
Minimum Width for Universal Approximation using Squashable Activation Functions

Jonghyun Shin¹ Namjun Kim² Geonho Hwang³ Sejun Park²

Abstract

The exact minimum width that allows for universal approximation of unbounded-depth networks is known only for ReLU and its variants. In this work, we study the minimum width of networks using general activation functions. Specifically, we focus on *squashable* functions that can approximate the identity function and binary step function by alternatively composing with affine transformations. We show that for networks using a squashable activation function to universally approximate L^p functions from $[0, 1]^{d_x}$ to \mathbb{R}^{d_y} , the minimum width is $\max\{d_x, d_y, 2\}$ unless $d_x = d_y = 1$; the same bound holds for $d_x = d_y = 1$ if the activation function is monotone. We then provide sufficient conditions for squashability and show that all non-affine analytic functions and a class of piecewise functions are squashable, i.e., our minimum width result holds for those general classes of activation functions.

1. Introduction

Understanding what neural networks can or cannot do is a fundamental problem in deep learning theory. The classical universal theorem states that two-layer networks can approximate any continuous function if an activation function is non-polynomial (Cybenko, 1989; Hornik et al., 1989; Leshno et al., 1993; Pinkus, 1999). Likewise, several studies on memorization show that neural networks can fit arbitrary finite training dataset (Baum, 1988; Huang and Babri, 1998). These results guarantee the existence of networks that can perform tasks in various practical applications such as computer vision (He et al., 2016), natural language processing (Vaswani, 2017; Brown et al., 2020), and science (Jumper et al., 2021).

¹Department of Mathematics Education, Korea University
²Department of Artificial Intelligence, Korea University
³Department of Mathematical Sciences, GIST. Correspondence to: Sejun Park <sejun.park000@gmail.com>.

The minimum size of networks that can universally approximate or memorize has also been studied. For example, classical results show that the minimum depth for both universal approximation and memorization is exactly two (Pinkus, 1999; Baum, 1988). The minimum number of parameters depends on the depth of networks. For universal approximation using ReLU networks, it is known that shallow wide architectures require more parameters than deep narrow ones (Yarotsky, 2018), where similar results are also known for memorization (Park et al., 2021a; Vardi et al., 2022). While these results show the benefits of depth, they also imply the existence of the minimum width enabling universal approximation and memorization.

There have been extensive research efforts to characterize such a minimum width. The minimum width for memorization is constantly bounded (i.e., independent of the input dimension) since any finite set of inputs can be mapped to distinct scalar values by projecting them (Park et al., 2021a). Intriguingly, the minimum width for universal approximation depends on the input dimension d_x and the output dimension d_y . Several works have shown that the minimum width lies between d_x and $d_x + d_y + \alpha$ where $\alpha \geq 0$ is some constant depending on the activation function and target functions space; however, the exact minimum width is known only for approximating L^p functions when the activation function is ReLU or its variants (Park et al., 2021b; Cai, 2023; Kim et al., 2024).

1.1. Related works

The minimum width for universal approximation has been studied for two function spaces $C(\mathcal{X}, \mathcal{Y})$ and $L^p(\mathcal{X}, \mathcal{Y})$: $C(\mathcal{X}, \mathcal{Y})$ denotes the space of continuous functions from \mathcal{X} to \mathcal{Y} endowed with the supremum norm $\sup_{x \in \mathcal{X}} \|f(x)\|_\infty$ and $L^p(\mathcal{X}, \mathcal{Y})$ denotes the space of L^p functions from \mathcal{X} to \mathcal{Y} endowed with the L^p -norm $\|f\|_{L^p} \triangleq (\int_{\mathcal{X}} \|f\|_p^p d\mu_{d_x})^{1/p}$ for $p \geq 1$. Recent studies on the minimum width (say w_{\min}) was initiated by Lu et al. (2017). They show that $d_x + 1 \leq w_{\min} \leq d_x + 4$ for universally approximating $L^1(\mathbb{R}^{d_x}, \mathbb{R})$ using ReLU networks. Hanin and Sellke (2017) consider universally approximating $C([0, 1]^{d_x}, \mathbb{R}^{d_y})$ using ReLU networks and prove $d_x + 1 \leq w_{\min} \leq d_x + d_y$. Johnson (2019) proves the lower bound $w_{\min} \geq d_x + 1$ for

Table 1: A summary of known bounds on the minimum width for universal approximation.

Reference	Function class	Activation σ	Upper/lower bounds
Lu et al. (2017)	$L^1(\mathbb{R}^{d_x}, \mathbb{R})$	RELU	$d_x + 1 \leq w_{\min} \leq d_x + 4$
Hanin and Sellke (2017)	$C([0, 1]^{d_x}, \mathbb{R}^{d_y})$	RELU	$d_x + 1 \leq w_{\min} \leq d_x + d_y$
Johnson (2019)	$C([0, 1]^{d_x}, \mathbb{R}^{d_y})$	uniformly conti.	$d_x + 1 \leq w_{\min}$
Kidger and Lyons (2020)	$C([0, 1]^{d_x}, \mathbb{R}^{d_y})$	conti. nonpoly. [†]	$w_{\min} \leq d_x + d_y + 1$
	$C([0, 1]^{d_x}, \mathbb{R}^{d_y})$	nonaffine poly.	$w_{\min} \leq d_x + d_y + 2$
Park et al. (2021b)	$L^p(\mathbb{R}^{d_x}, \mathbb{R}^{d_y})$	RELU	$w_{\min} = \max\{d_x + 1, d_y\}$
	$L^p([0, 1]^{d_x}, \mathbb{R}^{d_y})$	conti. nonpoly. [†]	$w_{\min} \leq \max\{d_x + 2, d_y + 1\}$
Cai (2023)	$L^p([0, 1]^{d_x}, \mathbb{R}^{d_y})$	Leaky-RELU	$w_{\min} = \max\{d_x, d_y, 2\}$
Kim et al. (2024)	$L^p([0, 1]^{d_x}, \mathbb{R}^{d_y})$	RELU-LIKE ^{‡*}	$w_{\min} = \max\{d_x, d_y, 2\}$
Ours (Theorem 2)	$L^p([0, 1]^{d_x}, \mathbb{R}^{d_y})$	Squashable ^{§*}	$w_{\min} = \max\{d_x, d_y, 2\}$

^{||} requires that σ is uniformly approximated by a sequence of one-to-one functions.

[†] requires that σ is continuously differentiable at some point z , with $\sigma'(z) \neq 0$.

[‡] denotes RELU, leaky-RELU, ELU, SOFTPLUS, CELU, SELU, GELU, SiLU, and MISH.

[§] includes all analytic functions and a class of piecewise functions such as leaky-RELU (see Sections 3.1 and 3.3).

* $d_x + d_y \geq 3$ is required for non-monotone activation functions.

activation functions that can be uniformly approximated by a sequence of one-to-one functions, which has been later extended to Lipschitz continuous and monotone functions when $d_x \geq d_y$ (Rochau et al., 2024). Kidger and Lyons (2020) show that for $C([0, 1]^{d_x}, \mathbb{R}^{d_y})$, $w_{\min} \leq d_x + d_y + 1$ if an activation function is continuous, non-polynomial, and continuously differentiable at some point with non-zero derivative. For non-affine polynomial activation functions, they also show $w_{\min} \leq d_x + d_y + 2$. However, the upper bounds in these results are at least $d_x + d_y$, which has a large gap compared to the lower bound $d_x + 1$. Such limitation arises from their universal approximator constructions that use d_x neurons to preserve the d_x -dimensional input and $d_y + \alpha$ neurons to compute the d_y -dimensional output.

The exact minimum width was first characterized by Park et al. (2021b). They show $w_{\min} = \max\{d_x + 1, d_y\}$ for RELU networks to universally approximate $L^p(\mathbb{R}^{d_x}, \mathbb{R}^{d_y})$ and for networks using both RELU and the binary step function $\text{STEP}(x)$ ¹ to universally approximate $C([0, 1]^{d_x}, \mathbb{R}^{d_y})$. For $L^p([0, 1]^{d_x}, \mathbb{R}^{d_y})$, they also show $w_{\min} \leq \max\{d_x + 2, d_y + 1\}$ for a class of continuous non-polynomial activation functions. Specifically, Park et al. (2021b) construct two networks called the *encoder* and *decoder* to approximate a target function $f^* : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_y}$: the encoder maps a d_x -dimensional input vector x to some scalar value c containing the information of x and the decoder maps c to a neighborhood of the d_y -dimensional target vector $f^*(x)$. Then, the concatenation of the encoder and decoder approximates f^* using an autoencoder structure with an internal width one. Cai (2023) use a different approach to show $w_{\min} = \max\{d_x, d_y, 2\}$ for leaky-RELU networks to uni-

versally approximate $L^p([0, 1]^{d_x}, \mathbb{R}^{d_y})$. Specifically, they derive the bound using the two results: (1) neural ODE can approximate any L^p functions (Li et al., 2022) and (2) Leaky-RELU networks can approximate any neural ODEs using width $\max\{d_x, d_y, 2\}$ (Duan et al., 2022).

The *coding scheme* proposed by Park et al. (2021b) has also been used to characterize w_{\min} for different problem setups. For example, Cai (2023) show $w_{\min} = \max\{d_x, d_y, 2\}$ for networks using both RELU and FLOOR, Kim et al. (2024) show $w_{\min} = \max\{d_x, d_y, 2\}$ for variants of RELU (see the footnote [‡] in Table 1), and Rochau et al. (2024) show $w_{\min} = \max\{d_x, d_y, 2\}$ for leaky-ReLU. Nevertheless, existing proof techniques for tight minimum width highly rely on the property of RELU and/or its variants, and the minimum width for general activation functions was unknown.

1.2. Summary of contributions

In this work, we study the minimum width enabling universal approximation of $L^p([0, 1]^{d_x}, \mathbb{R}^{d_y})$ using general activation functions. Specifically, we consider activation functions σ such that an alternative composition of σ and affine transformations can approximate the identity function and binary step function; we call such functions *squashable* (see Definition 1). Using the squashability of an activation function σ , we show that the minimum width of σ networks to universally approximate $L^p([0, 1]^{d_x}, \mathbb{R}^{d_y})$ is $\max\{d_x, d_y\}$ unless $d_x = d_y = 1$ (Theorem 2). We also show $w_{\min} = 2$ when $d_x = d_y = 1$ if the squashable function σ is monotone.

Our result can be used to characterize the minimum width for a general class of practical activation functions, by showing their squashability. For example, we show that *any non-affine analytic function* (e.g., non-affine polynomial,

¹ $\text{STEP}(x) = 1$ if $x \geq 0$ and $\text{STEP}(x) = 0$ otherwise.

SIGMOID, tanh, sin, exp, etc.) is squashable (Lemma 4). Furthermore, we also show that a wide class of piecewise continuously differentiable functions including leaky-RELU and HARDSWISH are also squashable (Lemma 5). Hence, our result significantly extends the prior exact minimum width results for RELU and its variants.

Even if an activation is not analytic or piecewise continuously differentiable, it can be squashable, i.e., our minimum width result can be applicable. To check the squashability of general functions, we also provide a sufficient condition for the squashability: σ is squashable if and only if there exists an alternative composition f of σ and affine transformations such that f is strictly increasing and has a locally sigmoidal shape on some proper interval (Lemma 3).

We use the coding scheme (Park et al., 2021b) to prove our result. In particular, we construct our decoder f_{dec} as a curve that densely fills the codomain of a target function so that $\sup_{x \in f^*([0,1]^{d_x})} \inf_{y \in f_{\text{dec}}([0,1])} \|x - y\|_\infty$ is small. We then construct our encoder to map each $x \in [0,1]^{d_x}$ to a neighborhood of $f_{\text{dec}}^{-1}(z)$ for some $z \approx f^*(x)$. Although we use the coding scheme, our constructions largely differ from existing ones (Park et al., 2021b; Cai, 2023; Kim et al., 2024; Rochau et al., 2024), which highly rely on properties (e.g., shape and discontinuity) of specific activation functions (e.g., RELU, FLOOR, STEP, and their variants). Namely, existing proofs do not easily extend to general activation functions. See Section 4 for our detailed constructions.

1.3. Organization

We first introduce notations and the problem setup in Section 2. We then formally define the squashability of activation functions, describe our main result on minimum width for universal approximation, and provide sufficient conditions for the squashability in Section 3. We prove our main result in Section 4 and conclude the paper in Section 5. Proofs of technical lemmas are deferred to Appendix.

2. Problem setup and notations

In this section, we introduce notations and our problem setup. For $n \in \mathbb{N}$, we use $[n]$ to denote $\{1, \dots, n\}$. For $\mathcal{S}, \mathcal{T} \subset \mathbb{R}^d$, we use $\text{diam}(\mathcal{S}) \triangleq \sup_{x, y \in \mathcal{S}} \|x - y\|_\infty$ and $\text{dist}(\mathcal{S}, \mathcal{T}) \triangleq \inf_{x \in \mathcal{S}, y \in \mathcal{T}} \|x - y\|_\infty$. If \mathcal{S} is a singleton set, (i.e., $\mathcal{S} = \{s\}$), we use $\text{dist}(s, \mathcal{T})$ to denote $\text{dist}(\{s\}, \mathcal{T})$. For $y \in \mathbb{R}^d$ and $\mathcal{S} \subset \mathbb{R}^d$, $\mathcal{B}_r(y) \triangleq \{x \in \mathbb{R}^d : \text{dist}(x, y) \leq r\}$ and $\mathcal{B}_r(\mathcal{S}) \triangleq \{x \in \mathbb{R}^d : \text{dist}(x, \mathcal{S}) \leq r\}$. For a function $f : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$, $f(x)_i$ denotes the i -th coordinate of $f(x)$. For $n \in \mathbb{N}$, we use f^n to denote the n times composition of f . We use $\iota : \mathbb{R} \rightarrow \mathbb{R}$ to denote the identity function ($\iota(x) = x$) and STEP to denote the binary step function ($\text{STEP}(x) = 0$ if $x < 0$ and $\text{STEP}(x) = 1$ otherwise). We note that all intervals in this paper are proper: they are

neither empty nor degenerate (e.g. $[a, a] = \{a\}$).

2.1. Fully-connected networks

Throughout this paper, we consider fully-connected neural networks. Formally, given a set of activation functions Σ , we define an L -layer neural network f with input dimension $d_0 = d_x$, output dimension $d_L = d_y$, and hidden layer dimensions d_1, \dots, d_{L-1} as

$$f \triangleq t_L \circ \tilde{\sigma}_{L-1} \circ t_{L-1} \circ \dots \circ \tilde{\sigma}_1 \circ t_1$$

where $t_\ell : \mathbb{R}^{d_{\ell-1}} \rightarrow \mathbb{R}^{d_\ell}$ is an affine transformation and $\tilde{\sigma}_\ell(x_1, \dots, x_{d_\ell}) = (\sigma_{\ell,1}(x_1), \dots, \sigma_{\ell,d_\ell}(x_{d_\ell}))$ for some $\sigma_{\ell,1}, \dots, \sigma_{\ell,d_\ell} \in \Sigma$ for all $\ell \in [L]$. We denote a neural network using a single activation function σ (i.e., $\Sigma = \{\sigma\}$) by a “ σ network” and a neural network using a two activation functions σ_1, σ_2 (i.e., $\Sigma = \{\sigma_1, \sigma_2\}$) by a “ (σ_1, σ_2) network”. Here, the *width* w of f is defined as the maximum over the hidden dimensions d_1, \dots, d_{L-1} .

We say “ σ networks of width w are dense in $L^p(\mathcal{X}, \mathcal{Y})$ ” if for any $f^* \in L^p(\mathcal{X}, \mathcal{Y})$ and $\varepsilon > 0$, there exists a σ network of width w such that $\|f^* - f\|_{L^p} \leq \varepsilon$. Given an activation function σ and $d_x, d_y \in \mathbb{N}$, we use w_{σ, d_x, d_y} to denote the minimum $w \in \mathbb{N}$ satisfying the following: σ networks of width w are dense in $L^p([0,1]^{d_x}, \mathbb{R}^{d_y})$ but σ networks of width $w - 1$ are not dense. We often drop d_x, d_y and use w_σ if d_x, d_y are clear from the context.

3. Main results

3.1. Squashable activation functions

To formally state our main result, we first introduce a class of activation functions σ that we mainly focus on.

Condition 1. There exists $z \in \mathbb{R}$ such that σ is continuously differentiable at z and $\sigma'(z) \neq 0$.

Condition 2. σ is continuous and for any compact set $\mathcal{K} \subset \mathbb{R}$ and for any $\varepsilon, \zeta > 0$, there exists a σ -network $\rho_{\varepsilon, \zeta} : \mathbb{R} \rightarrow \mathbb{R}$ of width 1 such that

- $\max_{x \in \mathcal{K} \setminus (-\zeta, \zeta)} |\rho_{\varepsilon, \zeta}(x) - \text{STEP}(x)| \leq \varepsilon$,
- $\rho_{\varepsilon, \zeta}$ is strictly increasing on \mathcal{K} , and
- $\rho_{\varepsilon, \zeta}(\mathcal{K}) \subset [0, 1]$.

Condition 1 is that an activation function σ has a continuously differentiable point with a nonzero derivative. This property enables us to approximate the identity function on a compact domain by composing σ with affine transformations as stated in the following lemma.

Lemma 1 (Lemma 4.1 in (Kidger and Lyons, 2020)). *For any $\varepsilon > 0$, $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ satisfying Condition 1, and compact set $\mathcal{K} \subset \mathbb{R}$, there exist affine transformations $h_1, h_2 : \mathcal{K} \rightarrow \mathbb{R}$ such that*

$$\sup_{x \in \mathcal{K}} \|h_2 \circ \sigma \circ h_1(x) - x\| \leq \varepsilon.$$

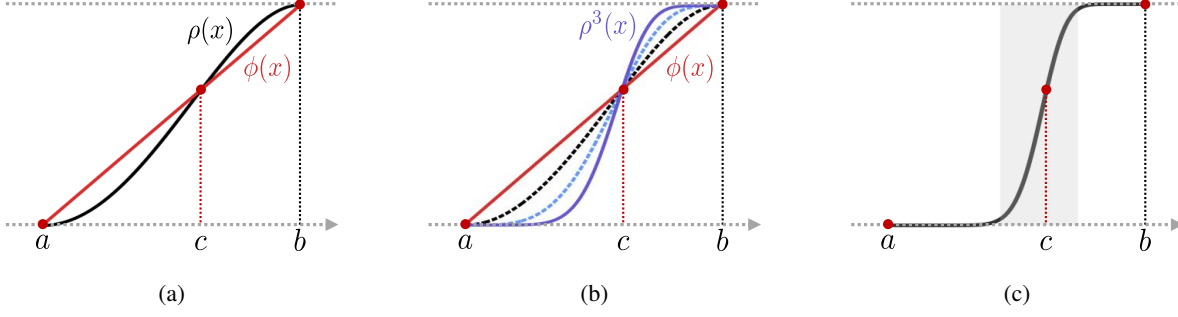


Figure 1: Illustration of construction of squashable function using a σ network ρ of width 1 that has a sigmoidal shape when $\phi(x) = x$. The intersections of $\rho(x)$ and $\phi(x)$ serve as *fixed points*. Thus, σ can achieve the squashability by iteratively composing ρ : $\rho^n(x) \rightarrow a$ for $x \in (a, c)$ and $\rho^n(x) \rightarrow b$ for $x \in (c, b)$ as $n \rightarrow \infty$ while ρ^n is strictly monotone.

Condition 2 assumes the continuity of σ and the existence of a σ network of width 1 (i.e., an alternative composition of affine transformations and σ) that can approximate the binary threshold function (i.e., STEP) on any compact set, except for a small neighborhood of zero (i.e., $(-\zeta, \zeta)$). One important property in Condition 2 is that $\rho_{\varepsilon, \zeta}$ should be strictly increasing on \mathcal{K} . This allows $\rho_{\varepsilon, \zeta}$ to preserve the information of inputs in \mathcal{K} since it is bijective on \mathcal{K} .

Using these conditions, we now define the *squashability* of an activation function.

Definition 1. A function $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ is “squashable” if σ satisfies Conditions 1 and 2.

One can observe that width-1 networks using a squashable activation function can approximate the identity function on any compact domain and the STEP function on any compact domain except for a small open neighborhood.

A class of squashable activation functions covers a wide range of practical functions. Condition 1 can be easily satisfied: e.g., any piecewise differentiable function with a non-constant piece satisfies Condition 1. Furthermore, we prove that any analytic activation function (e.g., SIGMOID, exp, sin.) and a class of piecewise continuously differentiable functions (e.g., leaky-RELU and HARDSWISH) satisfy Condition 2. We formally state these results and easily verifiable conditions for Condition 2 in Section 3.3.

3.2. Minimum width with squashable functions

We are now ready to introduce our main theorem on the minimum width for universal approximation.

Theorem 2. Let σ be a squashable function. Then, $w_\sigma = \max\{d_x, d_y\}$ if $d_x \geq 2$ or $d_y \geq 2$ and $w_\sigma \in \{1, 2\}$ if $d_x = d_y = 1$. Furthermore, if σ is monotone, then $w_\sigma = 2$ if $d_x = d_y = 1$.

Theorem 2 characterizes the exact minimum width enabling universal approximation for squashable activation functions:

$w_{\sigma, d_x, d_y} = \max\{d_x, d_y\}$ unless the input/output dimensions are both one. Furthermore, it fully characterizes w_{σ, d_x, d_y} for all d_x, d_y if an activation function is squashable and monotone. The proof of Theorem 2 is in Section 4.

To the best of our knowledge, the exact minimum width enabling universal approximation has been discovered only for a few RELU-LIKE activation functions such as RELU, leaky-RELU, SOFTPLUS, GELU (Park et al., 2021b; Cai, 2023; Kim et al., 2024). Furthermore, the best known upper bound for a general class of activation functions was $w_\sigma \leq \max\{d_x + 2, d_y + 1\}$ when σ is continuous non-polynomial and continuously differentiable at some point with non-zero derivative (Park et al., 2021b). Our result extends prior exact minimum width results to a general class of activation functions (i.e., squashable) including all analytic functions (e.g., SIGMOID, tanh, sin, exp, polynomial) and a class of piecewise continuously differentiable functions (e.g., HARDSWISH). See Lemmas 4 and 5 in Section 3.3 for more details on squashable functions.

3.3. Easily verifiable conditions for Condition 2

In Theorem 2, we have observed that w_{σ, d_x, d_y} can be characterized if σ is squashable. However, checking whether a given activation function is squashable, especially whether it satisfies Condition 2, can be non-trivial. In this section, we provide easily verifiable conditions for Condition 2 based on the following lemma.

Lemma 3. A continuous function $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ satisfies Condition 2 if there exist a σ network ρ of width 1 and $a, b \in \mathbb{R}$ with $a < b$ satisfying the following:

- ρ is strictly increasing on $[a, b]$ and
- there exists $c \in (a, b)$ such that

$$\rho(x) < \phi(x) \quad \forall x \in (a, c), \quad \rho(x) > \phi(x) \quad \forall x \in (c, b)$$

where $\phi(x)$ is a line passing $(a, \rho(a))$ and $(b, \rho(b))$.

Lemma 3 provides a sufficient condition for Condition 2: if

we can make a σ network of width 1 that has a “sigmoidal shape” on some compact domain (e.g., see Figure 1a), then σ satisfies Condition 2. We can easily approximate the STEP function using a function with the sigmoidal shape by composing the function and some affine transformations (see Figures 1b and 1c). For a more formal argument, see the proof of Lemma 3 in Appendix A.2.1.

Such a sigmoidal shape (or its symmetric variants) exists in various smooth activation functions such as GELU, SIGMOID, tanh, and sin. In addition, for any non-affine analytic function σ , we can always make a σ network of width 1 that has the sigmoidal shape. Since all non-constant analytic functions are continuously differentiable and have a non-zero derivative at some point, all non-affine analytic functions satisfy Conditions 1 and 2, i.e., they are squashable. The proof of Lemma 4 is presented in Appendix A.2.2.

Lemma 4. *All non-affine analytic functions from \mathbb{R} to \mathbb{R} satisfy Condition 2.*

In addition, a class of piecewise functions also satisfies the condition in Lemma 3. We defer the proof of Lemma 5 to Appendix A.2.3.

Lemma 5. *A continuous function $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ satisfies Condition 2 if there exist $c \in \mathbb{R}$ and $\delta > 0$ such that*

- σ is continuously differentiable on $(c - \delta, c + \delta) \setminus \{c\}$,
- $v^+ = \lim_{x \rightarrow c^-} \sigma'(x)$ and $v^- = \lim_{x \rightarrow c^+} \sigma'(x)$ exist, $v^+ \neq v^-$, and $v^+ v^- > 0$.

Lemma 5 states that if an activation function σ contains a point such that the left limit of the derivative and the right limit of derivative at that point are different but have the same sign, then σ satisfies Condition 2. We note that piecewise functions such as leaky-RELU and HARDSWISH satisfy the condition in Lemma 5; for those functions, one can choose the point c in Lemma 5 as some break point between consecutive pieces.

While we provide easily verifiable sufficient conditions (Lemmas 4 and 5) for Condition 2, we note that Theorem 2 covers any activation function satisfying Conditions 1 and 2, even if that activation function does not satisfy conditions in Lemmas 4 and 5. We also present additional sufficient conditions for Condition 2 in Appendix A.3.

4. Proof of Theorem 2

We now present the proof of Theorem 2. Theorem 2 is a direct corollary of the following lemmas.

Lemma 6. *Let σ be a squashable function, $\varepsilon > 0$, $f^* \in C([0, 1]^{d_x}, [0, 1]^{d_y})$, and $p \geq 1$. Then, there exists a σ network $f : [0, 1]^{d_x} \rightarrow \mathbb{R}^{d_y}$ of width $\max\{d_x, d_y, 2\}$ such that*

$$\|f - f^*\|_{L^p} \leq \varepsilon.$$

Lemma 7 (Lemmas 21 and 22 in (Kim et al., 2024)). *For any $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ and $d_x, d_y \in \mathbb{N}$, $w_\sigma \geq \max\{d_x, d_y\}$. Furthermore, if σ is monotone, then $w_\sigma \geq 2$.*

Lemma 6 implies that for any squashable activation function σ , $w_\sigma \leq \max\{d_x, d_y, 2\}$. This is because (1) continuous functions on $[0, 1]^{d_x}$ are dense in $L^p([0, 1]^{d_x}, \mathbb{R}^{d_y})$ (Rudin, 1987) and (2) $g([0, 1]^{d_x})$ is compact for all $g \in C([0, 1]^{d_x}, \mathbb{R}^{d_y})$, i.e., we can scale the range of g to be in $[0, 1]^{d_y}$. Lemma 7 provides the lower bound of the minimum width. Here, the lower bound $w_{\min} \geq \max\{d_x, d_y, 2\}$ naturally follows since the first/last operations in networks are affine transformations. Hence, combining Lemmas 6 and 7 results in Theorem 2. In the rest of this section, we prove Lemma 6.

4.1. Proof of Lemma 6

To illustrate our main idea for proving Lemma 6, we first define a δ -filling curve.

Definition 2. Let $d \in \mathbb{N}$ and $\delta > 0$. We say a continuous function $f : \mathbb{R} \rightarrow \mathbb{R}^d$ is a “ δ -filling curve” of $\mathcal{D} \subset \mathbb{R}^d$ if

$$\sup_{y \in \mathcal{D}} \text{dist}(y, f([0, 1])) \leq \delta.$$

A δ -filling curve of $\mathcal{D} \subset \mathbb{R}^d$ can be considered as a weaker version of a space-filling curve of \mathcal{D} (Sagan, 2012). While the range of the space-filling curve contains \mathcal{D} but the δ -filling curve covers \mathcal{D} within δ distance.

Suppose that we can implement a δ -filling curve h of $[0, 1]^{d_y}$ using a σ network for some small $\delta > 0$, i.e., for each $y \in [0, 1]^{d_y}$, there is $z \in [0, 1]$ such that $h(z) \approx y$. Hence, if we can design a σ network g that maps each $x \in [0, 1]^{d_x}$ to some z_x such that $h(z_x) \approx f^*(x)$, then the σ network $h \circ g$ approximates f^* . Here, g and h can be considered as an *encoder* and *decoder*: g encodes a d_x -dimensional vector x to a scalar value z_x that contains the information of $f^*(x)$ and h decodes z_x to a d_y -dimensional vector $h(z_x)$ that approximates $f^*(x)$.

We explicitly construct networks that approximate the encoder and decoder. To this end, we introduce the following lemma where the proof is deferred to Appendix C.

Lemma 8. *Let σ be a squashable function, $d, w \in \mathbb{N}$, and $\mathcal{K} \subset \mathbb{R}^d$ be a compact set. Then, for any $\varepsilon > 0$ and (σ, ι) network f of width w , there exists a σ network g of width w such that*

$$\sup_{x \in \mathcal{K}} \|f(x) - g(x)\|_\infty < \varepsilon.$$

Here, $\iota : \mathbb{R} \rightarrow \mathbb{R}$ denotes the identity function (see Section 2). Lemma 8 implies that constructing a (σ, ι) network of width $\max\{d_x, d_y, 2\}$ that approximates f^* is sufficient to prove Lemma 6. Hence, we focus on approximating the encoder and decoder using (σ, ι) networks.

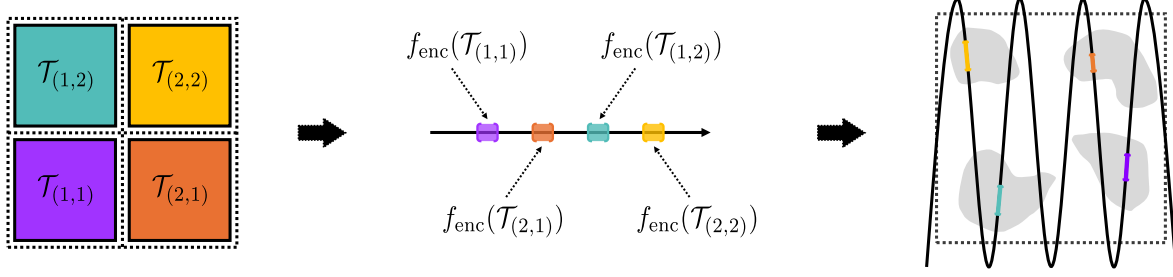


Figure 2: Illustration of $f_{\text{dec}} \circ f_{\text{enc}}$ when $d_x = 2$, $d_y = 2$, and $N = 2$. f_{enc} first encodes each \mathcal{T}_{ν} to a bounded interval $f_{\text{enc}}(\mathcal{T}_{\nu})$. Then, f_{dec} implements δ -filling curve of $[0, 1]^2$, represented by the black curve, to decode each $f_{\text{enc}}(\mathcal{T}_{\nu})$ (colored) that approximates $f^*(\mathcal{T}_{\nu})$ (represented by the light gray area).

We first show that the decoder can be implemented using a (σ, ι) network of width d_y . The proof of Lemma 9 is in Section 4.2.

Lemma 9. *Let σ be a squashable function and $\delta > 0$. Then, there exists a (σ, ι) network $f_{\text{dec}} : [0, 1] \rightarrow [0, 1]^{d_y}$ of width d_y that is a δ -filling curve of $[0, 1]^{d_y}$.*

Lemma 9 states that for any $\delta > 0$, we can always implement a δ -filling curve of $[0, 1]^{d_y}$ using a (σ, ι) network f_{dec} of width d_y . Further, the implemented network satisfies $f_{\text{dec}}([0, 1]) \subset [0, 1]^{d_y}$ regardless of δ .

We also show that the encoder can be approximated by a (σ, ι) network of width $\max\{d_x, 2\}$. The proof of Lemma 10 is in Section 4.3.

Lemma 10. *Let σ be a squashable function, $N \in \mathbb{N}$ and $\gamma \in (0, 0.5)$. For each $\nu \in [N]^{d_x}$, let $\mathcal{T}_{\nu} = \prod_{i=1}^{d_x} [\frac{\nu_i-1+\gamma}{N}, \frac{\nu_i-\gamma}{N}]$ and $c_{\nu} \in [0, 1]$. Then, there exists a (σ, ι) network $f_{\text{enc}} : [0, 1]^{d_x} \rightarrow [0, 1]$ of width $\max\{d_x, 2\}$ such that for each $\nu \in [N]^{d_x}$,*

$$f_{\text{enc}}(\mathcal{T}_{\nu}) \subset \mathcal{B}_{\gamma}(c_{\nu}).$$

Here, $\mathcal{B}_{\gamma}(c_{\nu})$ denotes the ℓ_{∞} -ball of radius γ centered at c_{ν} (see Section 2). The collection of \mathcal{T}_{ν} in Lemma 10 can be regarded as an approximate partition of $[0, 1]^{d_x}$: its elements are disjoint and it covers almost all parts of the domain with a small enough $\gamma > 0$. By choosing a large enough N , the diameter of $f^*(\mathcal{T}_{\nu})$ can be arbitrarily small, i.e., $f^*(x) \approx f^*(x')$ for all $x, x' \in \mathcal{T}_{\nu}$. Under this setup, choose c_{ν} for each ν so that $f_{\text{dec}}(c_{\nu}) \approx f^*(\mathcal{T}_{\nu})$. Then, f_{enc} in Lemma 10 maps each element \mathcal{T}_{ν} in the approximate partition to some small ball centered at c_{ν} , with diameter γ . Since f_{dec} is continuous, this implies that for each $x \in \mathcal{T}_{\nu}$, $f_{\text{dec}} \circ f_{\text{enc}}(x) \approx f^*(x)$ with small enough δ for f_{dec} and small enough γ , large enough N for f_{enc} . See Figure 2 for the illustration. Here, we note that $f_{\text{dec}} \circ f_{\text{enc}}$ is a (σ, ι) network of width $\max\{d_x, d_y, 2\}$.

For $x \notin \bigcup_{\nu} \mathcal{T}_{\nu}$, we have $f_{\text{dec}} \circ f_{\text{enc}}(x) \in [0, 1]^{d_y}$ (i.e., bounded) by Lemmas 9 and 10. Since $\mu_{d_x}([0, 1]^{d_x} \setminus$

$\bigcup_{\nu} \mathcal{T}_{\nu}) \rightarrow 0$ as $\gamma \rightarrow 0$, one can observe that for any $\varepsilon > 0$, there exist small enough γ, δ and large enough N such that $\|f_{\text{dec}} \circ f_{\text{enc}} - f^*\|_{L^p} \leq \varepsilon$. Namely, a (σ, ι) network $f = f_{\text{dec}} \circ f_{\text{enc}}$ has width $\max\{d_x, d_y, 2\}$ and completes the proof. Given $\varepsilon > 0$, our explicit choices of δ, γ, N and the detailed derivation of $\|f_{\text{dec}} \circ f_{\text{enc}} - f^*\|_{L^p} \leq \varepsilon$ can be found in Appendix B.

4.2. Proof of Lemma 9

In this section, we prove Lemma 9 by showing the following: for each $N, d \in \mathbb{N}$, there exists a (σ, ι) network of width d_y that is a $(1/N)$ -filling curve of $[0, 1]^d$. In particular, we inductively construct a $(1/N)$ -filling curve of $[0, 1]^d$ from $d = 1$. Here, the base case $d = 1$ is trivial: a (σ, ι) network $f(x) = \iota(x)$ is a $(1/N)$ -filling curve of $[0, 1]$ for all $N \in \mathbb{N}$.

We prove the general case ($d \geq 2$) using the inductive step described in the following lemma, whose formal proof is in Appendix D. Here, for $N, d \in \mathbb{N}$, we use $\mathcal{C}_{N,d,\nu} \triangleq \prod_{i=1}^d [\frac{\nu_i-1}{N}, \frac{\nu_i}{N}]$ and $\nu = (\nu_1, \dots, \nu_d) \in [N]^d$.

Lemma 11. *Let $N, d \in \mathbb{N}$ and σ be a squashable function. Suppose that there exist disjoint open intervals $\mathcal{I}_{\nu} \subset [0, 1]$ for all $\nu \in [N]^d$ and a (σ, ι) network $f : [0, 1] \rightarrow [0, 1]^d$ of width d such that for each $x \in [0, 1]$ and $\nu \in [N]^d$,*

$$f(x)_1 = x \text{ and } f(\mathcal{I}_{\nu}) \subset \mathcal{C}_{N,d,\nu}.$$

Then, there exist disjoint open intervals $\mathcal{J}_{\tilde{\nu}} \subset [0, 1]$ for all $\tilde{\nu} \in [N]^{d+1}$ and a (σ, ι) network $\tilde{f} : [0, 1] \rightarrow [0, 1]^{d+1}$ of width $d + 1$ such that for each $x \in [0, 1]$ and $\tilde{\nu} \in [N]^{d+1}$,

$$\tilde{f}(x)_1 = x \text{ and } \tilde{f}(\mathcal{J}_{\tilde{\nu}}) \subset \mathcal{C}_{N,d+1,\tilde{\nu}}.$$

One can observe that (σ, ι) networks f and \tilde{f} in Lemma 11 are $(1/N)$ -filling curves of $[0, 1]^d$ and $[0, 1]^{d+1}$, respectively. Furthermore, our filling curve construction $f(x) = \iota(x)$ for the base case satisfies the assumption in Lemma 11 with $\mathcal{I}_{\nu} = (\frac{\nu-1}{N}, \frac{\nu}{N})$ for all $N \in \mathbb{N}$ and $\nu \in [N]$. Hence, by Lemma 11, we can conclude that for each $N, d \in \mathbb{N}$,

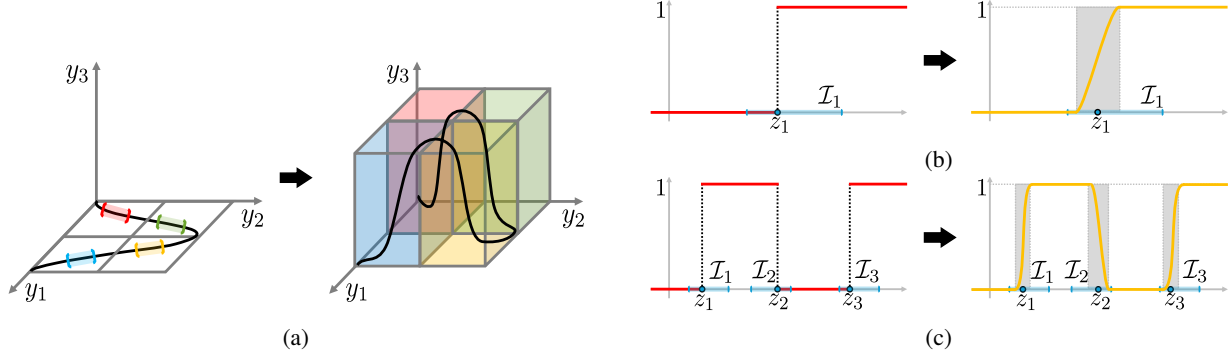


Figure 3: (a) Illustration of a $(1/N)$ -filling curve \tilde{f} of $[0, 1]^3$. \tilde{f} maps each open interval \mathcal{I}_ν , represented by the colored brackets (left), to be intersected with the corresponding cube of the same color (right). (b) and (c) illustrates our network ρ satisfying the properties of ϕ when $N = 1$ and $N = 3$, respectively.

there exists a (σ, ι) network that is a $(1/N)$ -filling curve of $[0, 1]^d$; this proves Lemma 9.

We now briefly illustrate our main idea for constructing \tilde{f} in Lemma 11 given f . Suppose that disjoint open intervals \mathcal{I}_ν for all $\nu \in [N]^d$ and corresponding (σ, ι) network f of width d in Lemma 11 are given. Then, to prove Lemma 11, it suffices to construct a (σ, ι) network \tilde{f} of width $d+1$ such that for each $i \in [d]$ and $\nu \in [N]^d$,

$$\tilde{f}(x)_i = f(x)_i \text{ and } [\frac{1}{2N}, 1 - \frac{1}{2N}] \subset \tilde{f}(\mathcal{I}_\nu)_{d+1}.$$

This implies that if we can construct a (σ, ι) network $\phi : [0, 1] \rightarrow \mathbb{R}^2$ of width 2 such that for each $\nu \in [N]^d$,

$$\phi(x)_1 = x \text{ and } [\frac{1}{2N}, 1 - \frac{1}{2N}] \subset \phi(\mathcal{I}_\nu)_2, \quad (1)$$

then we can construct \tilde{f} in Lemma 11 by choosing

$$\begin{aligned} \tilde{f}(x)_1 &= \phi(f(x)_1)_1, \quad \tilde{f}(x)_{d+1} = \phi(f(x)_1)_2, \text{ and} \\ \tilde{f}(x)_i &= \iota \circ \dots \circ \iota(f(x)_i) \text{ for all } i \in \{2, \dots, d\}. \end{aligned}$$

See Figure 3a for the illustration. We can construct such ϕ using the squashability of σ . For example, suppose that $N = 1$ and $d = 1$ (i.e., there is exactly one \mathcal{I}_ν). By Definition 1, for any $\varepsilon, \zeta > 0$ and compact $\mathcal{K} \subset \mathbb{R}$ with $[-\zeta, \zeta] \subset \mathcal{K}$, there is a width-1 σ network ρ such that

$$\max_{x \in \mathcal{K} \setminus (-\zeta, \zeta)} |\rho(x) - \text{STEP}(x)| \leq \varepsilon.$$

Then, by the intermediate value theorem, we have

$$[\varepsilon, 1 - \varepsilon] \subset \rho([-\zeta, \zeta]).$$

This implies that by choosing $\tilde{\rho}(x) = \rho(x - z_\nu)$ for some $z_\nu \in \mathcal{I}_\nu$ and \mathcal{K} containing \mathcal{I}_ν with small enough $\varepsilon, \zeta > 0$, it holds that $[\frac{1}{2N}, 1 - \frac{1}{2N}] \subset \tilde{\rho}(\mathcal{I}_\nu)$ (see Figure 3b). In this case, we can choose a width-2 (σ, ι) network ϕ satisfying Eq. (1) as $\phi(x)_1 = x$ and $\phi(x)_2 = \tilde{\rho}(x)$.

Such a construction also extends to an arbitrary number of \mathcal{I}_ν by composing ρ (i.e., an approximation of STEP). For example, let $\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3 \subset [0, 1]$ be disjoint open intervals and let $z_i \in \mathcal{I}_i$. Then, we have

$$\begin{aligned} \psi(x) &= \text{STEP}(x - z_1 + (z_1 - z_3) \times \text{STEP}(x - z_2)) \\ &= \begin{cases} 0 & \text{if } x \leq z_1 \text{ or } z_2 \leq x < z_3 \\ 1 & \text{otherwise} \end{cases}. \end{aligned}$$

Namely, by replacing STEP by ρ in ψ with small enough $\varepsilon, \zeta > 0$ (and denoting that function by $\tilde{\psi}$), we have $[\frac{1}{2N}, 1 - \frac{1}{2N}] \subset \tilde{\psi}(\mathcal{I}_i)$ by the intermediate value theorem (see Figure 3c). We present a more detailed argument for general N, d in the proof of Lemma 18 in Appendix D.1.

4.3. Proof of Lemma 10

We now prove Lemma 10. Our construction of f_{enc} consists of two (σ, ι) networks: $f_1 : [0, 1]^{d_x} \rightarrow \mathbb{R}$ of width d_x and $f_2 : \mathbb{R} \rightarrow \mathbb{R}$ of width 2. Here, f_1 maps each \mathcal{T}_ν to a disjoint compact interval $f_1(\mathcal{T}_\nu)$ and f_2 is designed to satisfy $f_2(f_1(\mathcal{T}_\nu)) \subset \mathcal{B}_\gamma(c_\nu)$ for each ν . Namely, $f_{\text{enc}} = f_2 \circ f_1$ satisfies $f(\mathcal{T}_\nu) \subset \mathcal{B}_\gamma(c_\nu)$.

Construction of f_2 . The following lemma shows the existence of f_2 such that $f_2(f_1(\mathcal{T}_\nu)) \subset \mathcal{B}_\gamma(c_\nu)$ for each $\nu \in [N]^{d_x}$.

Lemma 12. *Let $\mathcal{K} \subset \mathbb{R}$ be a compact interval and $\mathcal{I}_1, \dots, \mathcal{I}_N \subset \mathcal{K}$ be disjoint closed subintervals. Then, for any $\varepsilon > 0$, squashable σ , and $c_1, \dots, c_N \in \mathbb{R}$, there exists a (σ, ι) network $f : \mathcal{K} \rightarrow [0, 1]$ of width 2 such that for each $k \in [N]$,*

$$\sup_{x \in \mathcal{I}_k} |f(x) - c_k| \leq \varepsilon.$$

We prove Lemma 12 by explicitly constructing a (σ, ι) network that approximates a piecewise constant function which maps each interval \mathcal{I}_k to c_k . The formal proof of Lemma 12 is in Appendix E.

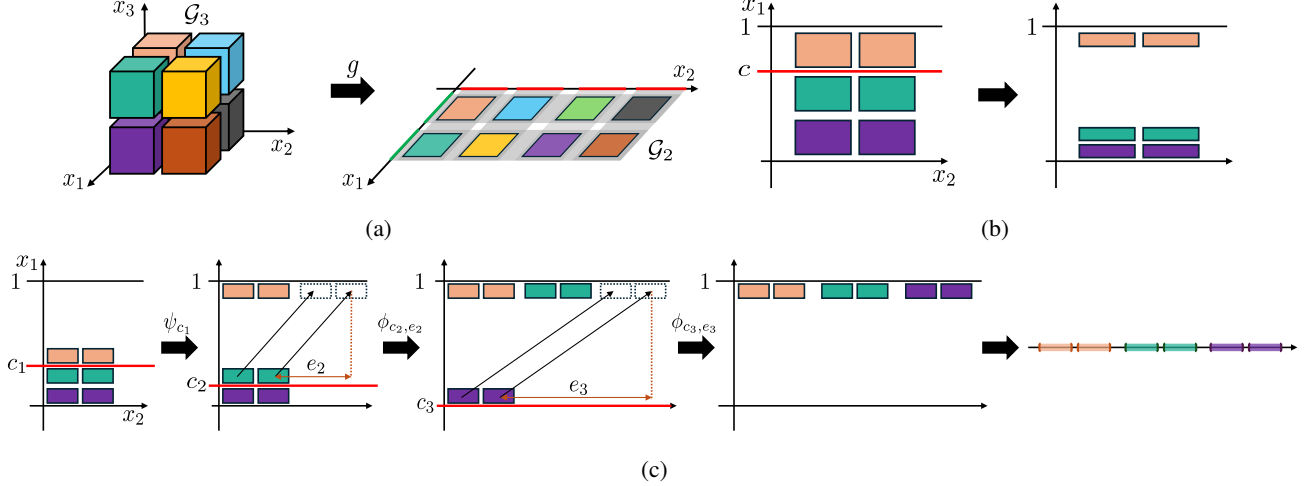


Figure 4: (a) Illustration of a function $g : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ that maps sets in a 3-grid \mathcal{G}_3 of size $(2, 2, 2)$ to distinct sets in 2-grid \mathcal{G}_2 of size $(2, 4)$. (b) Illustration of $\psi_c : \mathbb{R}^2 \rightarrow \mathbb{R}^2$. Here, the first coordinate of $\phi_c(x)$ is approximately 1 or 0 depending on whether x_1 exceeds c or not while the second coordinate is x_2 . (c) Illustration of our construction of f when \mathcal{G} is a 2-grid of size $(3, 2)$ and $e_2, e_3 > 0$ are chosen so that all sets in \mathcal{G} are disjoint in the second coordinate.

Construction of f_1 . In the remainder of this section, we construct a (σ, ι) network f_1 of width d_x that maps each \mathcal{T}_ν to a disjoint compact interval $f_1(\mathcal{T}_\nu)$. Here, we assume $d_x \geq 2$; if $d_x = 1$, we choose $f_1(x) = \iota(x)$. To describe our construction we define a d -grid.

Definition 3. A collection of sets $\mathcal{G} \subset 2^{\mathbb{R}^d}$ is a “ d -grid” of size $(n_1, \dots, n_d) \in \mathbb{N}^d$ if there exist disjoint compact intervals $\mathcal{I}_{i,1}, \dots, \mathcal{I}_{i,n_i} \subset \mathbb{R}$ for each $i \in [d]$ such that

$$\mathcal{G} = \{\mathcal{I}_{i,j_1} \times \dots \times \mathcal{I}_{i,j_d} : j_i \in [n_i], \forall i \in [d]\}.$$

One can observe that any finite set of disjoint intervals is a 1-grid and \mathcal{T}_ν is a d_x -grid. We construct f_1 using the following lemma. The proof of Lemma 13 is in Appendix F.

Lemma 13. Let σ be a squashable function and \mathcal{G} be a 2-grid of size (n_1, n_2) . Then, there exist a (σ, ι) network $f : \mathcal{K} \rightarrow \mathbb{R}$ of width 2 such that $\{f(\mathcal{S}) : \mathcal{S} \in \mathcal{G}\}$ is an 1-grid of size $n_1 n_2$.

Lemma 13 implies that there exists a (σ, ι) network f of width 2 that maps sets in a 2-grid to sets in an 1-grid. This implies that for any distinct sets $\mathcal{S}, \mathcal{S}'$ in the 2-grid, $f(\mathcal{S}) \cap f(\mathcal{S}') = \emptyset$. We now construct f_1 by using (σ, ι) networks that reduce dimensions one by one while preserving the disjointness of each \mathcal{T}_ν .

We first show that for any $d \geq 2$ and d -grid \mathcal{G} of size (n_1, \dots, n_d) , we can construct a (σ, ι) network g of width d that maps sets in the grid to a $(d-1)$ -grid of size $(n_1, \dots, n_{d-2}, n_{d-1} n_d)$. Specifically, such g_d can be constructed by using Lemma 13. Let \mathcal{G}' be a 2-grid defined by considering the last two coordinates of sets in \mathcal{G} , i.e.,

$$\mathcal{G}' = \{\{(x_{d-1}, x_d) : (x_1, \dots, x_d) \in \mathcal{S}\} : \mathcal{S} \in \mathcal{G}\}.$$

Then, g can be constructed as

$$g(x)_i = \begin{cases} x_i & \text{if } i \leq d-2 \\ \phi(x_{d-1}, x_d)_1 & \text{if } i = d-1 \\ \phi(x_{d-1}, x_d)_2 & \text{if } i = d \end{cases}$$

where ϕ is a (σ, ι) network of width 2 in Lemma 13 that maps the 2-grid \mathcal{G}' of size (n_{d-1}, n_d) to some 1-grid of size $n_{d-1} n_d$; see Figure 5a for the illustration of g when $d = 3$.

Let $\mathcal{G}_{d_x} = \{\mathcal{T}_\nu : \nu \in [N]^{d_x}\}$ be a d_x -grid of size (N, \dots, N) . As in the construction of g , we recursively construct g_i for $i = d_x, d_x - 1, \dots, 2$ as a (σ, ι) network of width i that maps an i -grid \mathcal{G}_i of size $(N, \dots, N, N^{d_x-i+1})$ to some $(i-1)$ -grid \mathcal{G}_{i-1} of size $(N, \dots, N, N^{d_x-i+2})$. We then construct f_1 as $f_1 = g_2 \circ g_3 \circ \dots \circ g_{d_x}$. One can observe that f_1 has width d_x and maps sets in \mathcal{G}_{d_x} to distinct sets in some 1-grid.

Intuition behind Lemma 13. We now briefly describe our main proof idea for Lemma 13 where the formal proof is deferred to Appendix F. Our construction of f is based on the squashability of σ . Observe that by the definition of the squashability (Definition 1), for any compact set $\mathcal{K} \subset \mathbb{R}$, there exists a width-1 network ρ that is strictly increasing and approximates STEP on \mathcal{K} (see Condition 2).

Consider a width-2 network $\psi_c : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ defined as $\psi_c(x) = (\rho(x_1 - c), x_2)$ for some $c \in \mathbb{R}$. Then, by choosing a proper c and \mathcal{K} , ψ splits sets in \mathcal{G} into two parts depending on whether their first coordinate exceeds c or not. Here, $\psi_c(\mathcal{S})_1$ will be close to one if the first coordinate of \mathcal{S} exceeds c and $\psi_c(\mathcal{S})_1$ will be close to zero otherwise. We note that by the strict monotonicity of ρ , the order of the

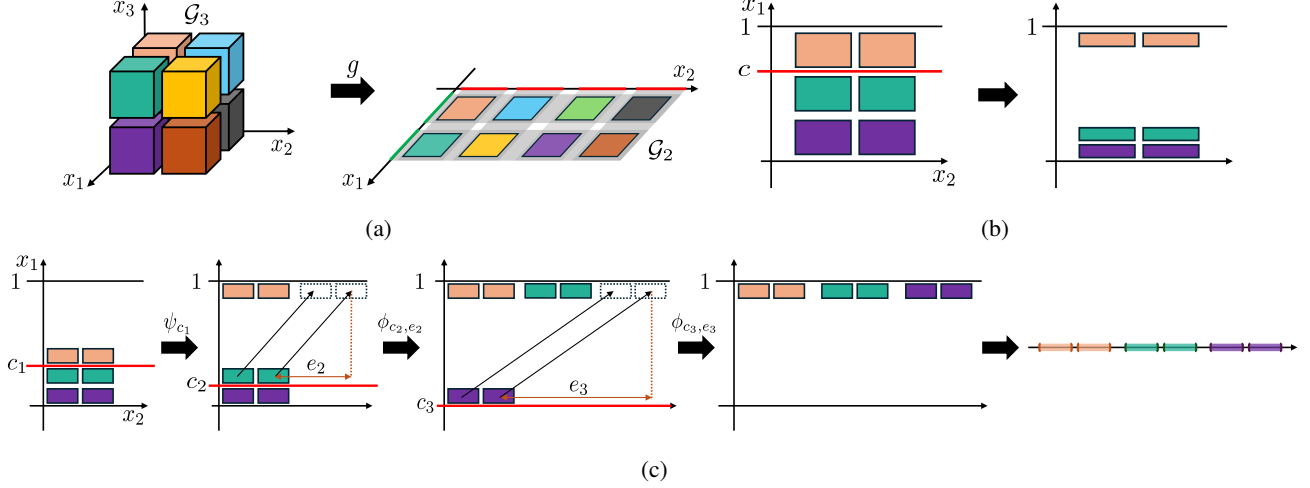


Figure 5: (a) Illustration of a function $g : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ that maps sets in a 3-grid \mathcal{G}_3 of size $(2, 2, 2)$ to distinct sets in 2-grid \mathcal{G}_2 of size $(2, 4)$. (b) Illustration of $\psi_c : \mathbb{R}^2 \rightarrow \mathbb{R}^2$. Here, the first coordinate of $\phi_c(x)$ is approximately 1 or 0 depending on whether x_1 exceeds c or not while the second coordinate is x_2 . (c) Illustration of our construction of f when \mathcal{G} is a 2-grid of size $(3, 2)$ and $e_2, e_3 > 0$ are chosen so that all sets in \mathcal{G} are disjoint in the second coordinate.

first coordinate of the sets does not change. See Figure 5b for the illustration.

Furthermore, we can also change the second coordinate while splitting the first coordinate. For any $e > 0$, by composing ψ_c with some invertible affine transformation $\kappa_e : \mathbb{R}^2 \rightarrow \mathbb{R}^2$, we can construct a width-2 network $\phi_{c,e} = \kappa_e^{-1} \circ \psi_c \circ \kappa_e$ so that

$$\phi_{c,e}(x) \approx \begin{cases} x & \text{if } x_1 \approx 1 \\ x & \text{if } x_1 \approx 0 \text{ and } x_1 < c \\ (1, x_2 + e) & \text{if } x_1 \approx 0 \text{ and } x_1 > c. \end{cases}$$

Using such ψ_c and $\phi_{c,e}$, we construct f by sequentially separating sets in \mathcal{G} based on their first coordinate. First, we apply some invertible affine transformation so that the first coordinate of all sets in \mathcal{G} is close to zero (as in the left of Figure 5c). We then split the sets of the largest first coordinate using ψ_c with some proper choice of c . After that, we sequentially split sets as in Figure 5c. Lastly, we apply a projection onto the second coordinate. For a more formal argument, see Appendix F.

5. Conclusion

In this work, we characterize the minimum width enabling universal approximation of $L^p([0, 1]^{d_x}, \mathbb{R}^{d_y})$. In particular, we consider a general class of activation functions, called squashable, whose alternative composition with affine transformations can approximate both the identity function and STEP on compact domains. We show that for networks using a squashable activation function, the minimum width is $\max\{d_x, d_y, 2\}$ unless $d_x = d_y = 1$; the same minimum

width holds for $d_x = d_y = 1$ if the squashable activation function is monotone. Since all non-affine analytic functions and a class of piecewise functions are squashable, our result covers almost all practical activation functions. We believe that our approach would contribute to a better understanding of the expressive power of deep and narrow networks.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government MSIT (RS-2019-II190079, Artificial Intelligence Graduate School Program, Korea University), the IITP-ITRC (Information Technology Research Center) grant funded by MSIT (IITP-2025-RS-2024-00436857), and the Culture, Sports, and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 (RS-2024-00345025 and 25% by RS-2024-00348469), the KIAS Individual Grant by the Center for AI and Natural Sciences at Korea Institute for Advanced Study (AP092801), and National Research Foundation of Korea grant funded by the MSIT (RS-2025-00515264 and RS-2024-00406127), and Global University Project grant funded by GIST in 2025.

Impact Statement

This paper investigates the theoretical properties of neural networks on the minimum width enabling universal approximation. We could not find notable potential societal consequences of our work.

References

- Eric B Baum. On the capabilities of multilayer perceptrons. *Journal of complexity*, 4(3):193–215, 1988.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Nee-lakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.
- Yongqiang Cai. Achieve the minimum width of neural networks for universal approximation. In *International Conference on Learning Representations (ICLR)*, 2023.
- George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.
- Yifei Duan, Li’ang Li, Guanghua Ji, and Yongqiang Cai. Vanilla feedforward neural networks as a discretization of dynamic systems. *arXiv preprint arXiv:2209.10909*, 2022.
- Boris Hanin and Mark Sellke. Approximating continuous functions by ReLU nets of minimal width. *arXiv preprint arXiv:1710.11278*, 2017.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.
- Guang-Bin Huang and Haroon A Babri. Upper bounds on the number of hidden neurons in feedforward networks with arbitrary bounded nonlinear activation functions. *IEEE transactions on neural networks*, 9(1):224–229, 1998.
- Jesse Johnson. Deep, skinny neural networks are not universal approximators. In *International Conference on Learning Representations (ICLR)*, 2019.
- John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Židek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.
- Patrick Kidger and Terry Lyons. Universal approximation with deep narrow networks. In *Conference on Learning Theory (COLT)*, 2020.
- Namjun Kim, Chanho Min, and Sejun Park. Minimum width for universal approximation using ReLU networks on compact domain. In *International Conference on Learning Representations (ICLR)*, 2024.
- Moshe Leshno, Vladimir Ya Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, 6(6):861–867, 1993.
- Qianxiao Li, Ting Lin, and Zuowei Shen. Deep learning via dynamical systems: An approximation perspective. *Journal of the European Mathematical Society*, 25(5): 1671–1709, 2022.
- Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The expressive power of neural networks: A view from the width. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2017.
- Sejun Park, Jaeho Lee, Chulhee Yun, and Jinwoo Shin. Provable memorization via deep neural networks using sub-linear parameters. In *Conference on Learning Theory (COLT)*, 2021a.
- Sejun Park, Chulhee Yun, Jaeho Lee, and Jinwoo Shin. Minimum width for universal approximation. In *International Conference on Learning Representations (ICLR)*, 2021b.
- Allan Pinkus. Approximation theory of the mlp model in neural networks. *Acta numerica*, 8:143–195, 1999.
- Dennis Rochau, Robin Chan, and Hanno Gottschalk. New advances in universal approximation with neural networks of minimal width. *arXiv preprint arXiv:2411.08735*, 2024.
- Walter Rudin. *Real and Complex Analysis*. McGraw-Hill, Inc., 1987.
- Hans Sagan. *Space-filling curves*. Springer Science & Business Media, 2012.
- Gal Vardi, Gilad Yehudai, and Ohad Shamir. On the optimal memorization power of relu neural networks. In *International Conference on Learning Representations (ICLR)*, 2022.
- A Vaswani. Attention is all you need. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2017.
- Dmitry Yarotsky. Optimal approximation of continuous functions by very deep ReLU networks. In *Conference on Learning Theory (COLT)*, 2018.

A. On activation functions

A.1. Definition of activation functions

- exp:

$$\exp(x) = e^x.$$

- SIGMOID:

$$\text{SIGMOID}(x) = \frac{1}{1 + \exp(-x)}.$$

- tanh:

$$\tanh(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)}.$$

- Leaky-RELU: for $\alpha \in (0, 1)$

$$\text{Leaky-RELU}(x; \alpha) = \begin{cases} x & \text{if } x > 0 \\ \alpha x & \text{if } x \leq 0. \end{cases}$$

- ELU: for $\alpha > 0$

$$\text{ELU}(x; \alpha) = \begin{cases} x & \text{if } x > 0 \\ \alpha(\exp(x) - 1) & \text{if } x \leq 0. \end{cases}$$

- SELU: for $\lambda > 1$ and $\alpha > 0$,

$$\text{SELU}(x; \lambda, \alpha) = \lambda \times \begin{cases} x & \text{if } x > 0 \\ \alpha(\exp(x) - 1) & \text{if } x \leq 0. \end{cases}$$

- GELU:

$$\text{GELU}(x) = x \times \Phi(x)$$

where Φ is the cumulative distribution function for the standard normal distribution.

- CELU: for $\alpha > 0$

$$\text{CELU}(x; \alpha) = \begin{cases} x & \text{if } x > 0 \\ \alpha(\exp(x/\alpha) - 1) & \text{if } x \leq 0. \end{cases}$$

- SOFTPLUS: for $\alpha > 0$,

$$\text{SOFTPLUS}(x; \alpha) = \frac{1}{\alpha} \log(1 + \exp(\alpha x)).$$

- SWISH:

$$\text{SWISH}(x) = x \times \text{SIGMOID}(x).$$

- MISH:

$$\text{MISH}(x) = x \times \tanh(\text{SOFTPLUS}(x; 1)).$$

- HARDSWISH:

$$\text{HARDSWISH}(x) = \begin{cases} 0 & \text{if } x \leq -3 \\ x & \text{if } x \geq 3 \\ x(x+3)/6 & \text{otherwise.} \end{cases}$$

A.2. Proofs related to squashable activation functions

In this section, we prove Lemmas 3–5 by constructing σ network of width 1 satisfying the conditions listed in Condition 2 where σ has the property in each lemma.

A.2.1. PROOF OF LEMMA 3

In this section, we prove Lemma 3. We first prove that if σ satisfies the conditions listed in Lemma 3, then σ is squashable by explicitly constructing a network of width 1 satisfying the Condition 2 using the activation σ that satisfies the conditions listed in Lemma 3. Namely, we now show that for any $\varepsilon, \zeta > 0$ and compact set \mathcal{K} , there exists a σ network $f : \mathbb{R} \rightarrow \mathbb{R}$ of width 1 such that $|f(x) - \text{STEP}(x)| < \varepsilon$ for all $x \in \mathcal{K} \setminus (-\zeta, \zeta)$. To this end, without loss of generality, we assume that $c = 0$, $\phi(x) = x$ and $\mathcal{K} = [-M, M]$ for some $M > 0$ and $[-M, M] \subset [a, b]$.

Then, we have $\rho([a, b]) \subset [a, b]$. For any $n \in \mathbb{N}$, define $\psi_n : \mathbb{R} \rightarrow \mathbb{R}$ by

$$\psi_n(x) = \rho^n(x).$$

Then, $\psi_n([a, b]) \subset [a, b]$ and ψ_n is strictly increasing on $[a, b]$. Furthermore, for any $n \in \mathbb{N}$, $\psi_n(x) < \psi_{n+1}(x)$ for $x \in (0, b)$ and $\psi_n(x) > \psi_{n+1}(x)$ for $x \in (a, 0)$. We now show that there exists $N \in \mathbb{N}$ such that if $n \geq N$,

$$a < \psi_n(-\zeta) < a + (b - a)\varepsilon, \quad b - (b - a)\varepsilon < \psi_n(\zeta) < b. \quad (2)$$

Then, since ψ_n is strictly increasing, $\psi(x) \in (a, a + (b - a)\varepsilon)$ for any $x \in [-M, -\zeta]$ and $\psi(x) \in (b - (b - a)\varepsilon, b)$ for any $x \in [\zeta, M]$. Then, define a σ network $f : \mathbb{R} \rightarrow \mathbb{R}$ of width 1 by

$$f(x) = \frac{1}{b - a}(\psi_N(x) - a).$$

Then, $f([-M, M]) \subset f([a, b]) \subset [0, 1]$ and f is strictly increasing, and $0 < f(x) \leq f(-\zeta) < \varepsilon$ for $x \in [-M, -\zeta]$ and $1 - \varepsilon < f(\zeta) \leq f(x) < 1$ for $x \in [\zeta, M]$. It implies that f is squashable and this completes the proof.

We now show the existence of $N \in \mathbb{N}$ such that ψ_n satisfies Eq. (2) if $n \geq N$. Let $a_n = \psi_n(\zeta)$. Then, $a_n < a_{n+1} < b$ for all $n \in \mathbb{N}$. Then, by the monotone convergence theorem, there exists $L \in \mathbb{R}$ such that $a < L \leq b$ and $\lim_{n \rightarrow \infty} a_n = L$. Here, if $L < b$, then

$$\lim_{n \rightarrow \infty} a_{n+1} = \lim_{n \rightarrow \infty} \rho(a_n) = \rho(L) > L$$

which is a contradiction. Hence, $L = b$ and this guarantees the existence of $N_1 \in \mathbb{N}$ such that if $n \geq N_1$, then $b - (b - a)\varepsilon < \psi_n(\zeta) < b$. Likewise, there exists $N_2 \in \mathbb{N}$ such that if $n \geq N_2$, then $a < \psi_n(-\zeta) < a + (b - a)\varepsilon$. If we choose $N > \max\{N_1, N_2\}$, then our σ network f of width 1 satisfies Condition 2.

A.2.2. PROOF OF LEMMA 4

In this section, we prove Lemma 4. To this end, it suffices to show the existence of the σ network $\rho : \mathbb{R} \rightarrow \mathbb{R}$ of width 1 such that

- ρ is strictly increasing on $[0, 1]$,
- $\rho(0) = 0$ and $\rho(1) = 1$, and
- $\rho'(0) < 1$ and $\rho'(1) < 1$.

Then, from the second and third line in the above conditions, one can observe that $\rho(x) < x$ if $x \in (0, \delta)$ and $\rho(x) > x$ if $x \in (1 - \delta, 1)$ for some $\delta > 0$. Then, by the intermediate value theorem, the equation $\rho(x) = x$ has at least one solution in $(0, 1)$. Here, since ρ is analytic, there are finitely many solutions $c_1, \dots, c_k \in (0, 1)$ such that $c_1 < \dots < c_k$ and $\rho(c_i) = c_i$ for $i \in [k]$. If $k = 1$, then ρ satisfies the conditions of Lemma 3 with $[0, 1]$ and $\phi(x) = x$. Otherwise, ρ satisfies the conditions of Lemma 3 with $[0, c_2]$ and $\phi(x) = x$. It completes the proof.

We now construct such a σ network ρ by considering the following cases: (1) there exists $a \in \mathbb{R}$ such that $\sigma'(a) = 0$ and (2) $\sigma'(x) \neq 0$ for all $x \in \mathbb{R}$.

We considered the case (1) in Lemma 15 in Appendix A.3. We now consider the case (2): $\sigma'(x) \neq 0$ for all $x \in \mathbb{R}$. Without loss of generality, $\sigma'(x) > 0$ for all $x \in \mathbb{R}$. To this end, we consider the following cases again: (2-1) there exists $c \in \mathbb{R}$ such that $\sigma''(x) > 0$ in $(c - \delta, c)$ and $\sigma''(x) < 0$ in $(c, c + \delta)$ for some $\delta > 0$ and (2-2) otherwise.

We considered the case (2-1) in Lemma 14 in Appendix A.3. We now consider the case (2-2). Specifically, it suffices to consider the case that there exists $a \in \mathbb{R}$ such that $\sigma''(a) > 0$ and $\sigma''(x) \geq 0$ for $x > a$. Otherwise, suppose that $\sigma''(x) \leq 0$ for all $x \in \mathbb{R}$. Then, we can make σ to convex function by taking an affine transformation: $\sigma_0(x) = -\sigma(-x)$.

Without loss of generality, assume that $a = 0$ and $\sigma(0) = 0$. Then, we define a σ network $\psi : \mathbb{R} \rightarrow \mathbb{R}$ such that

$$\psi(x) = \frac{1}{\sigma(b)} \sigma(bx)$$

for $b > 0$. We will assign an explicit value of b later. Then, we have $\psi(0) = 0$, $\psi(1) = 1$, and ψ is strictly increasing on $[0, 1]$. Then, we construct a σ network $\rho : \mathbb{R} \rightarrow \mathbb{R}$ of width 1 by

$$\rho(x) = 1 - \psi(1 - \psi(x)).$$

Then, $\rho(0) = 0$, $\rho(1) = 1$, and ρ is strictly increasing on $[0, 1]$. Furthermore, one can observe that

$$\rho'(0) = \rho'(1) = \psi'(0)\psi'(1) = \frac{b^2 \sigma'(b) \sigma'(0)}{\sigma(b)^2}.$$

We now show the existence of $b \in \mathbb{R}$ such that

$$\frac{b^2 \sigma'(b) \sigma'(0)}{\sigma(b)^2} < 1.$$

To this end, consider a function $g : (0, \infty) \rightarrow \mathbb{R}$ defined by

$$g(x) = \frac{1}{x} - \frac{\sigma'(0)}{\sigma(x)}.$$

Then, one can observe that

$$g'(x) = \frac{1}{x^2} \left(\frac{x^2 \sigma'(x) \sigma'(0)}{\sigma(x)^2} - 1 \right).$$

Since $x > 0$, it suffices to show the existence of $b > 0$ such that $g'(b) < 0$. Since $\sigma''(0) > 0$ and $\sigma(x) > 0$ for all $x > 0$, it can be easily shown that $\sigma(x) > \sigma'(0)x$ for all $x > 0$. It implies that $g(x) > 0$ for $x > 0$. Furthermore, since $\sigma(x) \rightarrow \infty$ as $x \rightarrow \infty$, it holds that $g(x) \rightarrow 0$ as $x \rightarrow \infty$. Then, there exists $M > 1$ such that $g(1) > g(M)$ since $g(x) \rightarrow 0$ as $x \rightarrow \infty$ and $g(1) > 0$. Then, by the mean value theorem, there exists $b \in (1, M)$ such that

$$\frac{g(M) - g(1)}{M - 1} = g'(b) < 0.$$

It completes the proof.

A.2.3. PROOF OF LEMMA 5

In this section, we prove Lemma 5. To this end, we first consider the case that the given activation is a piecewise linear function. Without loss of generality, we assume that

$$\sigma_1(x) = \begin{cases} ax & x \in [-1, 0) \\ x & x \in [0, 2] \end{cases} \quad (3)$$

where $0 < a < 1$. We now construct a σ network ρ of width 1 as

$$\rho(x) = 1 - \sigma_1(1 - \sigma_1(x)) = \begin{cases} ax & x \in [-1, 0) \\ x & x \in [0, 1) \\ ax + 1 - a & x \in [1, 2]. \end{cases}$$

Since $0 < a < 1$, it is easy to observe that σ_1 satisfies Condition 2 by Lemma 3.

We now consider the general case. Suppose that $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ satisfies the conditions listed in Lemma 5. We show this by constructing a σ network ψ of width 1 that approximates $\sigma_1(x)$ in Eq. (3) with $a = \sigma'(c_-)/\sigma'(c_+)$ within an arbitrary error for any $x \in [-1, 2]$. Then, we can easily verify that Lemma 3 can be applied to the same construction of σ network of width 1 as above, $1 - \psi(1 - \psi(x))$, and this completes the proof.

We now show the existence of such ψ . To this end, without loss of generality, we assume that $c = 0, \sigma(0) = 0$, $0 < \sigma'(c_-) < \sigma'(c_+)$, and σ is strictly increasing on $(c - \delta, c + \delta)$. For $r > 0$, construct a σ network ψ of width 1 as

$$\psi_r(x) = \frac{\sigma(rx)}{r}.$$

By the mean value theorem, for $-1 \leq x < 0$, there exists $d_r \in (rx, 0)$ such that $\psi_r(x) = x\sigma'(d_r)$ and for $0 < x \leq 2$, there exists $e_r \in (0, rx)$ such that $\psi_r(x) = x\sigma'(e_r)$. Since $\sigma'(x)$ is continuous on $(c - \delta, c + \delta)$, it holds that $\sigma'(d_r) \rightarrow \sigma'(c_-)$ and $\sigma'(e_r) \rightarrow \sigma'(c_+)$ as $r \rightarrow 0$, respectively. It implies that

$$\lim_{r \rightarrow 0} \psi_r(x) = \begin{cases} \sigma'(c_-)x & x \in [-1, 0) \\ \sigma'(c_+)x & x \in [0, 2]. \end{cases}$$

Thus, choosing $\psi(x) = \psi_r(x)/\sigma'(c_+)$ with sufficiently small $r > 0$ completes the proof.

A.3. Additional properties for functions to satisfy Condition 2

In this section, we suggest the additional properties for activation functions to satisfy Condition 2. Lemma 14 implies that an activation σ satisfies Condition 2 if there exists a point where the sign of σ'' converts from positive to negative.

Lemma 14. *Let $c \in \mathbb{R}$ and $\delta > 0$. Suppose that a function $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ such that σ is twice differentiable in $(c - \delta, c + \delta)$, $\sigma''(x) > 0$ in $(c - \delta, c)$ and $\sigma''(x) < 0$ in $(c, c + \delta)$. Then, σ satisfies Condition 2.*

Proof. To prove Lemma 14, we now choose appropriate $a, b \in \mathbb{R}$ and $\phi : \mathbb{R} \rightarrow \mathbb{R}$ and apply Lemma 3 with our a, b, c and ϕ . We consider a line passing $(c, \sigma(c))$ as ϕ . Since $\rho''(x) > 0$ if $x < c$ and $\rho''(x) < 0$ if $x > c$, we can choose a slope of ϕ so that ϕ and ρ meet once in $(c - \delta, c)$ and $(c, c + \delta)$, respectively. Let $\alpha = \max\{\sigma(c) - \sigma(c - \delta/2), \sigma(c + \delta/2) - \sigma(c)\}$ and $\phi(x) = \frac{\alpha}{\delta/2}(x - c) + \sigma(c)$. Here, one can easily observe that $\frac{\alpha}{\delta/2} < \sigma'(c)$. Without loss of generality, suppose that $\alpha = \sigma(c) - \sigma(c - \delta/2)$. Then, it holds that

$$\phi(c + \delta/2) = \sigma(c) - \sigma(c - \delta/2) + \sigma(c) \geq \sigma(c + \delta/2) - \sigma(c) = \sigma(c + \delta/2).$$

Then, by the intermediate value theorem, there exists $b \in (c, c + \delta/2]$ such that $\phi(b) = \sigma(b)$. Furthermore, since $\phi(c - \delta/2) = \sigma(c - \delta/2)$, choosing $a = c - \delta/2$ and applying Lemma 3 with our a, b, c and ϕ completes the proof. \square

Lemmas 15 and 16 imply that if σ satisfies a condition stronger than the analytic condition in a compact interval, then σ satisfies Condition 2.

Lemma 15. *Consider $a_1, a_2 \in \mathbb{R}$ such that $\sigma(x)$ is nonaffine analytic on $x \in [a_1, a_2]$. Suppose that there exists $c \in [a_1, a_2]$ such that $\sigma'(c) = 0$. Then, σ satisfies Condition 2.*

Proof. It suffices to show the existence of the σ network $\rho : \mathbb{R} \rightarrow \mathbb{R}$ of width 1 such that ρ is strictly increasing on $[0, 1]$, $\rho(0) = 0, \rho(1) = 1, \rho'(0) < 1$ and $\rho'(1) < 1$ (see Appendix A.2.2). Since σ is a nonaffine analytic function that has a zero derivative at some point, $b \in (c, a_2]$ such that σ is strictly monotone on $[c, b]$ with nonlinearity. Without loss of generality, assume that $c = 0, \sigma(0) = 0$ and $\sigma(x)$ is strictly increasing on $[0, b]$. Then, we define a σ network $\psi : \mathbb{R} \rightarrow \mathbb{R}$ such that

$$\psi(x) = \frac{1}{\sigma(b)}\sigma(bx).$$

Then, $\psi(0) = 0, \psi(1) = 1$, and ψ is strictly increasing on $[0, 1]$. We now construct a σ network ρ by

$$\rho(x) = 1 - \psi(1 - \psi(x)).$$

Then, $\rho(0) = 0$, $\rho(1) = 1$, and ρ is strictly increasing on $[0, 1]$. Furthermore, one can observe that

$$\rho'(x) = \psi'(1 - \psi(x))\psi'(x).$$

Then, we have $\rho'(0) = \rho'(1) = 0$ since $\psi'(0) = 0$. It completes the proof. \square

Lemma 16. Consider $a_1, a_2 \in \mathbb{R}$ such that $\sigma(x)$ is analytic on $x \in [a_1, a_2]$. Assume that there exists $x \in [a_1, a_2]$ such that

$$a_2 \geq \frac{2\sigma'(x)}{\sigma''(x)} + x.$$

Then, σ satisfies Condition 2.

Proof. In this proof, $\sigma^{(n)}(x)$ is defined as n -times derivative: $\sigma^{(n)}(x) = \frac{d^n \sigma(x)}{dx^n}$. We only need to consider the case $\sigma'(x) > 0$ and $\sigma^{(2)}(x) > 0$; see the case (1) and (2-1) in Appendix A.2.2.

Consider an arbitrary $x_0 \in (a_1, a_2)$. For $b \in (a_1 - x_0, a_2 - x_0)$, define $\psi : (0 - \epsilon, 1 + \epsilon) \rightarrow \mathbb{R}$ as

$$\psi(x) := \frac{1}{\sigma(b + x_0) - \sigma(x_0)}(\sigma(bx + x_0) - \sigma(x_0)).$$

Then, $\psi(0) = \psi(1) = 1$. Define ρ as

$$\rho(x) := 1 - \psi(1 - \psi(x)).$$

Then,

$$\rho'(0) = \rho'(1) = \psi'(0)\psi'(1) = \frac{b^2\sigma'(b + x_0)\sigma'(x_0)}{(\sigma(b + x_0) - \sigma(x_0))^2}.$$

It is sufficient to find a value b such that $\rho'(0) = \rho'(1) < 1$. Define g as

$$g(x) := \frac{1}{x} - \frac{\sigma'(x_0)}{\sigma(x + x_0) - \sigma(x_0)}.$$

Then, as

$$g'(x) = -\frac{1}{x^2} + \left(\frac{\sigma'(x_0)\sigma'(x + x_0)}{(\sigma(x + x_0) - \sigma(x_0))^2} \right) = \frac{1}{x^2} \left(\frac{x^2\sigma'(x + x_0)\sigma'(x_0)}{(\sigma(x + x_0) - \sigma(x_0))^2} - 1 \right),$$

it is sufficient to find a number x such that $g'(x) < 0$. Then, there exist smooth functions h, h_1, h_2 such that

$$\begin{aligned} g(x) &= \frac{1}{x} - \frac{\sigma'(x_0)}{\sigma(x + x_0) - \sigma(x_0)} = \frac{1}{x} - \frac{\sigma'(x_0)}{\sigma'(x_0)x + \sigma^{(2)}(x_0)\frac{x^2}{2} + \sigma^{(3)}(x_0)\frac{x^3}{6} + x^4h(x)} \\ &= \frac{1}{x} - \frac{1}{x + \frac{\sigma^{(2)}(x_0)}{\sigma'(x_0)}\frac{x^2}{2} + \frac{\sigma^{(3)}(x_0)}{\sigma'(x_0)}\frac{x^3}{6} + \frac{h(x)}{\sigma'(x_0)}x^4} \\ &= \frac{\frac{\sigma^{(2)}(x_0)}{2\sigma'(x_0)} \left(1 + \frac{\sigma^{(3)}(x_0)}{\sigma^{(2)}(x_0)}\frac{x}{3} + \frac{h(x)}{\sigma^{(2)}(x_0)}x^2 \right)}{1 + \frac{\sigma^{(2)}(x_0)}{\sigma'(x_0)}\frac{x}{2} + \frac{\sigma^{(3)}(x_0)}{\sigma'(x_0)}\frac{x^2}{6} + \frac{h(x)}{\sigma'(x_0)}x^3} = \frac{\sigma^{(2)}(x_0)}{2\sigma'(x_0)} \frac{1 + \frac{\sigma^{(3)}(x_0)}{\sigma^{(2)}(x_0)}\frac{x}{3} + h_2(x)x^2}{1 + \frac{\sigma^{(2)}(x_0)}{\sigma'(x_0)}\frac{x}{2} + h_1(x)x^2}. \end{aligned}$$

Then, $g'(x) < 0$ if

$$\frac{\sigma^{(2)}(x_0)}{2\sigma'(x_0)} > \frac{\sigma^{(3)}(x_0)}{2\sigma^{(2)}(x_0)}. \quad (4)$$

Assume that the above inequality is not satisfied for any $x_0 \in (a_1, a_2)$; that is, for any $x \in (a_1, a_2)$

$$\frac{\sigma^{(2)}(x)}{2\sigma'(x)} \leq \frac{\sigma^{(3)}(x)}{3\sigma^{(2)}(x)}.$$

Then, for any $a_1 < x_1 < y < a_2$,

$$\begin{aligned} \int_{x_1}^y \frac{\sigma^{(2)}(x)}{2\sigma'(x)} dx &\leq \int_{x_1}^y \frac{\sigma^{(3)}(x)}{3\sigma^{(2)}(x)} dx \iff \frac{3}{2} \log \left(\frac{\sigma'(y)}{\sigma'(x_1)} \right) \leq \log \left(\frac{\sigma^{(2)}(y)}{\sigma^{(2)}(x_1)} \right) \\ &\iff \frac{\sigma^{(2)}(x_1)}{\sigma'(x_1)^{\frac{3}{2}}} \leq \left(\frac{\sigma^{(2)}(y)}{\sigma'(y)^{\frac{3}{2}}} \right), \end{aligned}$$

which leads to

$$\frac{\sigma^{(2)}(x_1)}{\sigma'(x_1)^{\frac{3}{2}}} (z - x_1) \leq 2 \left(\frac{1}{\sigma'(x_1)^{\frac{1}{2}}} - \frac{1}{\sigma'(z)^{\frac{1}{2}}} \right),$$

for any $a_1 < x_1 < z < a_2$. Thus,

$$\frac{1}{\left(\frac{1}{\sigma'(x_1)^{\frac{1}{2}}} - \frac{\sigma^{(2)}(x_1)}{2\sigma'(x_1)^{\frac{3}{2}}} (z - x_1) \right)^2} \leq \sigma'(z).$$

□

We lastly present Lemma 17 which implies that if strictly monotone σ has a limit, then σ satisfies Condition 2.

Lemma 17. *A continuous function $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ satisfies Condition 2 if σ has strictly monotonicity and there exists $\lim_{x \rightarrow \infty} \sigma(x)$ or $\lim_{x \rightarrow -\infty} \sigma(x)$.*

Proof. Without loss of generality, we assume that $\sigma(x)$ is strictly increasing and $\lim_{x \rightarrow -\infty} \sigma(x) = 0$. We consider the two cases: (1) $\lim_{x \rightarrow \infty} \sigma(x) = \alpha < \infty$, and (2) $\lim_{x \rightarrow \infty} \sigma(x) = \infty$.

For the first case, we can easily verify that σ satisfies Condition 2 by composing affine functions before and after σ :

$$\rho(x) = \frac{1}{\alpha} \sigma(Mx)$$

where $M > 0$ is sufficiently large.

We now consider the second case. Suppose that $\lim_{x \rightarrow \infty} \sigma(x) = \infty$. We construct a σ network ψ of width 1 such that

$$\psi(x) = \frac{1}{\sigma(1)} \times (\sigma(1) - \sigma(1 - \sigma(x))).$$

Then, it is easy to observe that ψ is strictly increasing, $\lim_{x \rightarrow \infty} \psi(x) = 1$ and $\lim_{x \rightarrow -\infty} \psi(x) = 0$. Then, we can consider ϕ as in the first case. Hence, σ is squashable and this completes the proof. □

B. Our choice of δ, γ, N

We first choose a sufficiently small $\delta > 0$ so that $\delta \leq \varepsilon/(d_y^{1/p} \times 3^{1+1/p})$. And then, choose a small enough $\gamma > 0$ so that $\gamma \leq \varepsilon^p/(3d_x d_y)$ and $\omega_{f_{\text{dec}}}(\gamma) \leq \varepsilon/3^{1+1/p}$. Lastly, we choose large enough $N \in \mathbb{N}$ satisfying $\text{diam}(f^*(\mathcal{T}_\nu)) = \omega_{f^*}((1-2\gamma)/N) \leq \varepsilon/(d_y^{1/p} \times 3^{1+1/p})$ for each $\nu \in [N]^{d_x}$. Here, $\omega_{f_{\text{dec}}}$ and ω_{f^*} denote the modulus of continuity of given function f in the p -norm: $\|f(x) - f(x')\|_p \leq \omega_f(\|x - x'\|_p)$ for all $x, x' \in [0, 1]^{d_x}$. Then,

$$\begin{aligned}
 \|f_{\text{dec}} \circ f_{\text{enc}} - f^*\|_{L^p}^p &= \int_{[0,1]^{d_x}} \|f_{\text{dec}} \circ f_{\text{enc}}(x) - f^*(x)\|_p^p d\mu_{d_x} \\
 &\leq \int_{[0,1]^{d_x} \setminus \bigcup_{\nu \in [N]^{d_x}} \mathcal{T}_\nu} \|f_{\text{dec}} \circ f_{\text{enc}}(x) - f^*(x)\|_p^p d\mu_{d_x} + \int_{\bigcup_{\nu \in [N]^{d_x}} \mathcal{T}_\nu} \|f_{\text{dec}} \circ f_{\text{enc}}(x) - f^*(x)\|_p^p d\mu_{d_x} \\
 &\leq d_y \times \mu_{d_x} \left([0,1]^{d_x} \setminus \bigcup_{\nu \in [N]^{d_x}} \mathcal{T}_\nu \right) + \sum_{\nu \in [N]^{d_x}} \int_{\mathcal{T}_\nu} \|f_{\text{dec}} \circ f_{\text{enc}}(x) - f^*(x)\|_p^p d\mu_{d_x} \\
 &\leq d_y \times (1 - (1-2\gamma)^{d_x}) + \sum_{\nu \in [N]^{d_x}} \int_{\mathcal{T}_\nu} (\|f_{\text{dec}} \circ f_{\text{enc}}(x) - f_{\text{dec}}(c_\nu)\|_p + \|f_{\text{dec}}(c_\nu) - f^*(x)\|_p)^p d\mu_{d_x} \\
 &\leq 2d_x d_y \gamma + \sum_{\nu \in [N]^{d_x}} \int_{\mathcal{T}_\nu} (\omega_{f_{\text{dec}}}(\gamma) + d_y^{1/p} \times (\text{diam}(f^*(\mathcal{T}_\nu)) + \delta))^p d\mu_{d_x} \\
 &\leq 2d_x d_y \gamma + (\omega_{f_{\text{dec}}}(\gamma) + d_y^{1/p} \times (\text{diam}(f^*(\mathcal{T}_\nu)) + \delta))^p \leq \varepsilon^p
 \end{aligned}$$

where c_ν is chosen so that $\text{dist}(f_{\text{dec}}(c_\nu), f^*(\mathcal{T}_\nu)) \leq \delta$ for each $\nu \in [N]^{d_x}$. This leads us to the statement of Lemma 6.

C. Proof of Lemma 8

In this section, we prove Lemma 8. Since $f : \mathcal{K} \rightarrow \mathbb{R}^d$ is a (σ, ι) network of width w , we can express $f : \mathcal{K} \rightarrow \mathbb{R}^d$ as follows:

$$f = t_L \circ \phi_{L-1} \circ t_{L-1} \circ \cdots \circ \phi_1 \circ t_1$$

where $t_\ell : \mathbb{R}^{d_{\ell-1}} \rightarrow \mathbb{R}^{d_\ell}$ is an affine transformation, and $\phi_\ell(x) = (\rho_{\ell,1}(x), \dots, \rho_{\ell,d_\ell}(x))$ for $\rho_{\ell,1}, \dots, \rho_{\ell,d_\ell} \in \{\sigma, \iota\}$ for all $\ell \in [L]$. Since σ satisfies Condition 1, by Lemma 1, for arbitrary compact set \mathcal{C} and for any $\delta > 0$, there exist affine transformations $h_1 : \mathbb{R} \rightarrow \mathbb{R}$ and $h_2 : \mathbb{R} \rightarrow \mathbb{R}$ such that

$$|h_1 \circ \sigma \circ h_2(x) - \iota(x)| < \delta$$

for all $x \in \mathcal{C}$; we will assign explicit value to δ later. We denote $h_1 \circ \sigma \circ h_2$ as $\tilde{\sigma}$. We note that this lemma can be applied for any given compact set. Since we are considering a compact domain and a continuous activation function, the error arising from replacing ι with $\tilde{\sigma}$ can be reduced.

To this end, we choose a σ network g by applying same affine transformation t_1, \dots, t_L and $\tilde{\sigma}$:

$$g = t_L \circ \psi_{L-1} \circ t_{L-1} \circ \cdots \circ \psi_1 \circ t_1$$

where $\psi(x) = (\tilde{\rho}_{\ell,1}(x), \dots, \tilde{\rho}_{\ell,d_\ell}(x))$ with $\tilde{\rho}_{\ell,i} = \sigma$ if $\rho_{\ell,i} = \sigma$ and $\tilde{\rho}_{\ell,i} = \tilde{\sigma}$ if $\rho_{\ell,i} = \iota$ for $\ell \in [L]$ and $i \in [d_\ell]$.

We denote f_ℓ and g_ℓ by the first ℓ layers of f and g with the subsequent affine transformation t_ℓ , respectively. i.e.,

$$f_\ell = t_\ell \circ \phi_{\ell-1} \circ t_{\ell-1} \circ \cdots \circ \phi_1 \circ t_1 \quad \text{and} \quad g_\ell = t_\ell \circ \psi_{\ell-1} \circ t_{\ell-1} \circ \cdots \circ \psi_1 \circ t_1.$$

Then, for each $\ell \in [L] \setminus \{1\}$ and for any $x \in \mathcal{K}$, it holds that

$$\begin{aligned} \|f_\ell(x) - g_\ell(x)\|_\infty &= \|t_\ell \circ \phi_{\ell-1} \circ f_{\ell-1}(x) - t_\ell \circ \psi_{\ell-1} \circ g_{\ell-1}(x)\|_\infty \\ &\leq \omega_{t_\ell}(\|\phi_{\ell-1} \circ f_{\ell-1}(x) - \psi_{\ell-1} \circ g_{\ell-1}(x)\|_\infty) \\ &\leq \omega_{t_\ell}(\|\phi_{\ell-1} \circ f_{\ell-1}(x) - \phi_{\ell-1} \circ g_{\ell-1}(x)\|_\infty + \|\phi_{\ell-1} \circ g_{\ell-1}(x) - \psi_{\ell-1} \circ g_{\ell-1}(x)\|_\infty). \end{aligned}$$

Here, we note that for any $\ell \in [L]$, ω_{t_ℓ} is well-defined since t_ℓ is uniformly continuous on $\mathbb{R}^{d_{\ell-1}}$. Then, by the definition of $\psi_{\ell-1}$ and $\tilde{\sigma}$, it holds that

$$\|\phi_{\ell-1} \circ g_{\ell-1}(x) - \psi_{\ell-1} \circ g_{\ell-1}(x)\|_\infty \leq \max_{i \in [d_{\ell-1}]} |\tilde{\sigma}(g_{\ell-1}(x)_i) - \iota(g_{\ell-1}(x)_i)| < \delta.$$

Furthermore, since we are considering the compact domain and $\phi_{\ell-1}$ is continuous, $\omega_{\phi_{\ell-1}}$ is well-defined and

$$\|\phi_{\ell-1} \circ f_{\ell-1}(x) - \phi_{\ell-1} \circ g_{\ell-1}(x)\|_\infty \leq \omega_{\phi_{\ell-1}}(\|f_{\ell-1}(x) - g_{\ell-1}(x)\|_\infty)$$

Hence, we have

$$\begin{aligned} \|f_\ell(x) - g_\ell(x)\|_\infty &= \omega_{t_\ell}(\|\phi_{\ell-1} \circ f_{\ell-1}(x) - \phi_{\ell-1} \circ g_{\ell-1}(x)\|_\infty + \|\phi_{\ell-1} \circ g_{\ell-1}(x) - \psi_{\ell-1} \circ g_{\ell-1}(x)\|_\infty) \\ &\leq \omega_{t_\ell}(\omega_{\phi_{\ell-1}}(\|f_{\ell-1}(x) - g_{\ell-1}(x)\|_\infty) + \delta) \end{aligned} \tag{5}$$

for all $\ell \in [L] \setminus \{1\}$. By iteratively applying Eq. (5), we have

$$\begin{aligned} \|f(x) - g(x)\|_\infty &\leq \omega_{t_L}(\omega_{\phi_{L-1}}(\|f_{L-1}(x) - g_{L-1}(x)\|_\infty) + \delta) \\ &\vdots \\ &\leq \omega_{t_L}(\omega_{\phi_{L-1}}(\cdots (\omega_{t_3}(\omega_{\phi_2}(\omega_{t_2}(\delta) + \delta) + \delta) + \delta) \cdots) + \delta). \end{aligned}$$

Consequently, by choosing sufficiently small $\delta > 0$, we can reduce this within arbitrary error $\varepsilon > 0$ and this completes the proof.

D. Proof of Lemma 11

In this section, we prove Lemma 11. To show Lemma 11, we construct (σ, ι) network \tilde{f} of width $d + 1$ as follows: for each $i \in [d]$ and $\nu \in [N]^d$,

$$\tilde{f}(x)_i = f(x)_i \quad \text{and} \quad \left[\frac{1}{2N}, 1 - \frac{1}{2N} \right] \subset \tilde{f}(\mathcal{I}_\nu)_{d+1}. \quad (6)$$

Then, since \tilde{f} is continuous, for each $\nu \in [N]^d$ and $j \in [N]$, there exists $\mathcal{J}_{(\nu, j)} \subset \mathcal{I}_\nu$ such that

$$\tilde{f}(\mathcal{J}_{(\nu, j)})_{d+1} \subset \left[\frac{j-1}{N}, \frac{j}{N} \right].$$

Furthermore, since $\mathcal{J}_{(\nu, j)} \subset \mathcal{I}_\nu$ for each $\nu = (\nu_1, \dots, \nu_d) \in [N]^d$ and $j \in [N]$, it can be easily observed that

$$\tilde{f}(\mathcal{J}_{(\nu, j)})_i \subset \left[\frac{\nu_i - 1}{N}, \frac{\nu_i}{N} \right]$$

for all $i \in [d]$. It implies that $\tilde{f}(\mathcal{J}_{\nu'}) \subset \mathcal{C}_{N, d+1, \nu'}$ and this completes the proof.

We now construct a (σ, ι) network of width $d + 1$ satisfying Eq. (6). To this end, we first present the following lemma.

Lemma 18. *Let $\sigma \in \mathfrak{S}$ and $z_1, z_2, \dots, z_k \in [0, 1]$ such that $z_i \neq z_j$ for all $i \neq j$. Let $\gamma > 0$ such that $\gamma < \min_{i \neq j} |z_i - z_j|/2$. Then, there exists a (σ, ι) network $f : [0, 1] \rightarrow \mathbb{R}^2$ of width 2 satisfying the following:*

- $f(x)_1 = x$ on $[0, 1]$,
- $\left[\frac{1}{2N}, 1 - \frac{1}{2N} \right] \subset f(\mathcal{B}_\gamma(z_i))_2$ for all $i \in [k]$,
- $f([0, 1]) \subset [0, 1]^2$.

One can observe that Lemma 18 allows us to prove Lemma 11 directly. We choose $\gamma > 0$ and $z_\nu \in \mathcal{I}_\nu$ for each ν such that $\mathcal{B}_\gamma(z_\nu) \subset \mathcal{I}_\nu$. Applying Lemma 18 with our choices of z_ν 's and γ , we construct a (σ, ι) network $\phi : [0, 1] \rightarrow \mathbb{R}^2$ of width 2 satisfying the conditions listed in Lemma 18. Then, we complete the proof by constructing \tilde{f} in Eq. (6) as follows:

$$\begin{aligned} \tilde{f}(x)_1 &= \phi(f(x)_1)_1, \quad \tilde{f}(x)_{d+1} = \phi(f(x)_1)_2, \text{ and} \\ \tilde{f}(x)_i &= \iota \circ \dots \circ \iota(f(x)_i) \text{ for all } i \in \{2, \dots, d\}. \end{aligned}$$

D.1. Proof of Lemma 18

Without loss of generality, we assume $k = 2m$ for some $m \in \mathbb{N}$ and $0 = z_0 < z_1 < z_2 < \dots < z_{2m} < z_{2m+1} = 1$; otherwise, we can add an auxiliary $z_{k+1} \in \mathbb{R}$ such that $z_k < z_{k+1} < 1$. Let $\mathcal{X} = \{z_1, z_2, \dots, z_{2m}\}$, $\mathcal{D}_{\mathcal{X}, \gamma} = [0, 1] \setminus \bigcup_{i=1}^{2m} (z_i - \gamma, z_i + \gamma)$, and $\mathcal{A}_{\mathcal{X}} = \bigcup_{i=1}^m (z_{2i-1}, z_{2i}]$.

To construct f in Lemma 18 using (σ, ι) network, we use the Condition 2 that for any compact set \mathcal{C} , σ can approximate STEP except for the neighborhood of a breakpoint. We first construct (STEP, ι) network $h : [0, 1] \rightarrow \{0, 1\}$ of width 2 such that

$$h(x) = \begin{cases} 1 & \text{if } x \in \mathcal{A}_{\mathcal{X}} \\ 0 & \text{otherwise} \end{cases}, \quad (7)$$

and then we construct a (σ, ι) network $f : [0, 1] \rightarrow \mathbb{R}^2$ of width 2 such that $f(x)_1 = x$ and $|f(x)_2 - h(x)| < 1/2N$ except for the neighborhood of each $z_i \in \mathcal{X}$. Since f is a continuous function, one can observe that such f satisfies the conditions listed in Lemma 18.

We first construct h in Eq. (7) as follows: $h = h_{m+1}$ where $h_{m+1}(x)$ is recursively defined as

$$h_1(x) = \text{STEP}(x - z_{m+1}) \quad h_\ell(x) = \text{STEP}(x - z_{m-\ell+2} + (z_{m-\ell+2} - z_{m+\ell})h_{\ell-1}(x)). \quad (8)$$

From Eq. (8),

$$h_\ell(x) = \begin{cases} \text{STEP}(x - z_{m-\ell+2}) & h_{\ell-1}(x) = 0 \\ \text{STEP}(x - z_{m+\ell}) & h_{\ell-1}(x) = 1 \end{cases}$$

for any $\ell \in \{2, \dots, m+1\}$. One can observe that h_ℓ forms additional breakpoints $z_{m-\ell+2}$ and $z_{m+\ell}$, and for any $x \in [z_i, z_{i+1})$ where $i \in \{m-\ell+2, \dots, m+\ell\}$, the values of $h_\ell(x)$ alternates with 0 and 1 as ℓ increases. Hence, $h_\ell(x)$ in Eq. (8) can be rewritten by

$$h_\ell(x) = \begin{cases} 1 & x \in [z_{m-\ell+2k}, z_{m-\ell+2k+1}), \forall k \in [\ell-1] \text{ or } x \geq z_{m+\ell} \\ 0 & \text{otherwise} \end{cases}$$

for any $\ell \in \{2, \dots, m+1\}$, which implies that h_{m+1} is equal to h in Eq. (7).

We now construct a (σ, ι) network f of width 2 based on h . It suffices to show that for any $\varepsilon > 0$ and $\ell \in [m+1]$ there exists a (σ, ι) network $f_\ell : [0, 1] \rightarrow \mathbb{R}^2$ of width 2 such that

- C1. $f_\ell(x)_1 = x$ on $[0, 1]$,
- C2. $|f_\ell(x)_2 - h_\ell(x)| < \varepsilon$ for $x \in \mathcal{D}_{\mathcal{X}, \gamma}$,
- C3. $f_\ell([0, 1]) \subset [0, 1]^2$.

Then, choosing $f = f_{m+1}$ with $\varepsilon < 1/(2N)$ completes the proof: C1 and C3 directly imply the first and third conditions of Lemma 18, respectively, and C2 guarantees that f_{m+1} satisfies the second condition of Lemma 18 from the definition of $\mathcal{D}_{\mathcal{X}, \gamma}$ and h . We prove this via mathematical induction on ℓ . We first consider the base case, $\ell = 1$. Since σ satisfies Condition 2, there exists a σ network ρ of width 1 such that

$$|\rho(x) - \text{STEP}(x)| < \varepsilon$$

for all $x \in [0, 1] \setminus (-\gamma, \gamma)$ and $\rho([0, 1]) \subset [0, 1]$. Then, we construct a (σ, ι) network $f^{(1)} : [0, 1] \rightarrow \mathbb{R}^2$ of width 2 as

$$f_1(x)_1 = x, \quad f_1(x)_2 = \rho(x - z_{m+1}).$$

Then, one can easily observe that $f^{(1)}$ satisfies C1–3. We now consider the general case, $\ell \geq 2$. From the induction hypothesis, for any $\delta > 0$, there exists a (σ, ι) network $f_{\ell-1} : [0, 1] \rightarrow \mathbb{R}^2$ of width 2 such that $f_{\ell-1}(x)_1 = x$, $|f_{\ell-1}(x)_2 - h_{\ell-1}(x)| < \delta$ and $f_{\ell-1}([0, 1]) \subset [0, 1]^2$. Since σ satisfies Condition 2, for any compact set \mathcal{C} , there exists a σ network ρ of width 1 such that

$$|\rho(x) - \text{STEP}(x)| < \varepsilon/2$$

for all $\mathcal{C} \setminus (-\gamma, \gamma)$. We now construct $f_\ell : [0, 1] \rightarrow \mathbb{R}^2$ as

$$f_\ell(x)_1 = f_{\ell-1}(x)_1, \quad f_\ell(x)_2 = \rho(f_{\ell-1}(x)_1 - z_{m-\ell+2} + (z_{m-\ell+2} - z_{m+\ell})f_{\ell-1}(x)_2)$$

Here, by the induction hypothesis, $f_{\ell-1}(x)_1 = x$. Thus, we can simplify this to

$$f_\ell(x)_1 = x, \quad f_\ell(x)_2 = \rho(x - z_{m-\ell+2} + (z_{m-\ell+2} - z_{m+\ell})f_{\ell-1}(x)_2)$$

which is just the substitution of STEP in Eq. (8) by ρ . Here, one can observe that f_ℓ satisfies C1. Then, for any $x \in \mathcal{D}_{\mathcal{X}, \gamma}$

$$\begin{aligned} & |f_\ell(x)_2 - h_\ell(x)| \\ & \leq |\rho(x - z_{m-\ell+2} + (z_{m-\ell+2} - z_{m+\ell})f_{\ell-1}(x)_2) - \text{STEP}(x - z_{m-\ell+2} + (z_{m-\ell+2} - z_{m+\ell})h_{\ell-1}(x))| \\ & \leq |\rho(x - z_{m-\ell+2} + (z_{m-\ell+2} - z_{m+\ell})f_{\ell-1}(x)_2) - \rho(x - z_{m-\ell+2} + (z_{m-\ell+2} - z_{m+\ell})h_{\ell-1}(x))| \\ & \quad + |\rho(x - z_{m-\ell+2} + (z_{m-\ell+2} - z_{m+\ell})h_{\ell-1}(x)) - \text{STEP}(x - z_{m-\ell+2} + (z_{m-\ell+2} - z_{m+\ell})h_{\ell-1}(x))|. \end{aligned} \quad (9)$$

Here, we note that the second term of Eq. (9) is bounded by $\varepsilon/2$ since

$$x - z_{m-\ell+2} + (z_{m-\ell+2} - z_{m+\ell})h_{\ell-1}(x) \notin (-\gamma, \gamma)$$

for all $x \in \mathcal{D}_{\mathcal{X}, \gamma}$; since $h_{\ell-1}(x) = 0$ or 1 , then $x - z_{m-\ell+2} + (z_{m-\ell+2} - z_{m+\ell})h_{\ell-1}(x) = x - z_{m-\ell+2}$ or $x - z_{m-\ell}$. Hence, we have

$$|f_{\ell}(x)_2 - h_{\ell}(x)| \leq \omega_{\rho}(|(z_{m-\ell+2} - z_{m-\ell})(f_{\ell-1}(x)_2 - h_{\ell-1}(x))|) + \varepsilon/2 < \omega_{\rho}(|(z_{m-\ell+2} - z_{m-\ell})\delta|) + \varepsilon/2 < \varepsilon$$

by choosing sufficiently small $\delta > 0$. It implies that f_{ℓ} follows C2. Lastly, we can easily observe that $f_{\ell}(x)_1 = x \in [0, 1]$ and $f_{\ell}(x)_2 \in [0, 1]$ since $\rho(x) \in [0, 1]$ for all $x \in \mathcal{C}$. It implies that f_{ℓ} satisfies C3 and this completes the proof.

E. Proof of Lemma 12

In this section, we prove Lemma 12. To this end, without loss of generality, assume that $\mathcal{K} \subset [0, \infty)$ and $c_1, \dots, c_N \in (0, 1)$; if there exists c_i such that $c_i = 0$ or $c_i = 1$, then we can substitute $c_i^* = \varepsilon/2$ or $1 - \varepsilon/2$ respectively and approximate them within error $\varepsilon/2$. Let $\xi = \text{dist}(\{c_1, \dots, c_k\}, \{0, 1\})$. Then, one can observe that $\xi > 0$. In addition, we assume that for any $i \in [N-1]$, $x < y$ for all $x \in \mathcal{I}_i$ and $y \in \mathcal{I}_{i+1}$. Since \mathcal{I}_i 's are disjoint, for any $i \in [N-1]$, there exists $x^{(i)} \in \mathbb{R}$ such that $\sup \mathcal{I}_i < x^{(i)} < \inf \mathcal{I}_{i+1}$. Let $x^{(0)} = \min \mathcal{K}$, $x^{(N)} = \max \mathcal{K}$ and

$$\gamma = \min_{i \in [N-1]} \left\{ \text{dist}(x^{(i)}, \mathcal{I}_i), \text{dist}(x^{(i)}, \mathcal{I}_{i+1}) \right\}. \quad (10)$$

In this proof, we construct a (σ, ι) network $f : \mathcal{K} \rightarrow [0, 1]$ of width 2 such that for any $k \in [N]$,

$$\sup_{x \in \mathcal{I}_k} |f(x) - c_k| \leq \eta$$

where $\eta := \min\{\xi, \varepsilon\}$.

To this end, we construct two (σ, ι) networks $h_1 : \mathcal{K} \rightarrow \mathbb{R}$ and $h_2 : \mathbb{R} \rightarrow \mathbb{R}$ of width 2 such that

C1. for each $k \in [N]$, $\sup_{x \in \mathcal{I}_k} |h_1(x) - c_k| \leq \eta/2$,

C2. for any $x \in \bigcup_{k=1}^N \mathcal{I}_k$, $|h_2 \circ h_1(x) - h_1(x)| \leq \eta/2$ and $h_2 \circ h_1(\mathcal{K}) \subset [0, 1]$.

Then, one can observe that h_1 maps input x to near the corresponding c_k if $x \in \mathcal{I}_k$, and h_2 bounds the codomain of h_1 while the approximation for piecewise constant is preserved. If we choose $f = h_2 \circ h_1$, then such f satisfies the desired conditions.

We first construct h_1 satisfying C1 using the property of σ that can approximate STEP. To this end, we consider a (STEP, ι) network $g : \mathcal{K} \rightarrow \mathbb{R}$ of width 2 approximating the given piecewise constant function, and then we construct a (σ, ι) network h_1 of width 2 approximating g in $\bigcup_{k=1}^N \mathcal{I}_k$.

We now construct a (STEP, ι) network g approximating piecewise constant function. To construct such g , we compose (STEP, ι) networks $g_1, \dots, g_N : \mathbb{R} \rightarrow \mathbb{R}$ of width 2 such that each g_i shifts x by a sufficiently large length $L_i > 0$ if $x \in [x^{(i-1)}, x^{(i)}]$. Here, for each $i \in [N]$, L_i is defined as $a \times (c_i + b)$ where $a > \max\{1, 4x^{(N)}/\eta\}$ and $b = x^{(N)} - \min_{i \in [N]} c_i$ which implies that each $g_i(x) = x + L_i > x^{(N)}$ for $x \in [x^{(i-1)}, x^{(i)}]$. i.e., we construct each g_i such that

$$g_i \circ \dots \circ g_1(x) = \begin{cases} x + a \times (c_1 + b) & x \in [x^{(0)}, x^{(1)}] \\ x + a \times (c_2 + b) & x \in [x^{(1)}, x^{(2)}] \\ \vdots & \\ x + a \times (c_i + b) & x \in [x^{(i-1)}, x^{(i)}] \\ x & \text{otherwise} \end{cases} \quad (11)$$

for all $i \in [N]$. Then, we define g as follows: $g = g_{\text{cut}} \circ g_N \circ g_{N-1} \circ \dots \circ g_1$ where $g_{\text{cut}} : \mathbb{R} \rightarrow \mathbb{R}$ is defined as

$$g_{\text{cut}}(x) = \frac{1}{a}x - b.$$

Then, one can easily observe that

$$g(x) = \begin{cases} c_1 + \frac{x}{a} & x \in [x^{(0)}, x^{(1)}] \\ c_2 + \frac{x}{a} & x \in [x^{(1)}, x^{(2)}] \\ \vdots \\ c_N + \frac{x}{a} & x \in [x^{(N-1)}, x^{(N)}]. \end{cases}$$

Since $a > 4x^{(N)}/\eta$, it holds that $|x/a| < \eta/4$ for all $x \in \mathcal{K}$. Thus, g approximates the piecewise constant function within an error $\eta/4$.

We now construct (STEP, ι) networks g_1, \dots, g_N satisfying Eq. (11). For each $i \in [N]$, we define $g_i : \mathbb{R} \rightarrow \mathbb{R}$ as

$$g_i(x) = x + a \times (c_i + b) \text{STEP}(-(x - x^{(i)})).$$

One can observe that g_i shifts x by $a \times (c_i + b)$ if $x < x^{(i)}$. Here, we note that since $a \times (c_i + b) > x^{(N)}$, the values shifted by g_i for some $i \in [N]$ are not shifted again, resulting that g_i shifts only $x \in [x^{(i-1)}, x^{(i)}]$. Thus, our g can approximate a given piecewise function within an error $\eta/2$.

We now construct a (σ, ι) network h_1 of width 2 approximating g on $\bigcup_{i \in [N]} \mathcal{I}_i$. Since σ is squashable, then for any compact set \mathcal{C} and $\alpha > 0$, there exists a σ network $\rho : \mathbb{R} \rightarrow \mathbb{R}$ such that ρ is increasing on \mathcal{C} , $\rho(\mathcal{C}) \subset [0, 1]$, and

$$|\rho(x) - \text{STEP}(x)| < \alpha$$

for all $x \in \mathcal{C} \setminus (-\beta, \beta)$ where $0 < \beta < \min\{\gamma, \xi\}$ (Eq. (10)). We will give an explicit value to α later. We now construct a (σ, ι) network h_1 of width 2 as follows:

$$h_1 = g_{\text{cut}} \circ f_N \circ \dots \circ f_1 \text{ where} \\ f_i(x) = x + a \times (c_i + b) \rho(-(x - x^{(i)})) \quad \forall i \in [N].$$

Then, one can observe that $|f_i(x) - g_i(x)| = |c_i + b| |\rho(-(x - x^{(i)})) - \text{STEP}(-(x - x^{(i)}))| < |c_i + b| \alpha$ for all $x \in \mathcal{C} \setminus \mathcal{B}_\beta(x^{(i)})$ and $i \in [N]$. For the notational simplicity, we denote $\delta_i = |c_i + b| \alpha$. We note that for any $i, j \in [N]$ and $x \in \mathcal{I}_i$,

$$g_j \circ \dots \circ g_1(x) \notin \mathcal{B}_\beta(x^{(i)}). \quad (12)$$

Eq. (12) holds since $g_j \circ \dots \circ g_1$ maps x to $x \in \mathcal{I}_i$, or a value out of \mathcal{K} ($x + a \times (c_i + b) > x + x^{(N)}$ from the definition of a and b) and $\beta < \gamma$. Then, for any $i \in [N]$ and $x \in \mathcal{I}_i$, it holds that

$$\begin{aligned} |h_1(x) - g(x)| &= |g_{\text{cut}} \circ f_N \circ \dots \circ f_1(x) - g_{\text{cut}} \circ g_N \circ \dots \circ g_1(x)| \\ &\leq \omega_{g_{\text{cut}}}(|f_N \circ \dots \circ f_1(x) - g_N \circ \dots \circ g_1(x)|) \\ &\leq \omega_{g_{\text{cut}}}(|f_N \circ \dots \circ f_1(x) - f_N \circ g_{N-1} \circ \dots \circ g_1(x) + f_N \circ g_{N-1} \circ \dots \circ g_1(x) - g_N \circ \dots \circ g_1(x)|) \\ &\leq \omega_{g_{\text{cut}}}(\omega_{f_N}(|f_{N-1} \circ \dots \circ f_1(x) - g_{N-1} \circ \dots \circ g_1(x)|) + |f_N \circ g_{N-1} \circ \dots \circ g_1(x) - g_N \circ \dots \circ g_1(x)|) \end{aligned}$$

Here, $|f_N \circ g_{N-1} \circ \dots \circ g_1(x) - g_N \circ \dots \circ g_1(x)| < \delta_N$ from Eq. (12). Thus, by conducting this procedure iteratively, we have

$$\begin{aligned} |h_1(x) - g(x)| &\leq |\omega_{g_{\text{cut}}}(\omega_{f_N}(f_{N-1} \circ \dots \circ f_1(x) - g_{N-1} \circ \dots \circ g_1(x)) + \delta_N)| \\ &\leq |\omega_{g_{\text{cut}}}(\omega_{f_N}(\omega_{f_{N-1}}(f_{N-2} \circ \dots \circ f_1(x) - g_{N-2} \circ \dots \circ g_1(x)) + \delta_{N-1}) + \delta_N)| \\ &\vdots \\ &\leq |\omega_{g_{\text{cut}}}(\omega_{f_N}(\dots(\omega_{f_2}(\delta_1) + \delta_2) \dots) + \delta_N)| < \eta/4 \end{aligned}$$

by choosing sufficiently small $\alpha > 0$, which leads us to have sufficiently small δ_i for all $i \in [N]$. Consequently, for any $i \in [N]$ and $x \in \mathcal{I}_i$, we have

$$|h_1(x) - c_i| \leq |h_1(x) - g(x)| + |g(x) - c_i| < \eta/4 + \eta/4 = \eta/2. \quad (13)$$

Hence, our h_1 satisfies C1.

We now construct h_2 satisfying C2. We suppose that there exists $u \in \mathcal{K}$ such that $h_1(u) < 0$; we will discuss the case that there exists $v \in \mathcal{K}$ such that $h_1(v) > 1$ later. To this end, we consider a (σ, ι) network of width 2 that iteratively adds some constant to the region such that $h_1(x) < 0$. Namely, it suffices to show that for any $\varepsilon' > 0$, there exists a (σ, ι) network $\psi : \mathbb{R} \rightarrow \mathbb{R}$ of width 2 such that

- if $x \geq \eta/4$, then $|\psi(x) - x| \leq \varepsilon'$
- if $x \in (0, \eta/4)$, then $\psi(x) \in [0, 1]$, and
- if $x \leq 0$, then $\psi(x) - x \geq 1/2$.

Then, let $h_2 = \psi^{N_1}$ for some $N_1 \in \mathbb{N}$ such that $N_1/2 > |\min_{x \in \mathcal{K}} h_1(x)|$, then we obtain

$$h_2 \circ h_1(x) = \psi^{N_1} \circ h_1(x) \in [0, 1]$$

for all $x \in \mathcal{K}$ such that $h_1(x) < 0$.

Furthermore, since $\eta \leq \xi$ and h_1 satisfies C1, $h_1(x) \geq \eta/2$ for any $x \in \bigcup_{i=1}^N \mathcal{I}_i$. Thus, if we choose sufficiently small $\varepsilon' > 0$ such that $\varepsilon' < \eta/(4N_1)$, then

$$\begin{aligned} |h_2 \circ h_1(x) - h_1(x)| &= |\psi^{N_1} \circ h_1(x) - h_1(x)| \\ &\leq |\psi^{N_1} \circ h_1(x) - \psi^{N_1-1} \circ h_1(x)| + \dots + |\psi \circ h_1(x) - h_1(x)| \\ &\leq N_1 \varepsilon' \leq \eta/4 \leq \eta/2. \end{aligned}$$

Here, for each $i \in [N_1 - 1]$, $\psi^i \circ h_1(x) \geq \eta/4$ since $\psi^i \circ h_1(x) \subset \mathcal{B}_{i\varepsilon'}(h_1(x))$ and $h_1(x) \geq \eta/2$. It guarantees that $|\psi^i \circ h_1(x) - \psi^{i-1} \circ h_1(x)| \leq \varepsilon'$ for each $i \in [N_1]$. Hence, h_2 satisfies C2. If there exists $v \in \mathcal{K}$ such that $h_1(v) > 1$, then the same argument can be applied with the choice of $\psi_1(x) = 1 - \psi(1 - x)$.

We now construct such ψ using the property that σ network can approximate STEP. Since σ is squashable, for any $\delta' > 0$ and a compact set \mathcal{D} , there exists a σ network $\rho^* : \mathcal{D} \rightarrow \mathbb{R}$ such that

$$|\rho^*(x) - \text{STEP}(x)| < \delta'$$

for all $x \in \mathcal{D} \setminus (-\eta/8, \eta/8)$. We choose $\delta' > 0$ such that $\delta' \leq \min\{3\varepsilon'/2, 1/4\}$. Consider a (σ, ι) network ψ of width 2 defined as

$$\psi(x) = x + \frac{2}{3}\rho^*(-(x - \eta/8)).$$

Then, one can easily observe that $|\psi(x) - x| < 2\delta'/3 \leq \varepsilon'$ if $x \geq \eta/4$, $\psi(x) \in (0, \eta/4 + 2/3) \subset [0, 1]$ if $x \in (0, \eta/4)$, and $|\psi(x) - (x + 2/3)| < 2\delta'/3 \leq 1/6$ if $x \leq 0$ which implies $\psi(x) - x \geq 1/2$. It completes the proof.

F. Proof of Lemma 13

In this section, we prove Lemma 13. To this end, we construct a (σ, ι) network f of width 2 that maps for each $\mathcal{S} \in \mathcal{G}$ to a disjoint interval. Then, since f is continuous, $\{f(\mathcal{S}) : \mathcal{S} \in \mathcal{G}\}$ is a 1-grid of size $n_1 n_2$ and this completes the proof. Before we illustrate our proof, we define the additional notation used in this proof. Since \mathcal{G} is a 2-grid of size (n_1, n_2) , there exist compact intervals $[a_1, b_1], \dots, [a_{n_1}, b_{n_1}], [a'_1, b'_1], \dots, [a'_{n_2}, b'_{n_2}]$ satisfying the following:

- $b_i < a_{i+1}$ and $b'_j < a'_{j+1}$ for each $i \in [n_1 - 1]$ and $j \in [n_2 - 1]$, respectively,
- for any $\mathcal{S} \in \mathcal{G}$, there uniquely exist $i \in [n_1]$ and $j \in [n_2]$ such that $\mathcal{S} = [a_i, b_i] \times [a'_j, b'_j]$.

For each $i \in [n_1]$ and $j \in [n_2]$, let $\mathcal{U}_{ij} = [a_i, b_i] \times [a'_j, b'_j]$, $\mathcal{V}_i = \bigcup_j \mathcal{U}_{ij}$,

$$\eta = \min_{j \in [n_2-1]} \{a'_{j+1} - b'_j\}$$

and $L = b'_m - a'_1$. We write $e_1 = (1, 0)$ and $e_2 = (0, 1) \in \mathbb{R}^2$. For $i \in \{1, 2\}$ and $b \in \mathbb{R}$, we use $\mathcal{H}(e_i, b) \triangleq \{x \in \mathbb{R}^2 | x_i + b = 0\}$. We first consider a (σ, ι) network $h_1 : \mathcal{K} \rightarrow \mathbb{R}^2$ of width 2 defined as

$$h_1(x)_1 = \rho(x_1 - c_1), \quad h_1(x)_2 = \iota(x_2)$$

where $c_1 \in (b_{n-1}, a_n)$ and ρ is a σ network of width 1 such that $|\rho(x) - \text{STEP}(x)| < \zeta$ on $[a_1, b_n]$. We will assign an explicit value to ζ . Then, one can observe that

$$h_1(\mathcal{V}_n) \subset \mathcal{B}_\zeta(\mathcal{H}(e_1, -1)), \quad h_1(\mathcal{V}_i) \subset \mathcal{B}_\zeta(\mathcal{H}(e_1, 0)) \text{ for all } i \in [n-1]. \quad (14)$$

Furthermore, since $h_1(x)_1$ is strictly increasing on \mathcal{K} , the ordering of \mathcal{V}_i 's with respect to the first coordinate is preserved: if $i < j$, then $x_1 < y_1$ for all $x \in \mathcal{V}_i, y \in \mathcal{V}_j$. We then iteratively apply some (σ, ι) networks h_2, \dots, h_{n_1} so that for each $i \in [n_1]$, h_i maps \mathcal{V}_{n_1-i+1} to $\mathcal{B}_\zeta(\mathcal{H}(e_1, -1))$ and shifts \mathcal{V}_{n_1-i+1} by sufficiently large length such that the images of \mathcal{V}_i are disjoint for the second coordinate.

We now formally construct such (σ, ι) networks h_2, \dots, h_{n_1} . See the following lemma where the proof is deferred to Appendix F.1.

Lemma 19. *Let $\xi > 0$ and $r > 0$. Let $\mathcal{X}_0 \subset \mathcal{B}_\xi(\mathcal{H}(e_1, 0))$, $\mathcal{X}_1 \subset \mathcal{B}_\xi(\mathcal{H}(e_1, -1))$, and $\mathcal{Y} \subset \mathcal{B}_\xi(\mathcal{H}(e_1, 0))$ be compact sets in \mathbb{R}^2 such that $y_1 > x_1$ for all $x \in \mathcal{X}_0$ and $y \in \mathcal{Y}$. Then, there exists a (σ, ι) network $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ of width 2 satisfying the following properties:*

- for any $x \in \mathcal{X}_0 \cup \mathcal{X}_1$, $|f(x)_2 - x_2| < 2r\xi$,
- for any $y \in \mathcal{Y}$, $|f(y)_2 - (y_2 + r)| < 2r\xi$,
- $f(\mathcal{X}_0) \subset \mathcal{B}_\xi(\mathcal{H}(e_1, 0))$ and $f(\mathcal{X}_1), f(\mathcal{Y}) \subset \mathcal{B}_\xi(\mathcal{H}(e_1, -1))$,
- there exists strictly increasing $\phi : \mathbb{R} \rightarrow \mathbb{R}$ such that $f(x)_1 = \phi(x_1)$ for all $x \in \mathcal{X}_0$.

Lemma 19 implies that there exists a (σ, ι) network of width 2 that maps \mathcal{Y} to in $\mathcal{H}(e_1, -1)$ with approximately shift for the second coordinate by r . From Eq. (14), we can apply Lemma 19 with

$$\mathcal{X}_0 = \bigcup_{i \in [n_1-2]} h_1(\mathcal{V}_i), \quad \mathcal{X}_1 = h_1(\mathcal{V}_{n_1}), \quad \mathcal{Y} = h_1(\mathcal{V}_{n_1-1}),$$

$r = L + 1$ and $\xi = \zeta$. Then, there exists a (σ, ι) network h_2 of width 2 that maps the points of \mathcal{X}_0 and \mathcal{X}_1 approximately identically while shifting the second coordinate of \mathcal{Y} by $L + 1$. Here, one can observe that if we choose a sufficiently small $\zeta > 0$, then $h_2(h_1(\mathcal{V}_{n_1}))$ and $h_2(h_1(\mathcal{V}_{n_1-1}))$ are disjoint for the second coordinate by our choice of r . Furthermore, from the third and fourth lines of the properties listed in Lemma 19, Lemma 19 can be applied iteratively with the recursive choice of $\mathcal{X}_0, \mathcal{X}_1, \mathcal{Y}, r$ and ξ in Lemma 19. In particular, by the fourth line of the properties from Lemma 19, the ordering of \mathcal{V}_i 's with respect to the first coordinate is preserved while Lemma 19 is applied. Thus, among the sets contained in \mathcal{X}_0 , we can choose \mathcal{Y} as the set that is the highest with respect to the first coordinate.

We now construct such (σ, ι) networks $h_2, \dots, h_{n_1} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ as follows: for each $k \in [n_1] \setminus \{1\}$, h_k is from Lemma 19 with the choices of

- $\mathcal{X}_0 = \bigcup_{i \in [n_1-k]} h_{k-1} \circ \dots \circ h_1(\mathcal{V}_i)$,
- $\mathcal{X}_1 = \bigcup_{i \in [k-1]} h_{k-1} \circ \dots \circ h_1(\mathcal{V}_{n_1-i+1})$,
- $\mathcal{Y} = h_{k-1} \circ \dots \circ h_1(\mathcal{V}_{n_1-k+1})$,
- $r = r_k$ where $r_k = (k-1)(L+1)$ and $\xi = \zeta$.

Then, we construct a (σ, ι) network $f : \mathcal{K} \rightarrow \mathbb{R}$ of width 2 as

$$f(x) = p \circ h_{n_1} \circ \dots \circ h_1(x) \quad (15)$$

where $p : \mathbb{R}^2 \rightarrow \mathbb{R}$ is a projection onto the second coordinate: $p(x, y) = y$. We now prove that if we choose sufficiently small $\zeta > 0$ such that

$$\sum_{k=2}^{n_1} 2r_k \zeta < \min \left\{ \frac{\eta}{2}, \frac{1}{2} \right\},$$

then for each $i \in [n_1]$ and $j \in [n_2]$, $f(\mathcal{U}_{ij})$ is disjoint.

We first show that if $i, j \in [n_1]$ such that $i < j$, then $f(x) > f(y)$ for all $x \in \mathcal{V}_i$ and $y \in \mathcal{V}_j$, and then we prove that for each $i \in [n_1]$, if $j, j' \in [n_2]$ such that $j < j'$, then $f(x) < f(y)$ for all $x \in \mathcal{U}_{ij}$ and $y \in \mathcal{U}_{ij'}$.

We first consider $x \in \mathcal{V}_i$ and $y \in \mathcal{V}_j$. From our definition of f (Eq. (15)) and Lemma 19, one can observe that

$$|f(x) - (x_2 + r_{n_1-i+1})| < \sum_{k=2}^{n_1} 2r_k \zeta \leq \frac{1}{2}, \quad |f(y) - (y_2 + r_{n_1-j+1})| < \sum_{k=2}^{n_1} 2r_k \zeta \leq \frac{1}{2}.$$

Since $r_{n_1-i+1} - r_{n_1-j+1} \geq L+1$, the above equation implies that

$$f(x) - f(y) > r_{n_1-i+1} - r_{n_1-j+1} - (y_2 - x_2) - 1 \geq L+1 - L-1 = 0$$

We now consider $x \in \mathcal{U}_{ij}$ and $y \in \mathcal{U}_{ij'}$. As in above, we have

$$|f(x) - (x_2 + r_{n_1-i+1})| < \sum_{k=2}^{n_1} 2r_k \zeta \leq \frac{\eta}{2}, \quad |f(y) - (y_2 + r_{n_1-i+1})| < \sum_{k=2}^{n_1} 2r_k \zeta \leq \frac{\eta}{2}.$$

Since $y_2 - x_2 > \eta$ by the definition of η , we have

$$f(y) - f(x) > y_2 - x_2 - \eta > 0$$

and this completes the proof.

F.1. Proof of Lemma 19

In this section, we prove Lemma 19. Let $b \in \mathbb{R}$ such that $x_1 < b < y_1$ for all $x = (x_1, x_2) \in \mathcal{X}_0$ and $y = (y_1, y_2) \in \mathcal{Y}$ and

$$\eta = \min\{y_1 - b, b - x_1 | y \in \mathcal{Y}, x \in \mathcal{X}_0\}.$$

We note that such b is well-defined and $\eta > 0$ because $x_1 < y_1$ for all $x \in \mathcal{X}_0, y \in \mathcal{Y}$ and $\mathcal{X}_0, \mathcal{Y}$ are compact. Since σ is squashable, for any compact set \mathcal{K} , there exists a σ network $\rho : \mathbb{R} \rightarrow \mathbb{R}$ of width 1 such that

$$|\rho(x) - \text{STEP}(x)| < \xi$$

for all $x \in \mathcal{K} \setminus (-\eta, \eta)$. Let $A = \begin{bmatrix} 1 & 0 \\ -r & 1 \end{bmatrix}$. Then, one can easily observe that $A^{-1} = \begin{bmatrix} 1 & 0 \\ r & 1 \end{bmatrix}$.

We now define functions $f_1, f_2, f_3 : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ as

$$f_1(x) = Ax, \quad f_2(x) = (\rho(x_1 - b), \iota(x_2)), \quad f_3(x) = A^{-1}x$$

for all $x = (x_1, \dots, x_n) \in \mathbb{R}^n$, respectively. We now define a function $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ as

$$f(x) = (f_3 \circ f_2 \circ f_1)(x)$$

for all $x \in \mathbb{R}^n$. Then, f is a (σ, ι) network of width 2 and

$$f(x)_1 = \rho(x_1 - b), \quad f(x)_2 = x_2 + r(\rho(x_1 - b) - x_1). \quad (16)$$

We now show that our f satisfies the properties listed in Lemma 19. One can easily observe that f satisfies the fourth property of Lemma 19. Thus, we consider the first–third properties. From Eq. (16), we can classify the image regions corresponding to each input region.

We first consider $x \in \mathcal{X}_0$. Since $\mathcal{X}_0 \subset \mathcal{B}_\xi(\mathcal{H}(e_1, 0))$ and $x_1 < b - \eta$, we have $x_1 \in (-\xi, \xi)$ and $\rho(x_1 - b) \in (0, \xi)$. Thus, it holds that $f(x)_1 \in \mathcal{B}_\xi(\mathcal{H}(e_1, -1))$ and $|f(x)_2 - x_2| < 2r\xi$. We now consider $x \in \mathcal{X}_1$. Since $\mathcal{X}_1 \subset \mathcal{B}_\xi(\mathcal{H}(e_1, -1))$ and $x_1 > b + \eta$, we have $x_1 \in (1 - \xi, 1 + \xi)$ and $\rho(x_1 - b) \in (1 - \xi, 1)$. Thus, $f(x)_1 \in \mathcal{B}_\xi(\mathcal{H}(e_1, 0))$ and $|f(x)_2 - x_2| < 2r\xi$. Lastly, let $y \in \mathcal{Y}$. Since $\mathcal{Y}_1 \subset \mathcal{B}_\xi(\mathcal{H}(e_1, 0))$ and $y_1 > b + \eta$, we have $y_1 \in (-\xi, \xi)$ and $\rho(y_1 - b) \in (1 - \xi, 1)$. Thus, $f(y)_1 \in \mathcal{B}_\xi(\mathcal{H}(e_1, -1))$ and $|f(y)_2 - (y_2 + r)| < 2r\xi$. Conclusively, f satisfies all properties listed in Lemma 19 and this completes the proof.