Three Quantization Regimes for ReLU Networks

Weigutian Ou
wou@mins.ee.ethz.ch

Philipp Schenkel
schenkel@fzi.de

Helmut Bölcskei
hboelcskei@ethz.ch

Abstract. We establish the fundamental limits in the approximation of Lipschitz functions by deep ReLU neural networks with finite-precision weights. Specifically, three regimes, namely under-, over-, and proper quantization, in terms of minimax approximation error behavior as a function of network weight precision, are identified. This is accomplished by deriving nonasymptotic tight lower and upper bounds on the minimax approximation error. Notably, in the proper-quantization regime, neural networks exhibit memory-optimality in the approximation of Lipschitz functions. Deep networks have an inherent advantage over shallow networks in achieving memory-optimality. We also develop the notion of depth-precision tradeoff, showing that networks with high-precision weights can be converted into functionally equivalent deeper networks with low-precision weights, while preserving memory-optimality. This idea is reminiscent of sigma-delta analog-to-digital conversion, where oversampling rate is traded for resolution in the quantization of signal samples. We improve upon the best-known ReLU network approximation results for Lipschitz functions and describe a refinement of the bit extraction technique which could be of independent general interest.

1 Introduction

This paper is concerned with the fundamental limits in the approximation of Lipschitz functions by deep ReLU neural networks with finite-precision weights. Specifically, we consider fully connected networks and allow depth and width to be chosen independently. The main conceptual contribution resides in the identification of three different regimes in terms of minimax approximation error behavior as a function of the network weight precision, i.e., the number of bits b needed to store each of the network weights. This is accomplished by deriving nonasymptotic and, in particular, tight lower and upper bounds on the minimax error. In the under-quantization regime the minimax error exhibits exponential decay in b, in the proper-quantization regime the decay is polynomial, and in the over-quantization regime we get constant behavior. Notably, in the proper-quantization regime, neural networks approximate Lipschitz functions in a memory-optimal fashion. In addition, deep networks are found to exhibit an inherent advantage over shallow networks in achieving memory-optimality.

H. Bölcskei gratefully acknowledges support by the Lagrange Mathematics and Computing Research Center, Paris, France.

Besides the conceptual contribution of identifying the three quantization regimes, we report three technical contributions. First, we develop the notion of depth-precision tradeoff, showing that networks with high-precision weights can be converted into equivalent (in terms of input-output relation) deeper networks with low-precision weights, while preserving memory-optimality. The underlying network transformation is constructive. This idea is reminiscent of the concept of sigma-delta analog-to-digital conversion [1], where sampling rate is traded for resolution in the signal samples. Here, we trade network depth for network weight resolution.

The second technical contribution is an improvement of the best-known neural network approximation results for 1-Lipschitz functions on $[0,1]$. Specifically, for networks of sufficiently large, but otherwise arbitrary, network width W and depth L, as well as weight magnitude bounded by 1, we show that the minimax approximation error behaves according to $C(W^2L^2\log(W))^{-1}$, with C an absolute constant. There is a significant body of literature on neural network approximation of 1-Lipschitz functions on $[0,1]$, usually presented in the broader context of approximation of smooth functions on hypercubes. Specifically, the references [2, 3, 4, 5, 6, 7] consider approximation with ReLU networks whose depth grows at most poly-logarithmically in network width, hence, in contrast to our results, with depth and width coupled. On the other hand, the findings reported in [8, 9, 10] allow network depth to grow faster than network width. Notably, in [9] network width and depth can be chosen independently and the same approximation error behavior as in our case, namely $C(W^2L^2\log(W))^{-1}$, is obtained. However, the network constructions proposed in [8, 9, 10] all come with weight-magnitude growth that is at least exponential in network depth, in contrast to the weight-magnitude upper-bounded by 1 in our case. This constant weight-magnitude upper bound will turn out to be essential in establishing memory optimality.

The third technical contribution we report is an improvement of the bit extraction technique pioneered in [11, 12]. Bit extraction refers to the recovery—through ReLU networks—of binary strings encoded into real numbers. This concept was originally used to lower-bound the VC-dimension of ReLU networks [11, 12] and later employed in the context of neural network approximation [8, 9, 10, 13]. The legacy approach yields extraction networks whose weight magnitude grows exponentially in network depth and polynomially in network width. In contrast, the novel construction we present exhibits only (polynomial) weight magnitude dependence on network width. While this refinement is essential in establishing the second technical contribution mentioned above, the technique could also be of independent interest.

We finally note that the results in this paper are readily extended to neural network approximation of Lipschitz functions on d-dimensional hypercubes. For clarity of exposition we decided, however, to restrict ourselves to the one-dimensional case.

Notation. We denote the cardinality of a set X by $|X|$. $\mathbb{N}=\{1,2,\dots\}$ designates the natural numbers, $\mathbb{R}$ stands for the real numbers, $\mathbb{R}_+$ for the positive real numbers, and $\emptyset$ for the empty set. For $\mathbb{A}\subseteq\mathbb{R}$, we denote its maximum, minimum, supremum, and infimum, by $\max\mathbb{A}$, $\min\mathbb{A}$, $\sup\mathbb{A}$, and $\inf\mathbb{A}$, respectively. The indicator function 1_P for proposition P is equal to 1 if P is true and 0 else.

For a vector $b\in\mathbb{R}^d$, we write $\|b\|_\infty:=\max_{i=1,\dots,d}|b_i|$ and $\|b\|_0:=\sum_{i=1}^d 1_{b_i\neq 0}$. Similarly, for a matrix $A\in\mathbb{R}^{m\times n}$, we let $\|A\|_\infty=\max_{i=1,\dots,m,j=1,\dots,n}|A_{i,j}|$ and $\|A\|_0:=\sum_{i=1}^m\sum_{j=1}^n 1_{A_{i,j}\neq 0}$. 1_m and 0_m stand for the m-dimensional vector with all entries equal to 1 and 0, respectively. I_m refers to the $m\times m$ identity matrix. $1_{m\times n}$ and $0_{m\times n}$ denote the $m\times n$ matrix with all entries equal to 1 and 0, respectively. For matrices $A_1,\dots,A_n$, possibly of different dimensions, we designate the block-diagonal matrix with diagonal element-matrices $A_1,\dots,A_n$ by $\mathrm{diag}(A_1,\dots,A_n)$.

$\log(\cdot)$ and $\ln(\cdot)$ denote the logarithm to base 2 and base e, respectively. The ReLU activation function ρ is given by $\rho(x) := \max\{x, 0\}$, for $x \in \mathbb{R}$, and, when applied to vectors, acts elementwise. The sign function $\text{sgn} : \mathbb{R} \mapsto \{0, 1\}$ is defined according to $\text{sgn}(x) = 1$, for $x \geq 0$, and $\text{sgn}(x) = 0$, for $x < 0$. We use $S(A, b)$ to refer to the affine mapping $S(A, b)(x) = Ax + b, x \in \mathbb{R}^{n_2}$, with $A \in \mathbb{R}^{n_1 \times n_2}$, $b \in \mathbb{R}^{n_1}$. For $f_1 : \mathbb{R}^{d_0} \mapsto \mathbb{R}^{d_1}$ and $f_2 : \mathbb{R}^{d_0} \mapsto \mathbb{R}^{d_2}$, we define $(f_1, f_2) : \mathbb{R}^{d_0} \mapsto \mathbb{R}^{d_1+d_2}$ according to $(f_1, f_2)(x) = (f_1(x), f_2(x))$, $x \in \mathbb{R}^{d_0}$. If $\mathcal{F}$ is a family of functions and $a \in \mathbb{R}$, we write $a \cdot \mathcal{F} := \{af : f \in \mathcal{F}\}$. For $\mathbb{X} \subseteq \mathbb{R}^d$ and $f : \mathbb{X} \mapsto \mathbb{R}$, we define the $L^\infty(\mathbb{X})$-norm of f as $\|f\|_{L^\infty(\mathbb{X})} := \sup_{x \in \mathbb{X}} |f(x)|$. A constant will be called absolute if it does not depend on any variables or parameters. We may use the same letter for different absolute constants at different places in the paper.

1.1 Definition of key concepts and organization of the paper

The purpose of this section is to introduce the key concepts needed to formalize the main results of the paper. We start by defining the family of 1-Lipschitz functions on $[0, 1]$ according to

$$H^1([0, 1]) := \{f \in C([0, 1]) : |f(x)| \leq 1, |f(x) - f(y)| \leq |x - y|, \forall x, y \in [0, 1]\}.$$

Next, we provide the definition of neural networks.

Definition 1.1. *Let $L, N_0, N_1, \ldots, N_L \in \mathbb{N}$. A neural network configuration Φ is a sequence of matrix-vector tuples*

$$\Phi = ((A_i, b_i))_{i=1}^L,$$

where $A_i \in \mathbb{R}^{N_i \times N_{i-1}}$, $b_i \in \mathbb{R}^{N_i}$, $i = 1, \ldots, L$. We refer to N_i as the width of the i-th layer, $i = 0, \ldots, L$, and call the tuple $(N_0, \ldots, N_L)$ the architecture of the network configuration. $\mathcal{N}((d, d'))$ refers to the set of all neural network configurations with input dimension $N_0 = d$ and output dimension $N_L = d'$. The depth of the configuration Φ is $\mathcal{L}(\Phi) := L$, its width $\mathcal{W}(\Phi) := \max_{i=0,\ldots,L} N_i$, the weight set $\mathcal{C}(\Phi) := \bigcup_{i=1,\ldots,L}(\mathcal{C}(A_i) \bigcup \mathcal{C}(b_i))$, where $\mathcal{C}(A)$ and $\mathcal{C}(b)$ denote the value set of the entries of A and b, respectively, and the weight magnitude $\mathcal{B}(\Phi) := \max_{i=1,\ldots,L} \max\{\|A_i\|_\infty, \|b_i\|_\infty\}$.

We define, recursively, the neural network realization $R(\Phi) : \mathbb{R}^{N_0} \mapsto \mathbb{R}^{N_L}$, associated with the neural network configuration Φ, and the activation function ρ, according to

$$R(\Phi) = \begin{cases} S(A_L, b_L), & \text{if } L = 1, \\ S(A_L, b_L) \circ \rho \circ R(((A_i, b_i))_{i=1}^{L-1}), & \text{if } L \geq 2. \end{cases} \tag{1}$$

The family of network configurations with depth at most L, width at most W, weight magnitude at most B, where $B \in \mathbb{R}_+ \cup \{\infty\}$, weights taking values in $\mathbb{A} \subseteq \mathbb{R}$, d-dimensional input, and d'-dimensional output, for $d, d' \in \mathbb{N}$, $W, L \in \mathbb{N} \cup \{\infty\}$, with[1] $W \geq \max\{d, d'\}$, is defined as

$$\mathcal{N}_{\mathbb{A}}((d, d'), W, L, B) := \{\Phi \in \mathcal{N}((d, d')) | \mathcal{W}(\Phi) \leq W,\ \mathcal{L}(\Phi) \leq L,\ \mathcal{B}(\Phi) \leq B, \mathcal{C}(\Phi) \subseteq \mathbb{A}\}, \tag{2}$$

with the family of associated network realizations

$$\mathcal{R}_{\mathbb{A}}((d, d'), W, L, B) := \{R(\Phi) | \Phi \in \mathcal{N}_{\mathbb{A}}((d, d'), W, L, B)\}. \tag{3}$$

[1] The condition $W \geq \max\{d, d'\}$ is formally stated here so as to prevent the trivial case of $\mathcal{N}_{\mathbb{A}}((d, d'), W, L, B)$ being an empty set. It will be a standing assumption throughout the paper.

To simplify notation, we allow the omission of the argument (d, d') *in* $\mathcal{N}_{\mathbb{A}}((d, d'), W, L, B)$ *and* $\mathcal{R}_{\mathbb{A}}((d, d'), W, L, B)$ *when* $(d, d') = (1, 1)$. *When* $B = \infty$, *we omit the argument* B *in* $\mathcal{N}_{\mathbb{A}}((d, d'), W, L, B)$ *and* $\mathcal{R}_{\mathbb{A}}((d, d'), W, L, B)$. *Furthermore, for* $\mathbb{A} = \mathbb{R}$, *we allow omission of the argument* $\mathbb{A}$ *in* $\mathcal{N}_{\mathbb{A}}((d, d'), W, L, B)$ *and* $\mathcal{R}_{\mathbb{A}}((d, d'), W, L, B)$. *One specific incarnation of this policy that will be used frequently is* $\mathcal{N}(W, L) = \mathcal{N}_{\mathbb{R}}((1, 1), W, L, \infty)$ *and* $\mathcal{R}(W, L) = \mathcal{R}_{\mathbb{R}}((1, 1), W, L, \infty)$.

To clarify and prevent confusion, we note that configurations in $\mathcal{N}_{\mathbb{A}}((d, d'), W, L, B)$ can have depth $\ell \leq L$ and will correspondingly be designated by $(A_i, b_i)_{i=1}^{\ell}$. We also emphasize the importance of differentiating between network configurations and network realizations. As we shall see later in the paper, network configurations with different architectures and weights taking values in different sets may realize the same function. Nevertheless, whenever there is no potential for confusion, we will use the term network to collectively refer to both configurations and realizations.

Regarding the value sets of the network weights, we will typically consider sets of the form

$$\begin{aligned} \mathbb{Q}_b^a &:= (-2^{a+1}, 2^{a+1}) \cap 2^{-b}\mathbb{Z} \\ &= \left\{ \pm \sum_{i=-b}^{a} \theta_i 2^i : \theta_i \in \{0, 1\} \right\}, \end{aligned} \tag{4}$$

for some $a, b \in \mathbb{N}$, which is the set of all base-2 quantized numbers with $a + 1$ digits before and b digits after the binary point. Each element in $\mathbb{Q}_b^a$ can hence be described by $a + b + 2$ bits, taking into account that we need one bit to encode the sign, and we have $|\mathbb{Q}_b^a| \leq 2^{a+b+2}$. To simplify notation, we shall write

$$\mathcal{N}_b^a((d, d'), W, L) := \mathcal{N}_{\mathbb{Q}_b^a}((d, d'), W, L), \tag{5}$$

$$\mathcal{R}_b^a((d, d'), W, L) := \mathcal{R}_{\mathbb{Q}_b^a}((d, d'), W, L), \tag{6}$$

and will, as before, allow omission of the argument (d, d') whenever $(d, d') = (1, 1)$. We shall frequently use the shorthands $\mathcal{N}_b^a(W, L) = \mathcal{N}_b^a((1, 1), W, L)$ and $\mathcal{R}_b^a(W, L) = \mathcal{R}_b^a((1, 1), W, L)$.

Throughout the paper, approximation errors will be quantified in terms of the following concept.

Definition 1.2 (Minimax (approximation) error)**.** *Let* $(\mathcal{X}, \delta)$ *be a metric space and* $\mathcal{F}, \mathcal{G} \subseteq \mathcal{X}$. *We define the minimax error in the approximation of elements of* $\mathcal{F}$ *through elements of* $\mathcal{G}$ *according to*

$$\mathcal{A}(\mathcal{F}, \mathcal{G}, \delta) := \sup_{f \in \mathcal{F}} \inf_{g \in \mathcal{G}} \delta(f, g).$$

When $\delta = \|\cdot\|_{L^\infty([0,1])}$, *we shall write* $\mathcal{A}_\infty(\mathcal{F}, \mathcal{G})$ *instead of* $\mathcal{A}(\mathcal{F}, \mathcal{G}, \delta)$.

The main goal of this paper is to characterize the behavior of

$$\mathcal{A}_\infty(H^1([0, 1]), \mathcal{R}_b^1(W, L)),$$

for independent choices of $W, L, b \in \mathbb{N}$. Motivated by the fact that numbers in $\mathbb{Q}_b^1$ are specified by $b + 3$ bits, we shall refer to b as the precision of $\mathbb{Q}_b^1$, $\mathcal{N}_b^1(W, L)$, and $\mathcal{R}_b^1(W, L)$.

The remainder of the paper is organized as follows. In Section 2, we introduce the concepts of network memory consumption, memory optimality, and memory redundancy. Three

lower bounds on $\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}_b^1(W,L))$ are then presented, the first one incurred by minimum memory requirements, the second one based on VC dimension arguments, and the third one resulting from numerical precision limitations inherent to ReLU networks with quantized weights. These three bounds combine to a minimax error lower bound whose constituents are active in different regimes with regards to the choice of b.

An upper bound on $\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}_b^1(W,L))$ is derived in Section 3 by constructing an approximating network whose precision b is carefully chosen to depend on network width W and depth L. In Section 4, we show how this dependency can be relaxed through what we call the depth-precision tradeoff establishing—in a constructive manner—that network depth can be traded for network weight precision. Finally, Section 5 combines the minimax error lower and upper bounds to identify the three different quantization regimes and to prove memory optimality in the proper-quantization regime.

2 Minimax Error Lower Bounds

We first introduce and explore the concept of minimum memory requirement and then derive an associated minimax error lower bound. To set the stage, we commence with a brief review of the Kolmogorov-Donoho rate-distortion theory for neural network approximation as developed in [14], and present a non-asymptotic version thereof. The theory in [14] considers a metric space $(\mathcal{X}, \delta)$ along with a set $\mathcal{Y} \subseteq \mathcal{X}$. For each $\ell \in \mathbb{N} \cup \{0\}$, the set of binary length-ℓ encoders E of $\mathcal{Y}$ of length ℓ is defined as

$$\mathfrak{E}^\ell(\mathcal{Y}) := \{E : \mathcal{Y} \mapsto \{0,1\}^\ell\}$$

along with the set of binary decoders

$$\mathfrak{D}^\ell(\mathcal{X}) := \{D : \{0,1\}^\ell \mapsto \mathcal{X}\}.$$

We denote the empty string by ϕ and use the convention $\{0,1\}^0 = \{\phi\}$.

A quantity of central interest is the minimal length $\ell \in \mathbb{N} \cup \{0\}$ for which there exists an encoder-decoder pair $(E,D) \in \mathfrak{E}^\ell(\mathcal{Y}) \times \mathfrak{D}^\ell(\mathcal{X})$ such that $\sup_{y\in\mathcal{Y}} \delta(y, D(E(y))) \leq \varepsilon$; we refer to $\sup_{y\in\mathcal{Y}} \delta(y, D(E(y)))$ as the uniform error over the set $\mathcal{Y}$. In plain language, ℓ is the minimum number of bits needed to encode the elements in $\mathcal{Y}$ while guaranteeing that the corresponding decoding error does not exceed ε.

Definition 2.1. *Consider the metric space $(\mathcal{X}, \delta)$ and the set $\mathcal{Y} \subseteq \mathcal{X}$. For $\varepsilon > 0$, the minimax code length needed to achieve uniform error ε over the set $\mathcal{Y}$, is*

$$\ell(\varepsilon, \mathcal{Y}, (\mathcal{X}, \delta)) := \min\{\ell \in \mathbb{N} \cup \{0\} : \exists (E,D) \in \mathfrak{E}^\ell(\mathcal{Y}) \times \mathfrak{D}^\ell(\mathcal{X}) : \sup_{y\in\mathcal{Y}} \delta(y, D(E(y))) \leq \varepsilon\}.$$

We omit the argument $(\mathcal{X}, \delta)$ in ℓ whenever it is clear from the context.

The theory developed in [14] is built on the asymptotic behavior of the minimax code length, as characterized by the optimal exponent $\sup\{\gamma \in \mathbb{R} : \ell(\varepsilon, \mathcal{Y}, (\mathcal{X}, \delta)) \in \mathcal{O}(\varepsilon^{-1/\gamma}), \varepsilon \to 0\}$. Here, we shall instead work directly with the minimax code length $\ell(\varepsilon, \mathcal{Y}, (\mathcal{X}, \delta))$, $\varepsilon \in \mathbb{R}_+$, which yields a more refined and, in particular, non-asymptotic picture.

The minimax code length ℓ does not only measure the minimum memory, i.e., the minimum number of bits, required to encode elements in $\mathcal{Y}$ at an error of no more than ε, but also

quantifies the minimum memory needed to store a set of approximants of $\mathcal{Y}$. This insight follows from the observation that a finite set $\mathcal{G}$ of approximants for $\mathcal{Y}$ induces specific encoder-decoder pairs for $\mathcal{Y}$, as follows.

Proposition 2.2. *Let $(\mathcal{X}, \delta)$ be a metric space, $\mathcal{Y} \subseteq \mathcal{X}$, and $\varepsilon \in \mathbb{R}_+$. Every finite subset $\mathcal{G} \subseteq \mathcal{X}$ such that $\mathcal{A}(\mathcal{Y}, \mathcal{G}, \delta) \leq \varepsilon$, induces an encoder-decoder pair $(E : \mathcal{Y} \mapsto \{0,1\}^{\lceil \log(|\mathcal{G}|) \rceil}, D : \{0,1\}^{\lceil \log(|\mathcal{G}|) \rceil} \mapsto \mathcal{G})$ satisfying $\sup_{y \in \mathcal{Y}} \delta(y, D(E(y))) \leq \varepsilon$ and*

$$\lceil \log(|\mathcal{G}|) \rceil \geq \ell(\varepsilon, \mathcal{Y}, (\mathcal{X}, \delta)). \tag{7}$$

Proof. We first note that

$$\begin{aligned} \varepsilon &\geq \mathcal{A}(\mathcal{Y}, \mathcal{G}, \delta) && (8)\\ &= \sup_{y \in \mathcal{Y}} \inf_{g \in \mathcal{G}} \delta(y, g) && (9)\\ &= \sup_{y \in \mathcal{Y}} \min_{g \in \mathcal{G}} \delta(y, g), && (10) \end{aligned}$$

where in (8) we used the assumption $\mathcal{A}(\mathcal{Y}, \mathcal{G}, \delta) \leq \varepsilon$, and the equivalence of inf and min in (10) follows from the fact that $\mathcal{G}$ is finite by assumption. The inequality (8)-(10) implies that, for every $y \in \mathcal{Y}$, there exists an element in $\mathcal{G}$, which we denote by $A(y)$, such that $\delta(y, A(y)) \leq \varepsilon$. This, in turn, induces a mapping $A : \mathcal{Y} \mapsto \mathcal{G}$ satisfying

$$\delta(y, A(y)) \leq \varepsilon, \quad \text{for all } y \in \mathcal{Y}. \tag{11}$$

We proceed to construct the desired encoder-decoder pair (E, D) by building on (11). First, define an auxiliary function $\tilde{E} : \mathcal{G} \mapsto \{0,1\}^{\lceil \log(|\mathcal{G}|) \rceil}$, which maps every element in $\mathcal{G}$ to a unique bitstring of length $\lceil \log(|\mathcal{G}|) \rceil$. For $|\mathcal{G}| = 1$, set $\tilde{E} : \mathcal{G} \mapsto \{0,1\}^{\lceil \log(|\mathcal{G}|) \rceil} = \{\phi\}$ to be the mapping that takes the single element of $\mathcal{G}$ into the empty string. For $|\mathcal{G}| \geq 2$, we first label the elements in $\mathcal{G}$ as $\mathcal{G} = (x_i)_{i=1}^{|\mathcal{G}|}$, in an arbitrary manner. Then, we take $\tilde{E} : \mathcal{G} \mapsto \{0,1\}^{\lceil \log(|\mathcal{G}|) \rceil}$ such that, for $i = 1, \ldots, |\mathcal{G}|$, $\tilde{E}(x_i)$ is the bitstring of the binary representation of the integer $(i-1)$ with 0's added at the beginning so that the overall bitstring has length $\lceil \log(|\mathcal{G}|) \rceil$. For all $|\mathcal{G}|$, $\tilde{E}$ is an injection, ensuring the existence of a decoder $D : \{0,1\}^{\lceil \log(|\mathcal{G}|) \rceil} \mapsto \mathcal{G}$ such that

$$D(\tilde{E}(x)) = x, \quad \text{for all } x \in \mathcal{G}. \tag{12}$$

With the mapping $A : \mathcal{Y} \mapsto \mathcal{G}$ defined above, we now set $E = \tilde{E} \circ A : \mathcal{Y} \mapsto \{0,1\}^{\lceil \log(|\mathcal{G}|) \rceil}$, and note that, for all $y \in \mathcal{Y}$,

$$\begin{aligned} \delta(y, D(E(y))) &= \delta(y, D(\tilde{E}(A(y)))) && (13)\\ &= \delta(y, A(y)) && (14)\\ &\leq \varepsilon, && (15) \end{aligned}$$

where in (14) we used (12) with $A(y) \in \mathcal{G}$, and (15) follows from (11). In summary, (E, D) constitutes an encoder-decoder pair of length $\lceil \log(|\mathcal{G}|) \rceil$ achieving uniform error ε over the set $\mathcal{Y}$. By the minimality of $\ell(\varepsilon, \mathcal{Y}, (\mathcal{X}, \delta))$, we deduce that $\lceil \log(|\mathcal{G}|) \rceil \geq \ell(\varepsilon, \mathcal{Y}, (\mathcal{X}, \delta))$. □

Proposition 2.2 states that a finite set of approximants $\mathcal{G}$ achieving minimax error ε in the approximation of $\mathcal{Y}$, under the metric δ, requires at least $\ell(\varepsilon, \mathcal{Y}, (\mathcal{X}, \delta))$ bits to encode.

This insight allows us to quantify the redundancy of a set of approximants. Specifically, consider the approximation of the set $\mathcal{Y}$ by the set $\mathcal{G}$ with minimax error $\varepsilon := \mathcal{A}(\mathcal{Y}, \mathcal{G}, \delta)$. In the case $\ell(\varepsilon, \mathcal{Y}, (\mathcal{X}, \delta)) \geq 1$, we quantify redundancy in a multiplicative manner by defining it according to $\frac{\lceil \log(|\mathcal{G}|) \rceil}{\ell(\varepsilon, \mathcal{Y}, (\mathcal{X}, \delta))}$. When $\ell(\varepsilon, \mathcal{Y}, (\mathcal{X}, \delta)) = 0$, we have to work with an additive redundancy measure, which we take to be $\lceil \log(|\mathcal{G}|) \rceil - \ell(\varepsilon, \mathcal{Y}, (\mathcal{X}, \delta) = \lceil \log(|\mathcal{G}|) \rceil$. Instead of carrying along two separate redundancy measures, we will simply use $\frac{\lceil \log(|\mathcal{G}|) \rceil}{1+\ell(\varepsilon, \mathcal{Y}, (\mathcal{X}, \delta))}$ to quantify redundancy. To see that this makes sense, we note that both multiplicative and additive redundancy, within their corresponding applicability regimes, are sandwiched between[2] $\frac{\lceil \log(|\mathcal{G}|) \rceil}{1+\ell(\varepsilon, \mathcal{Y}, (\mathcal{X}, \delta))}$ and $\frac{2\lceil \log(|\mathcal{G}|) \rceil}{1+\ell(\varepsilon, \mathcal{Y}, (\mathcal{X}, \delta))}$.

Often we shall be dealing with families of approximants $\{\mathcal{G}_i\}_{i \in \mathcal{I}}$ parametrized by a, not necessarily ordered, index set $\mathcal{I}$. This concept will allow us to consider neural network families indexed by their architectures and weight sets, aiming for different levels of approximation error. Specifically, we shall frequently take $\mathcal{I} \subseteq \mathbb{N}^3$, and, for $(W, L, b) \in \mathcal{I}$, set $i = (W, L, b)$ and $\mathcal{G}_i = \mathcal{G}_{(W,L,b)} = \mathcal{R}_b^1(W, L)$.

We will say that $\{\mathcal{G}_i\}_{i \in \mathcal{I}}$ approximates $\mathcal{X}$ in a memory-optimal fashion if the approximation error can be made arbitrarily small while ensuring that the memory redundancy remains bounded, as formalized next.

Definition 2.3 (Memory redundancy and memory optimality)**.** *Let $(\mathcal{X}, \delta)$ be a metric space and $\mathcal{Y} \subseteq \mathcal{X}$. We define the memory redundancy in the approximation of $\mathcal{Y}$ by a subset $\mathcal{G} \subseteq \mathcal{X}$ as*

$$r(\mathcal{Y}, \mathcal{G}, \rho) := \frac{\lceil \log(|\mathcal{G}|) \rceil}{1 + \ell(\mathcal{A}(\mathcal{Y}, \mathcal{G}, \rho), \mathcal{Y}, (\mathcal{X}, \delta))}.$$

A family of finite subsets $\{\mathcal{G}_i\}_{i \in \mathcal{I}} \subseteq \mathcal{X}$ is said to achieve memory optimality in the approximation of $\mathcal{Y}$ if

$$\inf_{i \in \mathcal{I}} \mathcal{A}(\mathcal{Y}, \mathcal{G}_i, \rho) = 0, \quad \textit{and} \tag{16}$$

$$\sup_{i \in \mathcal{I}} r(\mathcal{Y}, \mathcal{G}_i, \rho) < \infty. \tag{17}$$

Recall that our main focus is the minimax error $\mathcal{A}_\infty(\mathcal{F}, \mathcal{G})$ with $\mathcal{F} = H^1([0,1])$ and $\mathcal{G} = \mathcal{R}_b^1(W, L)$, for $W, L, b \in \mathbb{N}$. To analyze the associated memory redundancy, we hence need to characterize $|\mathcal{R}_b^1(W, L)|$ and the minimax code length $\ell(\varepsilon, H^1([0,1]), (L^\infty([0,1]), \|\cdot\|_{L^\infty([0,1])}))$, $\varepsilon \in \mathbb{R}_+$, short-handed as $\ell(\varepsilon, H^1([0,1]))$.

2.1 Upper-bounding the cardinality of $\mathcal{R}_b^1(W, L)$

We shall first establish an upper bound on the cardinality of $\mathcal{R}_{\mathbb{A}}((d, d'), W, L)$ for general $\mathbb{A}$ and then particularize this bound for $\mathbb{A} = \mathbb{Q}_b^1$. The more general result does not demand any extra technical effort and makes for a more accessible exposition.

Let us start with some heuristic reasoning. We can store a network realization in $\mathcal{R}_{\mathbb{A}}((d, d'), W, L)$ by storing its corresponding network configuration in $\mathcal{N}_{\mathbb{A}}((d, d'), W, L)$. A given network configuration in $\mathcal{N}_{\mathbb{A}}((d, d'), W, L)$ has at most $W(W+1)L$ weights, and each weight needs $\lceil \log(|\mathbb{A}|) \rceil$ bits to represent it. It therefore takes at most $W(W+1)L\lceil \log(|\mathbb{A}|) \rceil$ bits to store all the weights in the network configuration. Storing the network depth and the widths of

[2] When $\ell(\varepsilon, \mathcal{Y}, (\mathcal{X}, \delta)) \geq 1$, we have $\frac{\lceil \log(|\mathcal{G}|) \rceil}{1+\ell(\varepsilon, \mathcal{Y}, (\mathcal{X}, \delta))} \leq \frac{\lceil \log(|\mathcal{G}|) \rceil}{\ell(\varepsilon, \mathcal{Y}, (\mathcal{X}, \delta))} \leq \frac{2\lceil \log(|\mathcal{G}|) \rceil}{1+\ell(\varepsilon, \mathcal{Y}, (\mathcal{X}, \delta))}$. For $\ell(\varepsilon, \mathcal{Y}, (\mathcal{X}, \delta)) = 0$, it follows that $\frac{\lceil \log(|\mathcal{G}|) \rceil}{1+\ell(\varepsilon, \mathcal{Y}, (\mathcal{X}, \delta))} = \lceil \log(|\mathcal{G}|) \rceil < 2\lceil \log(|\mathcal{G}|) \rceil = \frac{2\lceil \log(|\mathcal{G}|) \rceil}{1+\ell(\varepsilon, \mathcal{Y}, (\mathcal{X}, \delta))}$.

the individual layers, requires an extra $\lceil \log(L) \rceil$ bits and $L\lceil \log(W) \rceil$ bits, respectively, which can be absorbed by a constant multiplying $W(W+1)L\lceil \log(|\mathbb{A}|) \rceil$. In total we hence need at most $CW(W+1)L\lceil \log(|\mathbb{A}|) \rceil$ bits, with C an absolute constant. This intuitive reasoning is formalized in the following result.

Proposition 2.4. *For $d, d', W, L \in \mathbb{N}$ and a finite subset $\mathbb{A} \subseteq \mathbb{R}$ with $|\mathbb{A}| \geq 2$, we have*

$$\log(|\mathcal{R}_{\mathbb{A}}((d,d'),W,L)|) \leq \log(|\mathcal{N}_{\mathbb{A}}((d,d'),W,L)|) \leq 5W^2 L \log(|\mathbb{A}|). \tag{18}$$

In particular, for $\mathbb{A} = \mathbb{Q}_b^a$, $a, b \in \mathbb{N}$,

$$\log(|\mathcal{R}_b^a((d,d'),W,L)|) \leq \log(|\mathcal{N}_b^a((d,d'),W,L)|) \leq 10W^2 L(a+b). \tag{19}$$

Proof. By definition,

$$\begin{aligned}&\mathcal{N}_{\mathbb{A}}((d,d'),W,L)\\ &\subseteq \{(A_i,b_i)_{i=1}^{\ell} \in \mathcal{N}((d,d')) : \mathcal{W}((A_i,b_i)_{i=1}^{\ell}) \leq W, \ell \leq L, \mathcal{C}((A_i,b_i)_{i=1}^{\ell}) \subseteq \mathbb{A}\}.\end{aligned}$$

Recall that, for a given network configuration $(A_i,b_i)_{i=1}^{\ell} \in \mathcal{N}_{\mathbb{A}}((d,d'),W,L)$ with $A_i \in \mathbb{R}^{N_i \times N_{i-1}}$, $N_{i-1}, N_i \in \mathbb{N}$, $i = 1,\dots,\ell$, we call the tuple $(N_0,\dots,N_\ell)$ the architecture of $(A_i,b_i)_{i=1}^{\ell}$. For given ℓ, there are at most $W^{\ell+1}$ different architectures. As $\ell \in \{1,\dots,L\}$, the total number of possible architectures $(N_0,\dots,N_\ell)$ for network configurations in $\mathcal{N}_{\mathbb{A}}((d,d'),W,L)$ is hence upper-bounded by $\sum_{\ell=1}^{L} W^{\ell+1} \leq LW^{L+1}$. For a given network configuration Φ with architecture $(N_0,\dots,N_\ell)$, the number of weights satisfies $\sum_{i=1}^{\ell}(N_i\,N_{i-1}+N_i) \leq LW(W+1)$; therefore, the number of possible Φ of a given architecture is no more than $|\mathbb{A}|^{LW(W+1)}$, as each weight can take $|\mathbb{A}|$ different values. Putting everything together, we obtain

$$|\mathcal{N}_{\mathbb{A}}((d,d'),W,L)| \leq LW^{L+1}|\mathbb{A}|^{LW(W+1)},$$

which, in turn, implies

$$\begin{aligned}\log(|\mathcal{N}_{\mathbb{A}}((d,d'),W,L)|) &\leq \log(LW^{L+1}) + (L(W^2+W))\log(|\mathbb{A}|)\\ &\leq \log(L) + (L+1)\log(W) + 2W^2 L\log(|\mathbb{A}|)\\ &\leq 5W^2 L\log(|\mathbb{A}|).\end{aligned} \tag{20}$$

Noting that $|\mathcal{R}_{\mathbb{A}}((d,d'),W,L)| \leq |\mathcal{N}_{\mathbb{A}}((d,d'),W,L)|$, yields (18). Finally, (19) follows by using $\log(|\mathbb{Q}_b^a|) = \log(|\{\pm\sum_{i=-b}^{a}\theta_i 2^i : \theta_i \in \{0,1\}\}|) \leq a+b+2 \leq 2(a+b)$. □

Proposition 2.4 provides an upper bound on the memory required to store the network realizations in $\mathcal{R}_{\mathbb{A}}((d,d'),W,L)$. As this storage method mirrors how neural networks are stored on a computer, we term it the natural encoding and refer to (18) as the memory consumption upper bound under natural encoding. In contrast, [14] considers networks which are sparse in the sense of having a small number of nonzero weights, and stores only the nonzero weights and their respective locations as uniquely decodable bitstrings.

2.2 Lower-bounding the minimax code length $\ell(\varepsilon, H^1([0,1]))$

We next lower-bound the minimax code length by relating it to the covering number and the packing number defined next.

Definition 2.5 (Covering number and packing number)**.** *[15, Definitions 5.1 and 5.4] Let $(\mathcal{X}, \delta)$ be a metric space. An ε-covering of $\mathcal{X}$ is a finite set $\{x_1, \ldots, x_n\}$ of $\mathcal{X}$ such that for all $x \in \mathcal{X}$, there exists an $i \in \{1, \ldots, n\}$ so that $\delta(x, x_i) \leq \varepsilon$. The ε-covering number $N(\varepsilon, \mathcal{X}, \delta)$ is the cardinality of a smallest ε-covering of $\mathcal{X}$. An ε-packing of $\mathcal{X}$ is a finite subset $\{x_1, \ldots, x_n\}$ of $\mathcal{X}$ such that $\delta(x_i, x_j) > \varepsilon$, for all $i, j \in \{1, \ldots, n\}$ with $i \neq j$. The ε-packing number $M(\varepsilon, \mathcal{X}, \delta)$ is the cardinality of a largest ε-packing of $\mathcal{X}$.*

An important relation between the covering number and the packing number is the following.

Lemma 2.6. *[15, Lemma 5.5] For a metric space $(\mathcal{X}, \delta)$ and $\varepsilon \in \mathbb{R}_+$, it holds that*

$$M(2\varepsilon, \mathcal{X}, \delta) \leq N(\varepsilon, \mathcal{X}, \delta) \leq M(\varepsilon, \mathcal{X}, \delta).$$

The minimax code length can be related to the covering and the packing numbers as follows.

Lemma 2.7. *Let $(\mathcal{X}, \delta)$ be a metric space, $\mathcal{Y} \subseteq \mathcal{X}$, and $\varepsilon \in \mathbb{R}_+$. We have*

$$\log(M(2\varepsilon, \mathcal{Y}, \delta)) \leq \ell(\varepsilon, \mathcal{Y}, (\mathcal{X}, \delta)) \leq \lceil \log(N(\varepsilon, \mathcal{Y}, \delta)) \rceil. \tag{21}$$

Proof. We first prove the inequality $\ell(\varepsilon, \mathcal{Y}, (\mathcal{X}, \delta)) \geq \log(M(2\varepsilon, \mathcal{Y}, \delta))$. To this end, let $(E : \mathcal{Y} \mapsto \{0,1\}^{\ell(\varepsilon, \mathcal{Y}, (\mathcal{X}, \delta))}, D : \{0,1\}^{\ell(\varepsilon, \mathcal{Y}, (\mathcal{X}, \delta))} \mapsto \mathcal{X})$ be an encoder-decoder pair achieving uniform error ε over the set $\mathcal{Y}$, i.e.,

$$\delta(y, D(E(y))) \leq \varepsilon, \quad \text{for all } y \in \mathcal{Y}, \tag{22}$$

and let $\mathcal{P}$ be a largest (2ε)-packing of $\mathcal{Y}$, i.e., $|\mathcal{P}| = M(2\varepsilon, \mathcal{Y}, \delta)$. For $|\mathcal{P}| = 1$, the first inequality in (21) is trivially satisfied as $\ell(\varepsilon, \mathcal{Y}, (\mathcal{X}, \delta)) \geq 0$ by definition. In the case $|\mathcal{P}| \geq 2$, we have, for distinct $p_1, p_2 \in \mathcal{P}$,

$$\begin{aligned}
\delta(D(E(p_1)), D(E(p_2))) &\geq \delta(D(E(p_1)), p_2) - \delta(p_2, D(E(p_2))) && (23)\\
&\geq \delta(p_1, p_2) - \delta(D(E(p_1)), p_1) - \delta(p_2, D(E(p_2))) && (24)\\
&> 2\varepsilon - \varepsilon - \varepsilon && (25)\\
&= 0, && (26)
\end{aligned}$$

where (23) and (24) follow from the triangle inequality, and in (25) we used (22) and $\delta(p_1, p_2) > 2\varepsilon$ owing to $\mathcal{P}$ being a (2ε)-packing. We can hence conclude that $D(E(p_1)) \neq D(E(p_2))$ and have thereby established the injectivity of $D \circ E$ on $\mathcal{P}$. Consequentially, E must also be injective on $\mathcal{P}$. This, in turn, implies that the cardinality of the range of E is no less than the cardinality of $\mathcal{P}$, namely, $2^{\ell(\varepsilon, \mathcal{Y}, (\mathcal{X}, \delta))} \geq |\mathcal{P}| = M(2\varepsilon, \mathcal{Y}, \delta)$, and therefore $\log(M(2\varepsilon, \mathcal{Y}, \delta)) \leq \ell(\varepsilon, \mathcal{Y}, (\mathcal{X}, \delta))$.

It remains to show that $\ell(\varepsilon, \mathcal{Y}, (\mathcal{X}, \delta)) \leq \lceil \log(N(\varepsilon, \mathcal{Y}, \delta)) \rceil$. Let $\mathcal{C} = \{c_i\}_{i=1}^{N(\varepsilon, \mathcal{Y}, \delta)}$ be a minimal ε-covering of $\mathcal{Y}$. Hence, $\mathcal{A}(\mathcal{Y}, \mathcal{C}, \delta) = \sup_{y \in \mathcal{Y}} \inf_{c \in \mathcal{C}} \delta(y, c) \leq \varepsilon$. Application of Proposition 2.2 with $\mathcal{G} = \mathcal{C}$ yields $\lceil \log(|\mathcal{C}|) \rceil \geq \ell(\varepsilon, \mathcal{Y}, (\mathcal{X}, \delta))$, which together with $N(\varepsilon, \mathcal{Y}, \delta) = |\mathcal{C}|$ establishes $\ell(\varepsilon, \mathcal{Y}, (\mathcal{X}, \delta)) \leq \lceil \log(N(\varepsilon, \mathcal{Y}, \delta)) \rceil$. □

Remark. *With the insights provided by Lemma 2.7, Proposition 2.2 and Definition 2.3 could equivalently have been formulated in terms of packing and covering number. We decided, however, to work with the minimax code length so as to emphasize the implications of our results in terms of memory consumption.*

We can now lower-bound the minimax code length of $H^1([0,1])$ using a lower bound on the covering number of $H^1([0,1])$.

Lemma 2.8. *For $\varepsilon > 0$, there exist absolute constants $C, \varepsilon_0 > 0$, such that*

$$\ell(\varepsilon, H^1([0,1])) \geq C\,\varepsilon^{-1}, \quad \forall \varepsilon \in (0, \varepsilon_0]. \tag{27}$$

Proof. By [15, Example 5.10], we have

$$\log(N(\varepsilon, H^1([0,1]), \|\cdot\|_{L^\infty([0,1])})) \geq c\,\varepsilon^{-1}, \quad \forall \varepsilon \in (0, \varepsilon_1], \tag{28}$$

for absolute constants $c, \varepsilon_1 \in \mathbb{R}_+$. Set $\varepsilon_0 = \frac{1}{2}\varepsilon_1$ and $C = \frac{1}{2}c$. Then, for $\varepsilon \leq \varepsilon_0$, we have

$$\begin{aligned}
\ell(\varepsilon, H^1([0,1])) &\geq \log(M(2\varepsilon, H^1([0,1]), \|\cdot\|_{L^\infty([0,1])})) && (29)\\
&\geq \log(N(2\varepsilon, H^1([0,1]), \|\cdot\|_{L^\infty([0,1])})) && (30)\\
&\geq c\,(2\varepsilon)^{-1} && (31)\\
&\geq C\varepsilon^{-1}, && (32)
\end{aligned}$$

where in (29) we used Lemma 2.7, (30) follows from Lemma 2.6, and in (31) we applied (28). □

2.3 Lower bound incurred by minimum memory requirement

We now have all the ingredients to derive a lower bound on the minimax error incurred by neural network approximation of 1-Lipschitz functions. The specific result is as follows.

Proposition 2.9. *There exists an absolute constant c_m such that for all $W, L, b \in \mathbb{N}$, it holds that*

$$\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}_b^1(W,L)) \geq c_m(W^2Lb)^{-1}. \tag{33}$$

Proof. Let C, ε_0 be the absolute constants in Lemma 2.8 and set $c_m = \min\{\frac{C}{30}, \varepsilon_0\}$. Suppose, for the sake of contradiction, that

$$\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}_b^1(W,L)) < c_m(W^2Lb)^{-1}. \tag{34}$$

Set $\varepsilon = c_m(W^2Lb)^{-1}$. Applying Proposition 2.2 with $\mathcal{Y} = H^1([0,1])$, $\mathcal{G} = \mathcal{R}_b^1(W,L)$, $\delta = \|\cdot\|_{L^\infty([0,1])}$, upon noting that the prerequisites in Proposition 2.2 are satisfied as $\mathcal{G} = \mathcal{R}_b^1(W,L)$ and (34) implies $\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}_b^1(W,L)) < c_m(W^2Lb)^{-1} = \varepsilon$, we obtain

$$\lceil \log(|\mathcal{R}_b^1(W,L)|) \rceil \geq \ell(\varepsilon, H^1([0,1])),$$

which together with $\lceil \log(|\mathcal{R}_b^1(W,L)|) \rceil \leq 10W^2L(1+b) \leq 20W^2Lb$, thanks to Proposition 2.4, establishes

$$20W^2Lb \geq \ell(\varepsilon, H^1([0,1])). \tag{35}$$

On the other hand,

$$\begin{aligned}
20W^2Lb &= 20\,c_m\,\varepsilon^{-1} && (36)\\
&< C\varepsilon^{-1} && (37)\\
&\leq \ell(\varepsilon, H^1([0,1])), && (38)
\end{aligned}$$

where (37) follows from $20c_m = 20\min\{\frac{C}{30}, \varepsilon_0\} < C$, and in (38) we applied Lemma 2.8 with the prerequisite satisfied as $\varepsilon = c_m(W^2Lb)^{-1} \leq c_m \leq \varepsilon_0$. Since (35) contradicts the strict inequality (36)-(38), we must have

$$\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}_b^1(W,L)) \geq c_m(W^2Lb)^{-1}. \qquad \square$$

We shall refer to (33) as the minimax error lower bound incurred by the minimum memory requirement. Attaining this lower bound to within a multiplicative constant, implies memory optimality, as demonstrated next.

Proposition 2.10. *Let $\mathcal{I} \subseteq \mathbb{N}^3$ be an infinite set. Suppose that*

$$\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}_b^1(W,L)) \leq D(W^2Lb)^{-1}, \quad \forall (W,L,b) \in \mathcal{I}, \tag{39}$$

for some $D \in \mathbb{R}_+$ independent of W, L, b. Then,

$$\inf_{(W,L,b)\in\mathcal{I}} \mathcal{A}_\infty(H^1([0,1]), \mathcal{R}_b^1(W,L)) = 0, \tag{40}$$

$$\sup_{(W,L,b)\in\mathcal{I}} r(H^1([0,1]), \mathcal{R}_b^1(W,L), \|\cdot\|_{L^\infty([0,1])}) < \infty, \tag{41}$$

and hence $\{R_b^1(W,L) : (W,L,b) \in \mathcal{I}\}$ achieves memory optimality—in the sense of Definition 2.3—in the approximation of $H^1([0,1])$.

Proof. We first note that

$$\inf_{(W,L,b)\in\mathcal{I}} \mathcal{A}_\infty(H^1([0,1]), \mathcal{R}_b^1(W,L)) \leq \inf_{(W,L,b)\in\mathcal{I}} D(W^2Lb)^{-1} \tag{42}$$

$$= 0, \tag{43}$$

where (43) follows as $\mathcal{I}$ is an infinite set[3]. This establishes (40). To prove (41), we first fix a tuple $(W,L,b) \in \mathcal{I}$ and consider the memory redundancy

$$r(H^1([0,1]), \mathcal{R}_b^1(W,L), \|\cdot\|_{L^\infty([0,1])}) = \frac{\lceil \log(|\mathcal{R}_b^1(W,L)|)\rceil}{1 + \ell(\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}_b^1(W,L)), H^1([0,1]))} \tag{44}$$

$$\leq \frac{10W^2L(1+b)}{1 + \ell(\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}_b^1(W,L)), H^1([0,1]))}, \tag{45}$$

where (45) follows from Proposition 2.4. We now distinguish two cases. First, for $D(W^2Lb)^{-1} > \varepsilon_0$, where ε_0 is the absolute constant in Lemma 2.8, we have $W^2L(1+b) < 2W^2Lb < 2D\varepsilon_0^{-1}$, which together with (44)-(45) leads to the memory redundancy upper bound $r(H^1([0,1]), \mathcal{R}_b^1(W,L), \|\cdot\|_{L^\infty([0,1])}) \leq 10W^2L(1+b) \leq 20D\varepsilon_0^{-1}$. Second, for $D(W^2Lb)^{-1} \leq \varepsilon_0$, we have

$$\ell(\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}_b^1(W,L)), H^1([0,1])) \tag{46}$$

$$\geq \ell(D(W^2Lb)^{-1}, H^1([0,1])) \tag{47}$$

$$\geq CD^{-1}W^2Lb, \tag{48}$$

where in (47) we used (39) together with the fact that $\varepsilon \mapsto \ell(\varepsilon, H^1([0,1]))$ is a nonincreasing function, and (48) follows from Lemma 2.8, with C being the absolute constant from Lemma 2.8. Then, the memory redundancy can be upper-bounded according to

$$r(H^1([0,1]), \mathcal{R}_b^1(W,L), \|\cdot\|_{L^\infty([0,1])}) \leq \frac{10W^2L(1+b)}{1 + CD^{-1}W^2Lb} \leq 20DC^{-1}.$$

[3]For $\mathcal{I}$ infinite, we have $\sup_{(W,L,b)\in\mathcal{I}} \max\{W,L,b\} = \infty$, which implies $\inf_{(W,L,b)\in\mathcal{I}} D(W^2Lb)^{-1} \leq \inf_{(W,L,b)\in\mathcal{I}} D(\max\{W,L,b\})^{-1} = 0$.

Combining the two cases, we get

$$r(H^1([0,1]), \mathcal{R}_b^1(W,L), \|\cdot\|_{L^\infty([0,1])}) \leq \max\{20D\varepsilon_0^{-1}, 20DC^{-1}\} = 20D\max\{\varepsilon_0^{-1}, C^{-1}\}.$$

Recalling that ε_0 and C are absolute constants, and the tuple $(W,L,b) \in \mathcal{I}$ is fixed but arbitrary, it follows that

$$\sup_{(W,L,b)\in\mathcal{I}} r(H^1([0,1]), \mathcal{R}_b^1(W,L), \|\cdot\|_{L^\infty([0,1])}) \leq \sup_{(W,L,b)\in\mathcal{I}} 20D\max\{\varepsilon_0^{-1}, C^{-1}\} < \infty,$$

where we used $D < \infty$ and $\varepsilon_0, C > 0$. This validates (41) and thereby finalizes the proof. □

2.4 Two additional lower bounds

We proceed to establish two additional minimax error lower bounds. The first one adapts the technique in [9] for lower-bounding the minimax error for unquantized networks, i.e., networks with real-valued weights. Notably, [9] uses an upper bound on the VC dimension of ReLU networks reported in [12]. While this adaptation in itself is not substantial, we still feel that the underlying idea is worthy of recording, also in the sense of clarity and completeness of exposition.

Proposition 2.11. *There exists an absolute constant c_v such that for all $W \in \mathbb{N}$ and $L \in \mathbb{N}$ with $L \geq 2$, it holds that*

$$\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}(W,L)) \geq c_v(W^2L^2(\log(W)+\log(L)))^{-1}, \tag{49}$$

and, hence, for nonempty $\mathbb{A} \subseteq \mathbb{R}$,

$$\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}_\mathbb{A}(W,L)) \geq c_v(W^2L^2(\log(W)+\log(L)))^{-1}. \tag{50}$$

Proof. See Appendix A. □

We will refer to (50) as the minimax error lower bound incurred by the VC-dimension limit. Additionally, we point out that this bound indicates an advantage of deep networks over shallow networks. Specifically, fixing the number of network weights, which is on the order of $n := W^2L$, an increase in depth L leads to a decrease in the minimax error lower bound according to $c_v(W^2L^2(\log(W)+\log(L)))^{-1} = c_v(nL(\log(\sqrt{\frac{n}{L}})+\log(L)))^{-1} = c_v\big(nL\log\big(\sqrt{nL}\big)\big)^{-1}$. In contrast, fixing n and increasing W leads to an increase in the minimax error lower bound according to $c_v(\frac{n^2}{W^2}\log(\frac{n}{W}))^{-1}$. This advantage of deep over shallow networks will manifest itself in our final characterization of the three quantization regimes.

The second bound we present is based on the observation that ReLU networks with quantized weights face inherent limitations in their approximation capability. The nature of these limitations is such that deep networks exhibit a fundamental advantage over shallow networks. To illustrate this aspect, consider ReLU networks with quantized weights of a fixed number of fractional bits[4] and inputs $x \in [0,1]$ also of a fixed number of fractional bits. The corresponding network outputs will also exhibit a fixed number of fractional bits. More importantly, this number increases with increasing L, but remains constant as a function of W. This is a

[4]The number of fractional bits of $x \in \mathbb{R}$ refers to the number of digits after the binary point in the binary representation of x.

consequence of the multiplication of two real numbers corresponding to the convolution of their binary expansions and the length of the convolution of two sequences being given by the sum of their lengths. More informally, multiplying small numbers in $[0,1]$ (deep network case) leads to even smaller numbers whereas adding them (shallow network case) can make them only larger and therefore does not result in an increase in numerical resolution. Finally, realizing that $f \in H^1([0,1])$ can take arbitrary values, in particular, values with an infinite number of fractional bits, it follows that the approximation error in $L^\infty([0,1])$-norm for shallow networks will suffer from an inherent numerical precision limitation, an effect not shared by deep networks. We proceed to formalize these back-of-the-envelope arguments by first establishing the statement on the numerical precision of the outputs of quantized networks.

Lemma 2.12. *Let $a, b, W, L \in \mathbb{N}$ and $c \in \mathbb{N} \cup \{0\}$. For $f \in \mathcal{R}_b^a(W,L)$ and $x \in 2^{-c}\mathbb{Z}$, it holds that*

$$f(x) \in 2^{-Lb-c}\mathbb{Z}. \tag{51}$$

Proof. Fix $x = 2^{-c}K \in 2^{-c}\mathbb{Z}$ with $K \in \mathbb{Z}$ and an $f \in \mathcal{R}_b^a(W,L)$. By definition, there exists $\Phi_0 = ((A_i, b_i))_{i=1}^{\mathcal{L}(\Phi_0)} \in \mathcal{N}_b^a(W,L)$ with $\mathcal{L}(\Phi_0) \leq L$ and $\mathcal{C}(\Phi_0) \subseteq \mathbb{Q}_b^a$ such that $R(\Phi) = f$. Now, consider the scaled network $\widetilde{\Phi}_0 = ((2^b A_i, 2^b b_i))_{i=1}^{\mathcal{L}(\Phi_0)}$. It follows from $\mathcal{C}(\Phi_0) \subseteq \mathbb{Q}_b^a = ((-2^{a+1}, 2^{a+1}) \cap 2^{-b}\mathbb{Z}) \subseteq 2^{-b}\mathbb{Z}$ that $\mathcal{C}(\widetilde{\Phi}_0) = 2^b\,\mathcal{C}(\Phi_0) \subseteq \mathbb{Z}$. Hence, $R(\widetilde{\Phi}_0)(K)$, as the output of an integer-weight ReLU network with integer-valued input, must be integer-valued. Thanks to the positive homogeneity of the ReLU function, i.e., $\rho(\lambda x) = \lambda\rho(x)$, $\forall \lambda \geq 0$, $\forall x \in \mathbb{R}$, we have $2^c(2^b)^{\mathcal{L}(\Phi_0)} R(\Phi_0)(x) = 2^c R(\widetilde{\Phi}_0)(x) = R(\widetilde{\Phi}_0)(2^c x) = R(\widetilde{\Phi}_0)(K) \in \mathbb{Z}$, and therefore, $f(x) = R(\Phi_0)(x) \in (2^{-b})^{\mathcal{L}(\Phi_0)} 2^{-c}\mathbb{Z} \subseteq 2^{-Lb-c}\mathbb{Z}$, which concludes the proof. □

The next result quantifies the numerical precision advantage of deep networks over shallow networks announced above.

Proposition 2.13. *Let $a, b, W, L \in \mathbb{N}$. It holds that*

$$\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}_b^a(W,L)) \geq \frac{1}{2}\,2^{-Lb}. \tag{52}$$

Proof. Fix an $f \in \mathcal{R}_b^a(W,L)$. It follows from Lemma 2.12, with $c = 0$ and $x = 1$, that $f(1) \in 2^{-Lb}\mathbb{Z}$. Let $g \in H^1([0,1])$ be given by $g(x) = \frac{1}{2}\,2^{-Lb}$, $x \in [0,1]$. We then have

$$\|f - g\|_{[0,1]} \geq |f(1) - g(1)| = \left| f(1) - \frac{1}{2}\,2^{-Lb} \right| \geq \frac{1}{2}\,2^{-Lb}.$$

As the choice of $f \in \mathcal{R}_b^a(W,L)$ is arbitrary, we have $\|f - g\|_{[0,1]} \geq \frac{1}{2}\,2^{-Lb}$, for all $f \in \mathcal{R}_b^a(W,L)$, which together with $g \in H^1([0,1])$ implies $\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}_b^a(W,L)) \geq \frac{1}{2}\,2^{-Lb}$. □

The bound in (52) will henceforth be referred to as the minimax error lower bound incurred by the numerical-precision limit. As announced above, this lower bound does not depend on network width W and decreases (exponentially) in network depth L, indicating an advantage of deep networks over shallow networks, which, again, will manifest itself when we characterize the three quantization regimes. We hasten to add that the exponential behavior of the lower bound (52) is a consequence of the length of the convolution product of two (binary) sequences being given by the sum of the lengths of the individual sequences.

We finally put together the individual minimax error lower bounds incurred by the minimum memory requirement, Proposition 2.9, the VC-dimension limit, Proposition 2.11, and the numerical-precision limit, Proposition 2.13, to obtain a combined lower bound as follows.

Corollary 2.14. *There exists an absolute constant c_ℓ such that for all $W, L, b \in \mathbb{N}$ with $L \geq 2$, it holds that*

$$\begin{aligned} &\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}_b^1(W,L)) \\ &\geq c_\ell \max\big\{(W^2 L\, b)^{-1}, (W^2 L^2(\log(W)+\log(L)))^{-1}, 2^{-Lb}\big\}. \end{aligned} \tag{53}$$

Proof. Set $c_\ell = \min\{c_m, c_v, \frac{1}{2}\}$, where c_m and c_v are the absolute constants in Proposition 2.9 and Proposition 2.11, respectively. Then, (53) follows from (33) in Proposition 2.9, (50) in Proposition 2.11 with $\mathbb{A} = \mathbb{Q}_b^1$, and (52) in (2.13) with $a = 1$. □

We now characterize the behavior of the combined lower bound (53) by identifying the neural network configurations, specifically the tuples (W, L, b), that make any given individual component in the lower bound dominate the other two components. To simplify the discussion, we fix sufficiently large values for W and L, and let b range from 1 to infinity.

1. The term 2^{-Lb} can dominate the other two terms only for very small values of b, as it decreases exponentially in b, in comparison to polynomial decrease and constant behavior. Consider the extreme case $b = 1$. Then, the individual terms in (53) become $(W^2 L)^{-1}$, $(W^2 L^2(\log(W)+\log(L)))^{-1}$, and 2^{-L}, respectively, and 2^{-L} can dominate the other two terms only if L is logarithmically smaller than W, i.e., $L \leq C\log(W)$ for some $C \in \mathbb{R}_+$. We call the regime where 2^{-Lb} dominates, the *under-quantization regime*, and note that it might be empty, concretely when L is not logarithmically smaller than W.

2. The term $(W^2 L\, b)^{-1}$ dominates for medium values of b and L logarithmically smaller than W, and for small to medium values of b when L is not logarithmically smaller than W. We refer to this regime as the *proper-quantization regime*.

3. For large b, the term $(W^2 L^2(\log(W)+\log(L)))^{-1}$ will dominate, owing to the other two terms going to 0 when $b \to \infty$. As the quantization resolution increases when b grows, we call this regime the *over-quantization regime*.

In summary, as b increases from 1 to infinity, the lower bound (53) transitions from exponential decay in the *under-quantization regime* (which might be empty) to polynomial decay in the *proper-quantization regime*, and finally levels out at a constant value that is independent of b in the *over-quantization regime*.

In the following two sections, we derive minimax error upper bounds that, when combined, exhibit the same three-regime behavior as (53) and allow for a precise characterization of the boundaries between the regimes.

3 A Constructive Minimax Error Upper Bound

The first bound we establish is inspired by the two-step approach employed in [14]. In the first step, we approximate functions in $H^1([0,1])$ by networks contained in $\mathcal{R}(W, L, 1)$, resulting in an upper bound on $\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}(W,L,1))$. Here we only demand that the weights of the approximating networks have absolute values no greater than 1, but they need not be

quantized. In the second step, we then quantize the weights of the networks in $\mathcal{R}(W, L, 1)$ chosen in the first step, by rounding to the nearest neighbor in the set $\mathbb{Q}_b^1$. We then bound the error $\mathcal{A}_\infty(\mathcal{R}(W, L, 1), \mathcal{R}_b^1(W, L))$ incurred by quantization. Using the triangle inequality formalized in Lemma H.1, we finally obtain an upper bound on the minimax error $\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}_b^1(W, L))$ according to

$$\begin{aligned}&\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}_b^1(W, L))\\ &\leq \mathcal{A}_\infty(H^1([0,1]), \mathcal{R}(W, L, 1)) + \mathcal{A}_\infty(\mathcal{R}(W, L, 1), \mathcal{R}_b^1(W, L)).\end{aligned} \tag{54}$$

We proceed to detail the first step announced above, namely approximating functions in $H^1([0,1])$ by ReLU networks with weight-magnitude bounded by 1.

Theorem 3.1. *There exist absolute constants $C, D \in \mathbb{R}_+$ such that, for all $W, L \in \mathbb{N}$ with $W, L \geq D$,*

$$\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}(W, L, 1)) \leq C(W^2L^2\log(W))^{-1}. \tag{55}$$

Proof. See Appendix B. □

The proof of Theorem 3.1 is constructive in the sense that, for every function $f \in H^1([0,1])$, it specifies a network $g \in \mathcal{R}(W, L, 1)$ such that $\|f - g\|_{L^\infty([0,1])} \leq C(W^2L^2\log(W))^{-1}$.

Theorem 3.1 is in the spirit of a line of papers on the approximation of smooth functions by ReLU networks, notably [6, 8, 9, 10, 16, 17]. Smoothness, in these references, is quantified by parameters $n \in \mathbb{N} \cup \{0\}$ and $\alpha \in (0, 1]$, with the functions bounded, up to order n bounded differentiable, and the n-th order derivative Hölder continuous[5] of exponent α. Here, we are concerned with the special case $n = 0$ and $\alpha = 1$. The most recent development in this line of work [9, Corollary 1.3] deals with the case $n \in \mathbb{N} \cup \{0\}$ and $\alpha = 1$ and, when particularized to $H^1([0,1])$-functions, yields

$$\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}(W, L, f(W, L))) \leq \widetilde{C}(W^2L^2\log(W))^{-1}, \tag{56}$$

for sufficiently large $W, L \in \mathbb{N}$, an absolute constant $\widetilde{C} \in \mathbb{R}_+$, and a function $f : \mathbb{N}^2 \mapsto \mathbb{R}$. Notably, upon examination of the proof of [9, Corollary 1.3], one sees that $f(W, L) \geq W^{kL}$. This is markedly different from the constant weight-magnitude 1 in our Theorem 3.1, which turns out to be crucial for achieving memory optimality. The improvement in the weight-magnitude behavior we obtain is predominantly owed to the novel bit extraction technique developed in the proof. This technique is interesting in its own right and can readily be applied for general $n \in \mathbb{N} \cup \{0\}$ and $\alpha \in (0, 1]$. For conciseness of exposition, however, we limit ourselves to the case $n = 0, \alpha = 1$.

We proceed to quantize the weights of the approximating networks in Theorem 3.1 and bound the resulting quantization error. To this end, we first state an upper bound on the distance between two general ReLU network realizations, expressed in terms of W, L, and the distance between their associated configurations.

Lemma 3.2. *Let $W, L, \ell \in \mathbb{N}$ with $\ell \leq L$, and let*

$$\Phi^i = ((A_j^i, b_j^i))_{j=1}^\ell \in \mathcal{N}(W, L, 1), \quad i = 1, 2,$$

[5] A function $f : [0,1] \mapsto \mathbb{R}$ is said to be Hölder continuous with exponent α, if there exists $C \in \mathbb{R}_+$ such that $|f(x) - f(y)| \leq C|x - y|^\alpha$.

have the same architecture. Then,

$$\|R(\Phi^1) - R(\Phi^2)\|_{L^\infty([0,1])} \leq L(W+1)^L \|\Phi^1 - \Phi^2\|, \tag{57}$$

where

$$\|\Phi^1 - \Phi^2\| := \max_{j=1,\dots,\ell} \max\big\{\|A_j^1 - A_j^2\|_\infty, \|b_j^1 - b_j^2\|_\infty\big\}. \tag{58}$$

Proof. See Appendix E. □

This result now allows us to bound the error incurred by replacing real-valued network weights with values in a finite set $\mathbb{A} \subseteq \mathbb{R}$.

Proposition 3.3. *Let $W, L \in \mathbb{N}$ and consider the finite set $\mathbb{A} \subseteq \mathbb{R}$ with $\mathbb{A} \cap [-1,1] \neq \emptyset$. It holds that*

$$\begin{aligned} \mathcal{A}_\infty(\mathcal{R}(W,L,1), \mathcal{R}_{\mathbb{A}}(W,L)) &\leq L(W+1)^L \mathcal{A}([-1,1], \mathbb{A}\cap[-1,1], |\cdot|) && (59)\\ &\leq 2L(W+1)^L \mathcal{A}([-1,1], \mathbb{A}, |\cdot|). && (60) \end{aligned}$$

In particular, for all $a, b \in \mathbb{N}$, we have $\mathcal{A}([-1,1], \mathbb{Q}_b^a \cap [-1,1], |\cdot|) \leq 2^{-b}$, and

$$\mathcal{A}_\infty(\mathcal{R}(W,L,1), \mathcal{R}_b^a(W,L)) \leq L(W+1)^L 2^{-b}. \tag{61}$$

Proof. We start by proving (59)-(60) and note that

$$\begin{aligned} \mathcal{A}([-1,1], \mathbb{A}\cap[-1,1], |\cdot|) &= \sup_{x\in[-1,1]} \inf_{y\in\mathbb{A}\cap[-1,1]} |x-y| && (62)\\ &= \sup_{x\in[-1,1]} \min_{y\in\mathbb{A}\cap[-1,1]} |x-y|, && (63) \end{aligned}$$

where in (63) we used that $\mathbb{A}\cap[-1,1]$ is finite. Operationally, (62)-(63) says that every $x \in [-1,1]$ can be quantized into an element $q(x) \in \mathbb{A}\cap[-1,1]$, depending on x, such that $|x - q(x)| \leq \mathcal{A}([-1,1], \mathbb{A}\cap[-1,1], |\cdot|)$. This induces a mapping $q : [-1,1] \mapsto \mathbb{A}\cap[-1,1]$. We can now conclude that, for every $\Phi \in \mathcal{N}(W,L,1)$, application of the mapping q to each entry in Φ yields a corresponding quantized version $Q(\Phi) \in \mathcal{N}_{\mathbb{A}\cap[-1,1]}(W,L)$ such that the weights of $Q(\Phi)$ differ from those in Φ by no more than $\mathcal{A}([-1,1], \mathbb{A}\cap[-1,1], |\cdot|)$. This induces a mapping $Q : \mathcal{N}(W,L,1) \mapsto \mathcal{N}_{\mathbb{A}\cap[-1,1]}(W,L)$, satisfying

$$\|\Phi - Q(\Phi)\| \leq \mathcal{A}([-1,1], \mathbb{A}\cap[-1,1], |\cdot|), \tag{64}$$

with $\|\Phi - Q(\Phi)\|$ defined according to (58). We next establish (59) through the following chain of arguments

$$\begin{aligned} \mathcal{A}_\infty(\mathcal{R}(W,L,1), \mathcal{R}_{\mathbb{A}}(W,L)) &= \sup_{f\in\mathcal{R}(W,L,1)} \inf_{\tilde{f}\in\mathcal{R}_{\mathbb{A}}(W,L)} \|f - \tilde{f}\|_{L^\infty([0,1])} && (65)\\ &= \sup_{\Phi\in\mathcal{N}(W,L,1)} \inf_{\widetilde{\Phi}\in\mathcal{N}_{\mathbb{A}}(W,L)} \|R(\Phi) - R(\widetilde{\Phi})\|_{L^\infty([0,1])} && (66)\\ &\leq \sup_{\Phi\in\mathcal{N}(W,L,1)} \|R(\Phi) - R(Q(\Phi))\|_{L^\infty([0,1])} && (67)\\ &\leq \sup_{\Phi\in\mathcal{N}(W,L,1)} L(W+1)^L \|\Phi - Q(\Phi)\| && (68)\\ &\leq L(W+1)^L \mathcal{A}([-1,1], \mathbb{A}\cap[-1,1], |\cdot|), && (69) \end{aligned}$$

where in (67) we used $Q(\Phi) \in \mathcal{R}_{\mathbb{A}\cap[-1,1]}(W,L) \subseteq \mathcal{R}_{\mathbb{A}}(W,L)$, for all $\Phi \in \mathcal{N}(W,L,1)$, (68) follows from Lemma 3.2 with $\Phi^1 = \Phi$ and $\Phi^2 = Q(\Phi) \in \mathcal{R}_{\mathbb{A}\cap[-1,1]}(W,L) \subseteq \mathcal{R}(W,L,1)$, and in (69) we employed (64). Then, (60) follows by combining (59) with the relation

$$\mathcal{A}([-1,1], \mathbb{A}\cap[-1,1], |\cdot|) \leq 2\mathcal{A}([-1,1], \mathbb{A}, |\cdot|),$$

established in Lemma H.7.

We proceed to the derivation of (61). Fix $a, b \in \mathbb{N}$. Let $\tilde{q} : [-1,1] \mapsto \mathbb{Q}_b^a \cap [-1,1]$ be given by

$$\tilde{q}(x) = \begin{cases} 2^{-b}\lceil 2^b x\rceil, & \text{if } x \leq 0, \\ 2^{-b}\lfloor 2^b x\rfloor, & \text{if } x > 0, \end{cases}$$

and note that

$$|x - \tilde{q}(x)| \leq 2^{-b}, \text{ for all } x \in [-1,1]. \tag{70}$$

We have

$$\mathcal{A}([-1,1], \mathbb{Q}_b^a\cap[-1,1], |\cdot|) = \sup_{x\in[-1,1]} \inf_{y\in\mathbb{Q}_b^a\cap[-1,1]} |x-y| \leq \sup_{x\in[-1,1]} |x-\tilde{q}(x)| \leq 2^{-b}, \tag{71}$$

where in the first inequality we used that $\tilde{q}(x) \in \mathbb{Q}_b^a$, and the second inequality follows from (70). Evaluating (65)-(69) with $\mathbb{A} = \mathbb{Q}_b^a$, upon noting that the prerequisite $\mathbb{Q}_b^a \cap [-1,1] \neq \emptyset$ is satisfied as $0 \in \mathbb{Q}_b^a \cap [-1,1]$, and using (71), finally yields (61). □

The quantization error upper bound (61) does not depend on a as the weight magnitude of the networks we consider is bounded by 1.

We are now ready to characterize the minimax error upper bound in the approximation of $H^1([0,1])$ through ReLU networks with quantized weights.

Proposition 3.4. *There exist absolute constants $C_1, D_1, E_1 \in \mathbb{R}_+$, with $D_1 \geq 2$, such that for $W, L \in \mathbb{N}$, with $W, L \geq D_1$,*

$$\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}^1_{\lceil E_1 L\log(W)\rceil}(W,L)) \leq C_1(W^2L^2\log(W))^{-1}. \tag{72}$$

Proof. Let C, D be the absolute constants specified in Theorem 3.1. Set $C_1 = C+1$, $D_1 = \max\{2, D\}$, and $E_1 = 12$. Then, for $W, L \geq D_1$,

$$\begin{aligned}
&\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}^1_{\lceil E_1 L\log(W)\rceil}(W,L)) && (73)\\
&\leq \mathcal{A}_\infty(H^1([0,1]), \mathcal{R}(W,L,1)) + \mathcal{A}_\infty(\mathcal{R}(W,L,1), \mathcal{R}^1_{\lceil E_1 L\log(W)\rceil}(W,L)) && (74)\\
&\leq C(W^2L^2\log(W))^{-1} + L(W+1)^L 2^{-\lceil E_1 L\log(W)\rceil} && (75)\\
&\leq (C+1)(W^2L^2\log(W))^{-1}, && (76)
\end{aligned}$$

where (74) follows from the triangle inequality (54) with $b = \lceil E_1 L\log(W)\rceil$, in (75) we used Theorem 3.1 with $W, L \geq D_1 \geq D$ and Proposition 3.3, and (76) is by

$$2^{-\lceil E_1 L\log(W)\rceil} = 2^{-\lceil 12L\log(W)\rceil} \leq ((W^{2L})^6)^{-1} \leq (W^2L^3\log(W)(W+1)^L)^{-1}, \tag{77}$$

as $W^{2L} \geq W^2$, $(W^{2L})^3 \geq L^3$, $W^{2L} \geq \log(W)$, and $W^{2L} \geq (W+1)^L$, all owing to $W, L \geq D_1 \geq 2$. □

As $\mathcal{R}^1_{\lceil E_1 L\log(W)\rceil}(W,L)$ is finite, we have

$$\begin{aligned}\mathcal{A}_\infty(H^1([0,1]),\mathcal{R}^1_{\lceil E_1 L\log(W)\rceil}(W,L)) &= \sup_{g\in H^1([0,1])}\inf_{f\in\mathcal{R}^1_{\lceil E_1 L\log(W)\rceil}(W,L)}\|g-f\|_{L^\infty([0,1])}\\ &= \sup_{g\in H^1([0,1])}\min_{f\in\mathcal{R}^1_{\lceil E_1 L\log(W)\rceil}(W,L)}\|g-f\|_{L^\infty([0,1])},\end{aligned}$$

and, therefore, for every $g\in H^1([0,1])$, there exists a network realization $f\in\mathcal{R}^1_{\lceil E_1 L\log(W)\rceil}(W,L)$ such that $\|g-f\|_{L^\infty([0,1])}\leq C_1(W^2L^2\log(W))^{-1}$. Indeed, the explicit construction of this f can be inferred from the proofs of Theorem 3.1 and Proposition 3.3.

We conclude the discussion by arguing that the construction underlying Proposition 3.4 achieves memory optimality. To see this, note that starting from (72), we get

$$\begin{aligned}\mathcal{A}_\infty(H^1([0,1]),\mathcal{R}^1_{\lceil E_1 L\log(W)\rceil}(W,L)) &\leq C_1(W^2L^2\log(W))^{-1} &(78)\\ &= C_1\frac{\lceil E_1 L\log(W)\rceil}{L\log(W)}(W^2L\lceil E_1 L\log(W)\rceil)^{-1} &(79)\\ &\leq C_1(E_1+1)(W^2L\lceil E_1 L\log(W)\rceil)^{-1}, &(80)\end{aligned}$$

where in the last inequality we used $L\log(W)\geq 1$ and $\lceil E_1 L\log(W)\rceil\leq E_1 L\log(W)+1$. Hence, for $b=\lceil E_1 L\log(W)\rceil$, the minimax error lower bound in Proposition 2.9 incurred by the minimum memory requirement is attained to within a multiplicative factor. More formally, it follows from Proposition 2.10 with $D=C_1(E_1+1)$ and $\mathcal{I}=\{(W,L,\lceil E_1 L\log(W)\rceil)\in\mathbb{N}^3 : W,L\geq D_1\}$ that $\{\mathcal{R}^1_{\lceil E_1 L\log(W)\rceil}(W,L)) : W,L\geq D_1\}$ achieves memory optimality in the approximation of functions in $H^1([0,1])$.

4 Depth-Precision Tradeoff

The bound in Proposition 3.4 applies to a fixed choice of b as a function of network width W and depth L according to $b=\lceil E_1 L\log(W)\rceil$. We now relax this dependency with the aim of obtaining a more flexible, and for certain parameter choices tighter, upper bound. In the process, we address a question of practical interest, namely "Can we realize neural networks with high precision weights by equivalent deeper networks of lower precision weights? If so, what would the impact of such a transformation on a potential memory optimality of the initial network be?" This question is of significant interest as high-resolution quantization is difficult to realize in electronic circuits. The idea of sigma-delta conversion [1] exemplifies this principle, in the context of analog-to-digital (A/D) conversion, by trading oversampling rate for quantization resolution. Here, we can trade network depth for quantization resolution. The next result expresses the corresponding depth-precision tradeoff in a formal manner.

Proposition 4.1. *Let $W,L,k\in\mathbb{N}$. For all $a,b\in\mathbb{N}$, we have*

$$\mathcal{R}^{ka}_{kb}(W,L)\subseteq\mathcal{R}^a_b(16W,(k+2)L). \tag{81}$$

Proof. See Appendix F. □

Proposition 4.1 states that (high-precision) networks in $\mathcal{R}^{ka}_{kb}(W,L)$ can equivalently be realized by lower-precision networks, specifically by networks in $\mathcal{R}^a_b(16W,(k+2)L)$, at the expense

of an increase in network width and depth. We emphasize that while the increase in width is by a constant factor, a reduction in network weight precision by a factor of k leads to a k-fold increase in depth. The proof of Proposition 4.1 is constructive, meaning that for a given network configuration $\Phi_1 \in \mathcal{N}_{kb}^{ka}(W, L)$, we explicitly specify a $\Phi_2 \in \mathcal{N}_b^a(16W, (k+2)L)$ such that $R(\Phi_1) = R(\Phi_2)$. In addition, the proof applies to more general weight sets, input and output dimensions than those assumed in the statement of Proposition 4.1.

An important property of the depth-precision tradeoff just identified resides in the fact that it essentially preserves memory consumption behavior. Concretely, it follows from Proposition 2.4 that network realizations $f \in \mathcal{R}_{kb}^{ka}(W, L)$ are uniquely specified by no more than $10W^2Lk(a+b)$ bits while the corresponding equivalent realizations in $\mathcal{R}_b^a(16W, (k+2)L)$ require at most $10(16W)^2(k+2)L(a+b)$ bits. Replacing a given (high-precision) network by a deep lower-precision network hence comes at the cost, in the number of bits needed, of at most a multiplicative constant factor of $3 \cdot 16^2$, while the scaling behavior in W, L, k, a, b is preserved.

We now show that this insight allows us to conclude that the (constructive) transformation from high-precision to deeper low-precision networks effected by Proposition 4.1 preserves memory optimality. Specifically, we recall that, by Proposition 3.4, $\{\mathcal{R}^1_{\lceil E_1 L \log(W)\rceil}(W, L) : W, L \geq D_1\}$ achieves memory optimality in the approximation of functions in $H^1([0,1])$ with

$$\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}^1_{\lceil E_1 L \log(W)\rceil}(W, L)) \leq C_1(W^2L^2\log(W))^{-1}, \tag{82}$$

where C_1, D_1, and E_1 are the absolute constants specified in Proposition 3.4. Suppose now that we want to replace the (high-precision) weights of network configurations in $\mathcal{N}^1_{\lceil E_1 L \log(W)\rceil}(W, L)$, which realize functions in $\mathcal{R}^1_{\lceil E_1 L \log(W)\rceil}(W, L)$, by lower-precision weights, say in $\mathbb{Q}^1_b$, with $b < \lceil E_1 L \log(W)\rceil$. Specifically, with $k = \left\lceil \frac{\lceil E_1 L \log(W)\rceil}{b} \right\rceil \geq 2$, we get $kb \geq \lceil E_1 L \log(W)\rceil$, and therefore

$$\mathcal{R}^1_b(16W, (k+2)L) \supseteq \mathcal{R}^k_{kb}(W, L) \supseteq \mathcal{R}^1_{\lceil E_1 L \log(W)\rceil}(W, L), \tag{83}$$

where (reading from the left) the first inclusion is a consequence of Proposition 4.1 and the second follows directly from $kb \geq \lceil E_1 L \log(W)\rceil$ and $k \geq 2$. In summary, we obtain

$$\begin{aligned}
&\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}^1_b(16W, (k+2)L)) && (84)\\
&\leq \mathcal{A}_\infty(H^1([0,1]), \mathcal{R}^1_{\lceil E_1 L \log(W)\rceil}(W, L)) && (85)\\
&\leq C_1(W^2L^2\log(W))^{-1} && (86)\\
&= \frac{256C_1(k+2)b}{L\log(W)}((16W)^2(k+2)L\,b)^{-1} && (87)\\
&\leq \frac{2 \cdot 256C_1kb}{L\log(W)}((16W)^2(k+2)L\,b)^{-1} && (88)\\
&\leq 4 \cdot 256C_1\frac{\lceil E_1 L \log(W)\rceil}{L\log(W)}((16W)^2(k+2)L\,b)^{-1} && (89)\\
&\leq 4 \cdot 256C_1(E_1+1)((16W)^2(k+2)L\,b)^{-1} && (90)
\end{aligned}$$

where (86) follows from (82), in (88) we used $k \geq 2$, (89) is a consequence of $kb = \left\lceil \frac{\lceil E_1 L \log(W)\rceil}{b} \right\rceil b \leq 2\frac{\lceil E_1 L \log(W)\rceil}{b} b = 2\lceil E_1 L \log(W)\rceil$, as $\lceil x\rceil \leq 2x$ for $x \geq 1$, and (90) is by $\lceil E_1 L \log(W)\rceil \leq E_1 L \log(W) + 1 \leq (E_1+1)L\log(W)$ as $W, L \geq D_1 \geq 2$. It then follows from Proposition 2.10 with $D = 4 \cdot 256C_1(E_1+1)$ and $\mathcal{I} = \{(16W, (\left\lceil \frac{\lceil E_1 L \log(W)\rceil}{b} \right\rceil + 2)L, b) \in \mathbb{N}^3 : W, L \geq D_1 \text{ and } b <$

$\lceil E_1 L \log(W) \rceil\}$ that $\{\mathcal{R}_b^1(16W, (\lceil \frac{\lceil E_1 L \log(W) \rceil}{b} \rceil + 2)L)) : W, L \geq D_1$ and $b < \lceil E_1 L \log(W) \rceil\}$ achieves memory optimality in the approximation of $H^1([0,1])$.

This shows, as announced, that the transformation from high-to-low precision networks effected by the construction in the proof of Proposition 4.1, indeed, preserves memory optimality.

5 The Three Quantization Regimes

Putting the upper bounds we have obtained together, with some refinement for simplicity in presentation, yields the following combined minimax error upper bound exhibiting three different quantization regimes.

Theorem 5.1. *There exist absolute constants $D_2, C_2, E_{2,1}, E_{2,2}, \alpha \in \mathbb{R}_+$, with $\alpha > 1$, such that, for $W, L, b \in \mathbb{N}$, with $W, L \geq D_2$, we have $E_{2,2}\frac{\log(W)}{L} < E_{2,1} L \log(W)$, and the following statements hold:*

1. *In the under-quantization regime*[6]*, i.e., $b \in [1, E_{2,2}\frac{\log(W)}{L})$, we have*

$$\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}_b^1(W, L)) \leq C_2 \alpha^{-Lb}.$$

2. *In the proper-quantization regime, i.e., $b \in [E_{2,2}\frac{\log(W)}{L}, E_{2,1} L \log(W))$, we have*

$$\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}_b^1(W, L)) \leq C_2 (W^2 L b)^{-1}.$$

3. *In the over-quantization regime, i.e., $b \in [E_{2,1} L \log(W), \infty)$, we have*

$$\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}_b^1(W, L)) \leq C_2 (W^2 L^2 \log(W))^{-1}.$$

Proof. See Appendix G. □

Combining Theorem 5.1 with Corollary 2.14 now yields three different quantization regimes in terms of b as a function of network width W and depth L. These regimes exhibit markedly different minimax error behavior, namely, exponential decrease, polynomial decrease, and constant, mirroring what was already indicated by the combined minimax error lower bound in Corollary 2.14. While the delineation of the three regimes was left vague in the context of the lower bound, the fact that the upper bound in Theorem 5.1 exhausts the parameter range for b allows us to make the transition boundaries more precise.

1. The *under-quantization regime*: For $b \in [1, E_{2,2}\frac{\log(W)}{L})$, the minimax error satisfies

$$c_\ell 2^{-Lb} \leq \mathcal{A}_\infty(H^1([0,1]), \mathcal{R}_b^1(W, L)) \leq C_2 \alpha^{-Lb},$$

and hence falls into a band that decreases exponentially in b. This behavior emerges as a consequence of the main limiting factor in the neural network approximation of $H^1([0,1])$ being given by the numerical precision of the quantized network weights.

[6] We use the convention $\left[1, E_{2,2}\frac{\log(W)}{L}\right) = \emptyset$, if $E_{2,2}\frac{\log(W)}{L} \leq 1$, that is the *under-quantization regime* can be empty.

2. The *proper-quantization regime*: For $b \in [E_{2,2}\frac{\log(W)}{L}, E_{2,1}L\log(W))$, the minimax error satisfies
$$c_\ell(W^2Lb)^{-1} \leq \mathcal{A}_\infty(H^1([0,1]), \mathcal{R}_b^1(W,L)) \leq C_2(W^2Lb)^{-1},$$
and is therefore contained in a polynomially decreasing band. Notably, by Proposition 2.10, in this regime $\{\mathcal{R}_b^1(W,L)) : W, L \geq D_2, b \in [E_{2,2}\frac{\log(W)}{L}, E_{2,1}L\log(W))\}$ achieves memory optimality—in the sense of Definition 2.3—in the approximation of $H^1([0,1])$.

3. The *over-quantization regime*: For $b \in [E_{2,1}L\log(W), \infty)$, the minimax error satisfies
$$c_\ell(W^2L^2(\log(W)+\log(L)))^{-1} \leq \mathcal{A}_\infty(H^1([0,1]), \mathcal{R}_b^1(W,L)) \leq C_2(W^2L^2\log(W))^{-1},$$
and hence resides in a band between two constants (w.r.t. b).

These results also provide guidance on the choice of network architectures in practical applications. Specifically, assume that one operates under a total memory budget for the storage of approximating neural networks in $\mathcal{R}_b^1(W,L)$. Recalling that $b+3$ bits are required to store an individual network weight, the overall fixed bit budget is given by $n := W^2L(b+3)$. Seeking memory optimality, we now ask which choices of W and L maximize the size of the proper-quantization regime. With $L = \frac{n}{W^2(b+3)}$, it follows that $\left[E_{2,2}\frac{\log(W)}{L}, E_{2,1}L\log(W)\right) = \left[\frac{E_{2,2}\log(W)W^2(b+3)}{n}, \frac{E_{2,1}\log(W)n}{W^2(b+3)}\right)$, which shows that for n and b, and hence W^2L, fixed, deep networks will result in larger proper-quantization regimes than wide networks. This insight adds to the existing literature on depth-width tradeoffs in neural network approximation, see, e.g., [13, 14, 18].

A Proof of Proposition 2.11

We first describe the main ingredients of the proof, starting with the definition of VC dimension.

Definition A.1. *[12, Definition 1] Let H denote a class of functions mapping from $\mathcal{X}$ to $\{0,1\}$. Define the growth function as*
$$\Pi_H(m) := \max_{x_1,\ldots,x_m \in \mathcal{X}} |\{(h(x_1),\ldots,h(x_m)) : h \in H\}|, \quad \text{for } m \in \mathbb{N}.$$
For a given set $\{x_1,\ldots,x_m\} \subseteq \mathcal{X}$, if $|\{(h(x_1),\ldots,h(x_m)) : h \in H\}| = 2^m$, we say that H shatters $\{x_1,\ldots,x_m\}$. The Vapnik-Chervonenkis (VC) dimension of H, denoted by $VCdim(H)$, is the largest m such that $\Pi_H(m) = 2^m$. If there is no such largest m, we set $VCdim(H) = \infty$.

VC dimension upper bounds for certain families of ReLU networks are provided in [12]. The results in [12] apply, however, only to families of network realizations whose associated configurations have a fixed architecture, whereas $\mathcal{N}(W,L)$, the object of interest here, consists of network configurations with different architectures. The following result shows how [12, Eq. (2)] can be adapted to this setting.

Lemma A.2. *For all $W \in \mathbb{N}$ and $L \in \mathbb{N}$, with $L \geq 2$, we have*
$$VCdim\left(sgn \circ \mathcal{R}(W,L)\right) \leq C_h W^2L^2(\log(W)+\log(L)), \tag{91}$$
where $sgn \circ \mathcal{R}(W,L) := \{sgn \circ f : f \in \mathcal{R}(W,L)\}$ and C_h is an absolute constant.

Proof. Fix $W \in \mathbb{N}$ and $L \in \mathbb{N}$, with $L \geq 2$, throughout the proof. Consider the set $\mathcal{N}^*(2W, L) = \{(A_\ell, b_\ell)_{\ell=1}^L : A_1 \in \mathbb{R}^{2W\times 1}, b_1 \in \mathbb{R}^{2W}, A_L \in \mathbb{R}^{1\times 2W}, b_L \in \mathbb{R}^1, A_\ell \in \mathbb{R}^{2W\times 2W}, b_\ell \in \mathbb{R}^{2W}, \text{ for } \ell \in \{2, \dots, L-1\}\}$ consisting of all network configurations with the fixed architecture

$$(N_\ell)_{\ell=0}^L = (1, \underbrace{2W, \dots,}_{\text{repeats } (L-1) \text{ times}} 1). \tag{92}$$

The associated family of network realizations is $\mathcal{R}^*(2W, L) = \{R(\Phi) : \Phi \in \mathcal{N}^*(2W, L)\}$. We note that the network configurations in $\mathcal{N}^*(2W, L)$ have $n(2W, L) := 6W + 1 + (L-2)((2W)^2 + 2W)$ weights. As $\mathcal{R}^*(2W, L)$ consists of realizations of network configurations with the fixed architecture (92), we can apply the results in [12]. Specifically, it follows that

$$\begin{aligned} \mathrm{VCdim}(\mathrm{sgn} \circ \mathcal{R}^*(2W, L)) &\leq Cn(2W, L)L \log(n(2W, L)) && (93)\\ &\leq C(13W^2L)L \log(13W^2L) && (94)\\ &\leq 104\, CW^2L^2(\log(W) + \log(L)), && (95) \end{aligned}$$

where in (93) we used [12, Eq.(2)] with $C \in \mathbb{R}_+$ an absolute constant, (94) follows from $n(2W, L) \leq 13W^2L$, and (95) is owing to $\log(13W^2L) \leq \log((WL)^8) = 8(\log(W) + \log(L))$.

We continue by showing that $R(W, L) \subseteq R^*(2W, L)$, which will then allow us to conclude that (95) also upper-bounds $\mathrm{VCdim}(\mathrm{sgn} \circ \mathcal{R}(W, L))$. To this end, fix an $f \in R(W, L)$. It follows from Lemma H.2 that there exists a network configuration $\overline{\Phi} = (\overline{A}_\ell, \overline{b}_\ell)_{\ell=1}^L \in \mathcal{N}(\max\{W, 2\}, L) \subseteq \mathcal{N}(2W, L)$ such that $R(\overline{\Phi}) = f$. Next, we enlarge the layers of $\overline{\Phi}$ such that the resulting configuration has the architecture (92) while realizing the same function f. To this end, denote the architecture of $\overline{\Phi}$ by $(\overline{N}_\ell)_{\ell=0}^L$, and note that $\overline{N}_0 = 1 = N_0$, $\overline{N}_L = 1 = N_L$, and $\overline{N}_\ell \leq 2W = N_\ell$, for $\ell \in \{2, \dots, L-1\}$. Now augment the configuration Φ to the architecture in (92) according to $\widetilde{\Phi} = (\tilde{A}_\ell, \tilde{b}_\ell)_{\ell=1}^L \in \mathcal{N}^*(2W, L)$, with

$$\tilde{A}_\ell = \begin{pmatrix} \overline{A}_\ell & 0_{\overline{N}_\ell \times (N_{\ell-1} - \overline{N}_{\ell-1})} \\ 0_{(N_\ell - \overline{N}_\ell)\times \overline{N}_{\ell-1}} & 0_{(N_\ell - \overline{N}_\ell)\times(N_{\ell-1} - \overline{N}_{\ell-1})} \end{pmatrix}, \quad \tilde{b}_\ell = \begin{pmatrix} \overline{b}_\ell \\ 0_{N_\ell - \overline{N}_\ell} \end{pmatrix}, \quad \ell = 1, \dots, L.$$

We then have $R(\widetilde{\Phi}) = R(\overline{\Phi})$ and thereby $f = R(\widetilde{\Phi}) = R(\overline{\Phi}) \in R^*(2W, L)$. As the choice of $f \in R(W, L)$ was arbitrary, we have established that $R(W, L) \subseteq R^*(2W, L)$ as announced. We can finally conclude that

$$\mathrm{VCdim}(\mathrm{sgn} \circ \mathcal{R}(W, L)) \leq \mathrm{VCdim}(\mathrm{sgn} \circ \mathcal{R}^*(2W, L)) \leq 104CW^2L^2(\log(W) + \log(L)),$$

where in the last inequality we used (93)-(95). The proof is concluded by setting $C_h := 104C$. □

With the VC dimension upper bound (91), we are now ready to proceed to the proof of Proposition 2.11.

Proof of Proposition 2.11. Set $c_v = (4(C_h + 1))^{-1}$, where C_h is the constant in (91). Suppose, for the sake of contradiction, that the approximation error lower bound (49) does not hold with this c_v. This would then imply the existence of a $W \in \mathbb{N}$ and an $L \in \mathbb{N}$, with $L \geq 2$, such that

$$\mathcal{A}_\infty(H^1([0, 1]), \mathcal{R}(W, L)) < c_v(W^2L^2(\log(W) + \log(L)))^{-1}. \tag{96}$$

Let $N = \lceil C_h W^2 L^2 (\log(W) + \log(L)) \rceil$. We shall show that (96) implies that the family $\operatorname{sgn} \circ \mathcal{R}(W, L)$ shatters the set $\{\frac{0}{N}, \dots, \frac{N}{N}\}$, which will then lead to a contradiction to (91). To this end, fix $(\theta_0, \dots, \theta_N) \in \{0,1\}^{N+1}$, and let $f \in H^1([0,1])$ be given by

$$f(x) = (\theta_i - \theta_{i-1})\left(x - \frac{i-1}{N}\right) + \frac{2\theta_{i-1} - 1}{2N}, \quad x \in \left[\frac{i-1}{N}, \frac{i}{N}\right], \ i = 1, \dots, N,$$

such that, for $i = 0, \dots, N$,

$$f\left(\frac{i}{N}\right) = \frac{2\theta_i - 1}{2N} = \begin{cases} \frac{1}{2N}, & \text{if } \theta_i = 1, \\ -\frac{1}{2N}, & \text{if } \theta_i = 0. \end{cases} \tag{97}$$

Then, according to (96), there exists a ReLU network realization $g \in \mathcal{R}(W, L)$ such that $\|g - f\|_{L^\infty([0,1])} < c_v (W^2 L^2 (\log(W) + \log(L)))^{-1} = \frac{1}{4}((C_h + 1) W^2 L^2 (\log(W) + \log(L)))^{-1} \leq \frac{1}{4}(\lceil C_h W^2 L^2 (\log(W) + \log(L)) \rceil)^{-1} = \frac{1}{4N}$, which in combination with (97) implies that, for $i = 0, \dots, N$,

$$g\left(\frac{i}{N}\right) \begin{cases} > 0, & \text{if } \theta_i = 1, \\ < 0, & \text{if } \theta_i = 0, \end{cases}$$

and therefore $(\operatorname{sgn} \circ g)(\frac{i}{N}) = \theta_i$, for $i = 0, \dots, N$. Upon noting that $(\operatorname{sgn} \circ g) \in \operatorname{sgn} \circ \mathcal{R}(W, L)$, we have

$$(\theta_0, \dots, \theta_N) \in \left\{ \left(h\left(\frac{0}{N}\right), \dots, h\left(\frac{N}{N}\right) \right) : h \in \operatorname{sgn} \circ \mathcal{R}(W, L) \right\}.$$

Since the choice of $\{\theta_0, \dots, \theta_N\} \in \{0,1\}^{N+1}$ was arbitrary, we have, indeed, established that

$$\{0,1\}^{N+1} \subseteq \left\{ \left(h\left(\frac{0}{N}\right), \dots, h\left(\frac{N}{N}\right) \right) : h \in \operatorname{sgn} \circ \mathcal{R}(W, L) \right\},$$

and therefore $|\{(h(\frac{0}{N}), \dots, h(\frac{N}{N})) : h \in \operatorname{sgn} \circ \mathcal{R}(W, L)\}| = |\{0,1\}^{N+1}| = 2^{N+1}$. This proves that $\operatorname{sgn} \circ \mathcal{R}(W, L)$ shatters the set $\{\frac{0}{N}, \dots, \frac{N}{N}\}$, which in turn leads to

$$\operatorname{VCdim}(\operatorname{sgn} \circ \mathcal{R}(W, L)) \geq \left| \left\{ \frac{0}{N}, \dots, \frac{N}{N} \right\} \right| = N + 1 > C_h W^2 L^2 (\log(W) + \log(L)),$$

and thus stands in contradiction to the VC dimension upper bound (91). Therefore, (49) must hold.

Upon noting that $\mathcal{R}_{\mathbb{A}}(W, L) \subseteq \mathcal{R}(W, L)$, which follows from $\mathcal{N}_{\mathbb{A}}(W, L) \subseteq \mathcal{N}(W, L)$, we obtain (50) from (49). This concludes the proof. □

B Proof of Theorem 3.1

We prove Theorem 3.1 with $\mathcal{R}(W, L, 1)$ replaced by $\mathcal{R}(W, L, W^K)$, where $K \in \mathbb{N}$ is an absolute constant, i.e., for networks with weight magnitude growing polynomially in W, and then relax this polynomial dependency using Proposition H.4.

Proposition B.1. *There exist absolute constants $C_a, D_a \in \mathbb{R}_+$ and an absolute constant $K \in \mathbb{N}$, such that for all $W, L \in \mathbb{N}$ with $W, L \geq D_a$,*

$$\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}(W, L, W^K)) \leq C_a (W^2 L^2 \log(W))^{-1}. \tag{98}$$

Proof. See Appendix B.1 for preparatory material and then Appendix B.2 for the actual proof. □

The proof of Theorem 3.1 is now effected by applying Proposition H.4 along with Proposition B.1 as follows. Set $D = \max\{(4K+2)\lceil D_a \rceil, 10\}$, where D_a is the constant specified in Proposition B.1. Fix $W, L \in \mathbb{N}$ with $W, L \geq D$. Let $U = \lfloor \frac{L}{2K+1} \rfloor$, ensuring that $(2K+1)U \leq L$ and $U \geq \lfloor \frac{D}{2K+1} \rfloor \geq \lfloor \frac{(4K+2)\lceil D_a \rceil}{2K+1} \rfloor \geq D_a$. We have

$$\begin{aligned} \mathcal{R}(W, U, W^K) &\subseteq R(W, (2K+1)U, 1) && (99)\\ &\subseteq \mathcal{R}(W, L, 1), && (100) \end{aligned}$$

where (99) is a consequence of Proposition H.4 with (W, L, L', B, B') replaced by $(W, U, 2KU, W^K, 1)$ and the prerequisite satisfied as $\frac{(\lfloor W/2 \rfloor)^{2KU}}{(W^K)^U} \geq \frac{W^{KU}}{W^{KU}} = 1$ thanks to $(\lfloor x/2 \rfloor)^2 \geq x$, for $x = W \geq D = \max\{(4K+2)\lceil D_a \rceil, 10\} \geq 10$, and in (100) we used $(2K+1)U \leq L$. We now get

$$\begin{aligned} \mathcal{A}_\infty(H^1([0,1]), \mathcal{R}(W, L, 1)) &\leq \mathcal{A}_\infty(H^1([0,1]), \mathcal{R}(W, U, W^K)) && (101)\\ &\leq C_a (W^2 U^2 \log(W))^{-1} && (102)\\ &\leq C_a \Big(W^2 \Big(\frac{1}{2} \cdot \frac{L}{2K+1} \Big)^2 \log(W) \Big)^{-1} && (103)\\ &= (4K+2)^2 C_a (W^2 L^2 \log(W))^{-1}, && (104) \end{aligned}$$

where in (101) we used the inclusion (99)-(100), in (102) we applied Proposition B.1 with C_a as specified in Proposition B.1, and (103) follows from $U = \lfloor \frac{L}{2K+1} \rfloor \geq \frac{1}{2} \cdot \frac{L}{2K+1}$, upon noting that $U \geq D_a > 0$. The proof is finalized by taking $C = (4K+2)^2 C_a$.

B.1 Preparation for the Proof of Proposition B.1

The proof of Proposition B.1 is based on two specific ingredients, namely the realization of one-dimensional bounded piecewise linear functions by ReLU networks and the bit extraction technique.

B.1.1 Realizing One-Dimensional Bounded Piecewise Linear Functions by ReLU Networks

We start with the definition of one-dimensional bounded piecewise linear functions.

Definition B.2 (One-dimensional bounded piecewise linear functions)**.** *Let $M \in \mathbb{N}$, with $M \geq 3$, $E \in \mathbb{R}_+ \cup \{\infty\}$, and let $X = (x_i)_{i=0}^{M-1}$ be a strictly increasing sequence taking values in $\mathbb{R}$. Define the set of functions*

$$\begin{aligned} \Sigma(X, E) = \big\{ f \in C(\mathbb{R}) : \|f\|_{L^\infty(\mathbb{R})} \leq E, f \text{ is constant on } (-\infty, x_0] \text{ and } [x_{M-1}, \infty), \\ f \text{ is affine on } [x_i, x_{i+1}],\ i = 0, \dots, M-2 \big\}. \end{aligned}$$

For a function $f \in \Sigma(X, E)$, we call X the set of breakpoints of f, as the slope of f can change only at these points. We refer to the intervals $(-\infty, x_0], [x_i, x_{i+1}]$, $i = 0, \dots, M-2, [x_{M-1}, \infty)$ as the piecewise linear regions of f.

We now show how one-dimensional bounded piecewise linear functions can be realized through ReLU networks while retaining control over the networks' weight magnitude. To avoid dealing with tedious corner cases, we restrict ourselves to $|X| \geq 3$ and $X \subseteq [0,1]$, which, as seen later, suffices to cover what is needed in the proof of Proposition B.1.

Proposition B.3. *Let $M \in \mathbb{N}$, with $M \geq 3$, $E \in \mathbb{R}_+$, and let $X = (x_i)_{i=0}^{M-1}$ be a strictly increasing sequence taking values in $[0,1]$. For all $u, v \in \mathbb{N}$, and $w \in \mathbb{R}$ with $w \geq 1$, such that*

$$u^2 v \geq M, \tag{105}$$
$$w^{30v} \geq M^6 (R_m(X))^4 E, \tag{106}$$

with $R_m(X) := \max_{i=1,\dots,M-1}(x_i - x_{i-1})^{-1}$, we have

$$\Sigma(X,E) \subseteq \mathcal{R}(20u, 30v, 2w).$$

Proof. The proof, detailed in Appendix C, is constructive, in the sense of explicitly specifying a network realizing a given $f \in \Sigma(X,E)$. □

Proposition B.3 makes the requirements on the width, depth, and weight magnitude of networks realizing functions $f \in \Sigma(X,E)$ explicit, in particular in two aspects. First, by (105) it suffices to choose the number of network weights, given by $W^2 L$ with $W = 20u$ and $L = 30v$, to be on the order of M (the number of breakpoints of f). Second, (106) rewritten as $w \geq (M^6 (R_m(X))^4 E)^{1/(30v)}$ shows that the weight magnitude $2w$ can be taken to grow no faster than polynomial in M, $R_m(X)$, and E, and is allowed to decrease with respect to network depth, $30v$, in an inverse exponential manner. While the first requirement is identical to those reported in [19, 20, 21], the second one is novel and constitutes a relaxation relative to the constructions available in the literature. Specifically, [19] shows that $\Sigma(X,\infty) \subseteq \mathcal{R}(W, L, \infty)$, with $W, L \in \mathbb{N}$ depending on X, in a way that does not allow for general conclusions on how the weight magnitude of the network depends on the function $f \in \Sigma(X,\infty)$ to be realized. Scrutinizing the proofs in [20, 21], one finds that the weight magnitude of the networks constructed therein has to be at least exponential in the number of breakpoints of the piecewise linear function realized; this stands in stark contrast to our construction which requires polynomial growth only.

B.1.2 Bit Extraction

Another important ingredient for the proof of Proposition B.1 is the bit extraction technique, as first introduced in [11] to derive a lower bound on the VC dimension of ReLU networks. A refinement of this technique improving the lower bound in [11] was reported in [12]. Further variants developed in the context of function approximation through ReLU networks can be found in [8, 9, 10, 13, 17].

There are two fundamental components constituting the bit extraction technique in all its variants. The first one encodes a string of elements from a finite alphabet into a real number. For example, [11] encodes the string $(\theta_i)_{i=1}^s \in \{0,1\}^s$ into the number $E((\theta_i)_{i=1}^s) = \sum_{i=1}^s 2^{-i}\theta_i$. The references [9, 12, 13] also work with $\{0,1\}$-alphabets, while [8, 10, 17] employ k-ary alphabets, $k \in \mathbb{N}$, with $k > 2$, but otherwise follow the philosophy of [11]. Specifically, all these approaches encode strings of k-ary digits into k-ary numbers. The second component consists of a decoder D, realized by a ReLU network, which extracts either individual elements of the string encoded

into the real number or functions thereof. For example, the decoder $D : \mathbb{R}^2 \mapsto \mathbb{R}$ in [11] extracts the individual θ_ℓ, $\ell = 1, \ldots, s$, from $E((\theta_i)_{i=1}^s)$ according to

$$D(E((\theta_i)_{i=1}^s), \ell) = \theta_\ell, \quad \ell = 1, \ldots, s,$$

whereas in [9] the sum of leading elements of the string $(\theta_i)_{i=1}^s$ is recovered through[7]

$$D(E((\theta_i)_{i=1}^s), k) = \sum_{i=1}^{k} \theta_i, \quad k = 0, \ldots, s. \tag{107}$$

The bit extraction technique we develop here is also based on the two constituents just described, but provides improvements and refinements—to be discussed as we go along—in ways that are fundamental to our purposes. We start with definitions and notation regarding the choice of alphabet and the encoding procedure.

Definition B.4. *Let $\mathbb{T}$ be the set of ternary numbers with reduced alphabet $\{0, 1\}$ and possibly infinitely many digits, formally,*

$$\mathbb{T} := \left\{ \sum_{i=1}^{\infty} \theta_i 3^{-i} : \theta_i \in \{0, 1\}, i \in \mathbb{N} \right\}.$$

Set $T((\theta_1, \ldots, \theta_s)) := \sum_{i=1}^s \theta_i 3^{-i}$, for $s \in \mathbb{N}$, $\theta_i \in \{0, 1, 2\}$, $i = 1, \ldots, s$.

The basic idea underlying the first component of our variation of the bit extraction technique is to encode $\{0, 1\}$-strings $(\theta_i)_{i=1}^s$, $s \in \mathbb{N}$, into ternary numbers according to $T((\theta_i)_{i=1}^s) = \sum_{i=1}^s \theta_i 3^{-i} \in \mathbb{T}$. The ReLU network realizing the decoder we employ is specified in the following result, whose proof is constructive.

Proposition B.5. *Let $N, L \in \mathbb{N}$. There exists a function*

$$F_{N,L} \in \mathcal{R}((2, 1), 2^{N+4}, 5L, 3^{N+2}) \tag{108}$$

such that for all $\sum_{i=1}^{\infty} \theta_i 3^{-i} \in \mathbb{T}$ and $k \in \mathbb{N} \cup \{0\}$,

$$F_{N,L}\left(\sum_{i=1}^{\infty} \theta_i 3^{-i}, k \right) = \sum_{i=1}^{\min\{NL, k\}} \theta_i, \tag{109}$$

where we use the convention $\sum_{i=1}^0 \theta_i = 0$. In particular, for $s \in \mathbb{N}$ such that $s \leq NL$, $\theta_i \in \{0, 1\}$, $i = 1, \ldots, s$, we have

$$F_{N,L}(T((\theta_1, \ldots, \theta_s)), k) = \sum_{i=1}^{k} \theta_i, \ k = 0, \ldots, s. \tag{110}$$

Proof. See Appendix D. □

[7] We use the convention $\sum_{i=1}^0 \theta_i = 0$.

It is worth emphasizing two aspects of the decoding ReLU network $F_{N,L}$ in Proposition B.5. First, the width of the network is determined solely by the parameter N, while its depth depends only on the parameter L, and both of these parameters can be chosen freely, and, in particular, independently of each other. Moreover, the weight magnitude of the network is polynomial in network width and does not depend on network depth. This control over the weight magnitude sets our result apart from the existing literature, and is essential for the proof of Theorem B.1. Specifically, the decoding networks in [8, 10, 11, 13, 17] all impose constant network width and have weight magnitude that is exponential in network depth, those in [9, 12] allow for decoupled width and depth behavior, but exhibit weight magnitudes that are exponential in network depth and polynomial in network width. We note that the aspect of weight magnitude behavior is not explicitly discussed in these references, but can be uncovered by scrutinizing the proofs of the corresponding results.

The flexibility afforded by our construction can be attributed to the use of a reduced alphabet, i.e., $\{0,1\}$ in a ternary expansion. While this idea is novel in the context of bit extraction, its roots can be traced back to the study of the computational power of neural networks [22]. Specifically, [22] encodes $\{0,1\}$-strings as quaternary numbers. We next illustrate the philosophy underlying this idea by way of an example.

Example B.6. *Let $n \in \mathbb{N}$ with $n \geq 4$. We want to encode a $\{0,1\}$-string of length n as a real number and extract the first element of the string from this number. For concreteness, we consider the $\{0,1\}$-strings of length n given by $s_1 = (1,0,\ldots,0)$ and $s_2 = (0,1,\ldots,1)$, and compare the Lipschitz constants of the decoders associated with base-2 and base-3 encoding, respectively. The reason for studying the Lipschitz constant of the decoders resides in the fact that it determines the behavior of the weight magnitude of corresponding ReLU network realizations.*

We first perform base-2 encoding according to $E((\theta_i)_{i=1}^n) = \sum_{i=1}^n 2^{-i}\theta_i$ with $(\theta_i)_{i=1}^n \in \{0,1\}^n$. This yields $b_1 := E(s_1) = 2^{-1}$ and $b_2 := E(s_2) = 2^{-1} - 2^{-n}$. Let D be any decoder[8] that extracts the first bit, i.e., $D(E((\theta_i)_{i=1}^n)) = \theta_1$, for $(\theta_i)_{i=1}^n \in \{0,1\}^n$. The Lipschitz constant of this decoder is at least exponential in n, which can be seen by evaluating

$$\frac{|D(E(s_1)) - D(E(s_2))|}{|E(s_1) - E(s_2)|} = 2^n. \tag{111}$$

Alternatively, we can employ base-3 encoding according to Definition B.4 and take the corresponding decoder $\tilde{D}$ to be given by

$$\tilde{D}(x) = 9\rho(x - T((0,2))) - 9\rho(x - T((1,0))), \quad x \in \mathbb{R}.$$

We note that $\tilde{D}(x) = 0$, for $x \in [0, T((0,2))]$, $\tilde{D}(x) = 1$, for $x \in [T((1,0)), \infty)$, and therefore $\tilde{D}(T((\theta_i)_{i=1}^n)) = \theta_1$, for $(\theta_i)_{i=1}^n \in \{0,1\}^n$. The Lipschitz constant of this decoder is given by 9, and is hence independent of n.

Example B.6 provides a heuristic argument for reduced alphabets leading to better weight magnitude behavior of decoding ReLU networks. The procedure described in Example B.6 extracts the first element of the string under consideration; it was used in [11] together with bitshift operations to extract multiple elements, one by one in multiple rounds. On the other hand, the bit extraction technique in [9, 12] and our Proposition B.5, while also employing a

[8] There are infinitely many such decoders.

multi-round approach, deliver multiple elements of the string in each round. Specifically, in Proposition B.5 N elements are extracted in each of the L rounds. Notably, the weight magnitude of our decoding network $F_{N,L}$ is independent of the number of rounds L and depends on N only. In contrast, the weight magnitudes of the decoding networks reported in the literature all depend exponentially on the number of rounds, a consequence of what was illustrated in Example B.6. As the depth of the extracting networks is proportional to the number of rounds they are to carry out, the constructions reported previously in the literature all exhibit weight magnitude growth that is exponential in network depth.

B.2 Proof of Proposition B.1

The proof is constructive in the sense that, for each $g \in H^1([0,1])$, and given $W, L \geq D_a$, we explicitly specify a function

$$f \in \mathcal{R}(W, L, W^K) \tag{112}$$

such that

$$\|g - f\|_{L^\infty([0,1])} \leq C_a (W^2 L^2 \log(W))^{-1}, \tag{113}$$

with absolute constants $C_a, D_a \in \mathbb{R}$ and $K \in \mathbb{N}$ to be specified later. In the following, let $m, n, \ell \in \mathbb{N}$, with $\ell \geq 2$, to be determined later, and set

$$\Delta = \frac{1}{10 m^2 \ell^2 n}. \tag{114}$$

Throughout the proof, we will frequently consider the grid points $\{\frac{i}{m^2\ell^2 n} : i = 0, \ldots, m^2\ell^2 n - 1\}$, which we sometimes rewrite as

$$\left\{ \frac{i}{m^2\ell^2 n} : i = 0, \ldots, m^2\ell^2 n - 1 \right\} = \left\{ \frac{j}{m^2\ell} + \frac{k}{m^2\ell^2 n} : (j,k) \in \mathcal{I} \right\}, \tag{115}$$

where

$$\mathcal{I} = \{(j,k) : j \in \{0, \ldots, m^2\ell - 1\}, k \in \{0, \ldots, n\ell - 1\}\}. \tag{116}$$

For fixed $g \in H^1([0,1])$, the construction of the corresponding f proceeds in four steps as follows.

1. We specify a function $f_1 \in C(\mathbb{R})$ realized by a ReLU network and approximating g "well enough" on the grid points $\{\frac{i}{m^2\ell^2 n}\}_{i=0}^{m^2\ell^2 n - 1}$.

2. Then, based on f_1, we construct a function $f_2 \in C(\mathbb{R})$ realized by a ReLU network and approximating g "well enough" on the subdomain $\bigcup_{i=0}^{m^2\ell^2 n - 1} \left[\frac{i}{m^2\ell^2 n}, \frac{i+1}{m^2\ell^2 n} - \Delta\right]$.

3. Starting from f_2, we determine a function $f \in C(\mathbb{R})$ realized by a ReLU network and approximating g "well enough" on the entire domain $[0,1]$.

4. The depth and width of the ReLU network in Step 3 as well as the corresponding approximation error $\|f - g\|_{L^\infty([0,1])}$ depend explicitly on m, n, ℓ. Then, values for m, n, ℓ are chosen to make (112) and (113) hold for absolute constants C_a, D_a, K. We do announce that m will be linear in W, ℓ linear in L, and n logarithmic in W.

Step 1. This step is summarized in the form of the following result.

Lemma B.7. *For $g \in H^1([0,1])$, $m, n, \ell \in \mathbb{N}$, with $\ell \geq 2$, there exists a function*

$$f_1 \in \mathcal{R}(200m + 2^{n+5}, 37\ell, \max\{8mn, 3^{n+2}\}), \tag{117}$$

such that for $i = 0, \ldots, m^2\ell^2 n - 1$,

$$\left| g\Big(\frac{i}{m^2\ell^2 n}\Big) - f_1\Big(\frac{i}{m^2\ell^2 n}\Big) \right| \leq \frac{1}{m^2\ell^2 n}. \tag{118}$$

Proof. We use a two-stage approach to approximate g, starting by fitting a function to g on the grid points $\{\frac{j}{m^2\ell}\}_{j=0}^{m^2\ell-1}$, and then lifting this function to obtain a ReLU network that approximates g on the target grid points $\{\frac{i}{m^2\ell^2 n}\}_{i=0}^{m^2\ell^2 n-1}$.

In the first stage, we identify a bounded piecewise linear function h that is fit to g at the grid points $\{\frac{j}{m^2\ell}\}_{j=0}^{m^2\ell-1}$. This function will later be realized by a ReLU network using Proposition B.3. Specifically, we let $h : \mathbb{R} \mapsto \mathbb{R}$ be given by

$$h(x) := \begin{cases} g\Big(\frac{j}{m^2\ell}\Big), & x \in \Big[\frac{j}{m^2\ell}, \frac{j+1}{m^2\ell} - \Delta\Big], j \in \Big\{0, \ldots, m^2\ell - 1\Big\}, \\ \dfrac{g\Big(\frac{j+1}{m^2\ell}\Big) - g\Big(\frac{j}{m^2\ell}\Big)}{\Delta}\Big(x - \dfrac{j+1}{m^2\ell} + \Delta\Big) & \\ \quad + g\Big(\dfrac{j}{m^2\ell}\Big), & x \in \Big[\dfrac{j+1}{m^2\ell} - \Delta, \dfrac{j+1}{m^2\ell}\Big], j \in \big\{0, \ldots, m^2\ell - 2\big\}, \\ g\Big(\dfrac{m^2\ell - 1}{m^2\ell}\Big), & x \in [1 - \Delta, \infty), \\ g(0), & x \in (\infty, 0], \end{cases}$$

with $\Delta = \frac{1}{10m^2\ell^2 n}$ as specified in (114). We note that

$$h\left(\frac{j}{m^2\ell} + \frac{k}{m^2\ell^2 n}\right) = g\left(\frac{j}{m^2\ell}\right), \quad \text{for } (j,k) \in \mathcal{I}, \tag{119}$$

where the index set $\mathcal{I}$ was defined in (116).

For the second stage, let $t = g - h$ be the remainder of the approximation of g by h. We want to approximate t on the target grid points $\{\frac{i}{m^2\ell^2 n} : i = 0, \ldots, m^2\ell^2 n - 1\} = \{\frac{j}{m^2\ell} + \frac{k}{m^2\ell^2 n} : (j,k) \in \mathcal{I}\}$. This will be done by rounding $t\big(\frac{j}{m^2\ell} + \frac{k}{m^2\ell^2 n}\big)$, for all $(j,k) \in \mathcal{I}$, down to the nearest number in $\frac{1}{m^2\ell^2 n}\mathbb{Z}$. To this end, we define

$$t_j(k) = \left\lfloor m^2\ell^2 n \cdot t\Big(\frac{j}{m^2\ell} + \frac{k}{m^2\ell^2 n}\Big) \right\rfloor, \tag{120}$$

and note that

$$\left| \frac{t_j(k)}{m^2\ell^2 n} - t\left(\frac{j}{m^2\ell} + \frac{k}{m^2\ell^2 n}\right) \right| \leq \frac{1}{m^2\ell^2 n}. \tag{121}$$

Formally, the rounding operation will be effected by application of a function $p : \mathbb{R} \mapsto \mathbb{R}$ satisfying

$$p\Big(\frac{j}{m^2\ell} + \frac{k}{m^2\ell^2 n}\Big) = \frac{t_j(k)}{m^2\ell^2 n}, \quad \text{for } (j,k) \in \mathcal{I}. \tag{122}$$

We note that (122) determines the values of p on $\{\frac{j}{m^2\ell}+\frac{k}{m^2\ell^2n} : (j,k)\in\mathcal{I}\}$ only, and we still have to specify p on $\mathbb{R}\backslash\{\frac{j}{m^2\ell}+\frac{k}{m^2\ell^2n} : (j,k)\in\mathcal{I}\}$. This will be done such that p can be realized by a ReLU network. Once p has been determined, we let $f_1 = h+p$ and note that by (122) and (121) the desired approximation error bound holds according to

$$\left|g\Big(\frac{j}{m^2\ell}+\frac{k}{m^2\ell^2n}\Big)-f_1\Big(\frac{j}{m^2\ell}+\frac{k}{m^2\ell^2n}\Big)\right| \tag{123}$$

$$= \left|t\Big(\frac{j}{m^2\ell}+\frac{k}{m^2\ell^2n}\Big)-\frac{t_j(k)}{m^2\ell^2n}\right| \tag{124}$$

$$\leq \frac{1}{m^2\ell^2n}, \quad \text{for } (j,k)\in\mathcal{I}. \tag{125}$$

We proceed with the construction of p, which will be based on the bit extraction technique described in Appendix B.1.2. To this end, we first represent $t_j(k)$, $(j,k)\in\mathcal{I}$, in the form $\sum_{i=1}^k\theta_{j,i}^+-\sum_{i=1}^k\theta_{j,i}^-$, for some $(\theta_{j,i}^+)_{i=1}^{n\ell-1},(\theta_{j,i}^-)_{i=1}^{n\ell-1}\in\{0,1\}^{n\ell-1}$. Specifically, we note that, for $(j,k)\in\mathcal{I}$, with $k\geq 1$,

$$\left|t\Big(\frac{j}{m^2\ell}+\frac{k}{m^2\ell^2n}\Big)-t\Big(\frac{j}{m^2\ell}+\frac{k-1}{m^2\ell^2n}\Big)\right| \tag{126}$$

$$= \Big|g\Big(\frac{j}{m^2\ell}+\frac{k}{m^2\ell^2n}\Big)-h\Big(\frac{j}{m^2\ell}+\frac{k}{m^2\ell^2n}\Big) - g\Big(\frac{j}{m^2\ell}+\frac{k-1}{m^2\ell^2n}\Big)+h\Big(\frac{j}{m^2\ell}+\frac{k-1}{m^2\ell^2n}\Big)\Big| \tag{127}$$

$$= \left|g\Big(\frac{j}{m^2\ell}+\frac{k}{m^2\ell^2n}\Big)-g\Big(\frac{j}{m^2\ell}+\frac{k-1}{m^2\ell^2n}\Big)\right| \tag{128}$$

$$\leq \frac{1}{m^2\ell^2n}, \tag{129}$$

where (128) follows from $h\Big(\frac{j}{m^2\ell}+\frac{k}{m^2\ell^2n}\Big)=h\Big(\frac{j}{m^2\ell}+\frac{k-1}{m^2\ell^2n}\Big)=g(\frac{j}{m^2\ell})$ as a consequence of (119), and in (129) we used the 1-Lipschitz-continuity of g. Multiplication of (126)-(129) by $m^2\ell^2n$ then yields

$$\left|m^2\ell^2n\cdot t\Big(\frac{j}{m^2\ell}+\frac{k}{m^2\ell^2n}\Big)-m^2\ell^2n\cdot t\Big(\frac{j}{m^2\ell}+\frac{k-1}{m^2\ell^2n}\Big)\right|\leq 1. \tag{130}$$

As $|\lfloor x\rfloor-\lfloor y\rfloor|\leq 1$, for all $x,y\in\mathbb{R}$ such that $|x-y|\leq 1$, it follows from (130) that $|\lfloor m^2\ell^2n\cdot t(\frac{j}{m^2\ell}+\frac{k}{m^2\ell^2n})\rfloor-\lfloor m^2\ell^2n\cdot t(\frac{j}{m^2\ell}+\frac{k-1}{m^2\ell^2n})\rfloor|\leq 1$, which is $|t_j(k)-t_j(k-1)|\leq 1$. Moreover, as $t_j(k)$ and $t_j(k-1)$ are integers, we must have $t_j(k)-t_j(k-1)\in\{-1,0,1\}$. Next, we define, for $(j,k)\in\mathcal{I}$ with $k\geq 1$,

$$\begin{aligned}\theta_{j,k}^+ &= \max\{t_j(k)-t_j(k-1),0\}\in\{0,1\},\\ \theta_{j,k}^- &= \max\{-(t_j(k)-t_j(k-1)),0\}\in\{0,1\}.\end{aligned}$$

Then, for $(j,k)\in\mathcal{I}$, we can write

$$t_j(k)-t_j(k-1)=\theta_{j,k}^+-\theta_{j,k}^-, \tag{131}$$

and get

$$t_j(k) = t_j(0) + \sum_{i=1}^{k} (t_j(i) - t_j(i-1)) \tag{132}$$

$$= \sum_{i=1}^{k} \theta^+_{j,i} - \sum_{i=1}^{k} \theta^-_{j,i}, \tag{133}$$

where in (133) we used (131) along with $t_j(0) = \lfloor m^2\ell^2 n \cdot t(\frac{j}{m^2\ell}) \rfloor = \lfloor m^2\ell^2 n \cdot (g(\frac{j}{m^2\ell}) - h(\frac{j}{m^2\ell})) \rfloor = 0$ thanks to (119).

We are now ready to detail the construction of p. For $j = 0, \ldots, m^2\ell - 1$, we encode the $\{0,1\}$-strings $(\theta^+_{j,1}, \ldots, \theta^+_{j,n\ell-1})$ and $(\theta^-_{j,1}, \ldots, \theta^-_{j,n\ell-1})$ into ternary numbers with reduced alphabet according to

$$b^+_j := T((\theta^+_{j,1}, \ldots, \theta^+_{j,n\ell-1})),$$
$$b^-_j := T((\theta^-_{j,1}, \ldots, \theta^-_{j,n\ell-1})).$$

By Proposition B.5, it follows that there exists a decoder $F_{n,\ell} \in \mathcal{R}((2,1), 2^{n+4}, 5\ell, 3^{n+2})$ such that, for $(j,k) \in \mathcal{I}$, we have $F_{n,\ell}(b^+_j, k) = \sum_{i=1}^{k} \theta^+_{j,i}$ and $F_{n,\ell}(b^-_j, k) = \sum_{i=1}^{k} \theta^-_{j,i}$, which, combined with (132)-(133), implies

$$t_j(k) = F_{n,\ell}(b^+_j, k) - F_{n,\ell}(b^-_j, k). \tag{134}$$

Next, let $b^+, b^-, s \in C(\mathbb{R})$ be defined according to

$$b^+(x) = \begin{cases} b^+_j, & x \in \left[\frac{j}{m^2\ell}, \frac{j+1}{m^2\ell} - \Delta\right], \ j \in \{0, \ldots, m^2\ell - 1\}, \\ b^+_j + \frac{b^+_{j+1} - b^+_j}{\Delta}\left(x - \frac{j+1}{m^2\ell} + \Delta\right), & x \in \left[\frac{j+1}{m^2\ell} - \Delta, \frac{j+1}{m^2\ell}\right], \ j \in \{0, \ldots, m^2\ell - 2\}, \\ b^+_0, & x \in (-\infty, 0], \\ b^+_{m^2\ell-1}, & x \in [1-\Delta, \infty), \end{cases}$$

$$b^-(x) = \begin{cases} b^-_j, & x \in \left[\frac{j}{m^2\ell}, \frac{j+1}{m^2\ell} - \Delta\right], \ j \in \{0, \ldots, m^2\ell - 1\}, \\ b^-_j + \frac{b^-_{j+1} - b^-_j}{\Delta}\left(x - \frac{j+1}{m^2\ell} + \Delta\right), & x \in \left[\frac{j+1}{m^2\ell} - \Delta, \frac{j+1}{m^2\ell}\right], \ j \in \{0, \ldots, m^2\ell - 2\}, \\ b^-_0, & x \in (-\infty, 0], \\ b^-_{m^2\ell-1}, & x \in [1-\Delta, \infty), \end{cases}$$

and

$$s(x) = \begin{cases} m^2\ell^2 n\left(x - \frac{j}{m^2\ell}\right), & x \in \left[\frac{j}{m^2\ell}, \frac{j+1}{m^2\ell} - \Delta\right], \ j \in \{0, \ldots, m^2\ell - 1\}, \\ n\ell \frac{m^2\ell\Delta - 1}{\Delta}\left(x - \frac{j+1}{m^2\ell}\right), & x \in \left[\frac{j+1}{m^2\ell} - \Delta, \frac{j+1}{m^2\ell}\right], \ j \in \{0, \ldots, m^2\ell - 2\}, \\ 0, & x \in (-\infty, 0], \\ n\ell - m^2\ell^2 n\Delta, & x \in [1-\Delta, \infty). \end{cases}$$

These choices of b^+, b^-, and s guarantee that, for $(j,k) \in \mathcal{I}$,

$$b^+\Big(\frac{j}{m^2\ell} + \frac{k}{m^2\ell^2 n}\Big) = b_j^+, \tag{135}$$

$$b^-\Big(\frac{j}{m^2\ell} + \frac{k}{m^2\ell^2 n}\Big) = b_j^-, \tag{136}$$

$$s\Big(\frac{j}{m^2\ell} + \frac{k}{m^2\ell^2 n}\Big) = k, \tag{137}$$

as a consequence of $\frac{j}{m^2\ell} + \frac{k}{m^2\ell^2 n} \in [\frac{j}{m^2\ell}, \frac{j+1}{m^2\ell} - \Delta]$. We specify p according to

$$p := \frac{1}{m^2\ell^2 n}(F_{n,\ell} \circ (b^+, s) - F_{n,\ell} \circ (b^-, s)). \tag{138}$$

It hence follows that p has the desired property (122), as, for $(j,k) \in \mathcal{I}$,

$$p\Big(\frac{j}{m^2\ell} + \frac{k}{m^2\ell^2 n}\Big) \tag{139}$$

$$= \frac{1}{m^2\ell^2 n}\Big(F_{n,\ell}\Big(b^+\Big(\frac{j}{m^2\ell} + \frac{k}{m^2\ell^2 n}\Big), s\Big(\frac{j}{m^2\ell} + \frac{k}{m^2\ell^2 n}\Big)\Big) - F_{n,\ell}\Big(b^-\Big(\frac{j}{m^2\ell} + \frac{k}{m^2\ell^2 n}\Big), s\Big(\frac{j}{m^2\ell} + \frac{k}{m^2\ell^2 n}\Big)\Big)\Big) \tag{140}$$

$$= \frac{1}{m^2\ell^2 n}\Big(F_{n,\ell}(b_j^+, k) - F_{n,\ell}(b_j^-, k)\Big) \tag{141}$$

$$= \frac{1}{m^2\ell^2 n}\, t_j(k), \tag{142}$$

where (141) follows from (135), (136), and (137), and (142) is by (134).

It remains to show that $f_1 = h + p = h + \frac{1}{m^2\ell^2 n} \cdot (F_{n,\ell} \circ (b^+, s) - F_{n,\ell} \circ (b^-, s)) \in \mathcal{R}(200m + 2^{n+5}, 37\ell, \max\{8mn, 3^{n+2}\})$. To this end, we first consider ReLU network realizations of h, $F_{n,\ell}$, b^+, b^-, and s, and then put them together according to Lemma H.3. We start by noting that h, b^+, b^-, and s are all bounded piecewise linear functions with breakpoints $X_1 = (x_i)_{i=0}^{2m^2\ell-1}$, where $x_{2k} = \frac{k}{m^2\ell}$ and $x_{2k+1} = \frac{k+1}{m^2\ell} - \Delta$, for $k = 0, \ldots, m^2\ell - 1$. In addition, the $L^\infty(\mathbb{R})$-norm of h, b^+, b^-, and s is upper-bounded by $n\ell$, which can be verified by checking the values of these functions at their breakpoints and noting that bounded piecewise linear functions take on their maximum absolute values at breakpoints[9]. We therefore have

$$h, b^+, b^-, s \in \Sigma(X_1, n\ell). \tag{143}$$

Upon noting that $|X_1| = 2m^2\ell$ and $R_m(X_1) = \frac{1}{\Delta} = 10m^2\ell^2 n$, application of Proposition B.3 to $\Sigma(X_1, n\ell)$ with $M = |X_1| = 2m^2\ell$, $E = n\ell$, $u = 2m$, $v = \ell$, and $w = 4mn$ so that $u^2 v = 4m^2\ell \geq |X_1|$ and $w^{30v} = (4mn)^{30\ell} \geq 2^{30\ell} 2^{30} (mn)^{30} \geq \ell^{30} 2^{30} (mn)^{30} \geq (2m^2\ell)^6 (10m^2\ell^2 n)^4 n\ell = M^6 (R_m(X_1))^4 E$, yields

$$\Sigma(X_1, n\ell) \subseteq \mathcal{R}(40m, 30\ell, 8mn). \tag{144}$$

Putting (143) and (144) together shows that h, b^+, b^-, and s can be realized by ReLU networks such that

$$h, b^+, b^-, s \in \mathcal{R}(40m, 30\ell, 8mn). \tag{145}$$

[9] To be more precise, the $L^\infty(\mathbb{R})$-norms of h, b^+, and b^- are upper-bounded by 1 and that of s by $n\ell - m^2\ell^2 n\Delta$.

In addition, we recall that, according to Proposition B.5, the decoder $F_{n,\ell}$ can be realized by a ReLU network so that

$$F_{n,\ell} \in \mathcal{R}\left((2,1), 2^{n+4}, 5\ell, 3^{n+2}\right). \tag{146}$$

Application of Lemma H.3 now yields

$$\begin{aligned}
(b^+, s), (b^-, s) &\in \mathcal{R}((1,2), 80m, 30\ell, 8mn),\\
F_{n,\ell} \circ (b^+, s), F_{n,\ell} \circ (b^-, s) &\in \mathcal{R}(\max\{80m, 2^{n+4}\}, 35\ell, \max\{8mn, 3^{n+2}\}),\\
F_{n,\ell} \circ (b^+, s) + (-1) \cdot F_{n,\ell} \circ (b^-, s) &\in \mathcal{R}(2\max\{80m, 2^{n+4}\}, 35\ell + 1, \max\{8mn, 3^{n+2}\})\\
&\subseteq \mathcal{R}(160m + 2^{n+5}, 36\ell, \max\{8mn, 3^{n+2}\}),\\
f_1 &= h + \frac{1}{m^2\ell^2 n} \cdot (F_{n,\ell} \circ (b^+, s) - F_{n,\ell} \circ (b^-, s))\\
&\in \mathcal{R}(200m + 2^{n+5}, \max\{30\ell, 36\ell\} + 1, \max\{8mn, 3^{n+2}\})\\
&\subseteq \mathcal{R}(200m + 2^{n+5}, 37\ell, \max\{8mn, 3^{n+2}\}).
\end{aligned}$$

□

Step 2. We summarize this step in the following result.

Lemma B.8. *For $m, n, \ell \in \mathbb{N}$, with $\ell \geq 2$, there exists a function*

$$u \in \mathcal{R}(\max\{40m, 40n\}, 61\ell, 8mn), \tag{147}$$

such that

$$u(x) = \frac{i}{m^2\ell^2 n}, \quad x \in \left[\frac{i}{m^2\ell^2 n}, \frac{i+1}{m^2\ell^2 n} - \Delta\right], \text{ for } i = 0, \ldots, m^2\ell^2 n - 1. \tag{148}$$

For $g \in H^1([0,1])$, let f_1 be the function given by Lemma B.7, and let $f_2 = f_1 \circ u$. We have

$$f_2 \in \mathcal{R}(200m + 2^{n+5}, 98\ell, \max\{8mn, 3^{n+2}\}) \tag{149}$$

with

$$|f_2(x) - g(x)| \leq \frac{2}{m^2\ell^2 n}, \quad \text{for } x \in \bigcup_{i=0}^{m^2\ell^2 n - 1} \left[\frac{i}{m^2\ell^2 n}, \frac{i+1}{m^2\ell^2 n} - \Delta\right]. \tag{150}$$

Proof. We start by constructing u. Let $u_1, u_2 \in C(\mathbb{R})$ be given by

$$u_1(x) = \begin{cases}
x - \frac{j}{m^2\ell}, & x \in \left[\frac{j}{m^2\ell}, \frac{j+1}{m^2\ell} - \Delta\right], \ j \in \{0, \ldots, m^2\ell - 1\},\\
\frac{m^2\ell\Delta - 1}{m^2\ell\Delta}\left(x - \frac{j+1}{m^2\ell}\right), & x \in \left[\frac{j+1}{m^2\ell} - \Delta, \frac{j+1}{m^2\ell}\right], \ k \in \{0, \ldots, m^2\ell - 2\},\\
0, & x \in (-\infty, 0],\\
\frac{1}{m^2\ell} - \Delta, & x \in [1 - \Delta, \infty),
\end{cases} \tag{151}$$

and

$$u_2(x) = \begin{cases}
x - \frac{k}{m^2\ell^2 n}, & x \in \left[\frac{k}{m^2\ell^2 n}, \frac{k+1}{m^2\ell^2 n} - \Delta\right], \ k \in \{0, \ldots, n\ell - 1\},\\
\frac{m^2\ell^2 n\Delta - 1}{m^2\ell^2 n\Delta}\left(x - \frac{k+1}{m^2\ell^2 n}\right), & x \in \left[\frac{k+1}{m^2\ell^2 n} - \Delta, \frac{k+1}{m^2\ell^2 n}\right], \ j \in \{0, \ldots, n\ell - 1\},\\
0, & x \in (-\infty, 0],\\
\frac{1}{m^2\ell^2 n} - \Delta, & x \in \left[\frac{1}{m^2\ell} - \Delta, \infty\right),
\end{cases} \tag{152}$$

and let

$$u = \mathrm{Id} \; - \; u_2 \circ u_1, \tag{153}$$

where $\mathrm{Id}(x) = x$, $x \in \mathbb{R}$.

We first verify (148). To this end, we start by noting that, for every $i \in \{0, \ldots, m^2\ell^2 n - 1\}$, thanks to (96), there exists $(j(i), k(i)) \in \mathcal{I}$ with $\mathcal{I}$ as defined in (116), such that $\frac{i}{m^2\ell^2 n} = \frac{j(i)}{m^2\ell} + \frac{k(i)}{m^2\ell^2 n}$. Then, for $i \in \{0, \ldots, m^2\ell^2 n - 1\}$ and $x \in \left[\frac{i}{m^2\ell^2 n}, \frac{i+1}{m^2\ell^2 n} - \Delta\right]$, we have

$$x \in \left[\frac{j(i)}{m^2\ell} + \frac{k(i)}{m^2\ell^2 n}, \frac{j(i)}{m^2\ell} + \frac{k(i)+1}{m^2\ell^2 n} - \Delta\right],$$

and (148), follows upon noting that

$$\begin{aligned} u(x) &= x - u_2(u_1(x)) && (154)\\ &= x - u_2\Big(x - \frac{j(i)}{m^2\ell}\Big) && (155)\\ &= x - \Big(x - \frac{j(i)}{m^2\ell} - \frac{k(i)}{m^2\ell^2 n}\Big) && (156)\\ &= \frac{j(i)}{m^2\ell} + \frac{k(i)}{m^2\ell^2 n} && (157)\\ &= \frac{i}{m^2\ell^2 n}. && (158) \end{aligned}$$

We proceed to realize u by a ReLU network with the goal of establishing (147). This will be accomplished by realizing the constituents u_1, u_2, Id of u by suitable ReLU networks and combining them using Lemma H.3. It follows by inspection that $u_1 \in \Sigma(X_1, 1)$, with X_1 as defined in the paragraph after (142). Together with $\Sigma(X_1, 1) \subseteq \Sigma(X_1, n\ell)$ and $\Sigma(X_1, n\ell) \subseteq \mathcal{R}(40m, 30\ell, 8mn)$, thanks to (144), this then implies

$$u_1 \in \mathcal{R}(40m, 30\ell, 8mn). \tag{159}$$

Again, by inspection,

$$u_2 \in \Sigma(X_2, 1), \tag{160}$$

with $X_2 = (x_i)_{i=0}^{2n\ell-1}$, where $x_{2k} = \frac{k}{m^2\ell^2 n}$, $x_{2k+1} = \frac{k+1}{m^2\ell^2 n} - \Delta$, for $k = 0, \ldots, n\ell - 1$. Upon noting that $|X_2| = 2n\ell$ and $R_m(X_2) = \frac{1}{\Delta} = 10m^2\ell^2 n$, application of Proposition B.3 to $\Sigma(X_2, n\ell)$ with $M = |X_2| = 2n\ell$, $E = 1$, $u = 2n$, $v = \ell$, and $w = 4mn$, ensuring that $u^2 v = 4n^2\ell \geq M$ and $w^{30v} = (4mn)^{30\ell} \geq 2^{30\ell} 2^{30} (mn)^{30} \geq \ell^{30} 2^{30} (mn)^{30} \geq (2n\ell)^6 (10m^2\ell^2 n)^4 = M^6 (R_m(X_2))^4 E$, yields

$$\Sigma(X_2, 1) \subseteq \mathcal{R}(40n, 30\ell, 8mn). \tag{161}$$

With (160) this then implies

$$u_2 \in \mathcal{R}(40n, 30\ell, 8mn). \tag{162}$$

Next, trivially,

$$\mathrm{Id} \in \mathcal{R}(1, 1, 1). \tag{163}$$

Application of Lemma H.3 together with (159), (162), and (163) leads to

$$\begin{aligned} u_2 \circ u_1 &\in \mathcal{R}(\max\{40m, 40n\}, 60\ell, 8mn), && (164)\\ u = \mathrm{Id} \; + (-1) \cdot \; u_2 \circ u_1 &\in \mathcal{R}(\max\{40m, 40n\} + 2, 61\ell, 8mn). && (165) \end{aligned}$$

Regarding f_2, application of Lemma H.3 together with (117) and (165), yields

$$\begin{aligned} f_2 &= f_1 \circ u && (166)\\ &\in \mathcal{R}(\max\{200m + 2^{n+5}, \max\{40m, 40n\} + 2\}, 61\ell + 37\ell, \max\{8mn, 3^{n+2}\}) && (167)\\ &\subseteq \mathcal{R}(200m + 2^{n+5}, 98\ell, \max\{8mn, 3^{n+2}\}), && (168)\end{aligned}$$

which establishes (149). Moreover, for $i \in \{0, \ldots, m^2\ell^2 n - 1\}$ and $x \in \left[\frac{i}{m^2\ell^2 n}, \frac{i+1}{m^2\ell^2 n} - \Delta\right]$, we have

$$\begin{aligned} |f_2(x) - g(x)| &\leq \left| f_1(u(x)) - g\left(\frac{i}{m^2\ell^2 n}\right)\right| + \left|g\left(\frac{i}{m^2\ell^2 n}\right) - g(x)\right| && (169)\\ &\leq \left| f_1\left(\frac{i}{m^2\ell^2 n}\right) - g\left(\frac{i}{m^2\ell^2 n}\right)\right| + \left|\frac{i}{m^2\ell^2 n} - x\right| && (170)\\ &\leq \frac{2}{m^2\ell^2 n}, && (171)\end{aligned}$$

where (170) follows from (154)-(158) and the 1-Lipschitz continuity of g, and in (171) we used (118). This concludes the proof. □

Step 3. In this step, we construct $f : \mathbb{R} \mapsto \mathbb{R}$ such that $\|f - g\|_{L^\infty([0,1])} \leq \frac{3}{m^2\ell^2 n}$, which will be effected by application of the median kernel smoothing technique, introduced in [23], to the function f_2 built in Step 2. The construction is formalized as follows.

Lemma B.9. *For $g \in H^1([0,1])$, $m, n, \ell \in \mathbb{N}$, with $\ell \geq 2$, define $f : \mathbb{R} \mapsto \mathbb{R}$ according to*

$$f(x) := \mathit{median}(f_2(\rho(x - 2\Delta)), f_2(\rho(x - 4\Delta)), f_2(\rho(x - 6\Delta))), \quad x \in \mathbb{R}, \tag{172}$$

with $\Delta = \frac{1}{10m^2\ell^2 n}$, and $f_2 : \mathbb{R} \mapsto \mathbb{R}$ as in Lemma B.8. Here, for $x_1, x_2, x_3 \in \mathbb{R}$ with reordering from smallest to largest denoted by $x_{(1)}, x_{(2)}, x_{(3)}$,

$$\mathit{median}(x_1, x_2, x_3) := x_{(2)}.$$

We have

$$\|f - g\|_{L^\infty([0,1])} \leq \frac{3}{m^2\ell^2 n} \tag{173}$$

and

$$f \in \mathcal{R}(600m + 2^{n+7}, 101\ell, \max\{8mn, 3^{n+2}\}).$$

Proof. We start by upper-bounding $\|f - g\|_{L^\infty([0,1])}$. Fix $x \in [0,1]$ and note that at least two elements[10] of $\{\rho(x-2\Delta), \rho(x-4\Delta), \rho(x-6\Delta)\}$ are contained in the set $\bigcup_{i=0}^{m^2\ell^2 n-1}[\frac{i}{m^2\ell^2 n}, \frac{i+1}{m^2\ell^2 n} - \Delta]$. Specifically, with $\frac{1}{m^2\ell^2 n} = 10\Delta$, we have $\rho(x-4\Delta), \rho(x-6\Delta) \in \bigcup_{i=0}^{m^2\ell^2 n-1}[\frac{i}{m^2\ell^2 n}, \frac{i+1}{m^2\ell^2 n} - \Delta]$ if $x \in \bigcup_{i=0}^{m^2\ell^2 n-1}\left[\frac{i}{m^2\ell^2 n}, \frac{i}{m^2\ell^2 n} + 3\Delta\right]$, $\rho(x-2\Delta), \rho(x-6\Delta) \in \bigcup_{i=0}^{m^2\ell^2 n-1}[\frac{i}{m^2\ell^2 n}, \frac{i+1}{m^2\ell^2 n} - \Delta]$ if $x \in \bigcup_{i=0}^{m^2\ell^2 n-1}\left[\frac{i}{m^2\ell^2 n} + 3\Delta, \frac{i}{m^2\ell^2 n} + 5\Delta\right]$, and $\rho(x-2\Delta), \rho(x-4\Delta) \in \bigcup_{i=0}^{m^2\ell^2 n-1}[\frac{i}{m^2\ell^2 n}, \frac{i+1}{m^2\ell^2 n} - \Delta]$ if

[10] It is possible that the set $\{\rho(x-2\Delta), \rho(x-4\Delta), \rho(x-6\Delta)\}$ contains duplicates, e.g., when $x = 0$, we have $\rho(x-2\Delta) = \rho(x-4\Delta) = \rho(x-6\Delta) = 0$. We shall not account for such cases explicitly, but simply note that our exposition incorporates them.

$x \in \bigcup_{i=0}^{m^2\ell^2 n-1}\left[\frac{i}{m^2\ell^2 n}+5\Delta, \frac{i}{m^2\ell^2 n}+10\Delta\right]$. Therefore, there exist distinct numbers $a_1(x), a_2(x) \in \{2,4,6\}$, depending on x, such that

$$\rho(x-a_1(x)\Delta), \rho(x-a_2(x)\Delta) \in \bigcup_{i=0}^{m^2\ell^2 n-1}\left[\frac{i}{m^2\ell^2 n}, \frac{i+1}{m^2\ell^2 n}-\Delta\right],$$

and we define $a_3(x)$ to be the unique element given by $\{2,4,6\}\backslash\{a_1(x), a_2(x)\}$. If all elements of $\{\rho(x-2\Delta), \rho(x-4\Delta), \rho(x-6\Delta)\}$ are contained in $\bigcup_{i=0}^{m^2\ell^2 n-1}[\frac{i}{m^2\ell^2 n}, \frac{i+1}{m^2\ell^2 n}-\Delta]$, we take $a_1(x)=2$, $a_2(x)=4$, and $a_3(x)=6$. For $i=1,2$, we have

$$\begin{aligned}
&|f_2(\rho(x-a_i(x)\Delta))-g(x)| && (174)\\
\leq &|f_2(\rho(x-a_i(x)\Delta))-g(\rho(x-a_i(x)\Delta))|+|g(\rho(x-a_i(x)\Delta))-g(x)| && (175)\\
\leq &\frac{2}{m^2\ell^2 n}+|g(\rho(x-a_i(x)\Delta))-g(x)| && (176)\\
\leq &\frac{3}{m^2\ell^2 n}, && (177)
\end{aligned}$$

where in (176) we used (150), and (177) follows from the 1-Lipschitz continuity of g, combined with $6\Delta \leq \frac{1}{m^2\ell^2 n}$. To simplify notation, we set $y_i(x) = f_2(\rho(x-a_i(x)\Delta))$, for $i=1,2,3$. Then, we have

$$f(x) = \text{ median}(y_1(x), y_2(x), y_3(x)) \in [\min(y_1(x), y_2(x)), \max(y_1(x), y_2(x))], \tag{178}$$

and

$$\begin{aligned}
&|f(x)-g(x)| && (179)\\
\leq &\max(|\min(y_1(x), y_2(x))-g(x)|, |\max(y_1(x), y_2(x))-g(x)|) && (180)\\
= &\max(|y_1(x)-g(x)|, |y_2(x)-g(x)|) && (181)\\
= &\max(|f_2(\rho(x-a_1(x)\Delta))-g(x)|, |f_2(\rho(x-a_2(x)\Delta))-g(x)|) && (182)\\
\leq &\frac{3}{m^2\ell^2 n}, && (183)
\end{aligned}$$

where in (180) we used (178), and (183) follows from (174)-(177). As the choice of $x \in [0,1]$ was arbitrary, we have established that $\|f-g\|_{L^\infty([0,1])} \leq \frac{3}{m^2\ell^2 n}$.

It remains to show that f can be realized by a ReLU network such that $f \in \mathcal{R}(600m+2^{n+7}, 101\ell, \max\{8mn, 3^{n+2}\})$. This will be accomplished by realizing the individual components of f by suitable ReLU networks and then combining them according to Lemma H.3. For $z \in \mathbb{R}$, define $r_z : \mathbb{R} \mapsto \mathbb{R}$, $r_z(x) = \rho(x-z)$ and note that

$$f = \text{median} \circ (f_2 \circ r_{2\Delta}, f_2 \circ r_{4\Delta}, f_2 \circ r_{6\Delta}).$$

As $2\Delta, 4\Delta, 6\Delta \leq 1$, we get

$$r_{2\Delta}, r_{4\Delta}, r_{6\Delta} \in \mathcal{R}(1,2,1). \tag{184}$$

Moreover, thanks to Lemma B.10, we have

$$\text{median} \in \mathcal{R}((3,1),16,3,1). \tag{185}$$

Application of Lemma H.3 together with (184),(185), and (149) then yields

$$f_2 \circ r_{2\Delta}, f_2 \circ r_{4\Delta}, f_2 \circ r_{6\Delta} \in \mathcal{R}(200m + 2^{n+5}, 98\ell + 2, \max\{8mn, 3^{n+2}\}) \tag{186}$$
$$\subseteq \mathcal{R}(200m + 2^{n+5}, 99\ell, \max\{8mn, 3^{n+2}\}), \tag{187}$$
$$(f_2 \circ r_{2\Delta}, f_2 \circ r_{4\Delta}, f_2 \circ r_{6\Delta}) \in \mathcal{R}((1,3), 600m + 3 \cdot 2^{n+5}, 99\ell, \max\{8mn, 3^{n+2}\}), \tag{188}$$

and

$$f = median \circ (f_2 \circ r_{2\Delta}, f_2 \circ r_{4\Delta}, f_2 \circ r_{6\Delta}) \tag{189}$$
$$\in \mathcal{R}(600m + 3 \cdot 2^{n+5}, 99\ell + 3, \max\{8mn, 3^{n+2}\}) \tag{190}$$
$$\subseteq \mathcal{R}(600m + 2^{n+7}, 101\ell, \max\{8mn, 3^{n+2}\}), \tag{191}$$

where in (187) and (191) we used the assumption $\ell \geq 2$. □

Step 4. We are now ready to prove Proposition B.1.

Proof of Proposition B.1. Set $D_a = 2000$. For $W, L \geq D_a = 2000$, set $m = \left\lfloor \frac{W}{1000} \right\rfloor > 1$, $n = \left\lfloor \log(\frac{2W}{5}) \right\rfloor - 7 > 1$, and $\ell = \left\lfloor \frac{L}{101} \right\rfloor > 10$. Fix $g \in H^1([0,1])$. Application of Lemma B.9 to g yields the existence of an

$$f \in \mathcal{R}(600m + 2^{n+7}, 101\ell, \max\{8mn, 3^{n+2}\}) \tag{192}$$

such that

$$\|f - g\|_{L^\infty([0,1])} \leq \frac{3}{m^2\ell^2 n}. \tag{193}$$

Owing to

$$600m + 2^{n+7} \leq 600 \cdot \frac{W}{1000} + \frac{2}{5}W \leq W \tag{194}$$
$$101\ell \leq L, \tag{195}$$
$$\max\{8mn, 3^{n+2}\} \leq W^2, \tag{196}$$

where in (196) we used $m \leq \frac{W}{1000}$, $n \leq \log(\frac{2W}{5}) \leq \frac{2W}{5}$, and $3^{n+2} \leq (2^{n+7})^2 \leq W^2$, it follows from (192) that

$$f \in \mathcal{R}(W, L, W^2). \tag{197}$$

Next, note that there exists an absolute constant $c \in \mathbb{R}_+$ such that $m \geq cW$, $n \geq c\log(W)$, and $\ell \geq cL$. Hence, (193) implies

$$\|f - g\|_{L^\infty([0,1])} \leq \frac{3}{m^2\ell^2 n} \leq \frac{3}{c^5} \left(W^2 L^2 \log(W)\right)^{-1}. \tag{198}$$

Since the choice of $g \in H^1([0,1])$ was arbitrary, we have established

$$\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}(W, L, W^2)) \leq \frac{3}{c^5}(W^2 L^2 \log(W))^{-1}, \tag{199}$$

which, upon setting $C_a = \frac{3}{c^5}$ and $K = 2$, concludes the proof. □

B.3 Realization of the Median Function by ReLU Networks

A ReLU network realization of the median function was reported in [23]. For completeness, we provide a formal statement thereof here, but note that its proof follows exactly the construction in [23].

Lemma B.10. *Let* $median : \mathbb{R}^3 \mapsto \mathbb{R}$ *be given by*

$$median(x_1, x_2, x_3) := x_{(2)},$$

where for $x_1, x_2, x_3 \in \mathbb{R}$*, the reordering of* x_1, x_2, x_3 *from smallest to largest is denoted as* $x_{(1)}, x_{(2)}, x_{(3)} \in \mathbb{R}$*. It holds that*

$$median \in \mathcal{R}((3,1), 16, 3, 1).$$

Proof. Let $x_1, x_2, x_3 \in \mathbb{R}$. First, note that

$$\text{median}(x_1, x_2, x_3) = x_1 + x_2 + x_3 - \max(x_1, x_2, x_3) - \min(x_1, x_2, x_3). \tag{200}$$

We have

$$\begin{aligned}
&\max(x_1, x_2, x_3) && (201)\\
&= \max(x_1, \max(x_2, x_3)) && (202)\\
&= x_1 + \rho(\max(x_2, x_3) - x_1) && (203)\\
&= x_1 + \rho(x_2 + \rho(x_3 - x_2) - x_1) && (204)\\
&= \rho(x_1) - \rho(-x_1) + \rho(\rho(x_2) - \rho(-x_2) - \rho(x_1) + \rho(-x_1) + \rho(x_3 - x_2)), && (205)
\end{aligned}$$

where (203) and (204) follow from $\max(a, b) = a + \rho(b - a)$, for $a, b \in \mathbb{R}$, and in (205) we used $x = \rho(x) - \rho(-x)$, for $x \in \mathbb{R}$. Inserting (201)-(205) with (x_1, x_2, x_3) replaced by $(-x_1, -x_2, -x_3)$ into the relation $\min(x_1, x_2, x_3) = -\max(-x_1, -x_2, -x_3)$, $x_1, x_2, x_3 \in \mathbb{R}$, yields

$$\begin{aligned}
&\min(x_1, x_2, x_3) && (206)\\
&= -(\rho(-x_1) - \rho(x_1) + \rho(\rho(-x_2) - \rho(x_2) - \rho(-x_1) + \rho(x_1) + \rho(-x_3 + x_2))) && (207)\\
&= -\rho(-x_1) + \rho(x_1) - \rho(\rho(-x_2) - \rho(x_2) - \rho(-x_1) + \rho(x_1) + \rho(x_2 - x_3)). && (208)
\end{aligned}$$

Moreover,

$$x_1 + x_2 + x_3 = \rho(x_1 + x_2 + x_3) - \rho(-x_1 - x_2 - x_3). \tag{209}$$

Substituting (201)-(209) into (200) and using $\rho \circ \rho = \rho$ yields

$$\begin{aligned}
&\text{median}(x_1, x_2, x_3)\\
&= \rho(\rho(x_1 + x_2 + x_3)) - \rho(\rho(-x_1 - x_2 - x_3))\\
&\quad - \rho(\rho(x_1)) + \rho(\rho(-x_1)) - \rho(\rho(x_2) - \rho(-x_2) - \rho(x_1) + \rho(-x_1) + \rho(x_3 - x_2))\\
&\quad + \rho(\rho(-x_1)) - \rho(\rho(x_1)) + \rho(\rho(-x_2) - \rho(x_2) - \rho(-x_1) + \rho(x_1) + \rho(x_2 - x_3)),
\end{aligned}$$

which allows us to conclude that

$$\text{median} \in \mathcal{R}((3,1), 16, 3, 1). \qquad \square$$

C Proof of Proposition B.3

We start with an intermediate result.

Proposition C.1. *Let $M \in \mathbb{N}$ with $M \geq 3$, $E \in \mathbb{R}_+$, and let $X = (x_i)_{i=0}^{M-1}$ be a strictly increasing sequence taking values in $[0,1]$. Then, for all $u, v \in \mathbb{N}$ such that $u^2 v \geq M$, we have*

$$\begin{aligned}\Sigma(X,E) &\subseteq \mathcal{R}(20u, 30v, \max\{1, C_k M^6 R_m(X)(R_c(X))^3 E\}) \\ &\subseteq \mathcal{R}(20u, 30v, \max\{1, C_k M^6 (R_m(X))^4 E\}),\end{aligned}$$

for an absolute constant $C_k \in \mathbb{R}$ satisfying $2 \leq C_k \leq 10^5$, and where $R_m(X) := \max_{i=1,\dots,M}(x_i - x_{i-1})^{-1}$ and $R_c(X) := \frac{\max_{i=1,\dots,M-1}(x_i - x_{i-1})}{\min_{i=1,\dots,M-1}(x_i - x_{i-1})}$.

We now show how Proposition C.1 leads to the proof of Proposition B.3 and provide the proof of Proposition C.1 thereafter.

Proof of Proposition B.3. Application of Proposition C.1 to $\Sigma\big(X, \frac{1}{C_k M^6 (R_m(X))^4}\big)$, with C_k the constant in the statement of Proposition C.1, yields

$$\Sigma\left(X, \frac{1}{C_k M^6 (R_m(X))^4}\right) \subseteq \mathcal{R}(20u, 30v, 1). \tag{210}$$

We hence get

$$\begin{aligned}\mathcal{R}(20u, 30v, 2w) &\supseteq (2w)^{30v} \cdot \mathcal{R}(20u, 30v, 1) && (211)\\ &\supseteq (2w)^{30v} \cdot \Sigma\left(X, \frac{1}{C_k M^6 (R_m(X))^4}\right) && (212)\\ &\supseteq (2w)^{30v} \cdot \frac{1}{(2w)^{30v}} \cdot \Sigma\left(X, \frac{(2w)^{30v}}{C_k M^6 (R_m(X))^4}\right) && (213)\\ &\supseteq \Sigma(X,E), && (214)\end{aligned}$$

where in (211) we applied Proposition H.4 with $(W, L, L', B, B') = (20u, 30v, 0, 1, 2w)$, (212) follows from (210), and (214) is a consequence of (106) and $2^{30v} \geq 10^5 \geq C_k$. Further, in (213) we used that[11] $a \cdot \Sigma(X,b) \subseteq \Sigma(X, ab)$, for all $a, b \in \mathbb{R}_+$, which follows from the fact that for every $f \in \Sigma(X,b)$, $a \cdot f$ is a bounded piecewise linear function with breakpoints in X and $L^\infty(\mathbb{R})$-norm no greater than ab. □

It remains to prove Proposition C.1. The proof will be effected by representing the functions in $\Sigma(X,B)$ in terms of a specific basis for the linear space $\Sigma(X,\infty)$. Crucially, the elements of this basis will be realized by ReLU networks with suitable properties. Concretely, we shall work with the basis $\{\gamma_i : \mathbb{R} \mapsto \mathbb{R}\}_{i=0}^{M-1}$ given by

$$\gamma_0(x) = \begin{cases} 1, & x \in (-\infty, x_0], \\ 1 - \dfrac{x - x_0}{x_1 - x_0}, & x \in (x_0, x_1], \\ 0, & x \in (x_1, \infty), \end{cases} \tag{215}$$

[11] While the reverse inclusion is also valid, it will not be needed here.

for $i = 1, \ldots, M-2$,

$$\gamma_i(x) = \begin{cases} 0, & x \in (-\infty, x_{i-1}] \cup (x_{i+1}, \infty), \\ \dfrac{x - x_{i-1}}{x_i - x_{i-1}}, & x \in (x_{i-1}, x_i], \\ 1 - \dfrac{x - x_i}{x_{i+1} - x_i}, & x \in (x_i, x_{i+1}], \end{cases} \tag{216}$$

and

$$\gamma_{M-1}(x) = \begin{cases} 0, & x \in (-\infty, x_{M-2}], \\ \dfrac{x - x_{M-2}}{x_{M-1} - x_{M-2}}, & x \in (x_{M-2}, x_{M-1}], \\ 1, & x \in (x_{M-1}, \infty). \end{cases} \tag{217}$$

We note that $\gamma_0, \ldots, \gamma_{M-1} \in \Sigma(X, 1)$, and for $i = 0, \ldots, M-1$, $j = 0, \ldots, M-1$, we have the interpolation property

$$\gamma_i(x_j) = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases}$$

An illustration of the basis $\{\gamma_i : \mathbb{R} \mapsto \mathbb{R}\}_{i=0}^{M-1}$ is provided in Figure 1.

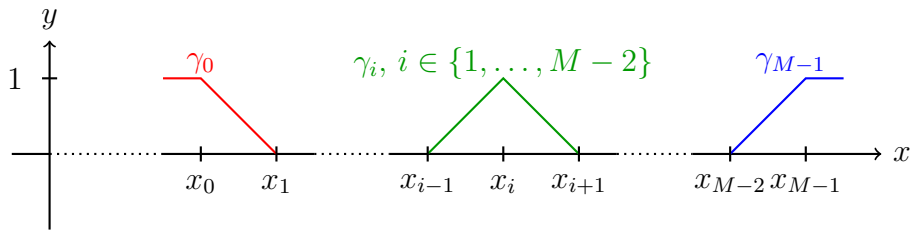

Figure 1: The basis $\{\gamma_i\}_{i=0}^{M-1}$ for $\Sigma(X, \infty)$.

For every $f \in \Sigma(X, \infty)$, we have

$$f(x) = \sum_{i=0}^{M-1} f(x_i)\gamma_i(x), \qquad \text{for all } x \in \mathbb{R}, \tag{218}$$

which is a consequence of the fact that f and $\sum_{i=0}^{M-1} f(x_i)\gamma_i$, by virtue of both being bounded piecewise linear functions with the same breakpoints and the same function values on these breakpoints, must be identical. This shows that $\{\gamma_i\}_{i=0}^{M-1}$ is, indeed, a basis for $\Sigma(X, \infty)$. Based on (218), we proceed to the next building block of our proof.

Lemma C.2. *Let $M \in \mathbb{N}$ with $M \geq 3$, $E \in \mathbb{R}_+$, and let $X = (x_i)_{i=0}^{M-1}$ be a strictly increasing sequence taking values in $[0, 1]$. Then, every function $f \in \Sigma(X, E)$ can be represented as*

$$f(x) = b + \sum_{i=0}^{M-1} a_i \, \rho(x - x_i), \quad x \in \mathbb{R},$$

for some $b, a_0, \ldots, a_{M-1} \in \mathbb{R}$ with $\max\{|b|, |a_0|, \ldots, |a_{M-1}|\} \leq 4R_m(X)E$, where $R_m(X) := \max_{i=1}^{M}(x_i - x_{i-1})^{-1}$. In particular, $f \in \mathcal{R}(M, 2, 4R_m(X)E)$.

Proof. The following representations of the functions $\{\gamma_0, \dots, \gamma_{M-1}\}$ can be read off directly from their definition

$$\begin{aligned}
\gamma_0(x) &= -\frac{1}{x_1 - x_0}\,\rho(x - x_0) + \frac{1}{x_1 - x_0}\,\rho(x - x_1) + 1, \\
\gamma_i(x) &= \frac{1}{x_i - x_{i-1}}\,\rho(x - x_{i-1}) \\
&\quad - \left(\frac{1}{x_i - x_{i-1}} + \frac{1}{x_{i+1} - x_i}\right)\rho(x - x_i) \\
&\quad + \frac{1}{x_{i+1} - x_i}\,\rho(x - x_{i+1}), \quad i = 1, \dots, M-2, \\
\gamma_{M-1}(x) &= \frac{1}{x_{M-1} - x_{M-2}}\,\rho(x - x_{M-2}) - \frac{1}{x_{M-1} - x_{M-2}}\,\rho(x - x_{M-1}),
\end{aligned} \tag{219}$$

for $x \in \mathbb{R}$ in all cases. Inserting (219) into (218) yields

$$\begin{aligned}
f(x) = f(x_0) &+ \left(-\frac{f(x_0)}{x_1 - x_0} + \frac{f(x_1)}{x_1 - x_0}\right)\rho(x - x_0) \\
&+ \sum_{i=1}^{M-2}\left(\frac{f(x_{i+1})}{x_{i+1} - x_i} - \frac{f(x_i)}{x_i - x_{i-1}} - \frac{f(x_i)}{x_{i+1} - x_i} + \frac{f(x_{i-1})}{x_i - x_{i-1}}\right)\rho(x - x_i) \\
&+ \left(\frac{f(x_{M-2})}{x_{M-1} - x_{M-2}} - \frac{f(x_{M-1})}{x_{M-1} - x_{M-2}}\right)\rho(x - x_{M-1}), \quad \text{for } x \in \mathbb{R}.
\end{aligned}$$

We conclude the proof by setting $b = f(x_0)$, $a_0 = -\frac{f(x_0)}{x_1 - x_0} + \frac{f(x_1)}{x_1 - x_0}$, $a_i = \frac{f(x_{i+1})}{x_{i+1} - x_i} - \frac{f(x_i)}{x_i - x_{i-1}} - \frac{f(x_i)}{x_{i+1} - x_i} + \frac{f(x_{i-1})}{x_i - x_{i-1}}$, for $i = 1, \dots, M-2$, and $a_{M-1} = \frac{f(x_{M-2})}{x_{M-1} - x_{M-2}} - \frac{f(x_{M-1})}{x_{M-1} - x_{M-2}}$, so that, indeed, $\max\{|b|, |a_0|, \dots, |a_{M-1}|\} \leq 4\|f\|_{L^\infty([0,1])} \max_{i=1,\dots,M-1}(x_i - x_{i-1})^{-1} \leq 4R_m(X)E$. □

We next provide a lemma describing ReLU network realizations of the basis functions $\{\gamma_i\}_{i=0}^{M-1}$. These constructions are inspired by [19] and [21].

Lemma C.3. *Let $t, u \in \mathbb{N}$ with $t \geq 8$, set $M = tu$ and*

$$\mathcal{I} := \{(k, \ell) : k \in \{0, \dots, u-1\}, \ell \in \{0, \dots, t-1\}\}. \tag{220}$$

We have

$$\{0, \dots, M-1\} = \{kt + \ell : (k, \ell) \in \mathcal{I}\}. \tag{221}$$

Let $X = (x_i)_{i=0}^{M-1}$ be a strictly increasing sequence taking values in $[0, 1]$ and let $\{\gamma_i\}_{i=0}^{M-1}$ be the basis for $\Sigma(X, \infty)$ defined in (215)-(217). Then, there exist $f_{k,\ell}^1, f_{k,\ell}^2, f_{k,\ell}^3 \in \Sigma(X, \infty)$, for all $(k, \ell) \in \mathcal{I}$, such that the following statements hold.

- *(Property 1) For all $(k, \ell) \in \mathcal{I}$, the basis function $\gamma_{kt+\ell}$ can be realized according to*

$$\gamma_{kt+\ell} = \rho \circ f_{k,\ell}^1 - \rho \circ f_{k,\ell}^2 + \rho \circ f_{k,\ell}^3.$$

- *(Property 2) Let $(z_i)_{i=0}^{8u-1}$ be the strictly increasing sequence obtained by sorting the elements in*

$$\{x_{kt+\ell} : k \in \{0, \dots, u-1\}, \ell \in \{0, 1, 2, 3, t-4, t-3, t-2, t-1\}\}. \tag{222}$$

For all $(k,\ell) \in \mathcal{I}$, and for $j \in \{1,2,3\}$, the function $f_{k,\ell}^j$ can be written as

$$f_{k,\ell}^j(x) = b + \sum_{i=0}^{8u-1} a_i \rho(x - z_i), \quad x \in \mathbb{R}, \tag{223}$$

for $b, a_0, \ldots, a_{8u-1} \in \mathbb{R}$ depending on k, ℓ, j and such that

$$|b|, |a_0|, \ldots, |a_{8u-1}| \leq 12t^2 R_m(X)(R_c(X))^3,$$

where $R_m(X) := \max_{i=1,\ldots,M-1}(x_i - x_{i-1})^{-1}$ and $R_c(X) := \frac{\max_{i=1,\ldots,M-1}(x_i - x_{i-1})}{\min_{i=1,\ldots,M-1}(x_i - x_{i-1})}$.

- *(Property 3) For every $(\ell, j) \in \{0, 1 \ldots, t-1\} \times \{1,2,3\}$, the functions in $\{f_{k,\ell}^j\}_{k=0}^{u-1}$ have pairwise disjoint supports*

Proof. See Appendix C.1. □

Further, we need the following technical lemma, which realizes a 3-layer ReLU network of a specific form by an equivalent deeper ReLU network. The result is inspired by the construction reported in [9, Lemma 4.2].

Lemma C.4. *Let $u, s, r \in \mathbb{N}$, $(z_i)_{i=0}^{r-1} \subseteq [0,1]$, and $T \in \mathbb{R}_+$. For $\ell = 0, \ldots, us-1$, let $h_\ell : \mathbb{R} \mapsto \mathbb{R}$ be given by*

$$h_\ell(x) = d_\ell + \sum_{i=0}^{r-1} c_{\ell,i} \rho(x - z_i), \tag{224}$$

with $d_\ell, c_{\ell,0}, \ldots, c_{\ell,r-1} \in \{x \in \mathbb{R} : |x| \leq T\}$, and let

$$H = \sum_{\ell=0}^{us-1} (\rho \circ h_\ell). \tag{225}$$

Then,

$$H \in \mathcal{R}(r + u + 1, s + 2, \max\{1, T\}).$$

Proof. See Appendix C.2. □

We proceed to the proof of Proposition C.1, which will be effected through Lemmata C.3 and C.4.

Proof of Proposition C.1. Let $g \in \Sigma(X, E)$ with $X = (x_\ell)_{\ell=0}^{M-1}$, $M \geq 3$, a strictly increasing sequence taking values in $[0,1]$ and $E \in \mathbb{R}_+$. We start with the special case $M = u^2 v$ and $uv \geq 8$, and will later reduce the other cases to this one. Write g as $H^+ - H^-$, for some H^+ and H^-, both of the form (225). Setting $t = uv \geq 8$, we can write $M = tu$. According to Lemma C.3, there exist $f_{k,\ell}^j$, $k = 0, \ldots, u-1$, $\ell = 0, \ldots, t-1$, $j = 1,2,3$, such that Properties 1-3 in the statement of Lemma C.3 hold. For $i = 0, \ldots, M-1$, let $y_i = g(x_i)$, $y_i^+ = \max\{y_i, 0\}$, $y_i^- = \max\{-y_i, 0\}$, and note that

$$y_i = y_i^+ - y_i^- \tag{226}$$

$$|y_i^+|, |y_i^-| \leq |y_i| \leq \|g\|_{L^\infty([0,1])} \leq E. \tag{227}$$

We then have, for $x \in \mathbb{R}$,

$$g(x) = \sum_{\ell=0}^{M-1} g(x_i)\gamma_i(x) \tag{228}$$

$$= \sum_{i=0}^{M-1} y_i \gamma_i(x) \tag{229}$$

$$= \sum_{\ell=0}^{t-1} \sum_{k=0}^{u-1} y_{kt+\ell} \gamma_{kt+\ell}(x) \tag{230}$$

$$= \sum_{\ell=0}^{t-1} \sum_{k=0}^{u-1} (y^+_{kt+\ell} - y^-_{kt+\ell}) \Big(\rho(f^1_{k,\ell}(x)) - \rho(f^2_{k,\ell}(x)) + \rho(f^3_{k,\ell}(x)) \Big) \tag{231}$$

$$= \sum_{\ell=0}^{t-1} \sum_{k=0}^{u-1} \Big(\rho(y^+_{kt+\ell} f^1_{k,\ell}(x)) + \rho(y^-_{kt+\ell} f^2_{k,\ell}(x)) + \rho(y^+_{kt+\ell} f^3_{k,\ell}(x)) \\ - \rho(y^-_{kt+\ell} f^1_{k,\ell}(x)) - \rho(y^+_{kt+\ell} f^2_{k,\ell}(x)) - \rho(y^-_{kt+\ell} f^3_{k,\ell}(x)) \Big) \tag{232}$$

$$= \sum_{\ell=0}^{t-1} \Bigg(\rho\Bigg(\sum_{k=0}^{u-1} y^+_{kt+\ell} f^1_{k,\ell}(x) \Bigg) + \rho\Bigg(\sum_{k=0}^{u-1} y^-_{kt+\ell} f^2_{k,\ell}(x) \Bigg) + \rho\Bigg(\sum_{k=0}^{u-1} y^+_{kt+\ell} f^3_{k,\ell}(x) \Bigg) \\ - \rho\Bigg(\sum_{k=0}^{u-1} y^-_{kt+\ell} f^1_{k,\ell}(x) \Bigg) - \rho\Bigg(\sum_{k=0}^{u-1} y^+_{kt+\ell} f^2_{k,\ell}(x) \Bigg) - \rho\Bigg(\sum_{k=0}^{u-1} y^-_{kt+\ell} f^3_{k,\ell}(x) \Bigg) \Bigg), \tag{233}$$

where (231) follows from Property 1 in Lemma C.3, in (232) we used the positive homogeneity of the ReLU function, i.e., $\rho(xy) = x\rho(y)$, for all $y \in \mathbb{R}$ and $x \in \mathbb{R}_+ \cup \{0\}$, and (233) is a consequence of the functions in $\{f^j_{k,\ell}\}_{k=0}^{u-1}$ being of pairwise-disjoint support, for all $\ell \in \{0, 1 \ldots, t-1\}$, $j \in \{1, 2, 3\}$, as guaranteed by Property 3 in Lemma C.3.

For $\ell = 0, \ldots, t-1$, now define

$$h^+_\ell = \sum_{k=0}^{u-1} y^+_{kt+\ell}\, f^1_{k,\ell}, \tag{234}$$

$h^+_{t+\ell} = \sum_{k=0}^{u-1} y^-_{kt+\ell} f^2_{k,\ell}$, $h^+_{2t+\ell} = \sum_{k=0}^{u-1} y^+_{kt+\ell} f^3_{k,\ell}$, $h^-_\ell = \sum_{k=0}^{u-1} y^-_{kt+\ell} f^1_{k,\ell}$, $h^-_{t+\ell} = \sum_{k=0}^{u-1} y^+_{kt+\ell} f^2_{k,\ell}$, $h^-_{2t+\ell} = \sum_{k=0}^{u-1} y^-_{kt+\ell} f^3_{k,\ell}$, and let

$$H^+ = \sum_{\ell=0}^{3t-1} \rho \circ h^+_\ell, \quad H^- = \sum_{\ell=0}^{3t-1} \rho \circ h^-_\ell.$$

By (228)-(233), we can therefore write

$$g = H^+ - H^-.$$

We next show that h^+_ℓ, h^-_ℓ, $\ell = 0, \ldots, 3t-1$ are all of the form (224) as required by Lemma C.4. First, consider h^+_0. Thanks to Property 2 in Lemma C.3, for all $(k, \ell) \in \mathcal{I}$ and $j \in \{1, 2, 3\}$, one has

$$f^j_{k,\ell}(x) = b^{(j,k,\ell)} + \sum_{i=0}^{8u-1} a_i^{(j,k,\ell)} \rho(x - z_i), \tag{235}$$

with the strictly increasing sequence $(z_i)_{i=0}^{8u-1}$ obtained by sorting the elements in $\{x_{kt+\ell} : k \in \{0,\dots,u-1\}, \ell \in \{0,1,2,3,t-4,t-3,t-2,t-1\}\}$, and $b^{(j,k,\ell)}, a_0^{(j,k,\ell)}, \dots, a_{8u-1}^{(j,k,\ell)} \in \mathbb{R}$ such that

$$|b^{(j,k,\ell)}|, \max_{i=0,\dots,8u-1} |a_i^{(j,k,\ell)}| \leq 12\, t^2 R_m(X)(R_c(X))^3, \tag{236}$$

where $R_m(X)$ and $R_c(X)$ are as defined in Lemma C.3. We hence obtain

$$h_0^+(x) = \sum_{k=0}^{u-1} y_{kt}^+ f_{k,0}^1(x) \tag{237}$$

$$= \sum_{k=0}^{u-1} y_{kt}^+ \left(b^{(1,k,0)} + \sum_{i=0}^{8u-1} a_i^{(1,k,0)} \rho(x - z_i) \right) \tag{238}$$

$$= d_0^+ + \sum_{i=0}^{8u-1} c_{0,i}^+ \rho(x - z_i),\ x \in \mathbb{R}, \tag{239}$$

where in (239) we set $d_0^+ := \sum_{k=0}^{u-1} y_{kt}^+ b^{(1,k,0)}$ and $c_{0,i}^+ := \sum_{k=0}^{u-1} y_{kt}^+ a_i^{(1,k,0)}$, for $i = 0,\dots,8u-1$. Moreover, we note that $|d_0^+| \leq \sum_{k=0}^{u-1} |y_{kt}^+||b^{(1,k,0)}| \leq 12ut^2 R_m(X)(R_c(X))^3 E$ and $|c_{0,i}^+| = \sum_{k=0}^{u-1} |y_{kt}^+||a_i^{(1,k,0)}| \leq 12ut^2 R_m(X)(R_c(X))^3 E$, $i = 0,\dots,8u-1$, where we used (227) and (236). Analogously, one can show that, for $\ell = 0,\dots,3t-1$,

$$h_\ell^+(x) = d_\ell^+ + \sum_{i=0}^{8u-1} c_{\ell,i}^+ \rho(x - z_i) \tag{240}$$

$$h_\ell^-(x) = d_\ell^- + \sum_{i=0}^{8u-1} c_{\ell,i}^- \rho(x - z_i), \tag{241}$$

with $d_\ell^+, d_\ell^-, c_{\ell,0}^+, \dots, c_{\ell,8u-1}^+, c_{\ell,0}^-, \dots, c_{\ell,8u-1}^- \in \mathbb{R}$ of absolute values not exceeding the term $12ut^2 R_m(X)(R_c(X))^3 E$. We recall that $H^+ = \sum_{\ell=0}^{3uv-1} \rho \circ h_\ell^+$ and $H^- = \sum_{\ell=0}^{3uv-1} \rho \circ h_\ell^-$, and apply Lemma C.4 to H^+ and H^-, individually, with (u,s,r) replaced by $(u,3v,8u)$ and $T = 12ut^2 R_m(X)(R_c(X))^3 E$, to get

$$H^+ \in \mathcal{R}(9u+1, 3v+2, \max\{1, 12ut^2 R_m(X)(R_c(X))^3 E\}),$$
$$H^- \in \mathcal{R}(9u+1, 3v+2, \max\{1, 12ut^2 R_m(X)(R_c(X))^3 E\}).$$

We conclude the proof for the special case $M = u^2 v$ and $uv \geq 8$ by noting that

$$g = H^+ - H^- \in \mathcal{R}\big(18u+2, 3v+3, \max\{1, 12ut^2 R_m(X)(R_c(X))^3 E\}\big) \tag{242}$$

$$\subseteq \mathcal{R}\big(18u+2, 3v+3, \max\{1, 12M^2 R_m(X)(R_c(X))^3 E\}\big), \tag{243}$$

where (242) follows from Lemma H.3, and (243) is by $ut^2 = u^3 v^2 \leq M^2$.

We proceed to the case of general $u, v, M \in \mathbb{N}$, with $M \geq 3$ and $u^2 v \geq M$. The proof will be effected by reducing to the special case just established. To this end, we shall find $\hat{u}, \hat{v}$ and a strictly increasing sequence $(\hat{x}_i)_{i=0}^{\hat{u}^2\hat{v}-1}$ taking values in $[0,1]$ such that $(x_i)_{i=0}^{M-1} \subseteq (\hat{x}_i)_{i=0}^{\hat{u}^2\hat{v}-1}$ and $g \in \Sigma((\hat{x}_i)_{i=0}^{\hat{u}^2\hat{v}-1}, E)$. Concretely, let $(\hat{u}, \hat{v})$ be a solution[12] to the following constrained optimization problem

$$(\hat{u}, \hat{v}) = \operatorname*{argmin}_{(e,f) \in \mathbb{G} \cap \mathbb{H}} (e + f), \tag{244}$$

[12] If there are multiple solutions, we simply pick any one of them.

with $\mathbb{G} := \{(e,f) \in \mathbb{N}^2 : e \le u, 8 \le f \le 8v\}$ and $\mathbb{H} = \{(e,f) \in \mathbb{N}^2 : e^2 f \ge M\}$, and set $\hat{M} = \hat{u}^2\hat{v}$. Here, the choice of $\mathbb{G}$ ensures that the condition $\hat{u}\hat{v} \ge 8$ is met while $\hat{v}$ is not much greater than v. Note that we enforce $\hat{u}\hat{v} \ge 8$ so as to, indeed, be able to apply the result for the special case dealt with above. Since $(u, 8v) \in \mathbb{G} \cap \mathbb{H}$, the feasible region $\mathbb{G} \cap \mathbb{H}$ is non-empty, which together with $\mathbb{G} \cap \mathbb{H}$ being finite guarantees the existence of a minimizer to (244). Now, let $\hat{X} = (\hat{x}_i)_{i=0}^{\hat{M}-1}$ with $\hat{x}_i = x_i$, for $i = 0, \dots, M-2$, and $\hat{x}_i = x_{M-2} + (i - M + 2) \cdot \frac{x_{M-1} - x_{M-2}}{\hat{M} - M + 1}$, for $i = M-1, \dots, \hat{M}-1$. As $\hat{x}_{\hat{M}-1} = x_{M-1}$, it follows that $X \subseteq \hat{X}$. Intuitively, $\hat{X}$ is obtained by adding $\hat{M} - M$ points in the interval $[x_{M-2}, x_{M-1}]$ to X chosen such that $[x_{M-2}, x_{M-1}]$ is partitioned into $\hat{M} - M + 1$ subintervals of equal length. It then follows that $g \in \Sigma(X,E) \subseteq \Sigma(\hat{X}, E)$. Using (242)-(243) with (X, u, v) replaced by $(\hat{X}, \hat{u}, \hat{v})$, we get

$$g \in \mathcal{R}\big(18\hat{u} + 2, 3\hat{v} + 3, \max\{1, 12\hat{M}^2 R_m(\hat{X})(R_c(\hat{X}))^3 E\}\big). \tag{245}$$

Next, we note that

$$\begin{aligned}
R_m(\hat{X}) &= \max_{i=1,\dots,\hat{M}} (\hat{x}_i - \hat{x}_{i-1})^{-1} \\
&= \max\left(\max_{i=1,\dots,M-2} (\hat{x}_i - \hat{x}_{i-1})^{-1}, \max_{i=M-1,\dots,\hat{M}-1} (\hat{x}_i - \hat{x}_{i-1})^{-1} \right) \\
&= \max\left(\max_{i=1,\dots,M-2} (x_i - x_{i-1})^{-1}, \left(\frac{x_{M-1} - x_{M-2}}{\hat{M} - M + 1} \right)^{-1} \right) \\
&\le (\hat{M} - M + 1) \max_{i=1,\dots,M-1} (x_i - x_{i-1})^{-1} \\
&= (\hat{M} - M + 1) R_m(X)
\end{aligned}$$

and

$$\begin{aligned}
R_c(\hat{X}) &= \frac{\max_{i=1,\dots,\hat{M}-1}(\hat{x}_i - \hat{x}_{i-1})}{\min_{i=1,\dots,\hat{M}-1}(\hat{x}_i - \hat{x}_{i-1})} \\
&= \frac{(\min_{i=1,\dots,\hat{M}-1}(\hat{x}_i - \hat{x}_{i-1})^{-1})^{-1}}{(\max_{i=1,\dots,\hat{M}-1}(\hat{x}_i - \hat{x}_{i-1})^{-1})^{-1}} \\
&= \frac{\max_{i=1,\dots,\hat{M}-1}(\hat{x}_i - \hat{x}_{i-1})^{-1}}{\min_{i=1,\dots,\hat{M}-1}(\hat{x}_i - \hat{x}_{i-1})^{-1}} \\
&\le \frac{(\hat{M} - M + 1)\max_{i=1,\dots,M-1}(x_i - x_{i-1})^{-1}}{\min_{i=1,\dots,M-1}(x_i - x_{i-1})^{-1}} \\
&= \frac{(\hat{M} - M + 1)\max_{i=1,\dots,M-1}(x_i - x_{i-1})}{\min_{i=1,\dots,M-1}(x_i - x_{i-1})} \\
&= (\hat{M} - M + 1) R_c(X),
\end{aligned}$$

which, when used in (245), yields

$$g \subseteq \mathcal{R}(18\hat{u} + 2, 3\hat{v} + 3, \max\{1, 12\hat{M}^2((\hat{M} - M + 1)R_m(X))((\hat{M} - M + 1)R_c(X))^3 E\}). \tag{246}$$

Finally, owing to $\hat{u} \le u$, $\hat{v} \le 8v$, and $\hat{M} - M + 1 \le \hat{M}$, it follows from (246) that

$$g \subseteq \mathcal{R}(18u + 2, 24v + 3, \max\{1, 12\hat{M}^6 R_m(X)(R_c(X))^3 E\}). \tag{247}$$

We next want to get rid of $\hat{M}$ in (247). This will be accomplished by establishing a quantitative relation between $\hat{M}$ and M organized into three different cases, which, taken together, are exhaustive in terms of the parameter range under consideration.

1. If $\hat{u} \geq 2$, it must hold that $(\hat{u}-1, \hat{v}) \notin \mathbb{H}$, i.e., $(\hat{u}-1)^2\hat{v} < M$, otherwise $(\hat{u}, \hat{v})$ can not be a solution to (244) as the objective function takes on a smaller value at $(\hat{u}-1, \hat{v}) \in \mathbb{G} \cap \mathbb{H}$. We therefore have $\hat{M} = \hat{u}^2\hat{v} \leq 4(\hat{u}-1)^2\hat{v} < 4M$.

2. If $\hat{v} \geq 9$, we must have $(\hat{u}, \hat{v}-1) \notin \mathbb{H}$, i.e., $\hat{u}^2(\hat{v}-1) < M$, as, again, otherwise $(\hat{u}, \hat{v})$ can not be a solution to (244). We then have $\hat{M} = \hat{u}^2\hat{v} \leq 2\hat{u}^2(\hat{v}-1) < 2M$.

3. If $\hat{u} = 1$ and $\hat{v} = 8$, we obtain $\hat{M} = \hat{u}^2\hat{v} = 8 \leq 4M$ owing to $M \geq 3$.

In all cases, one gets $\hat{M} \leq 4M$, so that (247) yields

$$g \in \mathcal{R}(18u+2, 24v+3, \max\{1, 12 \cdot 4^6 M^6 R_m(X)(R_c(X))^3 E\}).$$

The proof is concluded by setting $C_k = 12 \cdot 4^6 \leq 10^5$ and noting that

$$R_c(X) = \frac{\max_{i=1,\dots,M-1}(x_i - x_{i-1})}{\min_{i=1,\dots,M-1}(x_i - x_{i-1})} \leq \max_{i=1,\dots,M}(x_i - x_{i-1})^{-1} = R_m(X),$$

where the inequality follows from $(x_i)_{i=0}^{M-1} \in [0,1]$ and hence $\min_{i=1,\dots,M-1}(x_i - x_{i-1}) \leq 1$. □

C.1 Proof of Lemma C.3

The proof is constructive in the sense of explicitly specifying $f_{k,\ell}^1, f_{k,\ell}^2, f_{k,\ell}^3$, for $(k,\ell) \in \mathcal{I}$. We consider the cases $(k,\ell) \in \mathcal{I}^1$ with

$$\mathcal{I}^1 := \{(k,\ell) \in \mathcal{I} : \ell \in \{0, 1, 2, t-3, t-2, t-1\}\},$$

and $(k,\ell) \in \mathcal{I}^2$ with

$$\mathcal{I}^2 := \{(k,\ell) \in \mathcal{I} : \ell \in \{3, \dots, t-4\}\},$$

separately as the corresponding constructions are different. The sets $\mathcal{I}^1$ and $\mathcal{I}^2$ are nonempty and, owing to $t \geq 8$, disjoint. Further, $\mathcal{I}$ is the union of $\mathcal{I}^1$ and $\mathcal{I}^2$. The verification of the three properties to be established is conducted right after the respective constructions.

Fix $(k,\ell) \in \mathcal{I}^1$. Let $f_{k,\ell}^1 = \gamma_{kt+\ell}$, and let $f_{k,\ell}^2(x) = f_{k,\ell}^3(x) = 0$, $x \in \mathbb{R}$. Regarding Property 1, we have

$$\begin{aligned} \gamma_{kt+\ell}(x) &= \rho(\gamma_{kt+\ell}(x)) && (248)\\ &= \rho(f_{k,\ell}^1(x)) - \rho(f_{k,\ell}^2(x)) + \rho(f_{k,\ell}^3(x)), \quad x \in \mathbb{R}, && (249) \end{aligned}$$

where in (248) we used $\gamma_{kt+\ell}(x) \geq 0$, for $x \in \mathbb{R}$.

Property 2 is satisfied, for $j = 2, 3$, as $f_{k,\ell}^2(x) = f_{k,\ell}^3(x) = 0 = \sum_{i=0}^{8u-1} 0 \cdot \rho(x - z_i)$, for $x \in \mathbb{R}$. For $j = 1$, thanks to (219), we note that

1. $f_{k,\ell}^1(x) = -\frac{1}{x_1 - x_0}\rho(x - x_0) + \frac{1}{x_1 - x_0}\rho(x - x_1) + 1$, $x \in \mathbb{R}$, if $k = \ell = 0$,

2. $f_{k,\ell}^1(x) = \frac{1}{x_{ut-1} - x_{ut-2}}\rho(x - x_{ut-2}) - \frac{1}{x_{ut-1} - x_{ut-2}}\rho(x - x_{ut-1})$, $x \in \mathbb{R}$, if $k = u-1$, $\ell = t-1$,

3. $f^1_{k,\ell}(x) = \frac{1}{x_{kt+\ell}-x_{kt+\ell-1}}\rho(x-x_{kt+\ell-1})-\big(\frac{1}{x_{kt+\ell}-x_{kt+\ell-1}}+\frac{1}{x_{kt+\ell+1}-x_{kt+\ell}}\big)\rho(x-x_{kt+\ell})+\frac{1}{x_{kt+\ell+1}-x_{kt+\ell}}\cdot \rho(x-x_{kt+\ell+1}),\ x\in\mathbb{R}$, if $(k,\ell)\in\mathcal{I}^1\backslash\{(0,0),(u-1,t-1)\}$.

In each of these three cases, upon recalling the definition of $(z_i)_{i=0}^{8u-1}$ in (222), $f^1_{k,\ell}$ can be written as

$$f^1_{k,\ell}(x) = b + \sum_{\ell=0}^{8u-1} a_\ell \rho(x-z_i),\ x\in\mathbb{R},$$

for some $b, a_0, \ldots, a_{8u-1}\in\mathbb{R}$ such that $|b|, |a_0|, \ldots, |a_{8u-1}| \leq \max\{1, 2\max_{i=1,\ldots,M-1}(x_i-x_{i-1})^{-1}\}$ $= \max\{1, 2R_m(X)\} < 12t^2R_m(X)(R_c(X))^3$, where the inequality follows from $t\geq 1$, $R_c(X)\geq 1$, and $R_m(X)\geq 1$. Since the choice of $(k,\ell)\in\mathcal{I}^1$ was arbitrary, we have given the construction of $f^1_{k,\ell}, f^2_{k,\ell}, f^3_{k,\ell}$ and established Properties 1 and 2 for all $(k,\ell)\in\mathcal{I}^1$.

We proceed to verify Property 3 for $(k,\ell)\in\mathcal{I}^1$. To this end, we first note that for $(\ell,j)\in\{0,1,2,t-3,t-2,t-1\}\times\{2,3\}$, the functions $\{f^j_{k,\ell}\}_{k=0}^{u-1}$ are all identically equal to zero and hence trivially of pairwise disjoint support. For $(\ell,j)\in\{0,1,2,t-3,t-2,t-1\}\times\{1\}$, $\{f^1_{k,\ell}\}_{k=0}^{u-1} = \{\gamma_{kt+\ell}\}_{k=0}^{u-1}$ and the pairwise disjoint support is a consequence of (215)-(217).

We continue by constructing $f^1_{k,\ell}, f^2_{k,\ell}, f^3_{k,\ell}$ and verifying Properties 1-3 for $(k,\ell)\in\mathcal{I}^2$. Fix $(k,\ell)\in\mathcal{I}^2$. For $j=1,2,3$, let

$$f^j_{k,\ell}\in\Sigma((x_{kt}, x_{kt+1}, x_{kt+t-2}, x_{kt+t-1}), \infty) \tag{250}$$

be such that

$$f^j_{k,\ell}(x_{kt}) = 0, \tag{251}$$
$$f^j_{k,\ell}(x_{kt+\ell+j-2}) = 0, \tag{252}$$
$$f^j_{k,\ell}(x_{kt+t-2}) = y^j_{k,\ell}, \tag{253}$$
$$f^j_{k,\ell}(x_{kt+t-1}) = 0, \tag{254}$$

for a

$$y^j_{k,\ell} > 0 \tag{255}$$

to be specified later. We refer to Figure 2 for an illustration of the so-defined $f^j_{k,\ell}$, $j\in\{1,2,3\}$. Next, we remark that $f^j_{k,\ell}$, for $j=1,2,3$, is uniquely determined by $y^j_{k,\ell}$. To see this, first note that $f^j_{k,\ell}$ is affine on the interval $[x_{kt+1}, x_{kt+t-2}]$, by definition, and therefore the values of $f^j_{k,\ell}$ at the points $x_{kt+\ell+j-2}$ and x_{kt+t-2}, both of which are contained in the interval, determine the values of $f^j_{k,\ell}$ at the end points x_{kt+1}, x_{kt+t-2} of the interval. With (251) and (254), owing to (250), this determines $f^j_{k,\ell}$ itself. In preparation for the choice of $y^j_{k,\ell}$, we need to characterize $f^j_{k,\ell}$ and $\rho\circ f^j_{k,\ell}$ in more detail. Fix $j\in\{1,2,3\}$. We first investigate the sign of $f^j_{k,\ell}$ on $\mathbb{R}$. To this end, we use the fact that $f^j_{k,\ell}$ is affine on the interval $[x_{kt+1}, x_{kt+t-2}]$, by definition, and therefore satisfies the interpolation formula

$$f^j_{k,\ell}(x) = f^j_{k,\ell}(x_{kt+\ell+j-2}) + (x-x_{kt+\ell+j-2})\frac{f^j_{k,\ell}(x_{kt+t-2}) - f^j_{k,\ell}(x_{kt+\ell+j-2})}{x_{kt+t-2}-x_{kt+\ell+j-2}} \tag{256}$$
$$= y^j_{k,\ell}\frac{x-x_{kt+\ell+j-2}}{x_{kt+t-2}-x_{kt+\ell+j-2}},\quad x\in[x_{kt+1}, x_{kt+t-2}], \tag{257}$$

where we used (252) and (253). In particular, we have

$$f^j_{k,\ell}(x_{kt+1}) = y^j_{k,\ell}\frac{x_{kt+1} - x_{kt+\ell+j-2}}{x_{kt+t-2} - x_{kt+\ell+j-2}} < 0, \tag{258}$$

where we used that $\ell + j - 2 \in \{2, \dots, t-3\}$ and therefore $kt + 1 < kt + \ell + j - 2$ and $kt+t-2 > kt+\ell+j-2$. The values of $f^j_{k,\ell}$ at the points $\{x_{kt}, x_{kt+1}, x_{kt+\ell+j-2}, x_{kt+t-2}, x_{kt+t-1}\}$, as specified by (251), (258), (252), (253), (254), combined with (250), allow us to conclude that

$$f^j_{k,\ell}(x) \begin{cases} = 0, & x \in (-\infty, -x_{kt}] \cup \{x_{kt+\ell+j-2}\} \cup [x_{kt+t-1}, \infty), \\ > 0, & x \in (x_{kt+\ell+j-2}, x_{kt+t-1}), \\ < 0, & x \in (x_{kt}, x_{kt+\ell+j-2}), \end{cases} \tag{259}$$

which, in turn, implies

$$\rho(f^j_{k,\ell}(x)) = \begin{cases} f^j_{k,\ell}(x), & x \in (x_{kt+\ell+j-2}, x_{kt+t-1}), \\ 0, & x \in (-\infty, x_{kt+\ell+j-2}] \cup [x_{kt+t-1}, \infty). \end{cases} \tag{260}$$

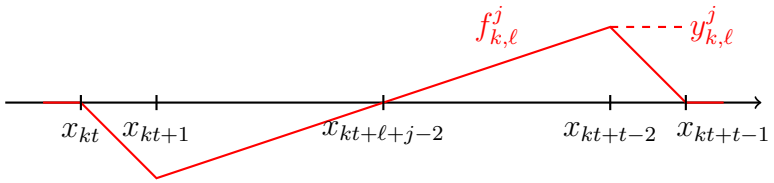

Figure 2: The function $f^j_{k,\ell}$.

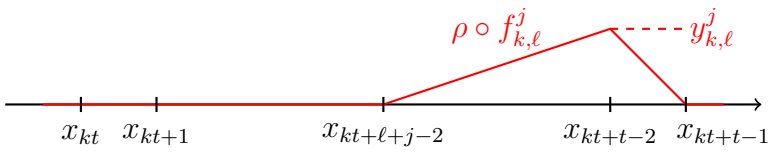

Figure 3: The function $\rho \circ f^j_{k,\ell}$.

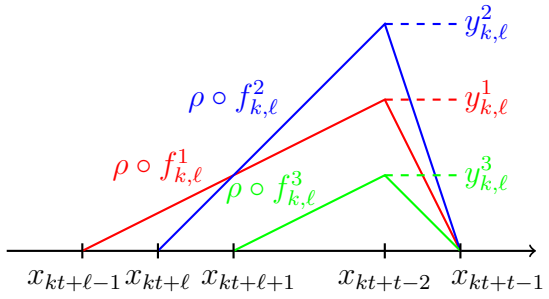

Figure 4: The functions $\rho \circ f^j_{k,\ell}$, $j = 1, 2, 3$.

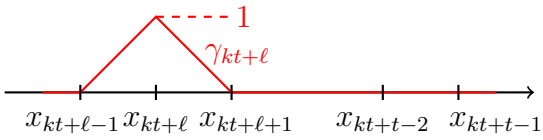

Figure 5: $\gamma_{kt+\ell} = \rho \circ f^1_{k,\ell} - \rho \circ f^2_{k,\ell} + \rho \circ f^3_{k,\ell}$

We refer to Figure 3 for an illustration of $\rho \circ f^j_{k,\ell}$. Owing to (250) and (260), we have

$$\rho \circ f^j_{k,\ell} \in \Sigma((x_{kt+\ell+j-2}, x_{kt+t-2}, x_{kt+t-1}), \infty). \tag{261}$$

As the choice of $j \in \{1,2,3\}$ was arbitrary, we have constructed $f^j_{k,\ell}$, $j = 1,2,3$, satisfying (250)-(261).

We proceed to determine $y^j_{k,\ell}$, $j = 1,2,3$, such that Property 1 is satisfied. As the functions $\gamma_{kt+\ell}$, $\rho \circ f^1_{k,\ell}$, $\rho \circ f^2_{k,\ell}$, and $\rho \circ f^3_{k,\ell}$ are all bounded piecewise linear functions with breakpoints $\{x_{kt+\ell-1}, x_{kt+\ell}, x_{kt+\ell+1}, x_{kt+t-2}, x_{kt+t-1}\}$, Property 1 is equivalent to

$$\gamma_{kt+\ell}(x_i) = \rho(f^1_{k,\ell}(x_i)) - \rho(f^2_{k,\ell}(x_i)) + \rho(f^3_{k,\ell}(x_i)), \tag{262}$$

for $i = kt+\ell-1, kt+\ell, kt+\ell+1, kt+t-2, kt+t-1$. As all of the functions $\gamma_{kt+\ell}$, $\rho \circ f^1_{k,\ell}$, $\rho \circ f^2_{k,\ell}$, $\rho \circ f^3_{k,\ell}$ take on the value 0 at $x_{kt+\ell-1}$ and x_{kt+t-1}, the relation (262) holds trivially for $i = kt+\ell-1, kt+t-1$ regardless of the choice of $y^j_{k,\ell}$, $j = 1,2,3$. Substituting (260) into (262), for $i = kt+\ell$, $kt+\ell+1$, $kt+t-2$ yields

$$1 = f^1_{k,\ell}(x_{kt+\ell}), \tag{263}$$
$$0 = f^1_{k,\ell}(x_{kt+\ell+1}) - f^2_{k,\ell}(x_{kt+\ell+1}), \tag{264}$$
$$0 = f^1_{k,\ell}(x_{kt+t-2}) - f^2_{k,\ell}(x_{kt+t-2}) + f^3_{k,\ell}(x_{kt+t-2}). \tag{265}$$

Next, using the interpolation formula (256)-(257) in (263)-(265), we can rewrite (263)-(265) in terms of $y^1_{k,\ell}, y^2_{k,\ell}, y^3_{k,\ell}$ according to

$$1 = y^1_{k,\ell} \frac{x_{kt+\ell} - x_{kt+\ell-1}}{x_{kt+t-2} - x_{kt+\ell-1}}, \tag{266}$$
$$0 = y^1_{k,\ell} \frac{x_{kt+\ell+1} - x_{kt+\ell-1}}{x_{kt+t-2} - x_{kt+\ell-1}} - y^2_{\ell,k} \frac{x_{kt+\ell+1} - x_{kt+\ell}}{x_{kt+t-2} - x_{kt+\ell}}, \tag{267}$$
$$0 = y^1_{k,\ell} - y^2_{k,\ell} + y^3_{k,\ell}, \tag{268}$$

which has the unique solution

$$y^1_{k,\ell} = \frac{x_{kt+t-2} - x_{kt+\ell-1}}{x_{kt+\ell} - x_{kt+\ell-1}}, \tag{269}$$
$$y^2_{k,\ell} = y^1_{k,\ell} \frac{x_{kt+\ell+1} - x_{kt+\ell-1}}{x_{kt+t-2} - x_{kt+\ell-1}} \cdot \frac{x_{kt+t-2} - x_{kt+\ell}}{x_{kt+\ell+1} - x_{kt+\ell}} \tag{270}$$
$$= \frac{x_{kt+\ell+1} - x_{kt+\ell-1}}{x_{kt+\ell} - x_{kt+\ell-1}} \frac{x_{kt+t-2} - x_{kt+\ell}}{x_{kt+\ell+1} - x_{kt+\ell}}, \tag{271}$$
$$y^3_{k,\ell} = \frac{x_{kt+\ell+1} - x_{kt+\ell-1}}{x_{kt+\ell} - x_{kt+\ell-1}} \frac{x_{kt+t-2} - x_{kt+\ell}}{x_{kt+\ell+1} - x_{kt+\ell}} - \frac{x_{kt+t-2} - x_{kt+\ell-1}}{x_{kt+\ell} - x_{kt+\ell-1}}. \tag{272}$$

It remains to verify that the specific choices for $y^1_{k,\ell}, y^2_{k,\ell}, y^3_{k,\ell}$ per (269)-(272), indeed, satisfy the positivity assumption (255). To this end, first note that, owing to $\ell \in \{3, \dots, t-4\}$, we have $kt+t-2 > kt+\ell > kt+\ell-1$, and hence (269)-(271) implies $y^1_{k,\ell}, y^2_{k,\ell} > 0$. Then, it follows from (270), exploiting the strictly increasing nature of $(x_i)_{i=0}^{M-1}$, that $\frac{x_{kt+\ell+1}-x_{kt+\ell-1}}{x_{kt+t-2}-x_{kt+\ell-1}} \cdot \frac{x_{kt+t-2}-x_{kt+\ell}}{x_{kt+\ell+1}-x_{kt+\ell}} = \frac{x_{kt+t-2}-x_{kt+\ell}}{x_{kt+t-2}-x_{kt+\ell-1}} \cdot \frac{x_{kt+\ell+1}-x_{kt+\ell-1}}{x_{kt+\ell+1}-x_{kt+\ell}} > 1$ and hence $y^2_{k,\ell} > y^1_{k,\ell}$. Together with (268) this yields $y^3_{k,\ell} = y^2_{k,\ell} - y^1_{k,\ell} > 0$. We refer to Figures 4 and 5 for an illustration of $\rho \circ f^j_{k,\ell}$, $j = 1, 2, 3$, and $\gamma_{kt+\ell} = \rho \circ f^1_{k,\ell} - \rho \circ f^2_{k,\ell} + \rho \circ f^3_{k,\ell}$, respectively.

Property 2 will be validated by upper-bounding $\|f^j_{k,\ell}\|_{L^\infty(\mathbb{R})}$, $j = 1, 2, 3$, followed by application of Lemma C.2. First, note that the functions $f^1_{k,\ell}, f^2_{k,\ell}, f^3_{k,\ell} \in \Sigma((x_{kt}, x_{kt+1}, x_{kt+t-2}, x_{kt+t-1}), \infty)$ are equal to zero on $(-\infty, x_{kt}] \cup [x_{kt+t-1}, \infty)$ and hence take on their maximum absolute values at the breakpoints x_{kt+1}, x_{kt+t-2}. Next, we provide a relation, which will be used repeatedly later, namely for integers m_1, m_2, m_3, m_4 with $0 \leq m_1 < m_2 < M-1, 0 \leq m_3 < m_4 \leq M-1$,

$$\left|\frac{x_{m_2} - x_{m_1}}{x_{m_4} - x_{m_3}}\right| = \left|\frac{\sum_{i=m_1+1}^{m_2}(x_i - x_{i-1})}{\sum_{i=m_3+1}^{m_4}(x_i - x_{i-1})}\right| \tag{273}$$

$$\leq \left|\frac{(m_2 - m_1)\max_{i=1,\dots,M-1}(x_i - x_{i-1})}{(m_4 - m_3)\min_{i=1,\dots,M-1}(x_i - x_{i-1})}\right| \tag{274}$$

$$= \frac{m_2 - m_1}{m_4 - m_3} R_c(X). \tag{275}$$

For $j = 1$, we have

$$|f^1_{k,\ell}(x_{kt+t-2})| = |y^1_{k,\ell}| \tag{276}$$

$$= \left|\frac{x_{kt+t-2} - x_{kt+\ell-1}}{x_{kt+\ell} - x_{kt+\ell-1}}\right| \tag{277}$$

$$\leq (t-2-(\ell-1))R_c(X) \tag{278}$$

$$\leq t R_c(X), \tag{279}$$

and

$$|f^1_{k,\ell}(x_{kt+1})| = \left|y^1_{k,\ell}\frac{x_{kt+1} - x_{kt+\ell-1}}{x_{kt+t-2} - x_{kt+\ell-1}}\right| \tag{280}$$

$$= \left|\frac{x_{kt+1} - x_{kt+\ell-1}}{x_{kt+\ell} - x_{kt+\ell-1}}\right| \tag{281}$$

$$\leq (\ell-2)R_c(X) \tag{282}$$

$$\leq t R_c(X), \tag{283}$$

where (276) follows from (253), (277) is by (269), in (278) we used (273)-(275), in (280) we employed (258), (281) is a consequence of (269), (282) is by (273)-(275), and for (283) we used $\ell \in \{3, \dots, t-4\}$ owing to $(k, \ell) \in \mathcal{I}^2$. Since $f^1_{k,\ell}$ takes on its maximum absolute value at x_{kt+1}

or x_{kt+t-2}, as noted above, we get $\|f^1_{k,\ell}\|_{L^\infty(\mathbb{R})} \le tR_c(X)$. Similarly, for $j=2$, it holds that

$$\begin{aligned}
|f^2_{k,\ell}(x_{kt+t-2})| &= |y^2_{k,\ell}| && (284)\\
&= \left|\frac{x_{kt+\ell+1}-x_{kt+\ell-1}}{x_{kt+\ell}-x_{kt+\ell-1}}\frac{x_{kt+t-2}-x_{kt+\ell}}{x_{kt+\ell+1}-x_{kt+\ell}}\right| && (285)\\
&= \left|\frac{x_{kt+\ell+1}-x_{kt+\ell-1}}{x_{kt+\ell}-x_{kt+\ell-1}}\right|\left|\frac{x_{kt+t-2}-x_{kt+\ell}}{x_{kt+\ell+1}-x_{kt+\ell}}\right| && (286)\\
&\le 2R_c(X)(t-2-\ell)R_c(X) && (287)\\
&\le 2t(R_c(X))^2, && (288)
\end{aligned}$$

and

$$\begin{aligned}
|f^2_{k,\ell}(x_{kt+1})| &= \left|y^2_{k,\ell}\frac{x_{kt+1}-x_{kt+\ell}}{x_{kt+t-2}-x_{kt+\ell}}\right| && (289)\\
&= \left|\frac{x_{kt+\ell+1}-x_{kt+\ell-1}}{x_{kt+\ell}-x_{kt+\ell-1}}\frac{x_{kt+1}-x_{kt+\ell}}{x_{kt+\ell+1}-x_{kt+\ell}}\right| && (290)\\
&\le 2R_c(X)(\ell-1)R_c(X) && (291)\\
&\le 2t(R_c(X))^2, && (292)
\end{aligned}$$

where (284) follows from (253), (285) is by (270)-(271), (287) is a consequence of (273)-(275), (288) is thanks to $\ell \in \{3,\dots,t-4\}$, in (289) we used (258), (291) follows from (273)-(275), and (292) is owing to $\ell \in \{3,\dots,t-4\}$. Again, as $f^2_{k,\ell}$ takes on its maximum absolute value at x_{kt+1} or x_{kt+t-2}, we get

$$\|f^2_{k,\ell}\|_{L^\infty(\mathbb{R})} \le 2t(R_c(X))^2.$$

For $j=3$, we have

$$\begin{aligned}
|f^3_{k,\ell}(x_{kt+t-2})| &= |y^3_{k,\ell}| && (293)\\
&= |y^2_{k,\ell}-y^1_{k,\ell}| && (294)\\
&\le |y^2_{k,\ell}|+|y^1_{k,\ell}| && (295)\\
&\le 2tR_c(X)^2+tR_c(X) && (296)\\
&\le 3t(R_c(X))^2, && (297)
\end{aligned}$$

where in (294) we used (268), and (296) follows from (276)-(279) and (284)-(288). Finally,

$$\begin{aligned}
|f^3_{k,\ell}(x_{kt+1})| &= \left|y^3_{k,\ell}\frac{x_{kt+1}-x_{kt+\ell+1}}{x_{kt+t-2}-x_{kt+\ell+1}}\right| && (298)\\
&\le 3t(R_c(X))^2\frac{\ell}{t-\ell-3}R_c(X) && (299)\\
&\le 3t^2(R_c(X))^3, && (300)
\end{aligned}$$

where (298) is a consequence of (258), and (299) follows from (273)-(275). We therefore have

$$\|f^3_{k,\ell}\|_{L^\infty(\mathbb{R})} \le 3t^2(R_c(X))^3.$$

In summary, we established that $\|f^1_{k,\ell}\|_{L^\infty(\mathbb{R})}, \|f^2_{k,\ell}\|_{L^\infty(\mathbb{R})}, \|f^3_{k,\ell}\|_{L^\infty(\mathbb{R})} \le 3t^2(R_c(X))^3$, which together with $f^j_{k,\ell} \in \Sigma((x_{kt},x_{kt+1},x_{kt+t-2},x_{kt+t-1}),\infty)$, implies

$$f^1_{k,\ell}, f^2_{k,\ell}, f^3_{k,\ell} \in \Sigma((x_{kt},x_{kt+1},x_{kt+t-2},x_{kt+t-1}), 3t^2(R_c(X))^3). \quad (301)$$

Application of Lemma C.2 to $f^j_{k,\ell}$, for $j \in \{1,2,3\}$, with $X = (x_{kt}, x_{kt+1}, x_{kt+t-2}, x_{kt+t-1})$ and $E = 3t^2(R_c(X))^3$, yields

$$f^j_{k,\ell}(x) = d + c_1\rho(x - x_{kt}) + c_2\rho(x - x_{kt+1}) + c_3\rho(x - x_{kt+t-2}) + c_4\rho(x - x_{kt+t-1}), \quad x \in \mathbb{R},$$

for some $d, c_1, c_2, c_3, c_4 \in \mathbb{R}$ such that $|d|, |c_1|, |c_2|, |c_3|, |c_4| \leq 12t^2 R_m(X)(R_c(X))^3$. The validity of Property 2 now follows upon noting that

$$(x_{kt}, x_{kt+1}, x_{kt+t-2}, x_{kt+t-1}) \subseteq (z_i)_{i=1}^{8u-1},$$

with $(z_i)_{i=1}^{8u-1}$ as defined in (222). Since the choice of $(k,\ell) \in \mathcal{I}^2$ was arbitrary, this completes the construction of $f^1_{k,\ell}, f^2_{k,\ell}, f^3_{k,\ell}$ and establishes Properties 1 and 2 for all $(k,\ell) \in \mathcal{I}^2$.

We conclude the proof by verifying Property 3. The statement follows directly as, for every $(\ell, j) \in \{3, \ldots, t-4\} \times \{1,2,3\}$, the functions $f^j_{k,\ell}$, $k \in \{0, \ldots, u-1\}$, are supported in (x_{kt}, x_{kt+t-1}) according to (259).

C.2 Proof of Lemma C.4

We start by defining several auxiliary quantities. Let $G(x) := (\rho(x - z_0), \ldots, \rho(x - z_{r-1}))^T$, $x \in \mathbb{R}$. For $j = 0, \ldots, s-1, x \in \mathbb{R}$, define,

$$U_j := \begin{pmatrix} c_{ju+0,0} & \cdots & c_{ju+0,r-1} \\ \vdots & \ddots & \vdots \\ c_{ju+u-1,0} & \cdots & c_{ju+u-1,r-1} \end{pmatrix}, \; v_j := \begin{pmatrix} d_{ju+0} \\ \vdots \\ d_{ju+u-1} \end{pmatrix}, \; w_j(x) := \begin{pmatrix} h_{ju+0}(x) \\ \vdots \\ h_{ju+u-1}(x) \end{pmatrix},$$

such that

$$U_j\, G(x) + v_j = w_j(x), \quad x \in \mathbb{R}. \tag{302}$$

Based on (302), we proceed to construct a ReLU network realization of H according to

$$H(x) = \sum_{\ell=0}^{us-1} \rho(h_\ell(x)) = \sum_{j=0}^{s-1} 1_u^T \rho(w_j(x)), \quad x \in \mathbb{R}, \tag{303}$$

where we recall that the ReLU function ρ acts componentwise. Specifically, let $\Phi = ((A_k, b_k))_{k=1}^{s+2}$ be given by

$$A_1 = 1_{r\times 1}, \qquad b_1 = (-z_0, \cdots, -z_{r-1})^T,$$
$$A_2 = \begin{pmatrix} I_r \\ U_0 \\ 0_{1\times r} \end{pmatrix}, \qquad b_2 = \begin{pmatrix} 0_r \\ v_0 \\ 0_1 \end{pmatrix},$$
$$A_{s+2} = \begin{pmatrix} 0_{1\times r} & 1_{1\times u} & 1_{1\times 1} \end{pmatrix}, \quad b_{s+2} = 0,$$

and for $k \in \mathbb{N}$ such that $3 \leq k \leq s+1$,[13]

$$A_k = \begin{pmatrix} I_r & 0_{r\times u} & 0_{r\times 1} \\ U_{k-2} & 0_{u\times u} & 0_{u\times 1} \\ 0_{1\times r} & 1_{1\times u} & 1_{1\times 1} \end{pmatrix}, \quad b_k = \begin{pmatrix} 0_r \\ v_{k-2} \\ 0_1 \end{pmatrix}.$$

[13] For $s = 1$, there is no $k \in \mathbb{N}$ satisfying the constraint and the assignment is thus skipped.

Note that $\mathcal{L}(\Phi) = s{+}2$, $\mathcal{W}(\Phi) = r{+}u{+}1$, and $\mathcal{B}(\Phi) \leq \max\{1, \max_{\ell,i} |c_{\ell,i}|, \max_\ell |d_\ell|, \max_i\{z_i\}\} \leq \max\{1, T\}$.

We proceed to show that $H = R(\Phi)$, which will then imply $H \in \mathcal{R}(r{+}u{+}1, s{+}2, \max\{1, T\})$, as desired. For $k = 1, \ldots, s+2$, set $y_k := R((A_\ell, b_\ell)_{\ell=1}^k)$. We have,

$$y_1(x) = A_1 x + b_1 = (x - z_0, \ldots, x - z_{r-1})^T, \quad x \in \mathbb{R}. \tag{304}$$

Next, it is proved by induction that for $k = 2, \ldots, s+1$,

$$y_k(x) = \begin{pmatrix} G(x) \\ w_{k-2}(x) \\ \sum_{j=0}^{k-3} 1_u^T \rho(w_j(x)) \end{pmatrix}, \quad x \in \mathbb{R}, \tag{305}$$

where we use the convention $\sum_{j=0}^{-1} 1_u^T \rho(w_j(x)) = 0$. The base case $k = 2$ follows from

$$\begin{aligned} y_2(x) &= A_2 \rho(y_1(x)) + b_2 && (306)\\ &= \begin{pmatrix} \rho((x - z_0, \ldots, x - z_{r-1})^T) \\ U_0 \rho((x - z_0, \ldots, x - z_{r-1})^T) + v_0 \\ 0_r \end{pmatrix} && (307)\\ &= \begin{pmatrix} G(x) \\ w_0(x) \\ 0_r \end{pmatrix} \quad x \in \mathbb{R}, && (308) \end{aligned}$$

where in (308) we used $U_0 G(x) + v_0 = w_0(x)$, for $x \in \mathbb{R}$. We proceed to prove the induction step $k - 1 \mapsto k$ with $3 \leq k \leq s+1$[14], noting that the induction assumption is given by

$$y_{k-1}(x) = \begin{pmatrix} G(x) \\ w_{k-3}(x) \\ \sum_{j=0}^{k-4} 1_u^T \rho(w_j(x)) \end{pmatrix}, \; x \in \mathbb{R}. \tag{309}$$

The result is immediate from

$$\begin{aligned} y_k(x) &= A_k \rho(y_{k-1}(x)) + b_k && (310)\\ &= \begin{pmatrix} I_r G(x) \\ U_{k-2} G(x) + v_{k-2} \\ 1_u^T \rho(w_{k-3}(x)) + \sum_{j=0}^{k-4} 1_u^T \rho(w_j(x)) \end{pmatrix} && (311)\\ &= \begin{pmatrix} G(x) \\ w_{k-2}(x) \\ \sum_{j=0}^{k-3} 1_u^T \rho(w_j(x)) \end{pmatrix}, \quad x \in \mathbb{R}, && (312) \end{aligned}$$

where in (311) we used the induction assumption (309) and the fact that $G(x), \sum_{j=0}^{k-4} 1_u^T \rho(w_j(x)) \geq$

[14]For $s = 1$, the induction step is not needed as the base case already complete the proof.

0, for all $x \in \mathbb{R}$. We conclude the overall proof by noting that, for $x \in \mathbb{R}$,

$$y_{s+2}(x) = A_{s+2}\rho(y_{s+1}(x)) + b_{s+2} \tag{313}$$

$$= 1_u^T \rho(w_{s-1}(x)) + \sum_{j=0}^{s-2} 1_u^T \rho(w_j(x)) \tag{314}$$

$$= \sum_{j=0}^{s-1} 1_u^T \rho(w_j(x)) \tag{315}$$

$$= H(x), \tag{316}$$

where (316) follows from (303).

D Proof of Proposition B.5

Let $N \in \mathbb{N}$ be fixed throughout. We start by providing intuition behind the construction of $F_{N,L}$, $L \in \mathbb{N}$, and then follow up with the proof. Consider the following decomposition, for all $L \in \mathbb{N}$, $\sum_{i=1}^{\infty} \theta_i 3^{-i} \in \mathbb{T}$, and $k \in \mathbb{N} \cup \{0\}$,

$$\sum_{i=1}^{\min\{N(L+1),k\}} \theta_i = \sum_{i=1}^{\min\{N,k\}} \theta_i + \sum_{i=1}^{\min\{NL,\max\{k-N,0\}\}} \theta_{N+i}, \tag{317}$$

where we recall the convention $\sum_{i=1}^{0} \theta_{N+i} = 0$. With $F_{N,L} : \mathbb{R} \mapsto \mathbb{R}$ according to (109), (317) implies the recursive relation

$$F_{N,L+1}\left(\sum_{i=1}^{\infty} \theta_i 3^{-i}, k\right) = \sum_{i=1}^{\min\{N,k\}} \theta_i + F_{N,L}\left(\sum_{i=1}^{\infty} \theta_{N+i} 3^{-i}, \max\{k-N, 0\}\right), \tag{318}$$

which will be seen below to inspire our proof.

We begin with two technical lemmata introducing the function g and the family of functions $\{G_{N,L}\}_{L\in\mathbb{N}}$ that serve as basic building blocks of our construction.

Lemma D.1. *There exists a function*

$$g \in \mathcal{R}((3,3), 2^{N+4}, 3, 3^{N+2}) \tag{319}$$

such that for all $X \in \mathbb{R}$, $\sum_{i=1}^{\infty} \theta_i 3^{-i} \in \mathbb{T}$, $k \in \mathbb{N} \cup \{0\}$,

$$g\left(X, \sum_{i=1}^{\infty} \theta_i\, 3^{-i}, k\right) = \left(\rho(X) + \sum_{i=1}^{\min\{N,k\}} \theta_i, \sum_{i=1}^{\infty} \theta_{N+i}\, 3^{-i}, \max\{k-N, 0\}\right). \tag{320}$$

Proof. See Appendix D.1. □

Define $G_{N,L} : \mathbb{R}^3 \mapsto \mathbb{R}^3$, $L \in \mathbb{N}$, recursively, according to

$$G_{N,L} = \begin{cases} g, & \text{for } L = 1, \\ g \circ G_{N,L-1}, & \text{for } L \geq 2, \end{cases} \tag{321}$$

where g is the function specified by Lemma D.1. The properties of $G_{N,L}$, $L \in \mathbb{N}$, are summarized in the following result.

Lemma D.2. *Let $L \in \mathbb{N}$. It holds that*

$$G_{N,L} \in \mathcal{R}((3,3), 2^{N+4}, 3L, 3^{N+2}), \tag{322}$$

and, for $\sum_{i=1}^{\infty} \theta_i 3^{-i} \in \mathbb{T}$, $k \in \mathbb{N} \cup \{0\}$,

$$G_{N,L}\left(0, \sum_{i=1}^{\infty} \theta_i 3^{-i}, k\right) = \left(\sum_{i=1}^{\min\{NL,k\}} \theta_i, \sum_{i=1}^{\infty} \theta_{NL+i}\, 3^{-i}, \max\{k-NL, 0\}\right). \tag{323}$$

Proof. See Appendix D.2. □

Fix $L \in \mathbb{N}$. We are now ready to define $F_{N,L}$ and verify the corresponding properties (108)-(110). Recall that $S(A,b)$ refers to the affine mapping given by $S(A,b)(x) = Ax + b, x \in \mathbb{R}_{n_2}$, for $A \in \mathbb{R}_{n_1 \times n_2}$, $b \in \mathbb{R}_{n_1}$, $n_1, n_2 \in \mathbb{N}$. Set

$$F_{N,L} = f_2 \circ G_{N,L} \circ f_1, \tag{324}$$

with

$$f_1 = S\left(\begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, 0_3\right) \in \mathcal{R}((2,3), 3, 1, 1), \tag{325}$$

$$f_2 = S((1\,0\,0), 0) \qquad \in \mathcal{R}((3,1), 3, 1, 1), \tag{326}$$

and note that

$$f_1(y,z) = (0,y,z), \quad \text{for } (y,z) \in \mathbb{R}^2, \tag{327}$$

$$f_2(x,y,z) = x, \quad \text{for } (x,y,z) \in \mathbb{R}^3. \tag{328}$$

Noting that $F_{N,L}$ is the composition of functions which can be realized by ReLU networks according to (322), (325), and (326), application of Lemma H.3 yields

$$\begin{aligned} F_{N,L} &\in \mathcal{R}((2,1), 2^{N+4}, 3L+2, 3^{N+2}) \\ &\subseteq \mathcal{R}((2,1), 2^{N+4}, 5L, 3^{N+2}), \end{aligned}$$

which establishes (108). Furthermore, for $\sum_{i=1}^{\infty} \theta_i 3^{-i} \in \mathbb{T}$, and $k \in \mathbb{N} \cup \{0\}$, we have

$$F_{N,L}\left(\sum_{i=1}^{\infty} \theta_i 3^{-i}, k\right) = f_2 \circ G_{N,L} \circ f_1\left(\sum_{i=1}^{\infty} \theta_i 3^{-i}, k\right) \tag{329}$$

$$= f_2 \circ G_{N,L}\left(0, \sum_{i=1}^{\infty} \theta_i 3^{-i}, k\right) \tag{330}$$

$$= f_2\left(\sum_{i=1}^{\min\{NL,k\}} \theta_i, \sum_{i=1}^{\infty} \theta_{NL+i} 3^{-i}, \max\{k-NL, 0\}\right) \tag{331}$$

$$= \sum_{i=1}^{\min\{NL,k\}} \theta_i, \tag{332}$$

where (330) follows from (327), in (331) we used (323), and (332) is by (328), thereby verifying property (109). We conclude the proof by noting that (110) follows directly from (109).

D.1 Proof of Lemma D.1

We start by constructing functions $g^1, g^2, g^3 \in \mathcal{R}((3,1))$ such that, for all $X \in \mathbb{R}$, $\sum_{i=1}^{\infty} \theta_i 3^{-i} \in \mathbb{T}$, $k \in \mathbb{N} \cup \{0\}$,

(i) $g^1(X, \sum_{i=1}^{\infty} \theta_i 3^{-i}, k) = \rho(X) + \sum_{i=1}^{\min\{N,k\}} \theta_i$,

(ii) $g^2(X, \sum_{i=1}^{\infty} \theta_i 3^{-i}, k) = \sum_{i=1}^{\infty} \theta_{N+i} 3^{-i}$,

(iii) $g^3(X, \sum_{i=1}^{\infty} \theta_i 3^{-i}, k) = \max\{k - N, 0\}$,

and then putting them together, through Lemma H.3, as $g = ((g_1, g_2), g_3)$[15] to obtain a ReLU network. We proceed to the construction of g^1, g^2, g^3.

(i) Construction of g^1. Fix $X \in \mathbb{R}$, $\sum_{i=1}^{\infty} \theta_i 3^{-i} \in \mathbb{T}$, $k \in \mathbb{N} \cup \{0\}$ throughout part (i). We first note that

$$\rho(X) + \sum_{i=1}^{\min\{N,k\}} \theta_i = \rho(X) + \sum_{\ell=1}^{N} 1_{\{\theta_\ell = 1\}} 1_{\{\ell \leq k\}}. \tag{333}$$

Next, we construct functions $h^\ell, m^\ell : \mathbb{R} \mapsto \mathbb{R}$ such that $h^\ell(\sum_{i=1}^{\infty} \theta_i 3^{-i}) = 1_{\{\theta_\ell = 1\}}$ and $m^\ell(k) = 1_{\{\ell \leq k\}}$, $\ell = 1, \ldots, N$.

(a) Construction of h^ℓ. We distinguish the cases $\ell = 1$ and $\ell \geq 2$. For $\ell = 1$, we let $h^1 : \mathbb{R} \mapsto \mathbb{R}$ be given by

$$h^1(y) = 9\rho(y - T((0,2))) - 9\rho(y - T((1,0))), \tag{334}$$

which takes on the value 0 on $[0, T((0,2))]$ and equals 1 on $[T((1,0)), \infty)$. Upon noting that $\sum_{i=1}^{\infty} \theta_i 3^{-i} \in \mathbb{T} \subseteq [0, T((0,2))] \cup [T((1,0)), \infty)$, we therefore have $h^1(\sum_{i=1}^{\infty} \theta_i 3^{-i}) = 1_{\{\theta_1 = 1\}}$. For $\ell \geq 2$, we first decompose $1_{\{\theta_\ell = 1\}}$ according to

$$1_{\{\theta_\ell = 1\}} = \sum_{\substack{(a_1, \ldots, a_{\ell-1}) \\ \in \{0,1\}^{\ell-1}}} 1_{\{\theta_1 = a_1, \ldots, \theta_{\ell-1} = a_{\ell-1}, \theta_\ell = 1\}}. \tag{335}$$

Next, note that, for $(a_1, \ldots, a_{\ell-1}) \in \{0,1\}^{\ell-1}$, the function $r^{(a_1,\ldots,a_{\ell-1})} : \mathbb{R} \mapsto \mathbb{R}$, given by

$$\begin{aligned} r^{(a_1,\ldots,a_{\ell-1})}(y) = {} & 3^{\ell+1}\rho(y - T((a_1, \ldots, a_{\ell-1}, 0, 2))) \\ & - 3^{\ell+1}\rho(y - T((a_1, \ldots, a_{\ell-1}, 1))) \\ & - 3^{\ell+1}\rho(y - T((a_1, \ldots, a_{\ell-1}, 2))) \\ & + 3^{\ell+1}\rho(y - T((a_1, \ldots, a_{\ell-1}, 2, 1))), \end{aligned} \tag{336}$$

satisfies $r^{(a_1,\ldots,a_{\ell-1})}(\sum_{i=1}^{\infty} \theta_i 3^{-i}) = 1_{\{\theta_1 = a_1, \ldots, \theta_{\ell-1} = a_{\ell-1}, \theta_\ell = 1\}}$. We refer to Figure 6 for an illustration of $r^{(a_1,\ldots,a_{\ell-1})}$. Summation over $r^{(a_1,\ldots,a_{\ell-1})}$ for $a_1, \ldots, a_{\ell-1} \in \{0,1\}$ yields

[15] Recall that for $f_1 : \mathbb{R}^d \mapsto \mathbb{R}^{d'}$ and $f_2 : \mathbb{R}^d \mapsto \mathbb{R}^{d''}$, $d, d', d'' \in \mathbb{N}$, $(f_1, f_2) : \mathbb{R}^d \mapsto \mathbb{R}^{d'+d''}$ is defined according to $(f_1, f_2)(x) = (f_1(x), f_2(x))$, for $x \in \mathbb{R}^d$.

the desired $h^\ell : \mathbb{R} \mapsto \mathbb{R}$ according to

$$h^\ell(y) := \sum_{\substack{(a_1,\ldots,a_{\ell-1})\\ \in\{0,1\}^{\ell-1}}} r^{(a_1,\ldots,a_{\ell-1})}(y) \tag{337}$$

$$= \sum_{\substack{(a_1,\ldots,a_{\ell-1})\\ \in\{0,1\}^{\ell-1}}} (3^{\ell+1}\rho(y - T((a_1,\ldots,a_{\ell-1},0,2))) \tag{338}$$

$$- 3^{\ell+1}\rho(y - T((a_1,\ldots,a_{\ell-1},1))) \tag{339}$$

$$- 3^{\ell+1}\rho(y - T((a_1,\ldots,a_{\ell-1},2))) \tag{340}$$

$$+ 3^{\ell+1}\rho(y - T((a_1,\ldots,a_{\ell-1},2,1)))), \tag{341}$$

which satisfies

$$h^\ell\left(\sum_{i=1}^{\infty}\theta_i 3^{-i}\right) = \sum_{\substack{(a_1,\ldots,a_{\ell-1})\\ \in\{0,1\}^{\ell-1}}} r^{(a_1,\ldots,a_{\ell-1})}\left(\sum_{i=1}^{\infty}\theta_i 3^{-i}\right) \tag{342}$$

$$= \sum_{\substack{(a_1,\ldots,a_{\ell-1})\\ \in\{0,1\}^{\ell-1}}} 1_{\{\theta_1=a_1,\ldots,\theta_{\ell-1}=a_{\ell-1},\theta_\ell=1\}} \tag{343}$$

$$= 1_{\{\theta_\ell=1\}}, \tag{344}$$

as desired. We finally note that h^ℓ, $\ell \in \mathbb{N}$, can be written as

$$h^\ell(y) = \sum_{j=1}^{2^{\ell+1}} u_{\ell,j}\rho(y - v_{\ell,j}), \quad y \in \mathbb{R}, \tag{345}$$

for some $u_{\ell,j}, v_{\ell,j} \in \mathbb{R}$ with $|u_{\ell,j}|, |v_{\ell,j}| \le 3^{\ell+1}$, $j = 1,\ldots,2^{\ell+1}$.

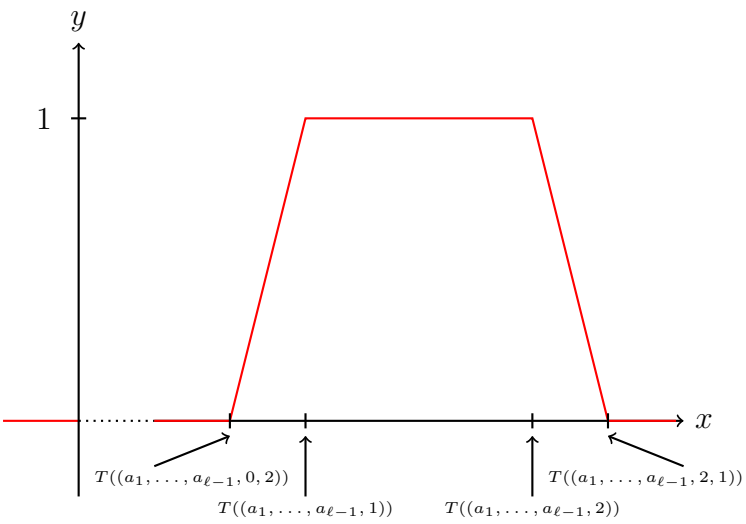

Figure 6: The function $r^{(a_1,\ldots,a_{\ell-1})}$.

(b) The ReLU networks realizing $m^\ell : \mathbb{R} \mapsto \mathbb{R}$ are given by $m^\ell(z) = \rho(z - (\ell - 1)) - \rho(z - \ell)$, $z \in \mathbb{R}$. It is readily seen that

$$m^\ell(k) = 1_{\{\ell \leq k\}}. \tag{346}$$

Now, define $g^1 : \mathbb{R}^3 \mapsto \mathbb{R}$ according to

$$g^1(x, y, z) := \rho(x) + \sum_{\ell=1}^{N} \rho(h^\ell(y) + m^\ell(z) - 1) \tag{347}$$

$$= \rho(\rho(x)) + \sum_{\ell=1}^{N} \rho\bigg(\sum_{j=1}^{2^{\ell+1}} u_{\ell,j}\rho(y - v_{\ell,j}) + \rho(z - (\ell - 1)) - \rho(z - \ell) - 1\bigg), \tag{348}$$

for $x, y, z \in \mathbb{R}$, where in (348) we used the ReLU network realizations of h^ℓ and m^ℓ, $\ell = 1, \ldots, N$, along with $\rho \circ \rho = \rho$. This shows that g^1 can be realized by a ReLU network of depth 3, input dimension 3, with $1 + \sum_{\ell=1}^{N}(2^{\ell+1} + 2) \leq 2^{N+3}$ nodes in the first layer, $N + 1$ nodes in the second layer, and, owing to $|u_{\ell,j}|, |v_{\ell,j}| \leq 3^{\ell+1} \leq 3^{N+1}$, $\ell = 1, \ldots, N$, $j = 1, \ldots, 2^{\ell+1}$, weight magnitude upper-bounded by 3^{N+1}. Formally, we have established that

$$g^1 \in \mathcal{R}((3, 1), 2^{N+3}, 3, 3^{N+1}). \tag{349}$$

We finally note that

$$g^1\bigg(X, \sum_{i=1}^{\infty} \theta_i 3^{-i}, k\bigg) = \rho(X) + \sum_{\ell=1}^{N} \rho\bigg(h^\ell\bigg(\sum_{i=1}^{\infty} \theta_i 3^{-i}\bigg) + m^\ell(k) - 1\bigg) \tag{350}$$

$$= \rho(X) + \sum_{\ell=1}^{N} \rho(1_{\{\theta_\ell = 1\}} + 1_{\{\ell \leq k\}} - 1) \tag{351}$$

$$= \rho(X) + \sum_{\ell=1}^{N} 1_{\{\theta_\ell = 1\}} 1_{\{\ell \leq k\}} \tag{352}$$

$$= \rho(X) + \sum_{i=1}^{\min\{N,k\}} \theta_i, \tag{353}$$

as desired, where in (351) we used (342)-(344) and (346), and (352) follows from $\rho(c + d - 1) = cd$, for $c, d \in \{0, 1\}$. As the choice of $X \in \mathbb{R}$, $\sum_{i=1}^{\infty} \theta_i 3^{-i} \in \mathbb{T}$, $k \in \mathbb{N} \cup \{0\}$ was arbitrary and the construction of g^1 does not depend on $X, \sum_{i=1}^{\infty} \theta_i 3^{-i}, k$, this concludes the argument.

(ii) Construction of g^2. We first show how to realize the bit shifting operation $\sum_{i=1}^{\infty} \theta_i 3^{-i} \in \mathbb{T} \mapsto \sum_{i=1}^{\infty} \theta_{N+i} 3^{-i} \in \mathbb{T}$ by a ReLU network. This will be accomplished by decomposing the mapping into submappings, realizing the individual submappings by ReLU networks, and then putting these networks together to obtain a ReLU network construction for the overall mapping. Specifically, we work with the decomposition

$$\sum_{i=1}^{\infty} \theta_{N+i}\, 3^{-i} = \sum_{(a_1,\ldots,a_N) \in \{0,1\}^N} \bigg(1_{\{\theta_1 = a_1, \ldots, \theta_N = a_N\}} \sum_{i=1}^{\infty} \theta_{N+i}\, 3^{-i}\bigg), \tag{354}$$

for $\sum_{i=1}^{\infty}\theta_i 3^{-i}\in\mathbb{T}$. Now, for $a_i\in\{0,1\}$, $i=1,\dots,N$, consider the function

$$f^{(a_1,\dots,a_N)}(y)=3^N\rho(y-T((a_1,\dots,a_N)))-7\cdot 3^N\rho(y-T((a_1,\dots,a_N,2)))\\ +6\cdot 3^N\rho(y-T((a_1,\dots,a_N,2,1))),\quad y\in\mathbb{R},$$

illustrated in Figure 7. For $\sum_{i=1}^{\infty}\theta_i 3^{-i}\in\mathbb{T}$, with $\theta_i=a_i$, for $i=1,\dots,N$, we have

$$\sum_{i=1}^{\infty}\theta_i 3^{-i}\in\left[\sum_{i=1}^{N}a_i 3^{-i},\sum_{i=1}^{N}a_i 3^{-i}+2\cdot 3^{-(N+1)}\right)=[T((a_1,\dots,a_N)),T((a_1,\dots,a_N,2))),$$

which together with the definition of $f^{a_1,\dots,a_N}$ implies

$$\begin{aligned}f^{(a_1,\dots,a_N)}\left(\sum_{i=1}^{\infty}\theta_i 3^{-i}\right)&=3^N\rho\left(\sum_{i=1}^{\infty}\theta_i 3^{-i}-T((a_1,\dots,a_N))\right)\\&=3^N\cdot\sum_{i=N+1}^{\infty}\theta_i 3^{-i}\\&=\sum_{i=1}^{\infty}\theta_{N+i}\,3^{-i}.\end{aligned}$$

Moreover, for numbers $\sum_{i=1}^{\infty}\theta_i 3^{-i}\in\mathbb{T}$ whose sequence of N-leading digits differ from $a_1,\dots,a_N$, we have $f^{(a_1,\dots,a_N)}(\sum_{i=1}^{\infty}\theta_i 3^{-i})=0$, as every number in the support of $f^{(a_1,\dots,a_N)}$, i.e., in $[T((a_1,\dots,a_N)),T(a_1,\dots,a_N,2,1)))$, has leading digits $a_1,\dots,a_N$ in its ternary representation. In summary, we therefore get, for $\sum_{i=1}^{\infty}\theta_i 3^{-i}\in\mathbb{T}$,

$$f^{(a_1,\dots,a_N)}\left(\sum_{i=1}^{\infty}\theta_i 3^{-i}\right)=1_{\{\theta_1=a_1,\dots,\theta_N=a_N\}}\sum_{i=1}^{\infty}\theta_{N+i}\,3^{-i}.$$

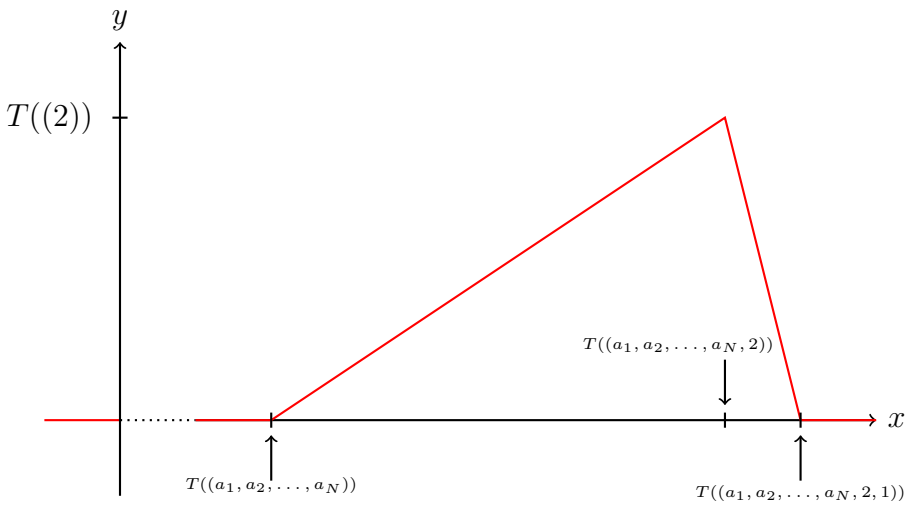

Figure 7: The function $f^{(a_1,\dots,a_N)}$.

Summing over all $(a_1, \ldots, a_N) \in \{0,1\}^N$, we finally obtain $g^2 : \mathbb{R}^3 \mapsto \mathbb{R}$, according to

$$g^2(x,y,z) := \sum_{(a_1,\ldots,a_N)\in\{0,1\}^N} f^{(a_1,\ldots,a_N)}(y) \tag{355}$$

$$= \sum_{(a_1,\ldots,a_N)\in\{0,1\}^N} \big(3^N \rho(y - T((a_1,\ldots,a_N))) \tag{356}$$

$$- 7\cdot 3^N \rho(y - T((a_1,\ldots,a_N,2))) \tag{357}$$

$$+ 6\cdot 3^N \rho(y - T((a_1,\ldots,a_N,2,1)))\big). \tag{358}$$

Then, for all $X \in \mathbb{R}$, $\sum_{i=1}^{\infty}\theta_i 3^{-i} \in \mathbb{T}$, $k \in \mathbb{N}\cup\{0\}$, we have

$$\begin{aligned} g^2\left(X, \sum_{i=1}^{\infty}\theta_i 3^{-i}, k\right) &= \sum_{(a_1,\ldots,a_N)\in\{0,1\}^N} f^{(a_1,\ldots,a_N)}\left(\sum_{i=1}^{\infty}\theta_i 3^{-i}\right) \\ &= \sum_{(a_1,\ldots,a_N)\in\{0,1\}^N} \left(1_{\{\theta_1=a_1,\ldots,\theta_N=a_N\}} \sum_{i=1}^{\infty}\theta_{N+i}3^{-i}\right) \\ &= \sum_{i=1}^{\infty}\theta_{N+i}3^{-i}, \end{aligned}$$

as desired. Moreover, it follows from (356)-(358) that g^2 can be realized by a 2-layer ReLU network with 3 nodes in the input layer, $3\cdot 2^N$ nodes in the first layer, and weight magnitude $7\cdot 3^N$, formally

$$g^2 \in ((3,1), 3\cdot 2^N, 2, 7\cdot 3^N). \tag{359}$$

(iii) Construction of g^3. Setting $g^3(x,y,z) := \rho(z-N)$, for $x,y,z \in \mathbb{R}$, we get $g^3(X, \sum_{i=1}^{\infty}\theta_i 3^{-i}, k) = \rho(k-N)$, for all $X \in \mathbb{R}$, $\sum_{i=1}^{\infty}\theta_i 3^{-i} \in \mathbb{T}$, $k \in \mathbb{N}\cup\{0\}$. Moreover, g^3 can be realized by a ReLU network according to

$$g^3 \in \mathcal{R}((3,1), 3, 2, N). \tag{360}$$

Finally, we note that $g = ((g^1, g^2), g^3)$ satisfies (320), and application of Lemma H.3, considering (349), (359), and (360), yields

$$\begin{aligned} g &\in \mathcal{R}((3,3), 2^{N+3} + 3\cdot 2^N + 3, 3, \max\{3^{N+1}, 7\cdot 3^N, N\}) \\ &\subseteq \mathcal{R}((3,3), 2^{N+4}, 3, 3^{N+2}). \end{aligned}$$

D.2 Proof of Lemma D.2

We prove (322) and (323) by induction on $L \in \mathbb{N}$. The base case $L = 1$ follows from $G_{N,1} = g$, with g per Lemma D.1 and the properties of g as specified in (319) and (320).

For the induction step $L-1 \mapsto L$, $L \geq 2$, we start with the induction assumption

$$G_{N,L-1} \in \mathcal{R}((3,3), 2^{N+4}, 3(L-1), 3^{N+2}), \tag{361}$$

and, for $\sum_{i=1}^{\infty}\theta_i 3^{-i}\in\mathbb{T}$, $k\in\mathbb{N}\cup\{0\}$,

$$G_{N,L-1}\left(0,\sum_{i=1}^{\infty}\theta_i 3^{-i},k\right)=\left(\sum_{i=1}^{\min\{N(L-1),k\}}\theta_i,\sum_{i=1}^{\infty}\theta_{N(L-1)+i}\,3^{-i},\max\{k-N(L-1),\,0\}\right). \tag{362}$$

Next, note that $G_{N,L}=g\circ G_{N,L-1}$, and both g and $G_{N,L-1}$ can be realized by ReLU networks according to (319) and (361). Application of Lemma H.3 yields

$$G_{N,L}\in\mathcal{R}((3,3),2^{N+4},3L,3^{N+2}).$$

Furthermore, for $\sum_{i=1}^{\infty}\theta_i 3^{-i}\in\mathbb{T}$, $k\in\mathbb{N}\cup\{0\}$, we have

$$\begin{aligned}
G_{N,L}\left(0,\sum_{i=1}^{\infty}\theta_i 3^{-i},k\right)&=g\left(G_{N,L-1}\left(0,\sum_{i=1}^{\infty}\theta_i 3^{-i},k\right)\right) &(363)\\
&=g\left(\sum_{i=1}^{\min\{N(L-1),k\}}\theta_i,\sum_{i=1}^{\infty}\theta_{N(L-1)+i}\,3^{-i},\max\{k-N(L-1),0\}\right) &(364)\\
&=\left(\rho\left(\sum_{i=1}^{\min\{N(L-1),k\}}\theta_i\right)+\sum_{i=1}^{\min\{N,\max\{k-N(L-1),0\}\}}\theta_{N(L-1)+i},\right.\\
&\quad\left.\sum_{i=1}^{\infty}\theta_{N+N(L-1)+i}\,3^{-i},\max\{\max\{k-N(L-1),0\}-N,0\}\right) &(365)\\
&=\left(\sum_{i=1}^{\min\{NL,k\}}\theta_i,\sum_{i=1}^{\infty}\theta_{NL+i}\,3^{-i},\max\{k-NL,0\}\right), &(366)
\end{aligned}$$

where (364) follows from the induction assumption (362), and in (365) we used (320). This finishes the proof.

E Proof of Lemma 3.2

Lemma 3.2 is a special case, with input dimension $d=1$ and weight magnitude $B=1$, of the more general result Lemma E.1 stated and proved here. We provide this more general result as its proof is no longer than that for $d=1$ and $B=1$, while offering deeper insights into the mechanisms at play.

Lemma E.1. *Let $d,W,L,\ell\in\mathbb{N}$ with $W\geq d$ and $\ell\leq L$, $B\in\mathbb{R}_+$ with $B\geq 1$, and let*

$$\Phi^i=((A_j^i,b_j^i))_{j=1}^{\ell}\in\mathcal{N}((d,1),W,L,B),\quad i=1,2,$$

have the same architecture. Then,

$$\|R(\Phi^1)-R(\Phi^2)\|_{L^\infty([0,1]^d)}\leq L(W+1)^L B^{L-1}\|\Phi^1-\Phi^2\|. \tag{367}$$

Proof. Fix $x\in[0,1]^d$. For $i=1,2$ and $k=1,\ldots,\ell$, let $y_k^i(x):=\Phi(((A_j^i,b_j^i))_{j=1}^k)(x)$, denote the output of the k-th layer of Φ^i.

We start with a preparatory result upper-bounding $\|y_k^1(x)\|_\infty$, for $k = 1, \ldots, \ell$. Specifically, we prove, by induction, that

$$\|y_k^1(x)\|_\infty \leq (W+1)^k B^k, \tag{368}$$

for $k = 1, 2, \ldots, \ell$. The base case $k = 1$ follows by noting that

$$\begin{aligned}
\|y_1^1(x)\|_\infty &= \|A_1^1 x + b_1^1\|_\infty && (369)\\
&\leq W\|A_1^1\|_\infty \|x\|_\infty + \|b_1^1\|_\infty && (370)\\
&\leq WB + B = (W+1)B, && (371)
\end{aligned}$$

where in (370) we used the fact that A_1^1 has at most W columns. We proceed to establish the induction step $k-1 \mapsto k$ with the induction assumption given by

$$\|y_{k-1}^1(x)\|_\infty \leq (W+1)^{k-1} B^{k-1}. \tag{372}$$

As

$$\begin{aligned}
\|y_k^1(x)\|_\infty &= \|A_k^1\, \rho(y_{k-1}^1(x)) + b_k^1\|_\infty && (373)\\
&\leq W\|A_k^1\|_\infty \|\rho(y_{k-1}^1(x))\|_\infty + \|b_k^1\|_\infty && (374)\\
&\leq W\|A_k^1\|_\infty \|y_{k-1}^1(x)\|_\infty + \|b_k^1\|_\infty && (375)\\
&\leq WB(W+1)^{k-1} B^{k-1} + B && (376)\\
&\leq (W+1)^k B^k, && (377)
\end{aligned}$$

where in (374) we used that A_k^1 has at most W columns, (375) follows from the 1-Lipschitz continuity of ρ, in (376) we employed the induction assumption (372), and (377) is by $B \geq 1$.

Next, we bound the difference $\|y_k^1(x) - y_k^2(x)\|_\infty$, for $k = 1, \ldots, \ell$. Specifically, we show, again by induction, that

$$\|y_k^1(x) - y_k^2(x)\|_\infty \leq k(W+1)^k B^{k-1} \|\Phi^1 - \Phi^2\|,$$

for $k = 1, \ldots, \ell$. The base case $k = 1$ follows according to

$$\begin{aligned}
\|y_1^1(x) - y_1^2(x)\|_\infty &= \|(A_1^1 - A_2^1)x + (b_1^1 - b_2^1)\|_\infty \\
&\leq W\|A_1^1 - A_2^1\|_\infty \|x\|_\infty + \|b_1^1 - b_2^1\|_\infty \\
&\leq W\|\Phi^1 - \Phi^2\| + \|\Phi^1 - \Phi^2\| \\
&\leq (W+1)\|\Phi^1 - \Phi^2\|.
\end{aligned} \tag{378}$$

To establish the induction step $k-1 \mapsto k$ starting from the induction assumption given by

$$\|y_{k-1}^1(x) - y_{k-1}^2(x)\|_\infty \leq (k-1)(W+1)^{k-1} B^{k-2} \|\Phi^1 - \Phi^2\|, \tag{379}$$

we note that

$$
\begin{aligned}
&\|y_k^1(x) - y_k^2(x)\|_\infty && (380)\\
&= \|A_k^1\rho(y_{k-1}^1(x)) + b_k^1 - A_k^2\rho(y_{k-1}^2(x)) - b_k^2\|_\infty && (381)\\
&\leq \|(A_k^1 - A_k^2)\rho(y_{k-1}^1(x))\|_\infty + \|b_k^1 - b_k^2\|_\infty \\
&+ \|A_k^2(\rho(y_{k-1}^1(x)) - \rho(y_{k-1}^2)(x))\|_\infty && (382)\\
&\leq W\|A_k^1 - A_k^2\|_\infty\|\rho(y_{k-1}^1(x))\|_\infty + \|b_k^1 - b_k^2\|_\infty \\
&+ W\|A_k^2\|_\infty\|\rho(y_{k-1}^1(x)) - \rho(y_{k-1}^2(x))\|_\infty && (383)\\
&\leq W\|\Phi^1 - \Phi^2\|(W+1)^{k-1}B^{k-1} + \|\Phi^1 - \Phi^2\| \\
&+ WB\|y_{k-1}^1(x) - y_{k-1}^2(x)\|_\infty && (384)\\
&\leq W\|\Phi^1 - \Phi^2\|(W+1)^{k-1}B^{k-1} + \|\Phi^1 - \Phi^2\| \\
&+ WB(k-1)(W+1)^{k-1}B^{k-2}\|\Phi^1 - \Phi^2\| && (385)\\
&\leq k(W+1)^k B^{k-1}\|\Phi^1 - \Phi^2\|, && (386)
\end{aligned}
$$

where (382) follows from the triangle inequality, in (384) we used (i) $\|(A_k^1 - A_k^2)\|_\infty, \|b_k^1 - b_k^2\|_\infty \leq \|\Phi^1 - \Phi^2\|$, (ii) $\|\rho(y_{k-1}^1(x))\|_\infty \leq \|y_{k-1}^1(x)\|_\infty \leq (W+1)^{k-1}B^{k-1}$ owing to (368), (iii) $\|A_k^2\|_\infty \leq B$, and (iv) $\|\rho(y_{k-1}^1(x)) - \rho(y_{k-1}^2(x))\|_\infty \leq \|y_{k-1}^1(x) - y_{k-1}^2(x)\|_\infty$ thanks to the 1-Lipschitz continuity of ρ, and (385) follows from the induction assumption (379) along with $W+1, B \geq 1$. In particular, we get

$$
\begin{aligned}
\|R(\Phi^1)(x) - R(\Phi^2)(x)\|_\infty &= \|y_\ell^1(x) - y_\ell^2(x)\|_\infty \\
&\leq \ell(W+1)^\ell B^{\ell-1}\|\Phi_1 - \Phi_2\| \\
&\leq L(W+1)^L B^{L-1}\|\Phi_1 - \Phi_2\|,
\end{aligned} \tag{387}
$$

where the last step follows from $\ell \leq L$ with $W+1, B \geq 1$. The proof is concluded by noting that (387) holds for all $x \in [0,1]^d$. □

F Proof of Proposition 4.1

We prove a result, Proposition F.1, that is more general than Proposition 4.1, namely we consider general weight sets and general input-output dimensions. This result is then particularized to the setting of Proposition 4.1.

Proposition F.1. *Let $d, d', W, L \in \mathbb{N}$, and let $\mathbb{A} \subseteq \mathbb{R}$ be a finite set satisfying $\{-1, 0, 1\} \subseteq \mathbb{A}$. Then, for every $k \in \mathbb{N}$ and all $u, v \in \mathbb{A} \cap \mathbb{R}_{\geq 0}$, it holds that*

$$
\mathcal{R}_{\mathcal{T}_1(\mathbb{A},u,v,k)}((d,d'),W,L) \subseteq \mathcal{R}_{\mathbb{A}}((d,d'),16W,(k+3)L)
$$

with

$$
\mathcal{T}_1(\mathbb{A},u,v,k) := \left\{\sum_{i=0}^{k}(u^i\alpha_i + v^i\beta_i) : |\alpha_i|, |\beta_i| \in \mathbb{A}, i = 0,\dots,k\right\}. \tag{388}
$$

Proof. See Appendix F.1. □

Proposition F.1 illustrates that a network with weights in the set $\mathcal{T}_1(\mathbb{A}, u, v, k)$ can equivalently be realized by networks with weights in the simpler underlying set $\mathbb{A}$, at the cost of increased network depth and width. We next demonstrate how Proposition 4.1 follows from Proposition F.1.

Proof of Proposition 4.1. Let $a, b \in \mathbb{N}$. For $k = 1$, (81) is trivially satisfied. For $k \geq 2$, we note that

$$\begin{aligned}
&\mathcal{T}_1(\mathbb{Q}_b^a, 2^{-b}, 2^a, k-1) && (389)\\
&= \left\{ \sum_{i=0}^{k-1} (2^{-bi}\alpha_i + 2^{ai}\beta_i) : |\alpha_i|, |\beta_i| \in \mathbb{Q}_b^a, i = 0, \dots, k-1 \right\} && (390)\\
&= \left\{ \sum_{i=0}^{k-1} (2^{-bi}\alpha_i + 2^{ai}\beta_i) : \alpha_i, \beta_i \in \mathbb{Q}_b^a, i = 0, \dots, k-1 \right\} && (391)\\
&\supseteq \left\{ \pm \sum_{i=-kb}^{ka} 2^i c_i : c_i \in \{0,1\} \right\} && (392)\\
&= \mathbb{Q}_{kb}^{ka}, && (393)
\end{aligned}$$

where in (392) we used $\mathbb{Q}_b^a = \{\pm \sum_{i=-b}^{a} \theta_i 2^i : \theta_i \in \{0,1\}\}$. Based on (389)-(393), it now follows that

$$\mathcal{R}_{kb}^{ka}(W, L) \subseteq \mathcal{R}_{\mathcal{T}_1(\mathbb{Q}_b^a, 2^{-b}, 2^a, k-1)}(W, L). \tag{394}$$

Application of Proposition F.1 with $d = d' = 1$, $u = 2^{-b}$, $v = 2^a$, $\mathbb{A} = \mathbb{Q}_b^a$, and k replaced by $k-1$, yields

$$\mathcal{R}_{\mathcal{T}_1(\mathbb{Q}_b^a, 2^{-b}, 2^a, k-1)}(W, L) \subseteq \mathcal{R}_b^a(16W, (k+2)L). \tag{395}$$

The proof is finalized by combining (394) and (395) to obtain (81). □

F.1 Proof of Proposition F.1

We start with a technical lemma, which shows that, for given $w \in \mathbb{A} \cap \mathbb{R}_{\geq 0}$, every affine mapping $S(A, b)$ with weight set $\mathcal{C}(A), \mathcal{C}(b)$[16] contained in

$$\mathcal{T}_2(\mathbb{A}, w, k) := \left\{ \sum_{i=0}^{k} w^i \alpha_i : \alpha_i \in \mathbb{A} \cap \mathbb{R}_{\geq 0}, i = 0, \dots, k \right\}, \tag{396}$$

can be realized by a composition of affine mappings with weights in $\mathbb{A} \cap \mathbb{R}_{\geq 0}$.

Lemma F.2. *Let $m, n, k \in \mathbb{N}$, $\mathbb{A} \subseteq \mathbb{R}$ with $\{0,1\} \subseteq \mathbb{A}$, and $w \in \mathbb{A} \cap \mathbb{R}_{\geq 0}$. Let $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ be such that $\mathcal{C}(A), \mathcal{C}(b) \subseteq \mathcal{T}_2(\mathbb{A}, w, k)$, with $\mathcal{T}_2(\mathbb{A}, w, k)$ as defined in (396). Then, there exists a neural network configuration $((G_i, h_i))_{i=1}^{k+1} \in \mathcal{N}_{\mathbb{A} \cap \mathbb{R}_{\geq 0}}((n, m), m+n, k+1)$ such that*

$$S(G_{k+1}, h_{k+1}) \circ \cdots \circ S(G_1, h_1) = S(A, b). \tag{397}$$

[16] Recall that $\mathcal{C}(A)$ denotes the set comprising all elements of the matrix A, while $\mathcal{C}(b)$ represents the set containing the entries of the vector b.

Proof of Lemma F.2. We first note that thanks to $\mathcal{C}(A), \mathcal{C}(b) \subseteq \mathcal{T}_2(\mathbb{A}, w, k)$, A and b can be written in the form

$$A = \sum_{i=0}^{k} w^i A_i, \quad b = \sum_{i=0}^{k} w^i b_i, \tag{398}$$

with $A_i \in \mathbb{R}^{m\times n}$, $b_i \in \mathbb{R}^m$, $\mathcal{C}(A_i), \mathcal{C}(b_i) \in \mathbb{A} \cap \mathbb{R}_{\geq 0}$, $i = 0, \ldots, k$. Next, set[17]

$$G_1 = \begin{pmatrix} I_n \\ A_k \end{pmatrix}, \qquad h_1 = \begin{pmatrix} 0_n \\ b_k \end{pmatrix}, \tag{399}$$

$$G_j = \begin{pmatrix} I_n & 0 \\ A_{k-j+1} & wI_m \end{pmatrix}, \qquad h_j = \begin{pmatrix} 0_n \\ b_{k-j+1} \end{pmatrix}, \quad j \in \{1, \ldots, k\}\backslash\{1\}, \tag{400}$$

$$G_{k+1} = \begin{pmatrix} A_0 & wI_m \end{pmatrix}, \qquad h_{k+1} = b_0. \tag{401}$$

We have $((G_i, h_i))_{i=1}^{k+1} \in \mathcal{N}_{\mathbb{A}\cap\mathbb{R}_{\geq 0}}((n, m), m + n, k + 1)$ as desired. It remains to verify (397). To simplify notation, we define $L(((G_i, h_i))_{i=1}^{j})$, $j = 1, \ldots, k + 1$, recursively, according to

$$L(((G_i, h_i))_{i=1}^{j}) = \begin{cases} S(G_1, h_1), & \text{if } j = 1, \\ S(G_j, h_j) \circ L(((G_i, h_i))_{i=1}^{j-1}), & \text{if } j \geq 2, \end{cases} \tag{402}$$

and note that $L((G_i, h_i)_{i=1}^{k+1}) = S(G_{k+1}, h_{k+1}) \circ \cdots \circ S(G_1, h_1)$. The verification of (397) will be effected by proving the following relation by induction. Specifically, for $j = 1, \ldots, k$,

$$L(((G_i, h_i))_{i=1}^{j})(x) = \begin{pmatrix} x \\ \sum_{i=k-j+1}^{k} w^{i-k+j-1} A_i x + \sum_{i=k-j+1}^{k} w^{i-k+j-1} b_i \end{pmatrix}, \quad x \in \mathbb{R}^n. \tag{403}$$

The base case $j = 1$ follows from

$$L(((G_i, h_i))_{i=1}^{1})(x) = S(G_1, h_1)(x) = \begin{pmatrix} x \\ A_k x + b_k \end{pmatrix}, \quad x \in \mathbb{R}^n.$$

If $k = 1$, the induction step is not needed. For $k \geq 2$, we prove the induction step $j - 1 \mapsto j$ with $2 \leq j \leq k$ starting from the induction assumption

$$L(((G_i, h_i))_{i=1}^{j-1})(x) = \begin{pmatrix} x \\ \sum_{i=k-j+2}^{k} w^{i-k+j-2} A_i x + \sum_{i=k-j+2}^{k} w^{i-k+j-2} b_i \end{pmatrix}, \; x \in \mathbb{R}^n, \tag{404}$$

through the following chain of arguments

$$L(((G_i, h_i))_{i=1}^{j})(x) \tag{405}$$

$$= S(G_j, h_j) \circ L(((G_i, h_i))_{i=1}^{j-1})(x) \tag{406}$$

$$= \begin{pmatrix} I_n & 0 \\ A_{k-j+1} & wI_m \end{pmatrix} \begin{pmatrix} x \\ \sum_{i=k-j+2}^{k} w^{i-k+j-2} A_i x + \sum_{i=k-j+2}^{k} w^{i-k+j-2} b_i \end{pmatrix} + \begin{pmatrix} 0_n \\ b_{k-j+1} \end{pmatrix} \tag{407}$$

$$= \begin{pmatrix} x \\ \sum_{i=k-j+1}^{k} w^{i-k+j-1} A_i x + \sum_{i=k-j+1}^{k} w^{i-k+j-1} b_i \end{pmatrix}, \quad x \in \mathbb{R}^n. \tag{408}$$

[17] Note that, if $k = 1$, then $\{1, \ldots, k\}\backslash\{1\} = \emptyset$, so that no assignment is made in (400).

The proof is concluded upon noting that

$$
\begin{aligned}
& S(G_{k+1}, h_{k+1}) \circ \cdots \circ S(G_1, h_1)(x) && (409)\\
&= S(G_{k+1}, h_{k+1}) \circ L(((G_i, h_i))_{i=1}^k)(x) && (410)\\
&= \begin{pmatrix} A_0 & wI_m \end{pmatrix} \begin{pmatrix} x \\ \sum_{i=1}^k w^{i-1} A_i x + \sum_{i=1}^k w^{i-1} b_i \end{pmatrix} + b_0 && (411)\\
&= \sum_{i=0}^k w^i A_i x + \sum_{i=0}^k w^i b_i && (412)\\
&= Ax + b && (413)\\
&= S(A, b)(x), \quad x \in \mathbb{R}^n, && (414)
\end{aligned}
$$

where in (411) we used (403) with $j = k$. □

We proceed by incorporating the effect of the ReLU activation function. Specifically, the following lemma is in the style of Lemma F.2, but for $S(A, b) \circ \rho$.

Lemma F.3. *Let $m, n, k \in \mathbb{N}$, $\mathbb{A} \subseteq \mathbb{R}$ with $\{-1, 0, 1\} \subseteq \mathbb{A}$, and $u, v \in \mathbb{A} \cap \mathbb{R}_{\geq 0}$. Let $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ be such that $\mathcal{C}(A), \mathcal{C}(b) \subseteq \mathcal{T}_1(\mathbb{A}, u, v, k)$, with $\mathcal{T}_1(\mathbb{A}, u, v, k)$ as defined in (388). Then, there exists a neural network configuration $((G_i, h_i))_{i=1}^{k+2}$ with $\mathcal{C}(((G_i, h_i))_{i=1}^{k+2}) \subseteq \mathbb{A}$ and $\mathcal{W}(((G_i, h_i))_{i=1}^{k+2}) \leq 4(m + n)$ such that*

$$
S(G_{k+2}, h_{k+2}) \circ \rho \circ \cdots \circ \rho \circ S(G_1, h_1) \circ \rho = S(A, b) \circ \rho. \tag{415}
$$

Proof of Lemma F.3. We first represent $S(A, b)$ in terms of affine mappings with coefficients in $\mathcal{T}_2(\mathbb{A}, u, k)$ or $\mathcal{T}_2(\mathbb{A}, v, k)$ and then apply Lemma F.2 to these mappings. To this end, we note that $x = \sum_{i=0}^k u^i \alpha_i + v^i \beta_i \in \mathcal{T}_1(\mathbb{A}, u, v, k)$ with $|\alpha_i|, |\beta_i| \in \mathbb{A}$, $i = 0, \ldots, k$, can be decomposed according to

$$
x = x^{(u,+)} - x^{(u,-)} + x^{(v,+)} - x^{(v,-)}, \tag{416}
$$

with $x^{(u,+)} = \sum_{i=0}^k u^i \max\{\alpha_i, 0\}$, $x^{(u,-)} = \sum_{i=0}^k u^i \max\{-\alpha_i, 0\}$, $x^{(v,+)} = \sum_{i=0}^k v^i \max\{\alpha_i, 0\}$, $x^{(v,-)} = \sum_{i=0}^k v^i \max\{-\alpha_i, 0\}$, such that

$$
\begin{aligned}
x^{(u,+)}, x^{(u,-)} &\in \mathcal{T}_2(\mathbb{A}, u, k), && (417)\\
x^{(v,+)}, x^{(v,-)} &\in \mathcal{T}_2(\mathbb{A}, v, k). && (418)
\end{aligned}
$$

Applying this decomposition entry-wise to the matrix A and the vector b yields

$$
\begin{aligned}
A &= A^{(u,+)} - A^{(u,-)} + A^{(v,+)} - A^{(v,-)}, && (419)\\
b &= b^{(u,+)} - b^{(u,-)} + b^{(v,+)} - b^{(v,-)}, && (420)
\end{aligned}
$$

where $A^{(u,+)}, A^{(u,-)}, A^{(v,+)}, A^{(v,-)} \in \mathbb{R}^{m \times n}$ and $b^{(u,+)}, b^{(u,-)}, b^{(v,+)}, b^{(v,-)} \in \mathbb{R}^m$ satisfy

$$
\begin{aligned}
\mathcal{C}(A^{(u,+)}), \mathcal{C}(A^{(u,-)}), \mathcal{C}(b^{(u,+)}), \mathcal{C}(b^{(u,-)}) &\in \mathcal{T}_2(\mathbb{A}, u, k), && (421)\\
\mathcal{C}(A^{(v,+)}), \mathcal{C}(A^{(v,-)}), \mathcal{C}(b^{(v,+)}), \mathcal{C}(b^{(v,-)}) &\in \mathcal{T}_2(\mathbb{A}, v, k). && (422)
\end{aligned}
$$

Consequently, we get

$$
S(A, b) = S(A^{(u,+)}, b^{(u,+)}) - S(A^{(u,-)}, b^{(u,-)}) + S(A^{(v,+)}, b^{(v,+)}) - S(A^{(v,-)}, b^{(v,-)}).
$$

For $\omega = (u,+),(u,-),(v,+),(v,-)$, application of Lemma F.2 to the affine mapping $S(A^\omega, b^\omega)$ now yields the existence of network configurations

$$((G_i^\omega, h_i^\omega))_{i=1}^{k+1} \in \mathcal{N}_{\mathbb{A}\cap\mathbb{R}_{\geq 0}}((n,m), m+n, k+1) \tag{423}$$

such that

$$S(G_{k+1}^\omega, h_{k+1}^\omega) \circ \cdots \circ S(G_1^\omega, h_1^\omega) = S(A^\omega, b^\omega). \tag{424}$$

Set

$$G_1 = \begin{pmatrix} G_1^{(u,+)} \\ G_1^{(u,-)} \\ G_1^{(v,+)} \\ G_1^{(v,-)} \end{pmatrix}, \quad h_1 = \begin{pmatrix} h_1^{(u,+)} \\ h_1^{(u,-)} \\ h_1^{(v,+)} \\ h_1^{(v,-)} \end{pmatrix}, \tag{425}$$

$$G_j = \operatorname{diag}\big(G_j^{(u,+)}, G_j^{(u,-)}, G_j^{(v,+)}, G_j^{(v,-)}\big), \quad h_j = \begin{pmatrix} h_j^{(u,+)} \\ h_j^{(u,-)} \\ h_j^{(v,+)} \\ h_j^{(v,-)} \end{pmatrix}, \tag{426}$$

for $j = 2, \ldots, k+1$, and

$$G_{k+2} = \begin{pmatrix} I_m & -I_m & I_m & -I_m \end{pmatrix}, \quad h_{k+2} = 0_m.$$

A direct calculation shows that

$$\begin{aligned}
&S(G_{k+2}, h_{k+2}) \circ \cdots \circ S(G_1, h_1) && (427)\\
&= S(G_{k+1}^{(u,+)}, h_{k+1}^{(u,+)}) \circ \cdots \circ S(G_1^{(u,+)}, h_1^{(u,+)}) \\
&\quad - S(G_{k+1}^{(u,-)}, h_{k+1}^{(u,-)}) \circ \cdots \circ S(G_1^{(u,-)}, h_1^{(u,-)}) \\
&\quad + S(G_{k+1}^{(v,+)}, h_{k+1}^{(v,+)}) \circ \cdots \circ S(G_1^{(v,+)}, h_1^{(v,+)}) \\
&\quad - S(G_{k+1}^{(v,-)}, h_{k+1}^{(v,-)}) \circ \cdots \circ S(G_1^{(v,-)}, h_1^{(v,-)}) \\
&= S(A^{(u,+)}, b^{(u,+)}) - S(A^{(u,-)}, b^{(u,-)}) && (428)\\
&\quad + S(A^{(v,+)}, b^{(v,+)}) - S(A^{(v,-)}, b^{(v,-)}) && (429)\\
&= S(A, b). && (430)
\end{aligned}$$

In turn, we directly get

$$S(G_{k+2}, h_{k+2}) \circ \cdots \circ S(G_1, h_1) \circ \rho = S(A, b) \circ \rho. \tag{431}$$

Moreover, as, by (423), G_j, h_j, for $j = 1, \ldots, k+1$, contains only non-negative entries, namely in $\mathbb{A} \cap \mathbb{R}_{\geq 0}$, the affine mapping $S(G_j, h_j)$ takes vectors with nonnegative entries into vectors with nonnegative entries, so that

$$S(G_j, h_j) \circ \rho = \rho \circ S(G_j, h_j) \circ \rho. \tag{432}$$

Substituting (432), for $j = 1, \ldots, k$, into (431) finally yields

$$S(G_{k+2}, h_{k+2}) \circ \rho \circ \cdots \circ \rho \circ S(G_1, h_1) \circ \rho = S(A, b) \circ \rho. \qquad \square$$

We are now ready to prove Proposition F.1.

Proof of Proposition F.1. Let

$$\Phi = ((A_j, b_j))_{j=1}^{\ell} \in \mathcal{N}_{\mathcal{T}_1(\mathbb{A},u,v,k)}((d, d'), W, L),$$

with $\ell \leq L$. Set

$$\tilde{A}_1 = \begin{pmatrix} I_d \\ -I_d \end{pmatrix}, \qquad \tilde{b}_1 = 0_{2d}, \tag{433}$$

$$\tilde{A}_2 = \begin{pmatrix} A_1 & -A_1 \end{pmatrix}, \quad \tilde{b}_2 = b_1, \tag{434}$$

and note that $\mathcal{C}(\tilde{A}_1), \mathcal{C}(\tilde{b}_1) \subseteq \{-1, 0, 1\} \subseteq \mathbb{A}$ and $\mathcal{C}(\tilde{A}_2), \mathcal{C}(\tilde{b}_2) \subseteq \mathcal{T}_1(\mathbb{A}, u, v, k)$. Next, using $x = \rho(x) - \rho(-x)$, $x \in \mathbb{R}$, we can write $A_1 x + b_1 = A_1\rho(x) - A_1\rho(-x) + b_1$, $x \in \mathbb{R}^d$, and therefore get

$$S(A_1, b_1) = S(\tilde{A}_2, \tilde{b}_2) \circ \rho \circ S(\tilde{A}_1, \tilde{b}_1). \tag{435}$$

Set[18]

$$\tilde{A}_{j+1} = A_j, \quad \tilde{b}_{j+1} = b_j, \quad \text{for every } j \in \{1, \dots, \ell\}\backslash\{1\}. \tag{436}$$

We have

$$\mathcal{C}(\tilde{A}_1), \mathcal{C}(\tilde{b}_1) \subseteq \mathbb{A}, \tag{437}$$

$$\mathcal{C}(\tilde{A}_j), \mathcal{C}(\tilde{b}_j) \subseteq \mathcal{T}_1(\mathbb{A}, u, v, k), \quad j \in \{2, \dots, \ell + 1\}, \tag{438}$$

$$\mathcal{W}(((\tilde{A}_j, \tilde{b}_j))_{j=1}^{\ell+1}) \leq 2W, \tag{439}$$

and, by (435) and (436),

$$\begin{aligned} R(\Phi) &= S(A_\ell, b_\ell) \circ \rho \circ \cdots \circ S(A_1, b_1) \\ &= S(\tilde{A}_{\ell+1}, \tilde{b}_{\ell+1}) \circ \rho \circ \cdots \circ S(\tilde{A}_2, \tilde{b}_2) \circ \rho \circ S(\tilde{A}_1, \tilde{b}_1). \end{aligned} \tag{440}$$

For $j = 2, \dots, \ell + 1$, we note that $S(\tilde{A}_j, \tilde{b}_j)$ is an affine mapping with $\tilde{A}_j \in \mathbb{R}^{n_j \times n_{j-1}}$, $\tilde{b}_j \in \mathbb{R}^{n_j}$ such that $\mathcal{C}(\tilde{A}_j), \mathcal{C}(\tilde{b}_j) \in \mathcal{T}_1(\mathbb{A}, u, v, k)$, and $n_j, n_{j-1} \leq 2W$. Application of Lemma F.3 to $S(\tilde{A}_j, \tilde{b}_j)$, $j = 2, \dots, \ell + 1$, therefore yields network configurations $((G_{j,i}, h_{j,i}))_{i=1}^{k+2}$ with

$$\mathcal{C}(((G_{j,i}, h_{j,i}))_{i=1}^{k+2}) \subseteq \mathbb{A}, \tag{441}$$

$$\mathcal{W}(((G_{j,i}, h_{j,i}))_{i=1}^{k+2}) \leq 4(n_j + n_{j-1}) \leq 16W, \tag{442}$$

such that

$$S(G_{j,k+2}, h_{j,k+2}) \circ \rho \circ \cdots \circ \rho \circ S(G_{j,1}, h_{j,1}) \circ \rho = S(\tilde{A}_j, \tilde{b}_j) \circ \rho. \tag{443}$$

Substituting (443), for $j = 2, \dots, \ell + 1$, into (440) then yields $R(\Phi) = R(\widetilde{\Phi})$, with

$$\begin{aligned} \widetilde{\Phi} :=& ((G_{\ell+1,k+2}, h_{\ell+1,k+2}), \dots, (G_{\ell+1,1}, h_{\ell+1,1}), \dots, \\ & (G_{2,k+2}, h_{2,k+2}), \dots, (G_{2,1}, h_{2,1}), (\tilde{A}_1, \tilde{b}_1)) \\ \in& \mathcal{N}((d, d'), 16W, (k + 2)\ell + 1) \\ \subseteq& \mathcal{N}((d, d'), 16W, (k + 3)L). \end{aligned}$$

[18] Note that, for $\ell = 1$, the set $\{1, \dots, \ell\}\backslash 1$ is empty and no $\tilde{A}_{j+1}$ or $\tilde{b}_{j+1}$ are assigned.

Since the network $\Phi \in \mathcal{N}_{\mathcal{T}_1(\mathbb{A},u,v,k)}((d,d'),W,L)$ was arbitrary, we have, indeed, established that

$$\{R(\Phi) : \Phi \in \mathcal{N}_{\mathcal{T}_1(\mathbb{A},u,v,k)}((d,d'),W,L)\} \subseteq \mathcal{R}((d,d'),16W,(k+3)L),$$

which is equivalent to

$$\mathcal{R}_{\mathcal{T}_1(\mathbb{A},u,v,k)}((d,d'),W,L) \subseteq \mathcal{R}((d,d'),16W,(k+3)L). \qquad \square$$

G Proof of Theorem 5.1

Let D_1, C_1, E_1 be the constants specified in Proposition 3.4. Set $E_{2,1} = \max\{E_1, 1\}$, $E_{2,2} = 16E_{2,1}(D_1+2)^2$, $C_2 = 2^{18}E_{2,1}C_1$, and $D_2 := \max\{16(D_1+2), 8(E_{2,2}+1)^2, 2^{15}\}$. Let $b, W, L \in \mathbb{N}$ with $W, L \geq D_2$. We have

$$\frac{E_{2,1}L\log(W)}{E_{2,2}\frac{\log(W)}{L}} = L^2\frac{E_{2,1}}{E_{2,2}} \geq D_2^2 \cdot \frac{1}{16(D_1+2)^2} > 1,$$

which, in turn, implies $E_{2,1}L\log(W) > E_{2,2}\frac{\log(W)}{L}$, and therefore the three regimes corresponding to Points 1-3 are well-defined and, indeed, pairwise disjoint. In addition, they exhaust $\mathbb{N}$.

In the *over-quantization regime*, i.e., for $b \geq E_{2,1}L\log(W)$, we have

$$\begin{aligned}
\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}_b^1(W,L)) &\leq \mathcal{A}_\infty(H^1([0,1]), \mathcal{R}^1_{\lceil E_1 L\log(W)\rceil}(W,L)) && (444)\\
&\leq C_1(W^2L^2\log(W))^{-1}, && (445)\\
&\leq C_2(W^2L^2\log(W))^{-1}, && (446)
\end{aligned}$$

where (444) follows from $b = \lceil b\rceil \geq \lceil E_{2,1}L\log(W)\rceil \geq \lceil E_1 L\log(W)\rceil$, in (445) we applied Proposition 3.4, and (446) is by $C_2 > C_1$.

In the *proper-quantization regime*, i.e., for $b \in \left[E_{2,2}\frac{\log(W)}{L}, E_{2,1}L\log(W)\right)$, we set

$$\widetilde{W} = \left\lfloor \frac{W}{16} \right\rfloor, \tag{447}$$

$$\widetilde{L} = \left\lfloor \sqrt{\frac{bL}{16E_{2,1}\log(\widetilde{W})}} \right\rfloor, \tag{448}$$

$$k = \left\lceil \frac{E_{2,1}\tilde{L}\log(\widetilde{W})}{b} \right\rceil. \tag{449}$$

These choices guarantee that $\widetilde{W} = \lfloor \frac{W}{16}\rfloor \geq \lfloor \frac{D_2}{16}\rfloor \geq \lfloor \frac{16(D_1+2)}{16}\rfloor \geq D_1+1$, $\tilde{L} = \left\lfloor \sqrt{\frac{bL}{16E_{2,1}\log(\widetilde{W})}} \right\rfloor \geq \left\lfloor \sqrt{\frac{L}{16E_{2,1}\log(\lfloor \frac{W}{16}\rfloor)}E_{2,2}\frac{\log(W)}{L}} \right\rfloor \geq \left\lfloor \sqrt{\frac{E_{2,2}}{16E_{2,1}}} \right\rfloor = \left\lfloor \sqrt{\frac{16E_{2,1}(D_1+2)^2}{16E_{2,1}}} \right\rfloor \geq D_1+1$, and $k \geq 1$. Application of Proposition 3.4 with W, L replaced by $\widetilde{W}, \tilde{L}$ then yields

$$\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}^1_{\lceil E_1\tilde{L}\log(\widetilde{W})\rceil}(\widetilde{W},\tilde{L})) \leq C_1(\widetilde{W}^2\tilde{L}^2\log(\widetilde{W}))^{-1}. \tag{450}$$

We next note that the family of approximants $\mathcal{R}^1_{\lceil E_1\tilde{L}\log(\widetilde{W})\rceil}(\widetilde{W},\tilde{L})$ in (450) satisfies

$$\mathcal{R}^1_{\lceil E_1\tilde{L}\log(\widetilde{W})\rceil}(\widetilde{W},\tilde{L}) \subseteq \mathcal{R}^k_{kb}(\widetilde{W},\tilde{L}) \tag{451}$$

$$\subseteq \mathcal{R}^1_b(16\widetilde{W},(k+2)\tilde{L}) \tag{452}$$

$$\subseteq \mathcal{R}^1_b(W,(k+2)\tilde{L}), \tag{453}$$

where (451) follows from $kb = \lceil kb\rceil = \left\lceil\left\lceil\frac{E_{2,1}\tilde{L}\log(\widetilde{W})}{b}\right\rceil b\right\rceil \geq \lceil E_{2,1}\tilde{L}\log(\widetilde{W})\rceil \geq \lceil E_1\tilde{L}\log(\widetilde{W})\rceil$, in (452) we used Proposition 4.1 with $a=1$ and (W,L) replaced by $(\widetilde{W},\tilde{L})$, and (453) is by $16\widetilde{W} = 16\lfloor\frac{W}{16}\rfloor \leq W$. We proceed to show that $(k+2)\tilde{L}\leq L$ which will then yield $\mathcal{R}^1_b(W,(k+2)\tilde{L}) \subseteq \mathcal{R}^1_b(W,L)$. This will be done by distinguishing the cases $\frac{E_{2,1}\tilde{L}\log(\widetilde{W})}{b}\geq 1$ and $\frac{E_{2,1}\tilde{L}\log(\widetilde{W})}{b} < 1$. For $\frac{E_{2,1}\tilde{L}\log(\widetilde{W})}{b}\geq 1$, we have $(k+2)\tilde{L} = (\lceil\frac{E_{2,1}\tilde{L}\log(\widetilde{W})}{b}\rceil+2)\tilde{L} \leq \left(\frac{E_{2,1}\tilde{L}\log(\widetilde{W})}{b}+3\right)\tilde{L} \leq \left(\frac{E_{2,1}\tilde{L}\log(\widetilde{W})}{b}+3\frac{E_{2,1}\tilde{L}\log(\widetilde{W})}{b}\right)\tilde{L} = \frac{4E_{2,1}\tilde{L}^2\log(\widetilde{W})}{b} \leq \frac{4E_{2,1}\log(\widetilde{W})}{b}\left(\sqrt{\frac{bL}{16E_{2,1}\log(\widetilde{W})}}\right)^2 \leq L$. For $\frac{E_{2,1}\tilde{L}\log(\widetilde{W})}{b} < 1$, we have $(k+2)\tilde{L} = (\lceil\frac{E_{2,1}\tilde{L}\log(\widetilde{W})}{b}\rceil+2)\tilde{L} = 3\tilde{L} = 3\left\lfloor\sqrt{\frac{bL}{16E_{2,1}\log(\widetilde{W})}}\right\rfloor \leq 3\sqrt{\frac{bL}{16E_{2,1}\log(\widetilde{W})}} \leq 3\sqrt{\frac{E_{2,1}L\log(W)L}{16E_{2,1}\log(\widetilde{W})}} = \sqrt{\frac{9\log(W)}{16\log(\lfloor W/16\rfloor)}}L \leq \sqrt{\frac{9\log(W)}{16\log((W/16)-1)}}L < L$, where in the last inequality we used that, for all $x\geq 2^{15}$, $\frac{\log(x)}{\log((x/16)-1)} \leq \frac{\log(x)}{\log(x/32)} = \frac{\log(x)}{\log(x)-5} \leq \frac{3}{2}$, with $x = W \geq D_2 = \max\{16(D_1+2), 8(E_{2,2}+1)^2, 2^{15}\} \geq 2^{15}$. Overall, we have $(k+2)\tilde{L}\leq L$, which together with (451)-(453) implies

$$\mathcal{R}^1_{\lceil E_1\tilde{L}\log(\widetilde{W})\rceil}(\widetilde{W},\tilde{L}) \subseteq \mathcal{R}^1_b(W,(k+2)\tilde{L}) \subseteq \mathcal{R}^1_b(W,L). \tag{454}$$

We then have

$$\mathcal{A}_\infty(H^1([0,1]),\mathcal{R}^1_b(W,L)) \leq \mathcal{A}_\infty(H^1([0,1]),\mathcal{R}^1_{\lceil E_1\tilde{L}\log(\widetilde{W})\rceil}(\widetilde{W},\tilde{L})) \tag{455}$$

$$\leq C_1(\widetilde{W}^2\tilde{L}^2\log(\widetilde{W}))^{-1} \tag{456}$$

$$\leq C_1\left(\left(\frac{W}{32}\right)^2\tilde{L}^2\log(\widetilde{W})\right)^{-1} \tag{457}$$

$$\leq C_1\left(\left(\frac{W}{32}\right)^2\frac{bL}{2^6\,E_{2,1}\log(\widetilde{W})}\log(\widetilde{W})\right)^{-1} \tag{458}$$

$$= 2^{16}E_{2,1}C_1(W^2Lb)^{-1} \tag{459}$$

$$\leq C_2(W^2Lb)^{-1}, \tag{460}$$

where in (455) we used the inclusion relation (454), (456) is (450), in (457) we employed $\widetilde{W} = \lfloor\frac{W}{16}\rfloor \geq \frac{W}{32}$, which is owing to $\lfloor x\rfloor \geq \frac{1}{2}x$, for $x\geq 1$, (458) follows from

$$\tilde{L}^2 = \left(\left\lfloor\sqrt{\frac{bL}{16E_{2,1}\log(\widetilde{W})}}\right\rfloor\right)^2$$

$$\geq \left(\frac{1}{2}\sqrt{\frac{bL}{16E_{2,1}\log(\widetilde{W})}}\right)^2$$

$$= \frac{bL}{2^6\,E_{2,1}\log(\widetilde{W})},$$

and in (460) we used the definition of C_2.

In the *under-quantization regime*, i.e., for $b \in [1, E_{2,2}\frac{\log(W)}{L})$, we reduce the problem to the *proper-quantization regime*. This will be done by finding a natural number $\overline{W}$ satisfying $D_2 \leq \overline{W} \leq W$ and $E_{2,2}\frac{\log(\overline{W})}{L} \leq b \leq E_{2,1}L\log(\overline{W})$ such that $\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}_b^1(\overline{W}, L))$ can be upper-bounded using the result from the *proper-quantization regime*. Specifically, set $\overline{W} = \left\lfloor 2^{\frac{Lb}{E_{2,2}}} \right\rfloor$, which implies $\overline{W} \leq 2^{\frac{Lb}{E_{2,2}}} \leq W$ and note that

$$\begin{aligned}
\overline{W} &\geq \left\lfloor 2^{\frac{D_2 b}{E_{2,2}}} \right\rfloor && (461)\\
&= \left\lfloor \left(2^{\frac{D_2 b}{2E_{2,2}}}\right)^2 \right\rfloor && (462)\\
&\geq \left\lfloor \left(\frac{D_2 b}{2E_{2,2}}\right)^2 \right\rfloor && (463)\\
&\geq \lfloor 2D_2 \rfloor && (464)\\
&\geq D_2, && (465)
\end{aligned}$$

where (461) follows from $L \geq D_2$, in (463) we used $2^x \geq x$, for $x \geq 1$, with $x = \frac{D_2 b}{2E_{2,2}} = \frac{\max\{16(D_1+2), 8(E_{2,2}+1)^2, 2^{15}\}b}{2E_{2,2}} \geq \frac{8(E_{2,2}+1)^2 b}{2E_{2,2}} \geq 1$, and (464) is by $D_2 = \max\{16(D_1+2), 8(E_{2,2}+1)^2, 2^{15}\} \geq 8(E_{2,2}+1)^2 \geq 8E_{2,2}^2$. We further note that $b \geq E_{2,2}\frac{\log(\overline{W})}{L}$ and

$$\begin{aligned}
b &\leq E_{2,2}\frac{\log(\overline{W}+1)}{L} && (466)\\
&= \frac{2E_{2,2}}{E_{2,1}L^2}\frac{\log(\overline{W}+1)}{2\log(\overline{W})} \cdot E_{2,1}L\log(\overline{W}) && (467)\\
&< E_{2,1}L\log(\overline{W}), && (468)
\end{aligned}$$

where (468) follows from $\frac{2E_{2,2}}{E_{2,1}L^2} = \frac{32E_{2,1}(D_1+2)^2}{E_{2,1}L^2} \leq \frac{32(D_1+2)^2}{D_2^2} = \frac{32(D_1+2)^2}{(\max\{16(D_1+2), 8(E_{2,2}+1)^2, 2^{15}\})^2} \leq \frac{32(D_1+2)^2}{(16(D_1+2))^2} \leq 1$. Then, application of the bound for the *proper-quantization regime*, specifically (459) with W replaced by $\overline{W}$, upon noting that $\overline{W}, L \geq D_2$, $E_{2,2}\frac{\log(\overline{W})}{L} \leq b < E_{2,1}L\log(\overline{W})$ as established in (461)-(468), yields

$$\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}_b^1(\overline{W}, L)) \leq 2^{16}E_{2,1}C_1\left(\overline{W}^2 Lb\right)^{-1}. \tag{469}$$

We finalize the proof for the *under-quantization regime* by noting that

$$\begin{aligned}
\mathcal{A}_\infty(H^1([0,1]), \mathcal{R}_b^1(W, L)) &\leq \mathcal{A}_\infty(H^1([0,1]), \mathcal{R}_b^1(\overline{W}, L)) && (470)\\
&\leq 2^{16}E_{2,1}C_1\left(\overline{W}^2 Lb\right)^{-1} && (471)\\
&\leq 2^{16}E_{2,1}C_1\overline{W}^{-2} && (472)\\
&= 2^{16}E_{2,1}C_1\left(\left\lfloor 2^{\frac{Lb}{E_{2,2}}} \right\rfloor\right)^{-2} && (473)\\
&\leq 2^{16}E_{2,1}C_1\left(\frac{1}{2}2^{\frac{Lb}{E_{2,2}}}\right)^{-2} && (474)\\
&= 2^{18}E_{2,1}C_1(2^{\frac{2}{E_{2,2}}})^{-Lb} && (475)\\
&\leq C_2\,\alpha^{-Lb}, && (476)
\end{aligned}$$

where in (470) we used $\overline{W} \leq W$, (471) is (469), (472) follows from $Lb \geq 1$, (474) is by $\lfloor x \rfloor \geq \frac{1}{2}x$, for $x \geq 1$, with $x = 2^{\frac{Lb}{E_{2,2}}} \geq 1$, and in (476) we set $\alpha := 2^{\frac{2}{E_{2,2}}} > 1$ and used $C_2 = 2^{18}E_{2,1}C_1$.

H Auxiliary results

H.1 Triangle Inequality

Lemma H.1. *Let $(\mathcal{X}, \delta)$ be a metric space and $\mathcal{F}, \mathcal{G}, \mathcal{H} \subseteq \mathcal{X}$. Then,*

$$\mathcal{A}(\mathcal{F}, \mathcal{H}, \delta) \leq \mathcal{A}(\mathcal{F}, \mathcal{G}, \delta) + \mathcal{A}(\mathcal{G}, \mathcal{H}, \delta). \tag{477}$$

Proof. Since δ is a metric, by the triangle inequality, it follows that

$$\delta(f, h) \leq \delta(f, g) + \delta(g, h), \text{ for all } f \in \mathcal{F}, g \in \mathcal{G}, h \in \mathcal{H}. \tag{478}$$

Taking $\inf_{h \in \mathcal{H}}$ in (478) yields

$$\begin{aligned} \inf_{h \in \mathcal{H}} \delta(f, h) &\leq \delta(f, g) + \inf_{h \in \mathcal{H}} \delta(g, h) && (479) \\ &\leq \delta(f, g) + \mathcal{A}(\mathcal{G}, \mathcal{H}, \delta), \text{ for all } f \in \mathcal{F}, g \in \mathcal{G}, && (480) \end{aligned}$$

which, upon taking $\inf_{g \in \mathcal{G}}$ on both sides, yields

$$\inf_{h \in \mathcal{H}} \delta(f, h) \leq \inf_{g \in \mathcal{G}} \delta(f, g) + \mathcal{A}(\mathcal{G}, \mathcal{H}, \delta), \text{ for all } f \in \mathcal{F}. \tag{481}$$

Taking $\sup_{f \in \mathcal{F}}$ on both sides of (481) finalizes the proof. □

H.2 Operations Over Functions Realized by ReLU Networks

This section is concerned with the construction of ReLU networks realizing the composition, linear combination, and "parallelization" of functions. We start with a technical lemma which shows how ReLU networks can be augmented to deeper networks while retaining their input-output relation. This result has been documented previously in [14], but we restate it here in our notation and provide a proof, for the sake of clarity of exposition and completeness.

Lemma H.2. *Let $d, d', W, L \in \mathbb{N}$ and $B \in \mathbb{R}_+$ with $B \geq 1$. Then, for $f \in \mathcal{R}((d, d'), W, L, B)$, there exists a network $\Phi \in \mathcal{N}((d, d'), \max\{W, 2d'\}, L, B)$ such that $\mathcal{L}(\Phi) = L$ and $R(\Phi) = f$.*

Proof. By definition, there exists a network $\widetilde{\Phi} = ((\tilde{A}_\ell, \tilde{b}_\ell))_{\ell=1}^{\tilde{L}} \in \mathcal{N}((d, d'), W, L, B)$, with $\tilde{L} \leq L$, such that $R(\widetilde{\Phi}) = f$. If $\tilde{L} = L$, the proof is finished by taking $\Phi = \widetilde{\Phi}$. For $L > \tilde{L}$, we set $\Phi := ((A_\ell, b_\ell))_{\ell=1}^{L}$ with

$$(A_\ell, b_\ell) := (\tilde{A}_\ell, \tilde{b}_\ell), \quad \text{for } 1 \leq \ell < \tilde{L},^{19} \tag{482}$$

[19] Here and in what follows, we use the convention that if there does not exist an ℓ satisfying the constraint, the assignment is skipped; in the present case, this would apply if $\tilde{L} = 1$. This convention is to unify the discussion on general cases and corner cases.

$A_{\tilde{L}} := \begin{pmatrix} \tilde{A}_{\tilde{L}} \\ -\tilde{A}_{\tilde{L}} \end{pmatrix}$, $b_{\tilde{L}} := \begin{pmatrix} \tilde{b}_{\tilde{L}} \\ -\tilde{b}_{\tilde{L}} \end{pmatrix}$, $A_\ell := I_{2d'}$, $b_\ell := 0_{2d'}$ for ℓ such that $\tilde{L} < \ell < L$, and $A_L := \begin{pmatrix} I_{d'} & -I_{d'} \end{pmatrix}$, $b_L := 0$. We have $\Phi \in \mathcal{R}((d, d'), \max\{W, 2d'\}, L, B)$ and $\mathcal{L}(\Phi) = L$. It remains to show that $R(\Phi) = f$. This will be effected by noting that

$$S(A_L, b_L) \circ \rho \circ \cdots \circ \rho \circ S(A_{\tilde{L}}, b_{\tilde{L}}) = S(A_L, b_L) \circ \rho \circ S(A_{\tilde{L}}, b_{\tilde{L}}) \tag{483}$$

$$= S(\tilde{A}_{\tilde{L}}, \tilde{b}_{\tilde{L}}), \tag{484}$$

where in (483) we used $\rho \circ S(A_\ell, b_\ell) = \rho \circ S(I_{2d'}, 0_{2d'}) = \rho$, for $\tilde{L} < \ell < L$, and $\rho \circ \rho = \rho$, and (484) is by $\left(S(A_L, b_L) \circ \rho \circ S(A_{\tilde{L}}, b_{\tilde{L}})\right)(x) = \rho(\tilde{A}_{\tilde{L}} x + \tilde{b}_{\tilde{L}}) - \rho(-\tilde{A}_{\tilde{L}} x - \tilde{b}_{\tilde{L}}) = \tilde{A}_{\tilde{L}} x + \tilde{b}_{\tilde{L}} = S(\tilde{A}_{\tilde{L}}, \tilde{b}_{\tilde{L}})(x)$, for $x \in \mathbb{R}_{d''}$, with d'' denoting the number of rows of $\tilde{A}_{\tilde{L}}$. Combining (483)-(484) and (482), we see that $R(\Phi) = R(\widetilde{\Phi}) = f$. □

We are now ready to state the main result of this section.

Lemma H.3. *Let $a \in \mathbb{R}$, $d, d', d'' \in \mathbb{N}$, $W_i, L_i, \in \mathbb{N}$, and $B_i \in \mathbb{R}_+$ with $B_i \geq 1$, $i = 1, 2, 3, 4$, and let*

$$f_1 \in \mathcal{R}((d, d'), W_1, L_1, B_1), \tag{485}$$

$$f_2 \in \mathcal{R}((d, d''), W_2, L_2, B_2), \tag{486}$$

$$f_3 \in \mathcal{R}((d', d'''), W_3, L_3, B_3), \tag{487}$$

$$f_4 \in \mathcal{R}((d, d'), W_4, L_4, B_4). \tag{488}$$

Then,

$$a \cdot f_1 \in \mathcal{R}((d, d'), W_1, L_1, \max\{|a| B_1, B_1\}), \tag{489}$$

$$(f_1, f_2) \in \mathcal{R}((d, d' + d''), \max\{W_1, 2d'\} + \max\{W_2, 2d''\}, \max\{L_1, L_2\}, \max\{B_1, B_2\}), \tag{490}$$

$$f_3 \circ f_1 \in \mathcal{R}((d, d'''), \max\{W_1, W_3, 2d'\}, L_1 + L_3, \max\{B_1, B_3\}), \tag{491}$$

$$f_1 + f_4 \in \mathcal{R}((d, d'), \max\{W_1, 2d'\} + \max\{W_4, 2d'\}, \max\{L_1, L_4\} + 1, \max\{B_1, B_4\}), \tag{492}$$

where, for $x \in \mathbb{R}^d$, $(a \cdot f_1)(x) := a \cdot f_1(x)$, $(f_1, f_2)(x) := (f_1(x), f_2(x))$, $f_3 \circ f_1 := f_3(f_1(x))$, and $(f_1 + f_4)(x) := f_1(x) + f_4(x)$.

Proof. We shall prove (489)-(492) individually as follows.

1. According to (485), there exists a $\Phi_1 = ((A_{1,\ell}, b_{1,\ell}))_{\ell=1}^{\mathcal{L}(\Phi_1)} \in \mathcal{R}((d, d'), W_1, L_1, B_1)$ such that $R(\Phi_1) = f_1$. Now, let $\widetilde{\Phi} = ((\tilde{A}_\ell, \tilde{b}_\ell))_{\ell=1}^{\mathcal{L}(\Phi_1)}$ with $(\tilde{A}_\ell, \tilde{b}_\ell) := (A_{1,\ell}, b_{1,\ell})$, for $1 \leq \ell < \mathcal{L}(\Phi_1)$, and $(\tilde{A}_{\mathcal{L}(\Phi_1)}, \tilde{b}_{\mathcal{L}(\Phi_1)}) = (a \cdot A_{1,\mathcal{L}(\Phi_1)}, a \cdot b_{1,\mathcal{L}(\Phi_1)})$. We hence have $\widetilde{\Phi} \in \mathcal{N}((d, d'), W_1, L_1, \max\{|a| B_1, B_1\})$ and $R(\widetilde{\Phi})(x) = a \cdot R(\Phi_1)(x) = a \cdot f_1(x)$, $x \in \mathbb{R}^d$, which establishes (489).

2. Application of Lemma H.2 to $f_1 \in \mathcal{R}((d, d'), W_1, L_1, B_1) \subseteq \mathcal{R}((d, d'), W_1, \max\{L_1, L_2\}, B_1)$ implies the existence of a network

$$\Phi_1 = ((A_{1,\ell}, b_{1,\ell}))_{\ell=1}^{\max\{L_1, L_2\}} \in \mathcal{N}((d, d'), \max\{W_1, 2d'\}, \max\{L_1, L_2\}, B_1)$$

such that $R(\Phi_1) = f_1$. Similarly, application of Lemma H.2 to $f_2 \in \mathcal{R}((d, d''), W_2, \max\{L_1, L_2\}, B_2)$ implies the existence of a $\Phi_2 = ((A_{2,\ell}, b_{2,\ell}))_{\ell=1}^{\max\{L_1, L_2\}} \in \mathcal{N}((d, d''),$

$\{W_2, 2d''\}, \max\{L_1, L_2\}, B_2)$ such that $R(\Phi_2) = f_2$. Now set $\widetilde{\Phi} := ((\tilde{A}_\ell, \tilde{b}_\ell))_{\ell=1}^{\max\{L_1,L_2\}}$ with $\tilde{A}_1 := \begin{pmatrix} A_{1,1} \\ A_{2,1} \end{pmatrix}$, $\tilde{b}_1 := \begin{pmatrix} b_{1,1} \\ b_{2,1} \end{pmatrix}$, and $\tilde{A}_\ell := \operatorname{diag}(A_{1,\ell}, A_{2,\ell})$, $\tilde{b}_\ell := \begin{pmatrix} b_{1,\ell} \\ b_{2,\ell} \end{pmatrix}$, for $1 < \ell \leq \max\{L_1, L_2\}$. We then get $R(\widetilde{\Phi})(x) = (R(\Phi_1)(x), R(\Phi_2)(x))$, $x \in \mathbb{R}^d$, and

$$\widetilde{\Phi} \in \mathcal{N}((d, d' + d''), \{W_1, 2d'\} + \{W_2, 2d''\}, \max\{L_1, L_2\}, \max\{B_1, B_2\}),$$

which establishes (490).

3. It follows from $f_1 \in \mathcal{R}((d, d'), W_1, L_1, B_1)$ that there exists a $\Phi_1 = ((A_{1,\ell}, b_{1,\ell}))_{\ell=1}^{\mathcal{L}(\Phi_1)} \in \mathcal{N}((d, d'), W_1, L_1, B_1)$ such that $R(\Phi_1) = f_1$. Similarly, it follows from $f_3 \in \mathcal{R}((d', d'''), W_3, L_3, B_3)$ that there exists a $\Phi_3 = ((A_{3,\ell}, b_{3,\ell}))_{\ell=1}^{\mathcal{L}(\Phi_3)} \in \mathcal{N}((d', d'''), W_3, L_3, B_3)$ so that $R(\Phi_3) = f_3$. Now, let $\widetilde{\Phi} := ((\tilde{A}_\ell, \tilde{b}_\ell))_{\ell=1}^{L_1+L_3}$ with

$$(\tilde{A}_\ell, \tilde{b}_\ell) := (A_{1,\ell}, b_{1,\ell}), \quad \text{for } 1 \leq \ell < \mathcal{L}(\Phi_1), \tag{493}$$

$$\tilde{A}_{\mathcal{L}(\Phi_1)} = \begin{pmatrix} A_{1,\mathcal{L}(\Phi_1)} \\ -A_{1,\mathcal{L}(\Phi_1)} \end{pmatrix}, \qquad \tilde{b}_{\mathcal{L}(\Phi_1)} = \begin{pmatrix} b_{1,\mathcal{L}(\Phi_1)} \\ -b_{1,\mathcal{L}(\Phi_1)} \end{pmatrix},$$
$$\tilde{A}_{\mathcal{L}(\Phi_1)+1} = \begin{pmatrix} A_{3,1} & -A_{3,1} \end{pmatrix}, \qquad \tilde{b}_{\mathcal{L}(\Phi_1)+1} = b_{3,1},$$

and

$$(\tilde{A}_\ell, \tilde{b}_\ell) = (A_{3,\ell-\mathcal{L}(\Phi_1)}, b_{3,\ell-\mathcal{L}(\Phi_1)}), \quad \text{for } \mathcal{L}(\Phi_1) + 1 < \ell \leq \mathcal{L}(\Phi_1) + \mathcal{L}(\Phi_3). \tag{494}$$

Next, note that

$$(S(\tilde{A}_{\mathcal{L}(\Phi_1)+1}, \tilde{b}_{\mathcal{L}(\Phi_1)+1}) \circ \rho \circ S(\tilde{A}_{\mathcal{L}(\Phi_1)}, \tilde{b}_{\mathcal{L}(\Phi_1)}))(x) \tag{495}$$
$$= \left(S(\begin{pmatrix} A_{3,1} & -A_{3,1} \end{pmatrix}, b_1^3) \circ \rho \circ S\left(\begin{pmatrix} A_{1,\mathcal{L}(\Phi_1)} \\ -A_{1,\mathcal{L}(\Phi_1)} \end{pmatrix}, \begin{pmatrix} b_{1,\mathcal{L}(\Phi_1)} \\ -b_{1,\mathcal{L}(\Phi_1)} \end{pmatrix} \right) \right)(x) \tag{496}$$
$$= A_{3,1}\rho(A_{1,\mathcal{L}(\Phi_1)}x + b_{1,\mathcal{L}(\Phi_1)}) - A_{3,1}\rho(-A_{1,\mathcal{L}(\Phi_1)}x - b_{1,\mathcal{L}(\Phi_1)}) + b_{3,1} \tag{497}$$
$$= A_{3,1}(A_{1,\mathcal{L}(\Phi_1)}x + b_{1,\mathcal{L}(\Phi_1)}) + b_{3,1} \tag{498}$$
$$= (S(A_{3,1}, b_{3,1}) \circ S(A_{1,\mathcal{L}(\Phi_1)}, b_{1,\mathcal{L}(\Phi_1)}))(x), \quad \text{for } x \in \mathbb{R}^d. \tag{499}$$

Combining (493), (494), and (495)-(499), we get $R(\widetilde{\Phi}) = R(\Phi_3) \circ R(\Phi_1) = f_3 \circ f_1$, which together with

$$\widetilde{\Phi} \in \mathcal{N}((d, d'''), \max\{W_1, W_3, 2d'\}, L_1 + L_3, \max\{B_1, B_3\})$$

establishes (491).

4. Let $f_5 = S(\begin{pmatrix} I_{d'} & I_{d'} \end{pmatrix}, 0_{d'}) \in \mathcal{R}((2d', d'), 2d', 1, 1)$ so that $f_5(y, z) = y + z$, for all $y, z \in \mathbb{R}^{d'}$. We have

$$f_1 + f_4 = f_5 \circ (f_1, f_4). \tag{500}$$

Application of (490) with f_2 replaced by f_4 yields

$$(f_1, f_4) \in \mathcal{R}((d, 2d'), \max\{W_1, 2d'\} + \max\{W_4, 2d'\}, \max\{L_1, L_4\}, \max\{B_1, B_4\}).$$

Finally, using (491) with f_3 replaced by f_5 and f_1 replaced by (f_1, f_4), we obtain

$$\begin{aligned} & f_1 + f_4 \\ = \; & f_5 \circ (f_1, f_4) \\ \in \; & \mathcal{R}((d, d'), \max\{W_1, 2d'\} + \max\{W_4, 2d'\}, \max\{L_1, L_4\} + 1, \max\{B_1, B_4\}). \end{aligned}$$

□

H.3 Depth-Weight-Magnitude Tradeoff

This section is concerned with the realization of given ReLU networks by corresponding deeper networks of smaller weight-magnitude. The main result is the following proposition.

Proposition H.4. *Let $W, L \in \mathbb{N}$ with $W \geq 2$, $L' \in \mathbb{N} \cup \{0\}$, and $B, B' \in \mathbb{R}$ with $B, B' \geq 1$. It holds that*

$$\frac{(B')^{L+L'} \lfloor W/2 \rfloor^{L'}}{B^L} \cdot \mathcal{R}(W, L, B) \subseteq \mathcal{R}(W, L + L', B'). \tag{501}$$

In particular, if $\frac{(B')^{L'+L} \lfloor W/2 \rfloor^{L'}}{B^L} \geq 1$, then

$$\mathcal{R}(W, L, B) \subseteq \mathcal{R}(W, L + L', B').$$

For the proof of Proposition H.4, we need the following two technical lemmata.

Lemma H.5. *Let $W, L \in \mathbb{N}$ with $W \geq 2$ and $B, B' \in \mathbb{R}$ with $B, B' \geq 1$. We have*

$$(B'/B)^L \cdot \mathcal{R}(W, L, B) = \mathcal{R}(W, L, B'). \tag{502}$$

Proof. Let $g \in \mathcal{R}(W, L, B)$. Application of Lemma H.2 yields the existence of a network configuration $\Phi = ((A_\ell, b_\ell))_{\ell=1}^L \in \mathcal{N}(W, L, B)$ with exactly L layers such that $R(\Phi) = g$. Now, let $\widetilde{\Phi} := ((\frac{B'}{B} A_\ell, \frac{B'}{B} b_\ell))_{\ell=1}^L \in \mathcal{R}(W, L, B')$. It then follows from the positive homogeneity of the ReLU function, i.e., $\rho(ax) = a\rho(x)$, for all $x \in \mathbb{R}$ and $a \in \mathbb{R}_+$, that $R(\widetilde{\Phi}) = (B'/B)^L \cdot \mathcal{R}(\Phi) = (B'/B)^L \cdot g$, which in turn implies $(B'/B)^L \cdot g \in \mathcal{R}(W, L, B')$. As the choice of $g \in \mathcal{R}(W, L, B)$ was arbitrary, we have established that

$$(B'/B)^L \cdot \mathcal{R}(W, L, B) \subseteq \mathcal{R}(W, L, B'). \tag{503}$$

Swapping the roles of B and B' in (503) yields

$$(B/B')^L \cdot \mathcal{R}(W, L, B') \subseteq \mathcal{R}(W, L, B). \tag{504}$$

Combining (503) with (504) establishes (502). □

Lemma H.6. *Let $W, L \in \mathbb{N}$ with $W \geq 2$ and $L' \in \mathbb{N} \cup \{0\}$. We have*

$$\lfloor W/2 \rfloor^{L'} \cdot \mathcal{R}(W, L, 1) \subseteq \mathcal{R}(W, L + L', 1). \tag{505}$$

Proof. For $L' = 0$, (505) is trivially satisfied. We continue with $L' \geq 1$. Let $g \in \mathcal{R}(W, L, 1)$. Application of Lemma H.2 then yields the existence of a network configuration $\Phi = ((A_\ell, b_\ell))_{\ell=1}^L \in \mathcal{N}(W, L, 1)$ with exactly L layers such that $R(\Phi) = g$. Let $D := \operatorname{diag}(1_{\lfloor W/2 \rfloor \times \lfloor W/2 \rfloor}, 1_{\lfloor W/2 \rfloor \times \lfloor W/2 \rfloor})$. Consider $\widetilde{\Phi} := ((\tilde{A}_\ell, \tilde{b}_\ell))_{\ell=1}^{L+L'}$ with

$$(\tilde{A}_\ell, \tilde{b}_\ell) := (A_\ell, b_\ell), \quad \text{for } 1 \leq \ell < L, \tag{506}$$

$$\tilde{A}_L = \begin{pmatrix} 1_{\lfloor W/2 \rfloor \times 1} \\ -1_{\lfloor W/2 \rfloor \times 1} \end{pmatrix} \cdot A_L = (\underbrace{A_L^T, \cdots, A_L^T}_{\lfloor W/2 \rfloor \text{ times}}, \underbrace{-A_L^T, \cdots, -A_L^T}_{\lfloor W/2 \rfloor \text{ times}})^T,$$

$$\tilde{b}_L = \begin{pmatrix} 1_{\lfloor W/2 \rfloor \times 1} \\ -1_{\lfloor W/2 \rfloor \times 1} \end{pmatrix} \cdot b_L = (\underbrace{b_L, \cdots, b_L}_{\lfloor W/2 \rfloor \text{ times}}, \underbrace{-b_L, \cdots, -b_L}_{\lfloor W/2 \rfloor \text{ times}})^T,$$

and

$$(\tilde{A}_\ell, \tilde{b}_\ell) = (D, 0_{2\lfloor W/2 \rfloor}), \quad \text{for } L < \ell < L + L', \tag{507}$$

$$(\tilde{A}_{L+L'}, \tilde{b}_{L+L'}) = ((1_{1\times\lfloor W/2\rfloor}, -1_{1\times\lfloor W/2\rfloor}), 0). \tag{508}$$

We have $R(\widetilde{\Phi}) \in \mathcal{R}(W, L + L', 1)$. Further, it follows from the definition of $(\tilde{A}_L, \tilde{b}_L)$ that

$$S(\tilde{A}_L, \tilde{b}_L) = S\left(\begin{pmatrix} 1_{\lfloor W/2\rfloor\times 1} \\ -1_{\lfloor W/2\rfloor\times 1}\end{pmatrix}, 0_{2\lfloor W/2\rfloor}\right) \circ S(A_L, b_L). \tag{509}$$

Putting everything together, we obtain

$$\begin{aligned} R(\widetilde{\Phi}) &= S\left(\begin{pmatrix} 1_{1\times\lfloor W/2\rfloor} & -1_{1\times\lfloor W/2\rfloor}\end{pmatrix}, 0\right) \circ \rho \circ \underbrace{S\left(D, 0_{2\lfloor W/2\rfloor}\right) \circ \cdots \circ \rho}_{\substack{(L'-1)\text{-fold self-composition} \\ \text{of } S(D, 0_{2\lfloor W/2\rfloor})\circ\rho}} \\ &\quad \circ S\left(\begin{pmatrix} 1_{\lfloor W/2\rfloor\times 1} \\ -1_{\lfloor W/2\rfloor\times 1}\end{pmatrix}, 0_{2\lfloor W/2\rfloor}\right) \circ R(\Phi) \\ &:= h \circ R(\Phi), \end{aligned}$$

where we used the convention that 0-fold self-composition of a function equals the identity function. A direct calculation yields

$$\begin{aligned} h(x) &= \begin{pmatrix} 1_{1\times\lfloor W/2\rfloor} & -1_{1\times\lfloor W/2\rfloor}\end{pmatrix} \cdot \begin{pmatrix} \lfloor W/2\rfloor^{L'-1}\rho(x)\cdot 1_{\lfloor W/2\rfloor} \\ \lfloor W/2\rfloor^{L'-1}\rho(-x)\cdot 1_{\lfloor W/2\rfloor}\end{pmatrix} \\ &= \lfloor W/2\rfloor^{L'} x, \quad x \in \mathbb{R}, \end{aligned}$$

which together with $R(\widetilde{\Phi}) = h \circ R(\Phi)$ implies $\lfloor W/2\rfloor^{L'} \cdot R(\Phi) = R(\widetilde{\Phi}) \in \mathcal{R}(W, L + L', 1)$. Since the choice of $g \in \mathcal{R}(W, L, 1)$ was arbitrary, we have established (505). □

We are now ready to prove Proposition H.4.

Proof of Proposition H.4. We have

$$\mathcal{R}(W, L + L', B') = (B')^{L'+L} \cdot \mathcal{R}(W, L + L', 1) \tag{510}$$

$$\supseteq (B')^{L'+L} \lfloor W/2\rfloor^{L'} \cdot \mathcal{R}(W, L, 1) \tag{511}$$

$$= \frac{(B')^{L'+L}\lfloor W/2\rfloor^{L'}}{B^L} \cdot \mathcal{R}(W, L, B), \tag{512}$$

where (510) and (512) follow from Lemma H.5, and in (511) we used Lemma H.6. This establishes the first part of the proposition, namely (501). Next, let $K := \frac{(B')^{L'+L}\lfloor W/2\rfloor^{L'}}{B^L} \geq 1$. For $g \in \mathcal{R}(W, L, B)$, we have $K \cdot g \in \mathcal{R}(W, L + L', B')$ thanks to (510)-(512). It then follows from Lemma H.3 that $K^{-1} \cdot (K \cdot g) \in \mathcal{R}(W, L + L', \max\{|K^{-1}|B', B'\}) = \mathcal{R}(W, L + L', B')$. Since the choice of $g \in \mathcal{R}(W, L, B)$ was arbitrary, we have established that $\mathcal{R}(W, L, B) \subseteq \mathcal{R}(W, L + L', B')$. □

H.4 Auxiliary Lemma on Minimax Error

Lemma H.7. *Let $[u,v] \subseteq \mathbb{R}$ and $\mathbb{A} \subseteq \mathbb{R}$ such that $\mathbb{A} \cap [u,v] \neq \emptyset$. Then,*

$$\mathcal{A}([u,v],\mathbb{A},|\cdot|) \leq \mathcal{A}([u,v],\mathbb{A}\cap[u,v],|\cdot|) \leq 2\mathcal{A}([u,v],\mathbb{A},|\cdot|). \tag{513}$$

Proof. We note that the first inequality in (513) follows directly upon noting that $\mathbb{A}\cap[u,v] \subseteq \mathbb{A}$, so we only have to prove the second inequality. Suppose first that $\mathbb{A}$ is a closed set. Let $u' = \inf(\mathbb{A}\cap[u,v])$ and $v' = \sup(\mathbb{A}\cap[u,v])$. The points u', v' are elements of $[u,v]\cap\mathbb{A}$, as $[u,v]\cap\mathbb{A}$ is closed and non-empty. We first note that, for $f:\mathbb{R}\mapsto\mathbb{R}$,

$$\sup_{x\in[u,v]} f(x) = \max\left\{\sup_{x\in[u,u']} f(x), \sup_{x\in[u',v']} f(x), \sup_{x\in[v',v]} f(x)\right\}.$$

Then, for $\mathbb{B}\subseteq\mathbb{R}$, we set $f(x) = \inf_{y\in\mathbb{B}}|x-y|$, $x\in\mathbb{R}$, and get

$$\mathcal{A}([u,v],\mathbb{B},|\cdot|) = \max\{\mathcal{A}([u,u'],\mathbb{B},|\cdot|),\mathcal{A}([u',v'],\mathbb{B},|\cdot|),\mathcal{A}([v',v],\mathbb{B},|\cdot|)\}. \tag{514}$$

The result will be established by showing that

$$\mathcal{A}([u,u'],\mathbb{A}\cap[u,v],|\cdot|) \leq 2\,\mathcal{A}([u,u'],\mathbb{A},|\cdot|), \tag{515}$$
$$\mathcal{A}([u',v'],\mathbb{A}\cap[u,v],|\cdot|) \leq 2\,\mathcal{A}([u',v'],\mathbb{A},|\cdot|), \tag{516}$$
$$\mathcal{A}([v',v],\mathbb{A}\cap[u,v],|\cdot|) \leq 2\,\mathcal{A}([v',v],\mathbb{A},|\cdot|) \tag{517}$$

and then combining these three inequalities with (514) for $\mathbb{B}=\mathbb{A}$ and $\mathbb{B}=\mathbb{A}\cap[u,v]$. We start by establishing (515). To this end, we note that

$$\begin{aligned}
\mathcal{A}([u,u'],\mathbb{A}\cap[u,v],|\cdot|) &\leq \mathcal{A}([u,u'],\{u'\},|\cdot|) && (518)\\
&= \sup_{x\in[u,u']}|x-u'| && (519)\\
&= |u'-u|, && (520)\\
\mathcal{A}([u,u'],\mathbb{A},|\cdot|) &\geq \mathcal{A}\left(\left\{\frac{u+u'}{2}\right\},\mathbb{A},|\cdot|\right) && (521)\\
&= \inf_{y\in\mathbb{A}}\left|\frac{u+u'}{2}-y\right| && (522)\\
&= \frac{1}{2}|u'-u|, && (523)
\end{aligned}$$

where in (521) we used that $\{\frac{u+u'}{2}\} \subseteq [u,u']$. Combining (518)-(520) and (521)-(523) then implies $\mathcal{A}([u,u'],\mathbb{A}\cap[u,v],|\cdot|) \leq |u'-u| \leq 2\mathcal{A}([u,u'],\mathbb{A},|\cdot|)$. A similar line of reasoning shows that $\mathcal{A}([v',v],\mathbb{A}\cap[u,v],|\cdot|) \leq |v-v'|$ and $\mathcal{A}([v',v],\mathbb{A},|\cdot|) \geq \frac{1}{2}|v-v'|$, which taken together yields $\mathcal{A}([v',v],\mathbb{A}\cap[u,v],|\cdot|) \leq 2\mathcal{A}([v',v],\mathbb{A},|\cdot|)$ and thereby establishes (517). It remains to prove (516). To this end, we first define the mapping $p:\mathbb{A}\mapsto\mathbb{A}\cap[u,v]$ according to

$$p(y) = \begin{cases} u', & \text{if } y\in(-\infty,u'],\\ y, & \text{if } y\in[u',v'],\\ v', & \text{if } y\in[v',\infty). \end{cases}$$

Next, for $x \in [u', v']$ and $y \in \mathbb{A}$, we have $|x - y| \geq |x - p(y)|$. (516) now follows from

$$\begin{aligned}
\mathcal{A}([u', v'], \mathbb{A}, |\cdot|) &= \sup_{x \in [u', v']} \inf_{y \in \mathbb{A}} |x - y| && (524)\\
&\leq \sup_{x \in [u', v']} \inf_{y \in \mathbb{A}} |x - p(y)| && (525)\\
&\leq \sup_{x \in [u', v']} \inf_{z \in \mathbb{A} \cap [u, v]} |x - z| && (526)\\
&= \mathcal{A}([u', v'], \mathbb{A} \cap [u, v], |\cdot|) && (527)\\
&\leq 2\mathcal{A}([u', v'], \mathbb{A} \cap [u, v], |\cdot|), && (528)
\end{aligned}$$

where (526) is by $p(y) \in \mathbb{A} \cap [u, v]$, for $y \in \mathbb{A}$.

For a general, not necessarily closed, set $\mathbb{A}$, we consider the closure of $\mathbb{A}$, denoted by $\bar{\mathbb{A}}$. As we have already established the second inequality in (513) for closed sets $\mathbb{A}$, we can conclude that

$$\mathcal{A}([u, v], \overline{\mathbb{A}} \cap [u, v], |\cdot|) \leq 2\mathcal{A}([u, v], \overline{\mathbb{A}}, |\cdot|). \tag{529}$$

The proof is finalized upon using $\mathcal{A}([u, v], \overline{\mathbb{A}}, |\cdot|) = \sup_{x \in [u,v]} \inf_{y \in \overline{\mathbb{A}}} |x-y| = \sup_{x \in [u,v]} \inf_{y \in \mathbb{A}} |x-y| = \mathcal{A}([u, v], \mathbb{A}, |\cdot|)$ and $\mathcal{A}([u, v], \overline{\mathbb{A}} \cap [u, v], |\cdot|) = \mathcal{A}([u, v], \overline{\mathbb{A} \cap [u, v]}, |\cdot|) = \mathcal{A}([u, v], \mathbb{A} \cap [u, v], |\cdot|)$ in (529) to obtain $\mathcal{A}([u, v], \mathbb{A} \cap [u, v], |\cdot|) \leq 2\mathcal{A}([u, v], \mathbb{A}, |\cdot|)$. □

References

[1] S. Tewksbury and R. Hallock, "Oversampled, linear predictive and noise-shaping coders of order $N > 1$," *IEEE Transactions on Circuits and Systems*, vol. 25, no. 7, pp. 436–447, 1978.

[2] J. Schmidt-Hieber, "Nonparametric regression using deep neural networks with ReLU activation function," *The Annals of Statistics*, vol. 48, no. 4, pp. 1875 – 1897, 2020.

[3] M. Chen, H. Jiang, W. Liao, and T. Zhao, "Nonparametric regression on low-dimensional manifolds using deep ReLU networks: Function approximation and statistical recovery," *Information and Inference: A Journal of the IMA*, vol. 11, no. 4, pp. 1203–1253, 03 2022.

[4] R. Nakada and M. Imaizumi, "Adaptive approximation and generalization of deep neural network with intrinsic dimensionality," *Journal of Machine Learning Research*, vol. 21, no. 174, pp. 1–38, 2020.

[5] I. Gühring and M. Raslan, "Approximation rates for neural networks with encodable weights in smoothness spaces," *Neural Networks*, vol. 134, pp. 107–130, 2021.

[6] J. Schmidt-Hieber, "Deep ReLU network approximation of functions on a manifold," *arXiv:1908.00695*, 2019.

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

[8] D. Yarotsky and A. Zhevnerchuk, "The phase diagram of approximation rates for deep neural networks," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 13 005–13 015.

[9] Z. Shen, H. Yang, and S. Zhang, "Optimal approximation rate of ReLU networks in terms of width and depth," *Journal de Mathématiques Pures et Appliquées*, vol. 157, pp. 101–135, 2022.

[10] M. Kohler and S. Langer, "On the rate of convergence of fully connected deep neural network regression estimates," *The Annals of Statistics*, vol. 49, no. 4, pp. 2231 – 2249, 2021.

[11] P. L. Bartlett, V. Maiorov, and R. Meir, "Almost linear VC-dimension bounds for piecewise polynomial networks," *Neural Computation*, vol. 10, no. 8, pp. 2159–2173, 1998.

[12] P. L. Bartlett, N. Harvey, C. Liaw, and A. Mehrabian, "Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks," *Journal of Machine Learning Research*, vol. 20, no. 63, pp. 1–17, 2019.

[13] G. Vardi, G. Yehudai, and O. Shamir, "Width is less important than depth in ReLU neural networks," in *Proceedings of Thirty Fifth Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, vol. 178. PMLR, 02–05 Jul 2022, pp. 1249–1281.

[14] D. Elbrächter, D. Perekrestenko, P. Grohs, and H. Bölcskei, "Deep neural network approximation theory," *IEEE Transactions on Information Theory*, vol. 67, no. 5, pp. 2581–2623, May 2021.

[15] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*, 2nd ed. Cambridge, UK: Cambridge University Press, 2019, vol. 48.

[16] D. Yarotsky, "Error bounds for approximations with deep ReLU networks," *Neural Networks*, vol. 94, pp. 103 – 114, 2017.

[17] ——, "Optimal approximation of continuous functions by very deep ReLU networks," in *Proceedings of the 31st Conference On Learning Theory*, ser. Proceedings of Machine Learning Research, vol. 75, 06–09 Jul 2018, pp. 639–649.

[18] M. Telgarsky, "Benefits of depth in neural networks," in *29th Annual Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, vol. 49, 23–26 Jun 2016, pp. 1517–1539.

[19] I. Daubechies, R. DeVore, S. Foucart, B. Hanin, and G. Petrova, "Nonlinear approximation and (deep) ReLU networks," *Constructive Approximation*, vol. 55, no. 1, pp. 127–172, 2022.

[20] Z. Shen, H. Yang, and S. Zhang, "Deep network approximation characterized by number of neurons," *Communications in Computational Physics*, vol. 28, no. 5, pp. 1768–1811, 2020.

[21] ——, "Nonlinear approximation via compositions," *Neural Networks*, vol. 119, pp. 74–84, 2019.

[22] H. T. Siegelmann and E. D. Sontag, "On the computational power of neural nets," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 1992, pp. 440–449.

[23] J. Lu, Z. Shen, H. Yang, and S. Zhang, "Deep network approximation for smooth functions," *SIAM Journal on Mathematical Analysis*, vol. 53, no. 5, pp. 5465–5506, 2021.