinates. Therefore, a two-layer ReLU network with $2n$ hidden neurons can realize the identity mapping in $\mathbb{R}^n$. Stacking such a hidden layer any number of times gives a deeper network that is still an identity mapping. Algorithm 5 follows from the above construction. Its complexity analysis is deferred to Table 5 in Appendix C. □

B.7 Proof of Lemma 4

Proof. Because a composition of two affine mappings is still affine, the first layer of either one of the two networks can be absorbed into the last layer of the other one if their dimensions are compatible. The resulting new network still satisfies Definition 4. The number of layers of the new network is $l_1 + l_2 - 1$. The number of hidden neurons of the new network is $n_1 + n_2$. The maximum width of the new network is at most $\max(w_1, w_2)$. Algorithm 4 follows from the above construction. Its complexity analysis is deferred to Table 4 in Appendix C. □

B.8 Proof of Lemma 5

Proof. For any positive even integer $k \geq 4$, it holds true that

$$r(k) - r(k-1) = \frac{3k}{2} + r\left(\frac{k}{2}\right) - 2 - \frac{3(k-2)}{2} - r\left(\frac{k}{2}\right) = 1. \tag{63}$$

For any positive odd integer k such that $k \geq 3$, we have

$$\begin{aligned} r(k) - r(k-1) &= 2 + \frac{3(k-1)}{2} + r\left(\frac{k+1}{2}\right) - \frac{3(k-1)}{2} - r\left(\frac{k-1}{2}\right) \\ &= \begin{cases} 3, & \text{if } \frac{k+1}{2} \text{ is even} \\ 2 + r\left(\frac{k+1}{2}\right) - r\left(\frac{k+1}{2} - 1\right), & \text{otherwise} \end{cases} \end{aligned} \tag{64}$$

which is strictly greater than zero. Note that (64) is greater than 0 because the equality in (64) can be applied over and over again to reach (63) or the base case $r(2) - r(1) = 3$. □

B.9 Proof of Lemma 6

Proof. By Lemma 5, $r(k)$ is a strictly increasing sequence. Then, it must be true that

$$r(k) = r\left(2^{\log_2 k}\right) \leq r\left(2^{\lceil \log_2 k \rceil}\right). \tag{65}$$

According to the recursion (23), it holds that

$$\begin{aligned} r\left(2^{\lceil \log_2 k \rceil}\right) &= \frac{3}{2} \sum_{i=1}^{\lceil \log_2 k \rceil} 2^i \\ &= \frac{3}{2}\left(2^{\lceil \log_2 k \rceil + 1} - 2\right) \\ &= 3\left(2^{\lceil \log_2 k \rceil} - 1\right) \\ &< 3\left(2^{(\log_2 k) + 1} - 1\right) \\ &= 3\left(2k - 1\right). \end{aligned} \tag{66}$$

□

B.10 Proof of Lemma 7

Proof. Two ReLU networks can be combined in parallel such that the new network shares the same input and the two output vectors from the two ReLU networks are concatenated together. To see this, we show that the weights of the new network can be found by the following operations. Let $\mathbf{W}_i^1$ and $\mathbf{b}_i^1$ be the weights of the i-th layer in $g_{(l_1,n_1,w_1)}$, and $\mathbf{W}_i^2$ and $\mathbf{b}_i^2$ are the weights of the i-th layer in $g_{(l_2,n_2,w_2)}$. Let $\mathbf{W}_i$ and $\mathbf{b}_i$ be the weights of the new network. Now, we find the weights for the new network. In the first layer, we construct

$$\mathbf{W}_1 = \begin{bmatrix} \mathbf{W}_1^1 \\ \mathbf{W}_1^2 \end{bmatrix} \tag{67}$$

and

$$\mathbf{b}_1 = \begin{bmatrix} \mathbf{b}_1^1 \\ \mathbf{b}_1^2 \end{bmatrix}. \tag{68}$$

For the i-th layer such that $1 < i \leq \min(l_1, l_2)$, we use

$$\mathbf{W}_i = \begin{bmatrix} \mathbf{W}_i^1 & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_i^2 \end{bmatrix} \tag{69}$$

and

$$\mathbf{b}_i = \begin{bmatrix} \mathbf{b}_i^1 \\ \mathbf{b}_i^2 \end{bmatrix}. \tag{70}$$

If $l_1 = l_2$, then the claim is proved. If $l_1 \neq l_2$, then we stack a network that implements the identity mapping to the shallower network such that the number of layers of the two networks are the same. Because the network $g_{(l_i,n_i,w_i)}$ is shallower than the other network, we append $|l_1 - l_2|$ hidden layers to $g_{(l_i,n_i,w_i)}$ such that the procedure in (69) and (70) can be used. By Lemma 3, there exists an $\big(|l_1 - l_2| + 1\big)$-layer ReLU network $g_{(|l_1-l_2|+1,2m_i|l_1-l_2|,2m_i)}$ with $2m_i|l_1 - l_2|$ hidden neurons and a maximum width bounded from above by $2m_i$ for representing the identity mapping in $\mathbb{R}^{m_i}$. By Lemma 4, there exists a network $g_{\big(l_i+|l_1-l_2|,n_i+2m_i|l_1-l_2|,\max(w_i,2m_i)\big)}$ that represents the composition of $g_{(|l_1-l_2|+1,2m_i|l_1-l_2|,2m_i)}$ and $g_{(l_i,n_i,w_i)}$. Now, (69) and (70) can be used to combine $g_{(l_j,n_j,w_j)}$ and $g_{\big(l_i+|l_1-l_2|,n_i+2m_i|l_1-l_2|,\max(w_i,2m_i)\big)}$ in parallel because the number of layers in network $g_{\big(l_i+|l_1-l_2|,n_i+2m_i|l_1-l_2|,\max(w_i,2m_i)\big)}$ is equal to l_j according to the fact that $l_i + |l_1 - l_2| = \max(l_1, l_2) = l_j$. Such a new network has $\max(l_1, l_2)$ layers and

$$n_j + n_i + 2m_i|l_1 - l_2| = n_1 + n_2 + 2m_i|l_1 - l_2| \tag{71}$$

hidden neurons. The maximum width of the new network is at most $w_j + \max(w_i, 2m_i)$. □

B.11 Proof of Lemma 8

Proof. The case $k = 1$ is trivial. The case $k = 2$ is proved by Lemma 7, which gives a tighter bound on the maximum width. The number of layers and hidden neurons of the claim agree with Lemma 7 when $k = 2$. The claim can be proved by following a similar procedure from the proof of Lemma 7. By Lemma 3, we can stack an identity mapping realized by an $(l - l_i + 1)$-layer ReLU network with $2m_i(l - l_i)$ hidden neurons and a maximum width of $2m_i$ on the i-th network for all $i \in [k]$ such that $l_i < l$. In other words, we increase the number of hidden layers for any network whose number of layers is less than l such that the cascade of the network and the corresponding identity mapping has l layers. For all $i \in [k]$ such that $l_i < l$, the extended network has $n_i + 2m_i(l - l_i)$ hidden neurons and a maximum width at most $\max(w_i, 2m_i)$ according to Lemma 4. Because all the networks now have the same number of layers, we can directly combine them in parallel. Hence, the resulting new network has $\max_{i\in[k]} l_i$ layers and

$$\sum_{i\in[k]} n_i + 2m_i(l - l_i) \tag{72}$$

hidden neurons and a maximum width at most

$$\sum_{i\in[k]} \max(w_i, 2m_i). \tag{73}$$

Algorithm 3 directly follows from the above construction. Its complexity analysis is deferred to Table 3 in Appendix C. □

B.12 Proof of Lemma 9

Proof. Because $\mathcal{S}^n_{f_1,f_2,\cdots,f_k}$ is a subset of $\mathfrak{S}(k)$ and $\big|\mathfrak{S}(k)\big| = k!$ is the number of permutations of k distinct objects, it follows that

$$\left|\mathcal{S}^n_{f_1,f_2,\cdots,f_k}\right| \leq k!. \tag{74}$$

On the other hand, the number of hyperplanes, or affine subspaces of $\mathbb{R}^n$ with dimension $n - 1$, induced by the distinct intersections between any two different affine functions is bounded from above by

$$\binom{k}{2}. \tag{75}$$

Let the arrangement of these hyperplanes be $\mathcal{A}$, and $|\mathcal{A}|$ be the number of hyperplanes in the arrangement. By Zaslavsky's Theorem [Zaslavsky, 1975], the number of connected components of the set

$$\mathbb{R}^n \setminus \bigcup_{H\in\mathcal{A}} H \tag{76}$$

is bounded from above by

$$\sum_{i=0}^{n} \binom{|\mathcal{A}|}{i} \tag{77}$$

Because there are at most $\binom{k}{2}$ hyperplanes in $\mathbb{R}^n$, it follows that

$$\left|\mathcal{S}^n_{f_1,f_2,\cdots,f_k}\right| \leq \sum_{i=0}^{n} \binom{\binom{k}{2}}{i}. \tag{78}$$

Combining (74) and (78) proves the claim. Notice that the ascending order does not change within a connected component. □

B.13 Proof of Lemma 10

Proof. Let $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_I$ be a family of subsets satisfying Definition 1 for a non-affine function. We prove the claim by contradiction. Suppose that there exists at least one nonempty closed subset, say $\mathcal{X}_i$, that is disjoint with every other closed subset $\mathcal{X}_j, j \in [I] \setminus i$. It follows that

$$\mathcal{X}_i \bigcap \bigcup_{j\in[I]\setminus i} \mathcal{X}_j = \emptyset \tag{79}$$

which implies

$$\left(\mathbb{R}^n \setminus \mathcal{X}_i\right) \bigcup \left(\mathbb{R}^n \setminus \bigcup_{j\in[I]\setminus i} \mathcal{X}_j\right) = \mathbb{R}^n. \tag{80}$$

Because the union of any finite collection of closed sets is closed, it must be true that $\bigcup_{j\in[I]\setminus i} \mathcal{X}_j$ is closed. Notice that $\mathcal{X}_i$ is never the whole space $\mathbb{R}^n$ because the CPWL function is assumed to be non-affine. $\bigcup_{j\in[I]\setminus i} \mathcal{X}_j$ must be nonempty due to Definition 1. Therefore, both $\mathbb{R}^n \setminus \mathcal{X}_i$ and $\mathbb{R}^n \setminus \bigcup_{j\in[I]\setminus i} \mathcal{X}_j$ are nonempty and open. Since $\mathbb{R}^n$ is connected, it cannot be represented as the union of two disjoint nonempty open subsets. It follows that the intersection between $\mathbb{R}^n \setminus \mathcal{X}_i$ and $\mathbb{R}^n \setminus \bigcup_{j\in[I]\setminus i} \mathcal{X}_j$ is nonempty. In other words, there exists an element of $\mathbb{R}^n$ that is not in $\mathcal{X}_i$ and $\bigcup_{j\in[I]\setminus i} \mathcal{X}_j$, contradicting Definition 1. □

B.14 Proof of Lemma 11(a)

Proof. If the CPWL function is affine, then there are no intersecting closed subsets because the only closed subset satisfying Assumption 1 is $\mathbb{R}^n$. On the other hand, if the CPWL function is non-affine, then there exist at least two intersecting closed subsets according to Lemma 10. For any two intersecting closed subsets, say $\mathcal{X}_i$ and $\mathcal{X}_j$, we first show that

$$\{\mathbf{x} \in \mathbb{R}^n \mid f_i(\mathbf{x}) = f_j(\mathbf{x})\} \neq \emptyset \tag{81}$$

where f_i and f_j are the affine functions corresponding to $\mathcal{X}_i$ and $\mathcal{X}_j$. We prove this statement by contradiction. Suppose that the intersection is empty, i.e., the linear equation $\left(\mathbf{a}_i - \mathbf{a}_j\right)^T \mathbf{x} + b_i - b_j = 0$ does not have a solution where $f_i(\mathbf{x}) = \mathbf{a}_i^T\mathbf{x} + b_i$ and $f_j(\mathbf{x}) = \mathbf{a}_j^T\mathbf{x} + b_j$ for $\mathbf{a}_i, \mathbf{a}_j \in \mathbb{R}^n$ and $b_i, b_j \in \mathbb{R}$. Then, it is necessary that $\mathbf{a}_i = \mathbf{a}_j$ and $b_i \neq b_j$. In other words, the two affine functions are parallel, implying that every point in $\mathcal{X}_i \cap \mathcal{X}_j$ gives two different values, which cannot be true for a valid function.

Next, we prove that there does not exist an intersection that is $\mathbb{R}^n$ by contradiction. Let us assume that there exists at least one intersection that is $\mathbb{R}^n$ between the affine functions corresponding to two intersecting closed subsets, say $\mathcal{X}_i$ and $\mathcal{X}_j$. Then, we can always replace $\mathcal{X}_i$ and $\mathcal{X}_j$ with their union. Such a replacement still satisfies Definition 1 but reduces the number of closed (connected) subsets by at least one, contradicting the fact that the number of closed subsets is a minimum. Because the two affine functions are identical if and only if the intersection is $\mathbb{R}^n$, the two affine functions must be different. □

B.15 Proof of Lemma 11(b)

Proof. The claim follows from Lemma 11(a). Because the two affine functions have nonempty intersection, their intersection must be $\mathbb{R}^n$ or an affine subspace of $\mathbb{R}^n$ with dimension $n-1$. However, the two affine functions must be different, implying that $\mathbb{R}^n$ is never the intersection. □

B.16 Proof of Lemma 11(c)

Proof. Let any given two intersecting subsets be $\mathcal{X}_i$ and $\mathcal{X}_j$. The intersection between their corresponding affine functions, say f_i and f_j, is given by $\mathcal{H}_{ij} = \{\mathbf{x} \in \mathbb{R}^n \mid f_i(\mathbf{x}) = f_j(\mathbf{x})\}$. Suppose that there exists a point $\mathbf{a} \in \mathcal{X}_i \bigcap \mathcal{X}_j$ such that $\mathbf{a} \notin \mathcal{H}_{ij}$, then it follows that $f_i(\mathbf{a}) \neq f_j(\mathbf{a})$. Such a result cannot be true for a valid function. We conclude that $\mathcal{X}_i \bigcap \mathcal{X}_j \subseteq \mathcal{H}_{ij}$. □

B.17 Proof of Lemma 11(d)

Proof. We prove the statement by contradiction. Suppose there exists a point $\mathbf{c} \in \mathbb{R}^n$ in the intersection of two intersecting closed connected subsets, say $\mathcal{X}_i$ and $\mathcal{X}_j$, such that $\mathbf{c}$ is an interior point of $\mathcal{X}_i$, then there exists an open ϵ-radius ball $B(\mathbf{c}, \epsilon)$ such that $\mathbf{x} \in \mathcal{X}_i, \forall \mathbf{x} \in B(\mathbf{c}, \epsilon)$ for some $\epsilon > 0$. By Lemma 11(b), the intersection between the two affine functions corresponding to $\mathcal{X}_i$ and $\mathcal{X}_j$ must be an affine subspace of $\mathbb{R}^n$ with dimension $n-1$. Let such an affine subsapce be denoted as $\mathcal{H}_{ij}$ and its corresponding linear subspace be denoted as $\mathcal{V}(\mathcal{H}_{ij})$. Then, there exists a nonzero vector $\mathbf{d} \in \mathbb{R}^n$ such that $\alpha\mathbf{d} \perp \mathbf{v}$ for all $\mathbf{v} \in \mathcal{V}(\mathcal{H}_{ij})$ and any $\alpha \neq 0$. Therefore, it follows that $\alpha\mathbf{d} + \mathbf{a} \notin \mathcal{H}_{ij}$ for any $\mathbf{a} \in \mathcal{H}_{ij}$ and any $\alpha \neq 0$. According to Lemma 11(c), $\mathcal{X}_i \cap \mathcal{X}_j \subseteq \mathcal{H}_{ij}$, so we have $\alpha\mathbf{d} + \mathbf{c} \notin \mathcal{X}_i \cap \mathcal{X}_j$ for any $\alpha \neq 0$. When $\alpha = \frac{\epsilon}{2\|\mathbf{d}\|_2}$ or $\alpha = \frac{-\epsilon}{2\|\mathbf{d}\|_2}$, $\alpha\mathbf{d} + \mathbf{c} \in B(\mathbf{c}, \epsilon)$. However, one of them must satisfy $\alpha\mathbf{d} + \mathbf{c} \notin \mathcal{X}_i$, contradicting the existence of a point in $\mathcal{X}_i \cap \mathcal{X}_j$ that is an interior point of $\mathcal{X}_i$. The same procedure can be applied to prove that there does not exist a point in $\mathcal{X}_i \cap \mathcal{X}_j$ such that it is an interior point of $\mathcal{X}_j$. We conclude that every element in $\mathcal{X}_i \cap \mathcal{X}_j$ is not an interior point of $\mathcal{X}_i$ or $\mathcal{X}_j$. □

B.18 Proof of Lemma 12(a)

Proof. Let $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_I$ be a finite number of closed subsets satisfying Assumption 1. The boundary or frontier of $\mathcal{X}_i$ is given by

$$\begin{aligned}
\mathrm{Fr}\mathcal{X}_i &= \overline{\mathcal{X}_i} \bigcap \overline{\mathbb{R}^n \setminus \mathcal{X}_i} \\
&= \mathcal{X}_i \bigcap \overline{\left(\bigcup_{k \in [I]} \mathcal{X}_k \right) \setminus \mathcal{X}_i} \\
&= \mathcal{X}_i \bigcap \overline{\bigcup_{k \in [I] \setminus i} \left(\mathcal{X}_k \setminus \mathcal{X}_k \bigcap \mathcal{X}_i \right)} \\
&= \mathcal{X}_i \bigcap \bigcup_{k \in [I] \setminus i} \overline{\left(\mathcal{X}_k \setminus \mathcal{X}_k \bigcap \mathcal{X}_i \right)} \\
&= \mathcal{X}_i \bigcap \bigcup_{k \in [I] \setminus i} \overline{\mathcal{X}_k} \\
&= \mathcal{X}_i \bigcap \bigcup_{k \in [I] \setminus i} \mathcal{X}_k \\
&= \bigcup_{k \in [I] \setminus i} \mathcal{X}_k \bigcap \mathcal{X}_i
\end{aligned} \tag{82}$$

where $\overline{\mathcal{A}}$ denotes the closure of a subset $\mathcal{A}$. We have used Lemma 11(d) for the equality between the 4-th and 5-th line of (82). Now, we prove that the interior of $\mathcal{X}_i$ is nonemtpy by contradiction. Suppose that the interior of $\mathcal{X}_i$ is empty, then it follows that $\mathcal{X}_i = \overline{\mathcal{X}_i} = \mathrm{Fr}\mathcal{X}_i$ because the closure of $\mathcal{X}_i$ is the union of the interior and the boundary of $\mathcal{X}_i$. Combining that with (82), we have $\mathcal{X}_i = \bigcup_{k \in [I] \setminus i} \mathcal{X}_k \bigcap \mathcal{X}_i$. which implies every element in $\mathcal{X}_i$ is at least covered by one of the other closed subset $\mathcal{X}_k$ for some $k \in [I] \setminus i$. In this case, we can delete $\mathcal{X}_i$ from $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_I$; and the remaining $I-1$ closed subsets still satisfy Definition 1. Such a valid deletion of $\mathcal{X}_i$ contradicts the fact that I is the minimum number of closed subsets. Hence, the interior of $\mathcal{X}_i$ must be nonempty. □

B.19 Proof of Lemma 12(b)

Proof. The statement is proved by (82) in Lemma 12(a). □

B.20 Proof of Lemma 12(c)

Proof. By Lemma 12(a), the interior of every subset is nonempty. Next, by Lemma 11(d), every point in the intersection between any two subsets is a boundary point of both subsets. It follows that the interiors of any two subsets are disjoint. □

B.21 Proof of Lemma 13

Proof. By Lemma 12(a), the interior of $\mathcal{X}_i$ is nonempty. Therefore, there exists an open ϵ-radius ball $B(\mathbf{c}_0, \epsilon)$ such that $\mathbf{x} \in \mathcal{X}_i, \forall \mathbf{x} \in B(\mathbf{c}_0, \epsilon)$ for some $\epsilon > 0$ and $\mathbf{c}_0 \in \mathcal{X}_i$. Let us consider the set

$$\bigcap_{j \in [k]} \left(B(\mathbf{c}_0, \epsilon) \bigcap \left(\mathcal{H}_j^+ \bigcup \mathcal{H}_j^- \right) \right) \tag{83}$$

where $\mathcal{H}_j^+$ and $\mathcal{H}_j^-$ are two open half spaces created by $\mathcal{H}_j$. It suffices to show that the nonemptyness of the set in (83) to prove the claim. If $\mathcal{H}_j$ and $B(\mathbf{c}_0, \epsilon)$ do not intersect, then $B(\mathbf{c}_0, \epsilon)$ completely belongs to $\mathcal{H}_j^+$ or $\mathcal{H}_j^-$. Without loss of generality, we can remove all j such that $\mathcal{H}_j$ does not intersect $B(\mathbf{c}_0, \epsilon)$ and assume there are k affine subspaces of $\mathbb{R}^n$ with dimension $n-1$ intersecting $B(\mathbf{c}_0, \epsilon)$. Let us sequentially carry out the intersection in (83). Every time before the operation of the j-th intersection between $B(\mathbf{c}_{j-1}, \frac{\epsilon}{2^{j-1}})$ and $\left(\mathcal{H}_j^+ \bigcup \mathcal{H}_j^-\right)$, there exists an open $\frac{\epsilon}{2^j}$-radius ball $B(\mathbf{c}_j, \frac{\epsilon}{2^j})$ for some $\mathbf{c}_j \in B(\mathbf{c}_{j-1}, \frac{\epsilon}{2^{j-1}})$ such that it does not intersect with $\mathcal{H}_j$. Therefore, at the end of the sequential process, there exists an open ball that does not intersect any of these k affine subspaces of $\mathbb{R}^n$ with dimension $n-1$. The set in (83) is nonempty, implying (35) holds true. □

B.22 Proof of Proposition 1

Proof. We prove the claim by contraposition. If the number of closed connected subsets is not a minimum, i.e., Assumption 1 is not satisfied, then such a number can be decreased by merging some of the intersecting closed connected subsets that have the same corresponding affine functions. Therefore, there exist at least two closed connected subsets that can be made larger.

On the other hand, if the closed connected subsets, say $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_I$, have at least one of subsets that can be made larger, then there exist at least two intersecting closed connected subsets, say $\mathcal{X}_i$ and $\mathcal{X}_j$, from $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_I$ such that their corresponding affine functions are the same. Otherwise, any closed connected subset cannot be made larger than itself. Therefore, $\mathcal{X}_i$ and $\mathcal{X}_j$ can be replaced with $\mathcal{X}_i \cup \mathcal{X}_j$ and these $I-1$ closed connected subsets still satisfy Definition 1, implying that I is not the minimum. □

C Algorithms and time complexities

Table 1: The running time of Algorithm 1 is upper bounded by poly(n, k, q).

Line	Operation count	Explanation
1	$\mathcal{O}\left(nq \max(n^2, q)\right)$	Algorithm 6 (see Table 6).
2	$\mathcal{O}(q)$	Repeat Line 3 to Line 9 q times.
3	$\mathcal{O}(1)$	Initialize an empty placeholder.
4	$\mathcal{O}(k)$	Repeat Line 5 to Line 7 k times.
5	poly(n, q)	Solve a linear program.
6	$\mathcal{O}(1)$	Add an index.
7	-	-
8	-	-
9	$\mathcal{O}\left(k^2 \max(k \log_2 k, n)\right)$	Algorithm 2 (see Table 2).
10	-	-
11	$\mathcal{O}\left(q \max(n, k)^2 \max(n, k, q) \log_2 k\right)$	Algorithm 3 (see Table 3).
12	$\mathcal{O}\left(q^3 \log_2 q\right)$	Algorithm 2 (see Table 2).
13	$\mathcal{O}\left(q^3 \max(n, k)^3 \log_2 q\right)$	Algorithm 4 (see Table 4).

Algorithm 2 Find a ReLU network that computes the extremum of affine functions

Input: Scalar-valued affine functions $f_1, f_2, \cdots, f_m$ on $\mathbb{R}^n$ and the type of extremum (max or min).
Output: Parameters of an l-layer ReLU network g computing $g(\mathbf{x}) = \max_{i\in[m]} . f_i(\mathbf{x})$ or $g(\mathbf{x}) = \min_{i\in[m]} . f_i(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$.

$\mathbf{A} \leftarrow \begin{bmatrix} -1 & 1 \\ 1 & 0 \\ -1 & 0 \end{bmatrix}, \mathbf{B} \leftarrow \begin{bmatrix} 1 & 1 & -1 \end{bmatrix}, \mathbf{C} \leftarrow \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ ▷ Constant matrices
$\mathbf{\Psi}(\mathbf{Y}, \mathbf{Z}) \leftarrow \begin{bmatrix} \mathbf{Y} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z} \end{bmatrix}$ ▷ A function generating a block diagonal matrix composed of $\mathbf{Y}$ and $\mathbf{Z}$
$\mathbf{\Phi}(\mathbf{Y}, s) \leftarrow \begin{bmatrix} \mathbf{Y}^{(1)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Y}^{(2)} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Y}^{(s)} \end{bmatrix}$ ▷ A block diagonal matrix with $\mathbf{Y}$ repeated s times
$l \leftarrow \lceil \log_2 m \rceil + 1, k_0 \leftarrow n, k_l \leftarrow 1, c_0 \leftarrow m$ ▷ l is the number of layers of g
for $i = 1, 2, \cdots, l-1$ **do**
  if c_{i-1} is even **then**
    $c_i \leftarrow \frac{c_{i-1}}{2}$
    $k_i \leftarrow 3c_i$ ▷ Output dimension of the i-th layer
  else
    $c_i \leftarrow \frac{c_{i-1}+1}{2}$
    $k_i \leftarrow 3c_i - 1$ ▷ Output dimension of the i-th layer
  end if
end for
$\mathbf{W}_1 \leftarrow \begin{bmatrix} \nabla f_1 & \nabla f_2 & \cdots & \nabla f_m \end{bmatrix}^\mathsf{T}, \mathbf{b}_1 \leftarrow \begin{bmatrix} f_1(0) & f_2(0) & \cdots & f_m(0) \end{bmatrix}^\mathsf{T}$
if $l > 1$ **then** ▷ Find the weights of input and output layers, if any
  if c_0 is even **then**
    $\mathbf{W}_1 \leftarrow \mathbf{\Phi}(\mathbf{A}, c_1)\mathbf{W}_1, \mathbf{b}_1 \leftarrow \mathbf{\Phi}(\mathbf{A}, c_1)\mathbf{b}_1$
  else
    $\mathbf{W}_1 \leftarrow \mathbf{\Psi}\left(\mathbf{\Phi}(\mathbf{A}, c_1 - 1), \mathbf{C}\right)\mathbf{W}_1, \mathbf{b}_1 \leftarrow \mathbf{\Psi}\left(\mathbf{\Phi}(\mathbf{A}, c_1 - 1), \mathbf{C}\right)\mathbf{b}_1$
  end if
  $\mathbf{W}_l \leftarrow \mathbf{B}, \mathbf{b}_l \leftarrow \mathbf{0}_{k_l}$
end if
if $l > 2$ **then** ▷ Find the weights of remaining layers, if any
  for $i = 2, 3, \cdots, l-1$ **do**
    if c_{i-1} is even **then**
      $\mathbf{T} \leftarrow \mathbf{\Phi}(\mathbf{A}, c_i)$
    else
      $\mathbf{T} \leftarrow \mathbf{\Psi}\left(\mathbf{\Phi}(\mathbf{A}, c_i - 1), \mathbf{C}\right)$
    end if
    if c_{i-2} is even **then**
      $\mathbf{W}_i \leftarrow \mathbf{T}\mathbf{\Phi}(\mathbf{B}, c_{i-1})$
    else
      $\mathbf{W}_i \leftarrow \mathbf{T}\mathbf{\Psi}\left(\mathbf{\Phi}(\mathbf{B}, c_{i-1} - 1), \mathbf{C}^\mathsf{T}\right)$
    end if
    $\mathbf{b}_i \leftarrow \mathbf{0}_{k_i}$
  end for
end if
if type of extremum is the minimum **then**
  $\mathbf{W}_1 \leftarrow -\mathbf{W}_1, \mathbf{b}_1 \leftarrow -\mathbf{b}_1$
  $\mathbf{W}_l \leftarrow -\mathbf{W}_l, \mathbf{b}_l \leftarrow -\mathbf{b}_l$
end if ▷ See Table 2 in Appendix C for complexity analysis

Algorithm 3 Find a ReLU network that concatenates a number of given ReLU networks

Input: Weights of k ReLU networks $g_1, g_2, \cdots, g_k$ denoted by $\{\mathbf{W}_i^j, \mathbf{b}_i^j\}_{i=1}^{l_j}$ for $j \in [k]$.

Output: Parameters of an l-layer ReLU network g computing $g(\mathbf{x}) = \begin{bmatrix} g_1(\mathbf{x}) \\ g_2(\mathbf{x}) \\ \vdots \\ g_k(\mathbf{x}) \end{bmatrix}, \forall \mathbf{x} \in \mathbb{R}^n$.

$l \leftarrow \max_{j \in [k]} l_j$

$\mathbf{W}_1 \leftarrow \begin{bmatrix} \mathbf{W}_1^1 \\ \mathbf{W}_1^2 \\ \vdots \\ \mathbf{W}_1^k \end{bmatrix}, \mathbf{b}_1 \leftarrow \begin{bmatrix} \mathbf{b}_1^1 \\ \mathbf{b}_1^2 \\ \vdots \\ \mathbf{b}_1^k \end{bmatrix}$ ▷ Weights of the input layer

for $j = 1, 2, \cdots, k$ **do**

  if $l_j < l$ **then** ▷ Append an identity mapping network to the network if it is shallower

    $m \leftarrow$ output dimsion of g_j

    $g_j^c \leftarrow$ run Algorithm 5 with an input dimension m and a number of layers $l - l_j + 1$

    $g_j' \leftarrow$ run Algorithm 4 with g_j and g_j^c

    $\{\mathbf{W}_i^j, \mathbf{b}_i^j\}_{i=1}^{l} \leftarrow$ weights of g_j'

  end if

end for

for $i = 2, 3, \cdots, l$ **do** ▷ Find the remaining weights

  $\mathbf{W}_i \leftarrow \begin{bmatrix} \mathbf{W}_i^1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_i^2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{W}_i^k \end{bmatrix}, \mathbf{b}_i \leftarrow \begin{bmatrix} \mathbf{b}_i^1 \\ \mathbf{b}_i^2 \\ \vdots \\ \mathbf{b}_i^k \end{bmatrix}$

end for ▷ See Table 3 in Appendix C for complexity analysis

Algorithm 4 Find a ReLU network computing a composition of two given ReLU networks

Input: Weights of two ReLU networks g_1 and g_2 denoted by $\{\mathbf{W}_i^1, \mathbf{b}_i^1\}_{i=1}^{l_1}$ and $\{\mathbf{W}_i^2, \mathbf{b}_i^2\}_{i=1}^{l_2}$.

Output: Parameters of an l-layer ReLU network g computing $g(\mathbf{x}) = g_2\left(g_1(\mathbf{x})\right), \forall \mathbf{x} \in \mathbb{R}^n$.

$l \leftarrow l_1 + l_2 - 1$

for $i = 1, 2, \cdots, l$ **do**

  if $i < l_1$ **then** ▷ The first $l_1 - 1$ layers are identical to the corresponding layers in g_1

    $\mathbf{W}_i \leftarrow \mathbf{W}_i^1, \mathbf{b}_i \leftarrow \mathbf{b}_i^1$

  else if $i = l_1$ **then** ▷ A composition of affine functions is still an affine function

    $\mathbf{W}_i \leftarrow \mathbf{W}_1^2 \mathbf{W}_{l_1}^1, \mathbf{b}_i \leftarrow \mathbf{W}_1^2 \mathbf{b}_{l_1}^1 + \mathbf{b}_1^2$

  else ▷ The last $l_2 - 1$ layers are identical to the corresponding layers in g_2

    $\mathbf{W}_i \leftarrow \mathbf{W}_{i-l_1+1}^2, \mathbf{b}_i \leftarrow \mathbf{b}_{i-l_1+1}^2$

  end if

end for ▷ See Table 4 in Appendix C for complexity analysis

Algorithm 5 Find a ReLU network that computes an identity mapping for a given depth

Input: The input dimension n and the number of layers l of the target ReLU network.
Output: Parameters of an l-layer ReLU network g computing $g(\mathbf{x}) = \mathbf{x}, \forall \mathbf{x} \in \mathbb{R}^n$.

$\mathbf{A} \leftarrow \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \mathbf{B} \leftarrow \begin{bmatrix} 1 & -1 \end{bmatrix}, \mathbf{C} \leftarrow \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$ ▷ Constant matrices
$\mathbf{\Phi}(\mathbf{Y}, s) = \begin{bmatrix} \mathbf{Y}^{(1)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Y}^{(2)} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Y}^{(s)} \end{bmatrix}$ ▷ A block diagonal matrix with $\mathbf{Y}$ repeated s times
$k_0 \leftarrow n, k_l \leftarrow n, \mathbf{b}_l \leftarrow \mathbf{0}_n$
for $i = 1, 2, \cdots, l-1$ **do**
    $k_i \leftarrow 2n$ ▷ The number of hidden neurons at the i-th hidden layer
    $\mathbf{b}_i \leftarrow \mathbf{0}_{k_i}$
end for
if $l = 1$ **then** ▷ Find the weights of input and output layers, if any
    $\mathbf{W}_1 \leftarrow \mathbf{I}_{n \times n}$ ▷ An identity matrix
else
    $\mathbf{W}_1 \leftarrow \mathbf{\Phi}(\mathbf{A}, k_0)$
    $\mathbf{W}_l \leftarrow \mathbf{\Phi}(\mathbf{B}, k_l)$
end if
if $l > 2$ **then** ▷ Find the weights of hidden layers, if any
    for $i = 2, 3, \cdots, l-1$ **do**
        $\mathbf{W}_i \leftarrow \mathbf{\Phi}(\mathbf{C}, n)$
    end for
end if ▷ See Table 5 in Appendix C for complexity analysis

Algorithm 6 Find all distinct linear components of a CPWL function

Input: An unknown CPWL function p whose output can be observed by feeding input from $\mathbb{R}^n$ to the function. A center $\mathbf{c}_i$ and radius $\epsilon_i > 0$ of any closed ϵ_i-radius ball $B(\mathbf{c}_i, \epsilon_i)$ such that $B(\mathbf{c}_i, \epsilon_i) \subset \mathcal{X}_i$ for $i = 1, 2, \cdots, q$ where $\{\mathcal{X}_i\}_{i \in [q]}$ are all pieces of p.
Output: All distinct linear components of p, denoted by $\mathcal{F}$.

$\mathcal{F} \leftarrow \emptyset$ ▷ Initialize the set of all distinct linear components
for $i = 1, 2, \cdots, q$ **do**
    $\mathbf{x}_0 \leftarrow \mathbf{c}_i$ ▷ select the center of $B(\mathbf{c}_i, \epsilon_i)$
    $y_0 \leftarrow p(\mathbf{x}_0)$
    $\begin{bmatrix} \mathbf{s}_1 & \mathbf{s}_2 & \cdots & \mathbf{s}_n \end{bmatrix} \leftarrow \epsilon_i \mathbf{I}_{n \times n}$ ▷ scale the standard basis of $\mathbb{R}^n$
    $\mathbf{S} \leftarrow \begin{bmatrix} \mathbf{s}_1 & \mathbf{s}_2 & \cdots & \mathbf{s}_n \end{bmatrix}$
    $\mathbf{z} \leftarrow \begin{bmatrix} p(\mathbf{s}_1 + \mathbf{x}_0) - y_0 \\ p(\mathbf{s}_2 + \mathbf{x}_0) - y_0 \\ \vdots \\ p(\mathbf{s}_n + \mathbf{x}_0) - y_0 \end{bmatrix}$
    $\mathbf{a} \leftarrow \mathbf{S}^{-\mathsf{T}} \mathbf{z}$ ▷ Find the linear map by solving a system of linear equations
    $b \leftarrow y_0 - \mathbf{a}^{\mathsf{T}} \mathbf{x}_0$ ▷ Find the translation
    $f \leftarrow \mathbf{x} \mapsto \mathbf{a}^{\mathsf{T}} \mathbf{x} + b$ ▷ The affine map on $\mathcal{X}_i$
    if $f \notin \mathcal{F}$ **then** ▷ Only add the affine map f to the set $\mathcal{F}$ if f is distinct to all elements of $\mathcal{F}$
        $\mathcal{F} \leftarrow \mathcal{F} \bigcup \{f\}$
    end if
end for ▷ See Table 6 in Appendix C for complexity analysis

Table 2: The time complexity of Algorithm 2 is $\mathcal{O}\left(m^2 \max(m \log_2 m, n)\right)$.

Line	Operation count	Explanation
1	$\mathcal{O}(1)$	Initialize constant matrices.
2	$\mathcal{O}\left(d_1^2\right)$	Let d_1 be the maximum dimension of $\mathbf{Y}$ and $\mathbf{Z}$.
3	$\mathcal{O}(s^2 d_2^2)$	Let d_2 be the maximum dimension of $\mathbf{Y}$.
4	$\mathcal{O}(1)$	Scalar assignments.
5	$\mathcal{O}(\log_2 m)$	Repeat Line 6 to Line 12 $\lceil \log_2 m \rceil$ times.
6	$\mathcal{O}(1)$	Check a scalar is even or not.
7	$\mathcal{O}(1)$	Compute a scalar.
8	$\mathcal{O}(1)$	Compute a scalar.
9	-	-
10	$\mathcal{O}(1)$	Compute a scalar.
11	$\mathcal{O}(1)$	Compute a scalar.
12	-	-
13	-	-
14	$\mathcal{O}(mn)$	Assign a matrix and a vector.
15	$\mathcal{O}(1)$	Check a scalar inequality.
16	$\mathcal{O}(1)$	Check a scalar is even or not.
17	$\mathcal{O}(m^2 n)$	Matrix creation and multiplication.
18	-	-
19	$\mathcal{O}(m^2 n)$	Matrix creation and multiplication.
20	-	-
21	$\mathcal{O}(1)$	Assign a constant matrix and vector.
22	-	-
23	$\mathcal{O}(1)$	Check a scalar inequality.
24	$\mathcal{O}(\log_2 m)$	Repeat Line 25 to Line 30 $\lceil \log_2 m \rceil - 1$ times.
25	$\mathcal{O}(1)$	Check a scalar is even or not.
26	$\mathcal{O}(m^2)$	Matrix creation.
27	-	-
28	$\mathcal{O}(m^2)$	Matrix creation.
29	-	-
30	$\mathcal{O}(1)$	Check a scalar is even or not.
31	$\mathcal{O}(m^3)$	Matrix creation and multiplication.
32	-	-
33	$\mathcal{O}(m^3)$	Matrix creation and multiplication.
34	-	-
35	$\mathcal{O}(m)$	Assign a vector whose length is at most $\lceil \frac{3m}{2} \rceil$.
36	-	-
37	-	-
38	$\mathcal{O}(1)$	Check the binary data type.
39	$\mathcal{O}(mn)$	Reverse the sign of $\mathbf{W}_1$ and $\mathbf{b}_1$.
40	$\mathcal{O}(1)$	Reverse the sign of a constant matrix.
41	-	-

Table 3: The time complexity of Algorithm 3 is $\mathcal{O}\left(d^2kl\max(d,k)\right)$ where d is the maximum dimension of all the weight matrices in $g_1, g_2, \cdots, g_k$ and $l = \max_{j\in[k]} l_j$.

Line	Operation count	Explanation
1	$\mathcal{O}(k)$	Find the maximum among k numbers.
2	$\mathcal{O}(d^2k)$	Matrix concatenation and assignment. Note that $n \leq k$.
3	$\mathcal{O}(k)$	Repeat Line 4 to Line 9 k times.
4	$\mathcal{O}(1)$	Check a scalar inequality.
5	$\mathcal{O}(1)$	A scalar assignment.
6	$\mathcal{O}(d^2l)$	Algorithm 5 (see Table 5).
7	$\mathcal{O}(d^3l)$	Algorithm 4 (see Table 4).
8	$\mathcal{O}(d^2l)$	Assign weights of the network.
9	-	-
10	-	-
11	$\mathcal{O}(l)$	Repeat Line 12 $l-1$ times.
12	$\mathcal{O}(d^2k^2)$	Assign a matrix and a vector.
13	-	-

Table 4: The time complexity of Algorithm 4 is $\mathcal{O}\left(d^3\max(l_1,l_2)\right)$ where d is the maximum dimension of all the weight matrices in g_1 and g_2.

Line	Operation count	Explanation
1	$\mathcal{O}(1)$	Assign a constant.
2	$\mathcal{O}(l)$	Repeat Line 3 to Line 9 l times.
3	$\mathcal{O}(1)$	Check a scalar inequality.
4	$\mathcal{O}(d^2)$	Assign a matrix and a vector (at most d^2+d elements).
5	$\mathcal{O}(1)$	Check a scalar equality.
6	$\mathcal{O}(d^3)$	Matrix multiplication and assignment.
7	-	-
8	$\mathcal{O}(d^2)$	Assign a matrix and a vector (at most d^2+d elements).
9	-	-
10	-	-

Table 5: The time complexity of Algorithm 5 is $\mathcal{O}(n^2l)$.

Line	Operation count	Explanation
1	$\mathcal{O}(1)$	Initialize constant matrices.
2	$\mathcal{O}(s^2d_1d_2)$	Create a block diagonal matrix from $\mathbf{Y} \in \mathbb{R}^{d_1\times d_2}$ and $s \in \mathbb{N}$.
3	$\mathcal{O}(n)$	Assign two constant scalars and one constant vector of length n.
4	$\mathcal{O}(l)$	Repeat Line 5 to line 6 l times.
5	$\mathcal{O}(1)$	Assign a scalar.
6	$\mathcal{O}(n)$	Assign a vector whose length k_i is equal to $2n$.
7	-	-
8	$\mathcal{O}(1)$	Check a scalar equality.
9	$\mathcal{O}(n^2)$	Assign a $2n$-by-n matrix.
10	-	-
11	$\mathcal{O}(n^2)$	Assign a $2n$-by-n block diagonal matrix.
12	$\mathcal{O}(n^2)$	Assign an n-by-$2n$ block diagonal matrix.
13	-	-
14	$\mathcal{O}(1)$	Check a scalar inequality.
15	$\mathcal{O}(l)$	Repeat Line 16 $l-2$ times.
16	$\mathcal{O}(n^2)$	Assign a $2n$-by-$2n$ block diagonal matrix.
17	-	-
18	-	-

Table 6: The time complexity of Algorithm 6 is $\mathcal{O}\left(nq\max(n^2, q)\right)$.

Line	Operation count	Explanation
1	$\mathcal{O}(1)$	Initialize an empty placeholder $\mathcal{F}$.
2	$\mathcal{O}(q)$	Repeat Line 3 to line 13 q times.
3	$\mathcal{O}(1)$	Select an interior point. Use the center of the ball.
4	$\mathcal{O}(1)$	Evaluate the function on the point.
5	$\mathcal{O}(n^2)$	Scale and assign an n-by-n matrix.
6	-	-
7	$\mathcal{O}(n)$	Translate, evaluate, and subtract n points.
8	$\mathcal{O}(n^3)$	Solve a system of n linear equations with n variables.
9	$\mathcal{O}(n)$	Solve the translation term in the affine map
10	-	-
11	$\mathcal{O}(nq)$	Each affine map has $n+1$ parameters and $\mathcal{F}$ has at most q elements.
12	$\mathcal{O}(1)$	Add a distinct affine map to $\mathcal{F}$.
13	-	-
14	-	-

D Open source implementation and run time of Algorithm 1

We implement Algorithm 1 in Python. Figure 3 shows that the run time of the algorithm is greatly affected by the number of pieces q.

Code is available at https://github.com/kjason/CPWL2ReLUNetwork.

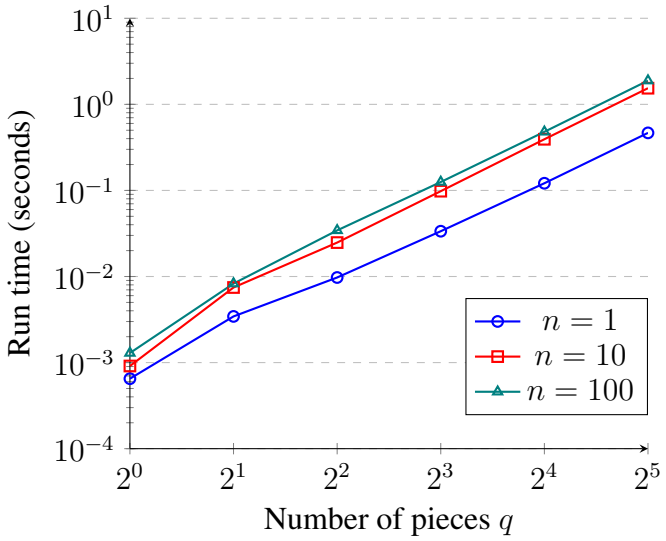

Figure 3: The run time of Algorithm 1 is an average of 50 trials. Every trial runs Algorithm 1 with a random CPWL function whose input dimension is n and number of pieces is q. The code provided in the above link is run on a computer (Microsoft Surface Laptop Studio) with the Intel Core i7-11370H.