Floating-Point Neural Networks Are Provably Robust Universal Approximators

Geonho Hwang[1⋆], Wonyeol Lee[2⋆], Yeachan Park[3], Sejun Park[4(✉)], and Feras Saad[5]

[1] GIST, Gwangju, Republic of Korea
hgh2134@gist.ac.kr
[2] POSTECH, Pohang, Republic of Korea
wonyeol.lee@postech.ac.kr
[3] Sejong University, Seoul, Republic of Korea
ychpark@sejong.ac.kr
[4] Korea University, Seoul, Republic of Korea
sejun.park000@gmail.com
[5] Carnegie Mellon University, Pittsburgh, PA, USA
fsaad@cmu.edu

Abstract. The classical universal approximation (UA) theorem for neural networks establishes mild conditions under which a feedforward neural network can approximate a continuous function f with arbitrary accuracy. A recent result shows that neural networks also enjoy a more general *interval* universal approximation (IUA) theorem, in the sense that the abstract interpretation semantics of the network using the interval domain can approximate the direct image map of f (i.e., the result of applying f to a set of inputs) with arbitrary accuracy. These theorems, however, rest on the unrealistic assumption that the neural network computes over infinitely precise real numbers, whereas their software implementations in practice compute over finite-precision floating-point numbers. An open question is whether the IUA theorem still holds in the floating-point setting.

This paper introduces the first IUA theorem for *floating-point* neural networks that proves their remarkable ability to *perfectly capture* the direct image map of any rounded target function f, showing no limits exist on their expressiveness. Our IUA theorem in the floating-point setting exhibits material differences from the real-valued setting, which reflects the fundamental distinctions between these two computational models. This theorem also implies surprising corollaries, which include (i) the existence of *provably robust* floating-point neural networks; and (ii) the *computational completeness* of the class of straight-line programs that use only floating-point additions and multiplications for the class of all floating-point programs that halt.

Keywords: Neural networks · Robust machine learning · Floating point · Universal approximation · Interval analysis · Abstract interpretation.

⋆Equal contribution. (✉)Corresponding author.

This manuscript is the full version of a conference paper appearing in CAV 2025.

1 Introduction

Background. Despite the remarkable success of neural networks on diverse tasks, these models often lack *robustness* and are subject to adversarial attacks. Slight perturbations to the network inputs can cause the network to produce significantly different outputs [23, 63], raising serious concerns in safety-critical domains such as healthcare [18], cybersecurity [57], and autonomous driving [16].

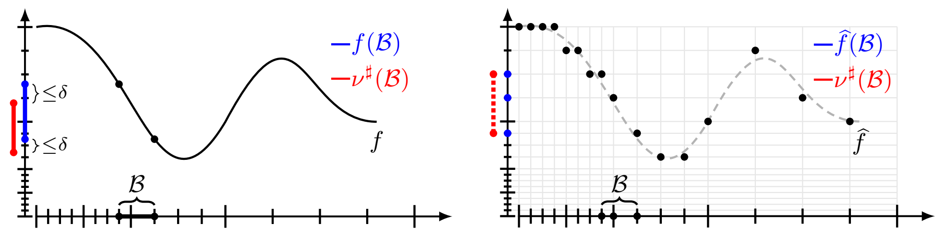
These issues have brought about significant advances in new algorithms for *robustness verification* [2, 34, 42], which prove the robustness of a given network; and *robust training* [24, 48, 56, 68], which train a network to be provably robust. But despite these advances, provably robust networks do not yet achieve state-of-the-art accuracy [38]. For example, on the CIFAR-10 image classification benchmark, non-robust networks achieve over 99% accuracy, whereas the best provably robust networks achieve less than 63% [37]. This performance gap has prompted researchers to explore whether there exists fundamental limits on the *expressiveness* of provably robust networks that restrict their accuracy [3].

Surprisingly, it has been proven that no such fundamental limit exists. Informally, for any continuous function $f : \mathbb{R}^d \to \mathbb{R}$ and compact set $\mathcal{K} \subset \mathbb{R}^d$, there exists a neural network $g : \mathbb{R}^d \to \mathbb{R}$ whose robustness properties are "sufficiently close" to those of f over $\mathcal{K}$ and easily provable using abstract interpretation [10] over the interval domain. This result, known as the *interval universal approximation* (IUA) theorem [4, 66], generalizes the classical universal approximation (UA) theorem [12, 27] from pointwise-values to intervals, and confirms that provably robust networks do not suffer from a fundamental loss of expressive power.

Key challenges. The IUA theorem in [4, 66] overlooks a critical aspect of real-world computation, which is the use of *floating-point arithmetic* instead of real arithmetic. It assumes that neural networks and interval analyses operate on arbitrary real numbers with exact operations. In reality, numerical implementations of neural networks use floating-point numbers and operations [22, §4.1], sometimes with extremely low-precision to speed-up performance [14, 29]. This discrepancy means that the existing IUA theorem does not directly apply to neural networks that are implemented in software and actually used in practice.

To our knowledge, no prior work has studied the robustness and expressiveness properties of floating-point neural networks or established an IUA theorem for them. The unique complexities of floating-point arithmetic introduce daunting challenges to any such theoretical study. For example, floating-point numbers are discretized and bounded, and their operations have rounding errors that become infinite in cases of overflow. Whereas the IUA proof over reals requires very large real numbers for network weights or intermediate computations, these values cannot be represented as floats. Naively rounding reals to floats causes approximation errors that invalidate many steps of the IUA proofs in [4, 66].

This work. We formally study the IUA theorem over floating point, as a step toward bridging the theory and practice of provably robust neural networks.



(a) $f : \mathbb{R}^d \to \mathbb{R}$ is a continuous target function; $\nu : \mathbb{R}^d \to \mathbb{R}$ is a neural network.

(b) $\widehat{f} : \overline{\mathbb{F}}^d \to \overline{\mathbb{F}}$ is a rounded target function; $\nu : \overline{\mathbb{F}}^d \to \overline{\mathbb{F}}$ is a neural network.

Fig. 1: Illustration and comparison of the IUA theorems. (a) In the real-valued setting, the neural network abstract interpretation $\nu^\sharp$ forms a δ-approximation to the image map of f. (b) In the floating-point setting, $\nu^\sharp$ exactly computes the upper and lower points of the image map of f: $\nu^\sharp(\mathcal{B}) = [\min \widehat{f}(\mathcal{B}), \max \widehat{f}(\mathcal{B})] \cap \overline{\mathbb{F}}$.

We first formulate a floating-point analog of the IUA theorem, considering the details of floating point. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a target function to approximate. Since all floating-point neural networks are functions between floating-point values, they can at-best approximate the rounded version $\widehat{f} : \overline{\mathbb{F}}^d \to \overline{\mathbb{F}}$ of f over floats, where $\overline{\mathbb{F}}$ denotes the set of all floats. The floating-point version of the IUA theorem asks the following: is there a floating-point neural network $\nu : \overline{\mathbb{F}}^d \to \overline{\mathbb{F}}$ whose *interval semantics* is arbitrarily close to the *direct image map* of the rounded target $\widehat{f}$ over $[-1,1]^d$? More formally, this property means that for any $\delta > 0$, there exists a neural network ν such that for all boxes $\mathcal{B} \subseteq [-1,1]^d \cap \overline{\mathbb{F}}^d$,

$$\left|\min \nu^\sharp(\mathcal{B}) - \min \widehat{f}(\mathcal{B})\right| \leq \delta, \qquad \left|\max \nu^\sharp(\mathcal{B}) - \max \widehat{f}(\mathcal{B})\right| \leq \delta. \tag{1}$$

In Eq. (1), $\nu^\sharp(\mathcal{B})$ is the result of abstract interpretation of $\mathcal{B}$ under ν (using the interval domain), and $\widehat{f}(\mathcal{B}) \coloneqq \{\widehat{f}(\mathbf{x}) \mid \mathbf{x} \in \mathcal{B}\} \subset \mathbb{R}$ is the image of $\mathcal{B}$ under $\widehat{f}$.

We prove that the IUA theorem holds for floating-point networks, despite all their numerical complexities. In particular, we show that for *any* target function f and a *large* class of activation functions σ, including most practical ones (e.g., ReLU, GELU, sigmoid), it is possible to find a floating-point network ν with σ whose interval semantics *exactly* matches the direct image map of the rounded target $\widehat{f}$ over $[-1,1]^d \cap \overline{\mathbb{F}}^d$ (Figure 1). This result implies that no fundamental limit exists on the expressiveness of provably robust floating-point neural networks.

Our result is considerably different from the previous IUA theorem over the reals in three key aspects. The previous theorem considers continuous target functions; requires a restricted class of so-called squashable activation functions; and finds networks that are arbitrarily close to target functions. In contrast, our result considers arbitrary target functions; allows almost all activation functions used in practice; and find networks that are precisely equal to (rounded) target functions. Our IUA theorem even holds for the *identity* activation function, which is not the case for the traditional IUA or UA theorems over real numbers, because any network that uses the identity activation is affine over the reals.

As a corollary of our main theorem, we prove the following existence of provably robust floating-point neural networks: given an ideal floating-point classifier $\widehat{f}$ (not necessarily a neural network) that is robust (not necessarily provably robust), we can find a floating-point neural network ν that is *identical* to $\widehat{f}$ and is *provably* robust with interval analysis. We also prove a nontrivial result about "floating-point completeness", as an unexpected byproduct of the main theorem. Specifically, we show that the class of straight-line floating-point programs that use only floating-point $+$ and $\times$ operations is *floating-point interval-complete*: it can simulate *any* terminating floating-point program that takes finite floats as input and returns arbitrary floats as output. The same statement holds under the interval semantics. To our knowledge, no prior work has identified such a small yet powerful class of floating-point programs, suggesting that this corollary is of significant independent interest to the extensive floating-point literature.

Contributions. This article makes the following contributions:

- We formalize a *floating-point* analog of the *interval universal approximation* (IUA) theorem, to bridge the theory and practice of *provably robust* neural networks (§2, §3). It asks if there is a floating-point network whose interval semantics is close to the direct image map of a given target function.
- We prove the floating-point version of the IUA theorem does hold, for *all* target functions and a *broad* class of activation functions that includes most of the activations used in practice (§3.1, §3.2, §5). This shows no fundamental limit exists on the expressiveness of provably robust networks over floats.
- We rigorously analyze the essential differences between the previous IUA theorem over reals and our IUA theorem over floats (§3.3). Unlike real-valued networks, floating-point networks can *perfectly* capture the behavior of *any* rounded target function, even with the *identity* activation function.
- We prove that if there exists an ideal robust floating-point classifier, then one can always find a *provably* robust floating-point network that makes *exactly* the same prediction as the classifier (§4.1).
- We prove that the set of straight-line floating-point programs with only $(+, \times)$ is *floating-point interval-complete*: it can simulate *any* terminating floating-point programs that take finite inputs and return finite/infinite outputs, under the usual floating-point semantics and interval semantics (§4.2).

2 Preliminaries

This section introduces floating-point arithmetic (§2.1), neural networks that compute over floating-point numbers (§2.2), and interval analysis for neural networks (§2.3). Throughout the paper, we define $\mathbb{N}$ to be the set of positive integers and let $[n] \coloneqq \{1, \ldots, n\}$ for each $n \in \mathbb{N}$.

2.1 Floating Point

Floating-point numbers. Let $E, M \in \mathbb{N}$. The set of *finite* floating-point numbers with E-bit exponent and $(M+1)$-bit significand is typically defined by

$$\mathbb{F}_M^E \coloneqq \left\{(-1)^b \times (s_0.s_1\dots s_M)_2 \times 2^e \,\middle|\, b, s_i \in \{0,1\}, e \in \{\mathfrak{e}_{\min}, \dots, \mathfrak{e}_{\max}\}\right\}, \quad (2)$$

where $\mathfrak{e}_{\min} \coloneqq -2^{E-1}+2$ and $\mathfrak{e}_{\max} \coloneqq 2^{E-1}-1$ [52]. The set of *all* floating-point numbers, including non-finite ones, is then defined by $\overline{\mathbb{F}}_M^E \coloneqq \mathbb{F}_M^E \cup \{-\infty, +\infty, \perp\}$, where $\perp$ denotes NaN (i.e., not-a-number). For brevity, we call a floating-point number simply a *float*, and write $\mathbb{F}_M^E$ and $\overline{\mathbb{F}}_M^E$ simply as $\mathbb{F}$ and $\overline{\mathbb{F}}$. In this paper, we assume $E \geq 5$ and $2^{E-1} \geq M \geq 3$, which hold for nearly all practical floating-point formats, including bfloat16 [1] and all the formats defined in the IEEE-754 standard [31] such as float16, float32, and float64.

We introduce several notations and terms related to finite floats. First, we define three key constants: the *smallest* positive float $\omega \coloneqq 2^{\mathfrak{e}_{\min}-M}$, the *largest* positive float $\Omega \coloneqq 2^{\mathfrak{e}_{\max}}(2-2^{-M})$, and the *machine epsilon* $\varepsilon \coloneqq 2^{-M-1}$. Next, consider a finite float $x \in \mathbb{F}$. We call x a *subnormal* number if $0 < |x| < 2^{\mathfrak{e}_{\min}}$, and a *normal* number otherwise. The *exponent* and *significand* of x are defined by $\mathfrak{e}_x \coloneqq \max\{\lfloor \log_2|x| \rfloor, \mathfrak{e}_{\min}\} \in [\mathfrak{e}_{\min}, \mathfrak{e}_{\max}]$ and $\mathfrak{s}_x \coloneqq |x|/2^{\mathfrak{e}_x} \in [0,2)$. We use $\mathfrak{s}_{x,0}, \dots, \mathfrak{s}_{x,M}$ to denote the binary expansion of $\mathfrak{s}_x$, i.e., $(\mathfrak{s}_{x,0}.\,\mathfrak{s}_{x,1}\dots\mathfrak{s}_{x,M})_2 = \mathfrak{s}_x$ with $\mathfrak{s}_{x,i} \in \{0,1\}$. The *predecessor* and *successor* of x in $\overline{\mathbb{F}}$ are written as $x^- \coloneqq \max\{y \in \overline{\mathbb{F}} \setminus \{\perp\} \mid x > y\}$ and $x^+ \coloneqq \min\{y \in \overline{\mathbb{F}} \setminus \{\perp\} \mid x < y\}$.

Floating-point operations. We define the *rounding function* $\mathrm{rnd} : \mathbb{R} \cup \{-\infty, +\infty\} \to \overline{\mathbb{F}}$ as follows: $\mathrm{rnd}(x) \coloneqq -\infty$ if $x \in [-\infty, -\Omega - c]$, $\mathrm{rnd}(x) \coloneqq \arg\min_{y \in \mathbb{F}} |y - x|$ if $x \in (-\Omega - c, \Omega + c)$, and $\mathrm{rnd}(x) \coloneqq +\infty$ if $x \in [\Omega + c, +\infty]$, where $c \coloneqq 2^{\mathfrak{e}_{\max}}\varepsilon$ and $\arg\min$ breaks ties by choosing a float y with $\mathfrak{s}_{y,M} = 0$. This function corresponds to the rounding mode "round to nearest (ties to even)", which is the default rounding mode in the IEEE-754 standard [31].

The floating-point *arithmetic operations* $\oplus, \ominus, \otimes : \overline{\mathbb{F}} \times \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ are defined via the rounding function: for finite floats $x, y \in \mathbb{F}$, $x \oplus y \coloneqq \mathrm{rnd}(x+y)$, $x \ominus y \coloneqq \mathrm{rnd}(x-y)$, and $x \otimes y \coloneqq \mathrm{rnd}(x \times y)$. We omit the definition for non-finite operands because they are unimportant in this paper, except that $x \oplus 0 = x \ominus 0 = x$ for all $x \in \{-\infty, +\infty\}$. For the full definition, refer to the IEEE-754 standard [31].

We introduce two more floating-point operations: $\mathrm{aff}_{W,\mathbf{b}}$ and $\mathrm{rnd}(f)$. First, we define the floating-point *affine transformation*: for a matrix $W = (w_{i,j})_{i \in [m], j \in [n]} \in \mathbb{F}^{m \times n}$ and a vector $\mathbf{b} = (b_1, \dots, b_m) \in \mathbb{F}^m$, $\mathrm{aff}_{W,\mathbf{b}} : \overline{\mathbb{F}}^n \to \overline{\mathbb{F}}^m$ is defined by

$$\mathrm{aff}_{W,\mathbf{b}}(x_1, \dots, x_n) \coloneqq \left(\left(\bigoplus_{j=1}^{n} x_j \otimes w_{1,j}\right) \oplus b_1, \;\dots,\; \left(\bigoplus_{j=1}^{n} x_j \otimes w_{m,j}\right) \oplus b_m\right). \quad (3)$$

Here, $\bigoplus$ denotes the floating-point summation defined in the left-associative way: $\bigoplus_{i=1}^{n} y_i \coloneqq (\cdots((y_1 \oplus y_2) \oplus y_3)\cdots) \oplus y_n$, where the order of $\oplus$ is important because $\oplus$ is not associative. Next, we define the *correctly rounded version* of a

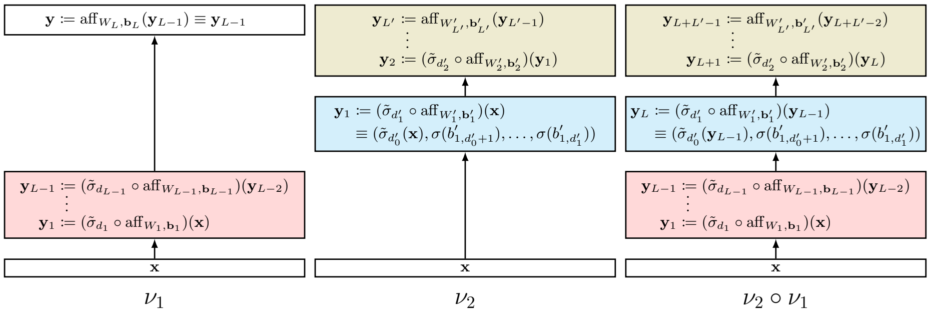


Fig. 2: Illustrations of a network ν_1 without the last affine layer (left), a network ν_2 without the first affine layer (middle), and their composition $\nu_2 \circ \nu_1$ (right). Note that $\mathrm{aff}_{W'_1,\mathbf{b}'_1} \circ \mathrm{aff}_{W_L,\mathbf{b}_L} = \mathrm{aff}_{W'_1,\mathbf{b}'_1}$ is a floating-point affine transformation.

real-valued function. For $f : \mathbb{R} \to \mathbb{R}$, the function $\mathrm{rnd}(f) : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ is defined by

$$\mathrm{rnd}(f)(x) \coloneqq \begin{cases} \mathrm{rnd}(f(x)) & \text{if } x \in (-\infty, +\infty) \\ \mathrm{rnd}\left(\lim_{t \to x} f(t)\right) & \text{if } x \in \{-\infty, +\infty\} \wedge \lim_{t \to x} f(t) \in \mathbb{R} \cup \{-\infty, +\infty\} \\ \bot & \text{otherwise.} \end{cases} \tag{4}$$

2.2 Neural Networks

A neural network typically refers to a composition of affine transformations and activation functions. Formally, for $L \in \mathbb{N}$ and $\sigma : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$, we call a function ν a *depth-L σ-neural network* (or a *neural network*) if ν is defined by

$$\nu : \overline{\mathbb{F}}^{d_0} \to \overline{\mathbb{F}}^{d_L}, \quad \nu \coloneqq \mathrm{aff}_{W_L,\mathbf{b}_L} \circ \tilde{\sigma}_{d_{L-1}} \circ \mathrm{aff}_{W_{L-1},\mathbf{b}_{L-1}} \circ \cdots \circ \tilde{\sigma}_{d_1} \circ \mathrm{aff}_{W_1,\mathbf{b}_1} \tag{5}$$

for some $d_\ell \in \mathbb{N}$, $W_\ell \in \mathbb{F}^{d_\ell \times d_{\ell-1}}$, and $\mathbf{b}_\ell \in \mathbb{F}^{d_\ell}$, where $\tilde{\sigma}_n : \overline{\mathbb{F}}^n \to \overline{\mathbb{F}}^n$ is the coordinatewise application of σ. Here, L denotes the number of layers, σ the floating-point activation function, d_0 and d_L the input and output dimensions, d_ℓ the number of hidden neurons in the ℓ-th layer ($\ell \in [L-1]$), and W_ℓ and $\mathbf{b}_\ell$ the parameters of the floating-point affine transformation in the ℓ-th layer ($\ell \in [L]$). We emphasize that a neural network in this paper is a function over *floating-point* values, defined in terms of *floating-point* activation function and arithmetic. For instance, a depth-1 neural network is a floating-point affine transformation.

Let ν be a neural network defined as Eq. (5). We say ν is *without the last affine layer* if $d_L = d_{L-1}$, W_L is the identity matrix, and $\mathbf{b}_L = \mathbf{0}$. Similarly, we say ν is *without the first affine layer* if $d_1 \geq d_0$, W_1 is a rectangular diagonal matrix whose diagonal entries are all 1, and $b_{1,i} = 0$ for all $i \in [d_0]$. The two definitions are not perfectly symmetric due to some technical details arising in our proofs. We note that a neural network can be constructed by composing networks without the first/last affine layer(s) and arbitrary networks (Figure 2). For example, consider arbitrary networks $\nu_1 : \overline{\mathbb{F}}^{n_0} \to \overline{\mathbb{F}}^{n_1}$ and $\nu_4 : \overline{\mathbb{F}}^{n_3} \to \overline{\mathbb{F}}^{n_4}$, a network without the first affine layer $\nu_2 : \overline{\mathbb{F}}^{n_1} \to \overline{\mathbb{F}}^{n_2}$, and a network without the first and last affine layers $\nu_3 : \overline{\mathbb{F}}^{n_2} \to \overline{\mathbb{F}}^{n_3}$. It is easily verified that the function $\nu : \overline{\mathbb{F}}^{n_0} \to \overline{\mathbb{F}}^{n_4}$ specified by $\nu(\mathbf{x}) = (\nu_4 \circ \cdots \circ \nu_1)(\mathbf{x})$ denotes a network, whose definition in the form of Eq. (5) can be obtained by "merging" the last layer of ν_1 and the first layer of ν_2, etc.

2.3 Interval Semantics

Interval analysis [11, 50] is a technique for analyzing the behavior of numerical programs soundly and efficiently, based on abstract interpretation [10]. It uses intervals to overapproximate the ranges of inputs and expressions, and propagates them through a program to overapproximate the output range. Interval analysis has been used to establish the robustness of practical neural networks [19, 24, 33, 43]. It can overapproximate the output range of a network over perturbed inputs, which is required to prove robustness; and it runs efficiently by performing only simple computations, which is required to analyze large-scale networks.

Interval domain and operations. We formalize interval analysis for neural networks as follows. We first define the *interval domain*

$$\mathbb{I} := \left\{ \langle a, b \rangle \,\middle|\, a, b \in \overline{\mathbb{F}} \setminus \{\bot\} \text{ with } a \le b \right\} \cup \{\top\}, \tag{6}$$

on which interval analysis operates. Here, $\langle a, b \rangle$ abstracts the floating-point interval $[a, b] \cap \overline{\mathbb{F}}$, and $\top$ abstracts the entire floating-point set $\overline{\mathbb{F}}$ including $\bot$. The concrete semantics of an *abstract interval* $\mathcal{I} \in \mathbb{I}$ and an *abstract box* $\mathcal{B} = (\mathcal{I}_1, \ldots, \mathcal{I}_d) \in \mathbb{I}^d$ are defined through the *concretization function* γ, where

$$\gamma : \cup_{d=1}^{\infty} \mathbb{I}^d \to \cup_{d=1}^{\infty} 2^{\overline{\mathbb{F}}^d}, \quad \gamma(\mathcal{I}) := \begin{cases} [a, b] \cap \overline{\mathbb{F}} & \text{if } \mathcal{I} = \langle a, b \rangle \\ \overline{\mathbb{F}} & \text{if } \mathcal{I} = \top \end{cases}, \quad \gamma(\mathcal{B}) := \prod_{i=1}^{d} \gamma(\mathcal{I}_i). \tag{7}$$

We say that an abstract box $\mathcal{B} \in \mathbb{I}^d$ *is in* a set $\mathcal{S} \subseteq \mathbb{R}^d$ if $\gamma(\mathcal{B}) \subseteq \mathcal{S}$.

For any function $\phi : \overline{\mathbb{F}}^d \to \overline{\mathbb{F}}$ over floats (which is not a neural network or a floating-point affine transformation), the *interval operation* $\phi^\sharp : \mathbb{I}^d \to \mathbb{I}$ extends ϕ to the interval domain as follows:

$$\phi^\sharp(\mathcal{B}) := \begin{cases} \langle \min \mathcal{S}, \max \mathcal{S} \rangle & \text{if } \bot \notin \mathcal{S} \\ \top & \text{if } \bot \in \mathcal{S} \end{cases}, \quad \text{where } \mathcal{S} := \phi(\gamma(\mathcal{B})). \tag{8}$$

In the special case that $\phi = \odot \in \{\oplus, \ominus, \otimes\}$ is a floating-point arithmetic operation, the above definition (using infix notation) is equivalent to the following:

$$\langle a, b \rangle \odot^\sharp \langle c, d \rangle := \begin{cases} \langle \min \mathcal{S}, \max \mathcal{S} \rangle & \text{if } \bot \notin \mathcal{S} \\ \top & \text{if } \bot \in \mathcal{S} \end{cases}, \quad \text{where } \mathcal{S} := \begin{Bmatrix} a \odot c, a \odot d, \\ b \odot c, b \odot d \end{Bmatrix}, \tag{9}$$

and $\odot^\sharp$ returns $\top$ if at least one of its operands is $\top$.[6] We remark that $\odot^\sharp$ can be efficiently computed, and so can $\phi^\sharp$ when $\phi : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ is piecewise-monotone with finitely many pieces, which holds for the correctly rounded versions of widely-used activation functions (e.g., ReLU, GELU, sigmoid). We then define the *interval affine transformation* $\mathrm{aff}^\sharp_{W,\mathbf{b}} : \mathbb{I}^n \to \mathbb{I}^m$, which extends its floating-point counterpart $\mathrm{aff}_{W,\mathbf{b}} : \overline{\mathbb{F}}^n \to \overline{\mathbb{F}}^m$: $\mathrm{aff}^\sharp_{W,\mathbf{b}}(\mathcal{I}_1, \dots, \mathcal{I}_n) \coloneqq \big((\bigoplus\nolimits_{j=1}^{n\,\sharp} \mathcal{I}_j \otimes^\sharp \langle w_{i,j}, w_{i,j}\rangle) \oplus^\sharp \langle b_i, b_i \rangle \big)_{i=1}^m$, where $\bigoplus^\sharp$ is the interval summation which uses $\oplus^\sharp$ instead of $\oplus$.

Interval semantics. The *interval semantics* $\nu^\sharp : \mathbb{I}^{d_0} \to \mathbb{I}^{d_L}$ of a neural network $\nu : \overline{\mathbb{F}}^{d_0} \to \overline{\mathbb{F}}^{d_L}$ is defined as the result of interval analysis on ν:

$$\nu^\sharp \coloneqq \mathrm{aff}^\sharp_{W_L,\mathbf{b}_L} \circ \tilde{\sigma}^\sharp_{d_{L-1}} \circ \mathrm{aff}^\sharp_{W_{L-1},\mathbf{b}_{L-1}} \circ \cdots \circ \tilde{\sigma}^\sharp_{d_1} \circ \mathrm{aff}^\sharp_{W_1,\mathbf{b}_1}, \tag{10}$$

where ν is assumed to be defined as Eq. (5) and $\tilde{\sigma}^\sharp_n : \mathbb{I}^n \to \mathbb{I}^n$ is the coordinatewise application of $\sigma^\sharp : \mathbb{I} \to \mathbb{I}$. It is easily verified that the interval semantics is sound with respect to the floating-point semantics:

$$\nu\big(\gamma(\mathcal{B})\big) \subseteq \gamma\left(\nu^\sharp(\mathcal{B})\right) \qquad\qquad (\mathcal{B} \in \mathbb{I}^{d_0}). \tag{11}$$

That is, the result of interval analysis $\nu^\sharp(\mathcal{B}) \in \mathbb{I}^{d_L}$ subsumes the set of all possible outputs of the network ν when the input is in the concrete box $\gamma(\mathcal{B}) \subseteq \overline{\mathbb{F}}^{d_0}$.

3 Interval Universal Approximation Over Floats

This section presents our main result on interval universal approximation (IUA) for floating-point neural networks. We first introduce conditions on activation functions for our result (§3.1), and then formally describe our result under these conditions (§3.2). We then compare our IUA theorem over floats with existing IUA theorems over reals, highlighting several nontrivial differences (§3.3).

3.1 Conditions on Activation Functions

Our IUA theorem is for floating-point neural networks that use activation functions satisfying the following conditions (Figure 3).

Condition 1. *An activation function $\sigma : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ satisfies the following conditions:*

(C1) *There exist $c_1, c_2 \in \mathbb{F}$ such that $\sigma(c_1) = 0$, $|\sigma(c_2)| \in [\frac{\varepsilon}{2} + 2\varepsilon^2, \frac{5}{4} - 2\varepsilon]$, $\max\{|c_1|, |c_2|\} \geq 2^{\mathfrak{e}_{\min}+1}$, and $\sigma(x)$ lies between $\sigma(c_1)$ and $\sigma(c_2)$ for all x between c_1 and c_2, where ε is the machine epsilon (see §2.1).*

[6] This definition of $\odot^\sharp$ differs slightly from the standard definition, as $\odot^\sharp$ uses "round to nearest" mode (implicit in $\odot$), whereas the more common mode is "round downward/upward" (e.g., $\langle a, b\rangle \oplus^\sharp \langle c, d\rangle \coloneqq \langle a \oplus_\downarrow c, b \oplus_\uparrow d\rangle$) [26, Section 5]. This choice is due to different goals to achieve: our definition overapproximates floating-point operations (e.g., $\oplus$), while the usual one overapproximates exact operations (e.g., $+$).

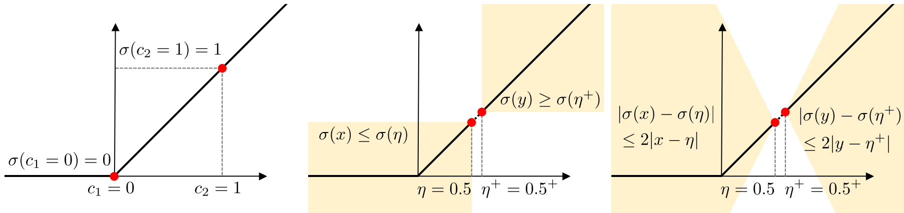


Fig. 3: Illustration of the first (left), second (middle), and third (right) conditions in Condition 1 for the ReLU activation function: $\sigma(x) \coloneqq \max\{x, 0\}$ for $x \in \overline{\mathbb{F}}$.

(C2) *There exists* $\eta \in \mathbb{F}$ *with* $|\eta| \in [2^{\mathfrak{e}_{\min}+5}, 4-8\varepsilon]$ *and* $|\sigma(\eta)|, |\sigma(\eta^+)| \in [2^{\mathfrak{e}_{\min}+5}, 2^{\mathfrak{e}_{\max}-6} \cdot |\eta|]$ *such that for any* $x, y \in \mathbb{F}$ *with* $x \leq \eta < \eta^+ \leq y$,

$$\sigma(x) \leq \sigma(\eta) < \sigma(\eta^+) \leq \sigma(y) \quad \textit{or} \quad \sigma(x) \geq \sigma(\eta) > \sigma(\eta^+) \geq \sigma(y). \tag{12}$$

(C3) *There exists* $\lambda \in [0, 2^{\mathfrak{e}_{\max}-7} \cdot \min\{|\sigma(\eta)|, 2^{M+3}\}]$ *such that for any* $x, y \in \mathbb{F}$ *with* $x \leq \eta < \eta^+ \leq y$,

$$|\sigma(x) - \sigma(\eta)| \leq \lambda|x - \eta| \quad \textit{and} \quad |\sigma(y) - \sigma(\eta^+)| \leq \lambda|y - \eta^+|. \tag{13}$$

The condition (C1) states that the activation function σ can output the exact zero (i.e., $\sigma(c_1)$) and some value whose magnitude is approximately in $[\frac{\varepsilon}{2}, \frac{5}{4}]$ (i.e., $\sigma(c_2)$); and its output is within $\sigma(c_1)$ and $\sigma(c_2)$ for all inputs between c_1 and c_2. The condition (C2) states that there exists some *threshold* η such that $\sigma(x)$ is either smaller or greater than $\sigma(\eta)$ or $\sigma(\eta^+)$, depending on whether x is on the left or right side of η. This condition holds automatically for all monotone activation functions that are non-constant on either $[2^{\mathfrak{e}_{\min}+5}, 4-8\varepsilon] \cap \mathbb{F}$ or $[-4+8\varepsilon, -2^{\mathfrak{e}_{\min}+5}] \cap \mathbb{F}$. The condition (C3) states that σ does not increase or decrease too rapidly from η and η^+, which implies that $\sigma(x)$ is finite for all finite floats $x \in \mathbb{F}$.

While Condition 1 is mild, verifying whether practical activation functions over floats satisfy Condition 1 can be cumbersome. Floating-point activation functions are typically implemented in complicated ways [7, 44, 51] (e.g., by intermixing floating-point operations with integer/bit-level operations and if-else branches), which makes it challenging to rigorously analyze such implementations [17, 36]. To bypass this issue, we focus on the correctly rounded version $\sigma : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ of a real-valued activation function $\rho : \mathbb{R} \to \mathbb{R}$ (i.e., $\sigma(x) \coloneqq \mathrm{rnd}(\rho(x))$), when verifying Condition 1. Correctly rounded versions of elementary mathematical functions have been actively developed in several software libraries [13, 39, 40, 58, 72].

Under the correct rounding assumption, we provide an easily verifiable sufficient condition for activation functions on reals that can be used to verify Condition 1 for their rounded versions. The proof of Lemma 1 is in §B.1.

Lemma 1. *For any activation function* $\rho : \mathbb{R} \to \mathbb{R}$*, the correctly rounded activation* $\mathrm{rnd}(\rho) : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ *satisfies Condition 1 if the following conditions hold:*

(C1′) *There exist* $c'_1, c'_2 \in \mathbb{F}$ *such that* $|\rho(c'_1)| \le \frac{\omega}{2}$, $|\rho(c'_2)| \in [\frac{\varepsilon}{2} + 2\varepsilon^2, \frac{5}{4} - 2\varepsilon]$, $\max\{|c'_1|, |c'_2|\} \ge 2^{\mathfrak{e}_{\min}+1}$, *and* $\rho(x)$ *lies between* $\rho(c'_1)$ *and* $\rho(c'_2)$ *for all* x *between* c'_1 *and* c'_2, *where* ω *is the smallest positive float (see §2.1).*

(C2′) *There exists* $\delta \in \mathbb{R}$ *with* $|\delta| \in [\frac{3}{8}, \frac{7}{8}]$ *such that*

- *for all* $x, y \in \mathbb{R}$ *satisfying* $x \le \delta - \frac{1}{8} < \delta + \frac{1}{8} \le y$,

$$\rho(x) \le \rho(\delta - \tfrac{1}{8}) < \rho(\delta + \tfrac{1}{8}) \le \rho(y) \;\; \text{or} \;\; \rho(x) \ge \rho(\delta - \tfrac{1}{8}) > \rho(\delta + \tfrac{1}{8}) \ge \rho(y),$$

- $|\rho(x)| \in [\frac{1}{4}, 1]$ *and* $|\rho(x) - \rho(y)| > \frac{1}{8}|x - y|$ *for all* $x, y \in [\delta - \frac{1}{8}, \delta + \frac{1}{8}]$.

(C3′) ρ *is* λ*-Lipschitz continuous for some* $\lambda \in [0, \frac{1}{5} \cdot 2^{\mathfrak{e}_{\max}-9}]$.

The conditions (C1′)–(C3′) in Lemma 1 correspond to the conditions (C1)–(C3) in Condition 1. The condition (C1′), corresponding to (C1), can be easily satisfied since modern activation functions are piecewise-monotone and either zero at zero (e.g., ReLU, GELU, softplus, tanh) or close to zero at $-\Omega$ or Ω (e.g., sigmoid). The condition (C2′) roughly states the existence of $\delta \in \mathbb{R}$ satisfying the following: (i) $\rho(\delta - \frac{1}{8})$ and $\rho(\delta + \frac{1}{8})$ are lower/upper bounds of ρ on $(-\infty, \delta - \frac{1}{8})$ and $(\delta + \frac{1}{8}, \infty)$; and (ii) ρ is bounded and strictly monotone on $[\delta - \frac{1}{8}, \delta + \frac{1}{8}]$. This condition guarantees the existence of $\eta \in \mathbb{F}$ in (C2). The condition (C3′), corresponding to (C3), can also be easily satisfied since $\lambda < 3$ for most practical activation functions. We note that Lemma 1 gives sufficient but not necessary conditions for a correctly rounded activation function to satisfy Condition 1.

The following corollary uses Lemma 1 to show that many prominent activation functions satisfy Condition 1. Its proof is in §B.2.

Corollary 1. *The correctly rounded implementations of the* ReLU*,* LeakyReLU*,* GELU*,* ELU*,* Mish*,* softplus*,* sigmoid*, and* tanh *activations satisfy Condition 1.*

3.2 Main Result

We are now ready to present our IUA theorem over floating-point arithmetic.

Theorem 1. *Let* $\sigma : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ *be an activation function satisfying Condition 1.*[7] *Then, for any target function* $f : \mathbb{R}^d \to \mathbb{R}$, *there exists a* σ*-neural network* $\nu : \overline{\mathbb{F}}^d \to \overline{\mathbb{F}}$ *such that*

$$\gamma\big(\nu^\sharp(\mathcal{B})\big) = \Big[\min \widehat{f}\big(\gamma(\mathcal{B})\big), \max \widehat{f}\big(\gamma(\mathcal{B})\big)\Big] \cap \overline{\mathbb{F}} \tag{14}$$

for $\widehat{f} = \mathrm{rnd}(f) : \overline{\mathbb{F}}^d \to \overline{\mathbb{F}}^d$ *and for all abstract boxes* $\mathcal{B}$ *in* $[-1, 1]^d$.[8]

[7] Condition 1 is sufficient for Theorem 1 but not necessary. E.g., Theorem 1 still holds under 8-bit floats (both E4M3 and E5M2 formats [46]) for the ReLU activation function; this corresponds to the case where (C1) and (C2) hold but (C3) is violated.

[8] In the literature on universal approximation theorems, it is typically assumed that the inputs are in $[0, 1]$ or in a compact subset of $\mathbb{R}$ (e.g., [4, 12, 66, 69]). Since the inputs are often normalized to $[-1, 1]$, we focus the theoretical analysis on $[-1, 1]^d$.

Theorem 1 states that for any activation function $\sigma : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ satisfying Condition 1 and any target function $f : \mathbb{R}^d \to \mathbb{R}$, there exists a σ-network ν whose interval semantics *exactly computes* the upper and lower points of the direct image map of the rounded target $\widehat{f} : \overline{\mathbb{F}}^d \to \overline{\mathbb{F}}$ on $[-1,1]^d \cap \mathbb{F}^d$. A special case of our IUA Theorem 1 is the following universal approximation (UA) theorem for floating-point neural networks:

$$\nu(\mathbf{x}) = \widehat{f}(\mathbf{x}) \qquad\qquad (\mathbf{x} \in [-1,1]^d \cap \mathbb{F}^d). \tag{15}$$

That is, floating-point neural networks using an activation function satisfying Condition 1 can represent any function $\widehat{f} : [-1,1]^d \cap \mathbb{F}^d \to \mathbb{F} \cup \{-\infty, +\infty\}$; or the rounded version of any real function $f : [-1,1]^d \to \mathbb{R}$. Moreover, Theorem 1 easily extends to any target function $f : \mathbb{R}^d \to \mathbb{R}^{d'}$ with multiple outputs.

As previous IUA results assume exact operations over reals, they do not extend to our setting of floating-point arithmetic (due to rounding errors, overflow, NaNs, discreteness, boundedness, etc.). As a simple example of these issues, consider the following subnetwork, which is used in the IUA proof of [4]:

$$\mu(x,y) = \frac{1}{2}\Big(\mathrm{ReLU}(x+y) - \mathrm{ReLU}(-x-y) - \mathrm{ReLU}(x-y) - \mathrm{ReLU}(y-x)\Big).$$

This subnetwork returns $\min\{x,y\}$ if all operations are exact. However, it does not under floating-point arithmetic due to the rounding error: if $(+, \times)$ is replaced by $(\oplus, \otimes)$, then $\mu(x,y) = 0 \neq \varepsilon = \min\{x,y\}$ for $x = 1$ and $y = \varepsilon$.

In addition, the network construction in [66, Theorem 4.10] requires multiplying a large number z that depends on the target error and the activation function, to the output of some neuron. However, because $\mathbb{F}$ is bounded and floating-point operations are subject to overflow, the number z and the result of the multiplication are not guaranteed to be within $\mathbb{F}$ when using a small target error (e.g., less than ω) or when using common activations functions (e.g., ReLU, softplus). To bypass these issues, we carefully analyze rounding errors and design a network without infinities in the intermediate layers, when proving Theorem 1.

We present the proof outline of Theorem 1 in §5, and the full proof in §D–§F. We implemented the proof (i.e., our network construction) in Python and made it available at https://github.com/yechanp/floating-point-iua-theorem.

3.3 Comparison With Existing Results Over Reals

Theorem 1, which gives an IUA theorem over floats, has notable differences from previous IUA theorems over the reals [4, Theorem 1.1]; [66, Theorem 3.7].

One difference is the class of target functions and the desired property of networks. Previous IUA theorems find a network that *sufficiently approximates* the direct image map of a *continuous* target function (i.e., $\delta > 0$ in Eq. (1)). In contrast, our IUA theorem finds a network that *exactly computes* the direct image map of an *arbitrary* rounded target function (i.e., $\delta = 0$ in Eq. (1)). This difference arises from the domains of the functions being approximated: the real-valued setting considers functions f over $[-1,1]^d$ (or a compact $\mathcal{K} \subset \mathbb{R}^d$); the floating-point setting considers functions $\widehat{f}$ over $[-1,1]^d \cap \mathbb{F}^d$.

- Since $[-1,1]^d$ is uncountable, exactly computing the direct image map of f requires a network to fit *uncountably* many input/output pairs and related box/interval pairs. This task is difficult to achieve, and indeed, recent works [3, 5, 47] prove that it is theoretically unachievable even for simple target functions (e.g., continuous piecewise linear functions).
- Since $[-1,1]^d \cap \mathbb{F}^d$ is finite, exactly computing the direct image map of $\widehat{f}$ requires a network to fit *finitely* many input/output and box/interval pairs. Our result proves that, despite all the complexities of floating-point computation, this task can be achieved for any rounded target function.

Another key difference is the class of activation functions. There are real-valued activation functions $\rho, \rho' : \mathbb{R} \to \mathbb{R}$ such that previous IUA theorems *cannot* hold for ρ but our IUA theorem *does* hold for $\mathrm{rnd}(\rho) : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$; and vice versa for ρ'.

- An example of ρ is the *identity* function: $\rho(x) = x$. No classical IUA or UA theorem can hold for ρ, since all *real-valued* ρ-networks $\mu : \mathbb{R}^d \to \mathbb{R}$ are *affine over the reals* (i.e., there exists $A \in \mathbb{R}^{1\times d}$ and $b \in \mathbb{R}$ such that $\mu(\mathbf{x}) = A\mathbf{x} + b$ for all $\mathbf{x} \in \mathbb{R}^d$). In contrast, our IUA theorem does hold for $\mathrm{rnd}(\rho)$, because $\mathrm{rnd}(\rho)$ satisfies all the conditions in Lemma 1 (with constants $c_1' = 0$, $c_2' = 1$, $\delta = 1/2$, and $\lambda = 1$). This counterintuitive result is made possible because *floating-point* $\mathrm{rnd}(\rho)$-networks $\nu : \overline{\mathbb{F}}^d \to \overline{\mathbb{F}}$ can be *non-affine over the reals* (i.e., there may not exist $A \in \mathbb{R}^{1\times d}$ and $b \in \mathbb{R}$ such that $\nu(\mathbf{x}) = A\mathbf{x} + b$ for all $\mathbf{x} \in \mathbb{F}^d$). This non-affineness arises from rounding errors: some floating-point affine transformations $\mathrm{aff}_{W,\mathbf{b}}$ are not actually affine over the reals due to rounding errors. An interesting implication of this result is discussed in §4.2.
- An example of ρ' is any function that is non-decreasing on $\mathbb{R}$, is constant on $[-\Omega, \Omega]$, and satisfies $\lim_{x\to-\infty} \rho'(x) < \lim_{x\to+\infty} \rho'(x)$, where the two limits exist in $\mathbb{R}$. The real-valued IUA theorem holds for ρ', because ρ' satisfies the condition in [66, Definition 2.3]. However, no floating-point IUA or UA theorem can hold for $\mathrm{rnd}(\rho')$, because all $\mathrm{rnd}(\rho')$-networks $\nu : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ must be monotone if its depth is 1, and must satisfy $\nu(0) = \nu(\omega)$ otherwise. The monotonicity holds when the depth is 1 since $\oplus, \otimes$ are monotone when an operand is a constant; and $\nu(0) = \nu(\omega)$ holds otherwise since $x \otimes a \oplus b \in [-\Omega, \Omega]$ for all $x \in \{0, \omega\}$ and $a, b \in \mathbb{F}$, and $\mathrm{rnd}(\rho')$ is constant on $[-\Omega, \Omega] \cap \mathbb{F}$.

4 Implications of IUA Theorem Over Floats

This section presents two important implications of our IUA theorem, on provable robustness and "floating-point completeness". We first prove the existence of a provably robust floating-point network, given an ideal robust floating-point classifier (§4.1). We then prove that floating-point $+$ and $\times$ are sufficient to simulate all halting programs that return finite/infinite floats when given finite floats (§4.2).

4.1 Provable Robustness of Neural Networks

Consider the task of classifying floating-point inputs $\mathbf{x} \in \mathcal{X}$ (e.g., images of objects) into $n \in \mathbb{N}$ classes (e.g., categories of objects), where $\mathcal{X} \coloneqq [-1,1]^d \cap \mathbb{F}^d$

denotes the space of inputs throughout this subsection. For this task, a function $f : \mathcal{X} \to \mathbb{F}^n$ is often viewed as a classifier in the following sense: f predicts $\mathbf{x}$ to be in the i-th class ($i \in [n]$), where $i \coloneqq \text{class}(f(\mathbf{x}))$ and $\text{class} : \mathbb{F}^n \to [n]$ is defined by $\text{class}(y_1, \ldots, y_n) \coloneqq \arg\max_{i\in[n]} y_i$ with an arbitrary tie-breaking rule.

A typical robustness property of a classifier f is that f should predict the same class for all neighboring inputs under the ℓ_∞ distance [38]. We formalize this notion of *robust* classifiers in a way similar to [66, Definition A.4].

Definition 1. *Let $\delta > 0$ and $\mathcal{D} \subseteq \mathcal{X}$. A classifier $f : \mathcal{X} \to \mathbb{F}^n$ is called δ-*robust *on $\mathcal{D}$ if for all $\mathbf{x}_0 \in \mathcal{D}$, $\mathbf{y}, \mathbf{y}' \in f(\mathcal{N}_\delta(\mathbf{x}_0))$ implies $\text{class}(\mathbf{y}) = \text{class}(\mathbf{y}')$, where $\mathcal{N}_\delta(\mathbf{x}_0) \coloneqq \{\mathbf{x} \in \mathcal{X} \mid \|\mathbf{x}_0 - \mathbf{x}\|_\infty \le \delta\}$ and $\|\cdot\|_\infty$ denotes the ℓ_∞-norm.*

Neural networks have been widely used as classifiers, but establishing the robustness properties of practical networks as in Definition 1 is intractable due to the enormous number of inputs to be checked (i.e., $|\mathcal{N}_\delta(\mathbf{x}_0)| \gg 1$ when $d \gg 1$). Instead, these properties have been proven often by using interval analysis, as mentioned in §2.3. We formalize the notion of such *provably robust* networks under interval analysis, in a way similar to [66, Definition A.5].

Definition 2. *Let $\delta > 0$ and $\mathcal{D} \subseteq \mathcal{X}$. A neural network $\nu : \overline{\mathbb{F}}^d \to \overline{\mathbb{F}}^n$ is called δ-*provably robust *on $\mathcal{D}$ if for all $\mathbf{x}_0 \in \mathcal{D}$, $\mathbf{y}, \mathbf{y}' \in \gamma(\nu^\sharp(\mathcal{B}))$ implies $\text{class}(\mathbf{y}) = \text{class}(\mathbf{y}')$, where $\mathcal{B} \in \mathbb{I}^d$ denotes the abstract box such that $\gamma(\mathcal{B}) = \mathcal{N}_\delta(\mathbf{x}_0)$.*

Under these definitions, we prove that given an ideal robust classifier f, we can always find a neural network ν (i) whose robustness property is *exactly* the same as that of f and is easily provable using only interval analysis, and (ii) whose predictions are *precisely* equal to those of f.

Theorem 2. *Let $f : \mathcal{X} \to \mathbb{F}^n$ be a classifier that is δ-robust on $\mathcal{D}$, and $\sigma : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ be an activation function satisfying Condition 1. Then, there exists a σ-neural network $\nu : \overline{\mathbb{F}}^d \to \overline{\mathbb{F}}^n$ that is δ-provably robust on $\mathcal{D}$ and makes the same prediction as f on $\mathcal{D}$ (i.e., $\text{class}(\nu(\mathbf{x})) = \text{class}(f(\mathbf{x}))$ for all $\mathbf{x} \in \mathcal{D}$).*

Proof sketch. We show this (i) by applying Theorem 1 to n target functions that are constructed from f, and (ii) by using the following observation: the network constructed in the proof of Theorem 1 has depth not depending on a target function (when d is fixed). The full proof is in §C.1. □

4.2 Floating-Point Interval-Completeness

To motivate our result, we recall the notion of Turing completeness. A computation model is called *Turing-complete* if for every Turing machine T, there exists a program in the model that can simulate the machine [6, 35, 49]. Extensive research has established the Turing completeness of numerous computation models: from untyped λ-calculus [8, 64] and μ-recursive functions [9, 20], to type systems (e.g., Haskell [67], Java [25]) and neural networks over the rationals (e.g., RNNs [59], Transformers [54]). These results identify simpler computation models as powerful as Turing machines, and shed light on the computational power of new models.

We ask an analogous question for *floating-point* computations instead of *binary* computations, where the former is captured by floating-point programs and the latter by Turing machines. That is, which small class of floating-point programs can simulate all (or nearly all) floating-point programs?

Formally, let $\mathcal{F}$ be the set of all terminating programs that take finite floats and return finite or infinite floats, where these programs can use any floating-point constants/operations (e.g., $-\infty$, $\otimes$) and language constructs (e.g., if-else, while). Then, $\mathcal{F}$ semantically denotes the set of all functions from $\mathbb{F}^n$ to $(\mathbb{F}\cup\{-\infty,+\infty\})^m$ for all $n, m \in \mathbb{N}$, because each such function can be expressed with if-else branches and floating-point constants. For this class of programs, we define the notion of *(interval-)simulation* and *floating-point (interval-)completeness* as follows.

Definition 3. *Let $P, Q \in \mathcal{F}$ be programs with arity n. We say Q* simulates *P if $Q(\mathbf{x}) = P(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{F}^n$, where $P(\mathbf{x})$ denotes the concrete semantics of P on $\mathbf{x}$. We say Q* interval-simulates *P if $\gamma(Q^\sharp(\mathcal{B})) = [\min P(\gamma(\mathcal{B})), \max P(\gamma(\mathcal{B}))] \cap \overline{\mathbb{F}}$ for all abstract boxes $\mathcal{B}$ in $\mathbb{F}^n$, where $Q^\sharp(\mathcal{B})$ denotes the interval semantics of Q on $\mathcal{B}$.*

Definition 4. *We say a class of programs $\mathcal{G} \subseteq \mathcal{F}$ is* floating-point (interval-)complete *if for every $P \in \mathcal{F}$, there exists $Q \in \mathcal{G}$ such that Q (interval-)simulates P.*

We prove that a surprisingly small class of programs is floating-point interval-complete (so floating-point complete). In particular, we show that only floating-point addition, multiplication, and constants are sufficient to interval-simulate *all* halting programs that output finite/infinite floats when given finite floats.

Theorem 3. *$\mathcal{F}_{\oplus,\otimes} \subset \mathcal{F}$ is floating-point interval-complete, where $\mathcal{F}_{\oplus,\otimes}$ denotes the class of straight-line programs that use only $\oplus$, $\otimes$, and floating-point constants.*

Proof sketch. We show this by extending the key lemma used in the proof of Theorem 1: there exist σ-networks that capture the direct image maps of indicator functions over $[-1,1]^n \cap \mathbb{F}^n$ (Lemma 2). In particular, we prove that $[-1,1]^n \cap \mathbb{F}^n$ can be extended to $\mathbb{F}^n$ if σ is the identity function. The full proof is in §C.2. □

To our knowledge, this is the first non-trivial result on floating-point (interval-)completeness. This result is an extension of our IUA theorem (Theorem 1) for the identity activation function σ_{id}, in that floating-point interval-completeness considers the input domain $\mathbb{F}^n$ (not $[-1,1]^n \cap \mathbb{F}^n$) and $\mathcal{F}_{\oplus,\otimes}$ includes all σ_{id}-networks (but no other σ-networks). Theorem 3, however, *cannot* be extended to the input domain $(\mathbb{F} \cup \{-\infty, +\infty\})^n$ (instead of $\mathbb{F}^n$), since no program in $\mathcal{F}_{\oplus,\otimes}$ can represent a non-constant function that maps an infinite float to a finite float—this is because $\oplus$ and $\otimes$ do not return finite floats when applied to $\pm\infty$.

5 Proof of IUA Theorem Over Floats

We now prove Theorem 1 by constructing a σ-neural network that computes the upper and lower points of the direct image map of a rounded target function $\widehat{f}$.

For $a, b \in \mathbb{R}$, we let $[a,b]_{\mathbb{F}} \coloneqq [a,b] \cap \mathbb{F}$ and $\mathbb{I}_{[a,b]} \coloneqq \{\mathcal{I} \in \mathbb{I} \mid \gamma(\mathcal{I}) \subseteq [a,b]\}$. With this notation, $(\mathbb{I}_{[a,b]})^d$ is the set of all abstract boxes in $[a,b]^d$.

We start with defining indicator functions for a set of floating-point values and for an abstract box, which play a key role in our proof.

Definition 5. *Let $d \in \mathbb{N}$. For $\mathcal{S} \subseteq \overline{\mathbb{F}}^d$, we define $\iota_{\mathcal{S}} : \overline{\mathbb{F}}^d \to \overline{\mathbb{F}}$ as $\iota_{\mathcal{S}}(\mathbf{x}) \coloneqq 1$ if $\mathbf{x} \in \mathcal{S}$, and $\iota_{\mathcal{S}}(\mathbf{x}) \coloneqq 0$ otherwise. For $a \in \mathbb{F}$, we define $\iota_{>a} : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ by $\iota_{\{x>a \,|\, x \in \mathbb{F}\}}$, and define $\iota_{\geq a}$, $\iota_{<a}$, $\iota_{\leq a}$ analogously. For $\mathcal{C} \in \mathbb{I}^d$, we define $\iota_{\mathcal{C}} : \overline{\mathbb{F}}^d \to \overline{\mathbb{F}}$ by $\iota_{\gamma(\mathcal{C})}$.*

Our proof of Theorem 1 consists of two parts. We first show the existence of σ-networks that precisely compute indicator functions under the interval semantics. We then construct a σ-network stated in Theorem 1 by composing the σ-networks for indicator functions and using the properties of indicator functions.

Both parts of our proof are centered around a new property of activation functions, which we call "$([a,b]_{\mathbb{F}}, \eta, K, L_\phi, L_\psi)$-separability" and define as follows.

Definition 6. *We say that $\sigma : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ is* $([a,b]_{\mathbb{F}}, \eta, K, L_\phi, L_\psi)$-separable *for $a, b, \eta, K \in \mathbb{F}$ and $L_\phi, L_\psi \in \mathbb{N}$ if the following hold:*

- *For every $z \in [a,b]_{\mathbb{F}}$, there exist depth-L_ϕ σ-networks $\phi_{\leq z}, \phi_{\geq z} : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ without the last affine layer such that $\phi^\sharp_{\leq z} = (K\iota_{\leq z})^\sharp$ and $\phi^\sharp_{\geq z} = (K\iota_{\geq z})^\sharp$ on $\mathbb{I}_{[a,b]}$.*
- *There exists a depth-L_ψ σ-network $\psi_{>\eta} : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ without the first and last affine layers such that $\psi^\sharp_{>\eta} = (K\iota_{>\eta})^\sharp$ on $\mathbb{I}_{[a,b]}$.*

The first condition in Definition 6 ensures the existence of σ-networks that perfectly implement scaled indicator functions $K\iota_{\leq z}$ and $K\iota_{\geq z}$ under the interval semantics, for all $z \in [a,b]_{\mathbb{F}}$. Since these networks should have the same depth L_ϕ without the last affine layer, a function $\nu : \overline{\mathbb{F}}^n \to \overline{\mathbb{F}}$ defined, e.g., by

$$\nu(x_1, \ldots, x_n) = \left(\bigoplus_{i=1}^{n} \alpha \otimes \phi_{\leq z_i}(x_i) \right) \oplus \beta \tag{16}$$

denotes a depth-L_ϕ σ-network for any $z_i \in [a,b]_{\mathbb{F}}$ and $\alpha, \beta \in \mathbb{F}$. The second condition in Definition 6 guarantees that another scaled indicator function $K\iota_{>\eta}$ can be precisely implemented by a depth-L_ψ σ-network $\psi_{>\eta}$ without the first and last affine layers. This implies, e.g., that $\psi_{>\eta} \circ \nu$ denotes a depth-$(L_\phi + L_\psi - 1)$ σ-network, where ν denotes the network presented in Eq. (16).

Using the separability property, we can formally state the two parts of our proof as Lemmas 2 and 3. Theorem 1 is a direct corollary of the two lemmas. We present the proofs of Lemmas 2 and 3 in the next subsections (§5.1 and §5.2).

Lemma 2. *Suppose that $\sigma : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ satisfies Condition 1 with constants $c_2, \eta \in \mathbb{F}$. Then, σ is $([-1,1]_{\mathbb{F}}, \eta, K, L_\phi, L_\psi)$-separable for some $L_\phi, L_\psi \in \mathbb{N}$, where η and $K \coloneqq \sigma(c_2)$ satisfy $|\eta| \in [2^{\mathfrak{e}_{\min}+5}, 4 - 8\varepsilon]$ and $|K| \in [\frac{\varepsilon}{2} + 2\varepsilon^2, \frac{5}{4} - 2\varepsilon]$.*

Lemma 3. *Suppose that $\sigma : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ is $([a,b]_{\mathbb{F}}, \eta, K, L_\phi, L_\psi)$-separable for some $a, b, \eta, K \in \mathbb{F}$ and $L_\phi, L_\psi \in \mathbb{N}$ with $|\eta| \in [2^{\mathfrak{e}_{\min}+5}, 4 - 8\varepsilon]$ and $|K| \in [\frac{\varepsilon}{2} + 2\varepsilon^2, \frac{5}{4} - 2\varepsilon]$. Then, for every $d \in \mathbb{N}$ and function $h : \overline{\mathbb{F}}^d \to \overline{\mathbb{F}} \setminus \{\bot\}$, there exists a σ-neural network $\nu : \overline{\mathbb{F}}^d \to \overline{\mathbb{F}}$ such that $\nu^\sharp(\mathcal{B}) = h^\sharp(\mathcal{B})$ for all abstract boxes $\mathcal{B}$ in $[a,b]^d$.*

To prove Lemma 2, we construct a σ-network for the scaled indicator function $K\iota_{\geq z}$ in two steps. We first construct a σ-network that maps all inputs smaller than z to some point x_1, and all other inputs to another point $x_2 \neq x_1$ (Lemmas 4 and 6), where we exploit round-off errors to obtain such "contraction" (Lemma 17 in §F). We then map x_1 to c_1 and x_2 to c_2, and apply σ to the result so that the final network maps all inputs smaller than z to $\sigma(c_1) = 0$ and all other inputs to $\sigma(c_2) = K$ (Lemma 5). We construct σ-networks for $K\iota_{\leq z}$ and $K\iota_{>\eta}$ analogously.

To prove Lemma 3, we construct σ-networks for the scaled indicator functions of every box in $([a, b]_{\mathbb{F}})^d$ (Lemma 7) and every subset of $([a, b]_{\mathbb{F}})^d$ (Lemma 8), using the indicator functions constructed in Lemma 2. We construct the final σ-network (i.e., universal interval approximator) as a floating-point linear combination of the σ-networks that represent the scaled indicator functions of the level sets of the target function (Lemma 9).

5.1 Proof of Lemma 2

To prove Lemma 2, we assume that the activation function $\sigma : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ satisfies Condition 1 with some constants $c_1, c_2, \eta \in \mathbb{F}$. By Condition 1, the constants η and $K \coloneqq \sigma(c_2)$ clearly satisfy the range condition in Lemma 2. Hence, it remains to show the $([-1, 1]_{\mathbb{F}}, \eta, K, L_\phi, L_\psi)$-separability of σ for some $L_\phi, L_\psi \in \mathbb{N}$. This requires us to construct σ-networks $\psi_{>\eta}$ and $\phi_{\leq z}, \phi_{\geq z}$ for every $z \in [-1, 1]_{\mathbb{F}}$ such that $\psi^\sharp_{>\eta} = (K\iota_{>\eta})^\sharp$, $\phi^\sharp_{\leq z} = (K\iota_{\leq z})^\sharp$, and $\phi^\sharp_{\geq z} = (K\iota_{\geq z})^\sharp$ on $\mathbb{I}_{[-1,1]}$ (Definition 6).

We first construct $\psi_{>\eta}$ using Lemmas 4 and 5 (Figure 4). The proofs of these lemmas, presented in §D.1 and §D.2, rely heavily on (C1)–(C3) of Condition 1.

Lemma 4. *There exists a σ-network $\mu : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ without the first affine layer such that $\mu^\sharp(\langle -\Omega, \eta\rangle) = \langle \eta, \eta\rangle$, $\mu^\sharp(\langle \eta^+, \Omega\rangle) = \langle \eta^+, \eta^+\rangle$, and $\mu^\sharp(\langle -\Omega, \Omega\rangle) = \langle \eta, \eta^+\rangle$.*

Lemma 5. *Let (θ, θ') be either (c_1, c_2) or (c_2, c_1). Then, there exists a depth-2 σ-network $\tau_{\theta,\theta'} : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ without the first affine layer such that $\tau^\sharp_{\theta,\theta'}(\langle \eta, \eta\rangle) = \langle \theta, \theta\rangle$, $\tau^\sharp_{\theta,\theta'}(\langle \eta^+, \eta^+\rangle) = \langle \theta', \theta'\rangle$, and $\tau^\sharp_{\theta,\theta'}(\langle \eta, \eta^+\rangle) = \langle \min\{\theta, \theta'\}, \max\{\theta, \theta'\}\rangle$.*

Lemma 4 states that we can construct a σ-network μ without the first affine layer, whose interval semantics maps all finite (abstract) intervals left of η to the singleton interval $\langle \eta, \eta\rangle$, all finite intervals right of η^+ to $\langle \eta^+, \eta^+\rangle$, and all the remaining finite intervals to $\langle \eta, \eta^+\rangle$. Similarly, Lemma 5 shows that there exists a σ-network $\tau_{\theta,\theta'}$ without the first affine layer, whose interval semantics maps $\langle \eta, \eta\rangle$ to $\langle \theta, \theta\rangle$, $\langle \eta^+, \eta^+\rangle$ to $\langle \theta', \theta'\rangle$, and $\langle \eta, \eta^+\rangle$ to the interval between θ and θ'. By composing these networks with σ, we construct $\psi_{>\eta}$ as

$$\psi_{>\eta} \coloneqq \sigma \circ \tau_{c_1,c_2} \circ \mu. \tag{17}$$

This function $\psi_{>\eta}$ is a σ-network without the first and last affine layers, since τ_{c_1,c_2} are μ are without the first affine layer. Moreover, $\psi^\sharp_{>\eta} = (K\iota_{>\eta})^\sharp$ on $\mathbb{I}_{[-1,1]}$ by the aforementioned properties of τ_{c_1,c_2} and μ, and by the next properties of σ from (C1) of Condition 1: $\sigma(c_1) = 0$, $\sigma(c_2) = K$, and $\sigma(x)$ lies between them for all x between c_1 and c_2. Lastly, we choose L_ψ as the depth of $\psi_{>\eta}$.

We next construct $\phi_{\leq z}$ and $\phi_{\geq z}$ using Lemma 6 (Figure 4). The proof of this lemma is provided in §D.3.

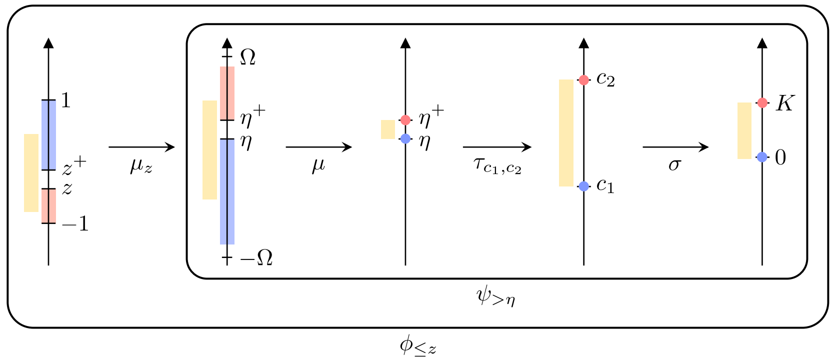


Fig. 4: Illustration of networks μ, τ_{c_1,c_2}, μ_z (Lemmas 4–6) and $\psi_{>\eta}$, $\phi_{\leq z}$ (Eqs. (17) and (18)), assuming (b) in Lemma 6. A box/dot denotes an abstract interval.

Lemma 6. *Let $z \in \mathbb{F}$ with $|z| \leq 1^+$. Then, there exists a depth-1 σ-network $\mu_z : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ such that one of the following holds.*

(a) $\gamma\big(\mu_z^\sharp(\langle -1, z\rangle)\big) \subset [-\Omega, \eta]$ *and* $\gamma\big(\mu_z^\sharp(\langle z^+, 1\rangle)\big) \subset [\eta^+, \Omega]$.
(b) $\gamma\big(\mu_z^\sharp(\langle -1, z\rangle)\big) \subset [\eta^+, \Omega]$ *and* $\gamma\big(\mu_z^\sharp(\langle z^+, 1\rangle)\big) \subset [-\Omega, \eta]$.

Lemma 6 ensures the existence of a depth-1 σ-network μ_z, whose interval semantics maps $\langle -1, z\rangle$ and $\langle z^+, 1\rangle$ to an interval left of η and an interval right of η^+. By composing μ_z with the previous networks $\tau_{\theta,\theta'}$ and μ, we construct $\phi_{\leq z}$ as

$$\phi_{\leq z} := \begin{cases} \sigma \circ \tau_{c_2,c_1} \circ \mu \circ \mu_z & \text{if (a) holds in Lemma 6} \\ \sigma \circ \tau_{c_1,c_2} \circ \mu \circ \mu_z & \text{if (b) holds in Lemma 6.} \end{cases} \tag{18}$$

By a similar argument used above, the function $\phi_{\leq z}$ is a σ-network without the last affine layer, and it satisfies the desired equation: $\phi_{\leq z}^\sharp = (K\iota_{\leq z})^\sharp$ on $\mathbb{I}_{[-1,1]}$. We construct $\phi_{\geq z}$ analogously, but using μ_{z^-} instead of μ_z. Since the depths of $\phi_{\leq z}$ and $\phi_{\geq z}$ are identical for all z, we denote this depth by L_ϕ. This completes the construction of $\psi_{>\eta}$, $\phi_{\leq z}$, and $\phi_{\geq z}$, finishing the proof of Lemma 2.

5.2 Proof of Lemma 3

To prove Lemma 3, we assume that the activation function σ is $([a,b]_{\mathbb{F}}, \eta, K, L_\phi, L_\psi)$-separable for some $\eta, K \in \mathbb{F}$ with $|\eta| \in [2^{\mathfrak{e}_{\min}+5}, 4 - 8\varepsilon]$ and $|K| \in [\frac{\varepsilon}{2} + 2\varepsilon^2, \frac{5}{4} - 2\varepsilon]$. Given this, we construct a σ-network whose interval semantics exactly computes that of the target function $h : \overline{\mathbb{F}}^d \to \overline{\mathbb{F}} \setminus \{\bot\}$ for all abstract boxes in $[a,b]^d$. In our construction, we progressively implement the following functions using σ-networks: (i) scaled indicator functions of arbitrary boxes, (ii) scaled indicator functions of arbitrary sets, and (iii) the target function.

We first construct a σ-network $\tilde{\nu}_{\mathcal{B}}$, for any abstract box $\mathcal{B}$ in $[a,b]^d$, that implements the scaled indicator function $K\iota_{\mathcal{B}}$ under the interval semantics.

Lemma 7. *For any* $\mathcal{B} \in (\mathbb{I}_{[a,b]})^d$*, there exists a depth-*L σ*-network* $\tilde{\nu}_{\mathcal{B}} : \overline{\mathbb{F}}^d \to \overline{\mathbb{F}}$ *without the last affine layer such that* $\tilde{\nu}_{\mathcal{B}}^\sharp = (K\iota_{\mathcal{B}})^\sharp$ *on* $(\mathbb{I}_{[a,b]})^d$*, where* $L := L_\phi + (L_\psi - 1)(\lceil \log_{2^M} d \rceil + 1)$.

In the proof of Lemma 7, we design $\tilde{\nu}_{\mathcal{B}}$ using the networks $\psi_{>\eta}$, $\phi_{\leq z}$, and $\phi_{\geq z}$ constructed in §5.1. Specifically, for an abstract box $\mathcal{B} = (\langle a_1, b_1 \rangle, \ldots, \langle a_d, b_d \rangle)$, we define a σ-network $\tilde{\nu}_i : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ as

$$\tilde{\nu}_i(x) := \psi_{>\eta}\Big(\Big(\alpha \otimes \phi_{\geq a_i}(x)\Big) \oplus \Big(\alpha \otimes \phi_{\leq b_i}(x)\Big) \oplus \beta\Big), \tag{19}$$

where $\alpha, \beta \in \mathbb{F}$ are constants such that $\beta \leq \eta$, $(\alpha \otimes K) \oplus \beta \leq \eta$, and $(\alpha \otimes K) \oplus (\alpha \otimes K) \oplus \beta > \eta$. Then, we can show that $\tilde{\nu}_i^\sharp = (K\iota_{\langle a_i, b_i \rangle})^\sharp$ on $\mathbb{I}_{[a,b]}$. When d is small (e.g., $d \leq 2^{M+1}$), we construct $\tilde{\nu}_{\mathcal{B}}$ using $\tilde{\nu}_i$ and $\psi_{>\eta}$, as follows:

$$\tilde{\nu}_{\mathcal{B}}(x_1, \ldots, x_d) := \psi_{>\eta}\left(\left(\bigoplus_{i=1}^{d} \alpha' \otimes \tilde{\nu}_i(x_i)\right) \oplus \beta'\right), \tag{20}$$

where $\alpha', \beta' \in \mathbb{F}$ are suitably chosen so that $\tilde{\nu}_{\mathcal{B}}^\sharp = (K\iota_{\mathcal{B}})^\sharp$ on $(\mathbb{I}_{[a,b]})^d$. When d is large (e.g., $d > 2^{M+1}$), this construction does not work since $\bigoplus_{i=1}^{d} \alpha' \otimes \tilde{\nu}_i(x_i)$ may not be computed as we want due to rounding errors (e.g., $\bigoplus_{i=1}^{n} 1 = 2^{M+1} < n$ for all $n > 2^{M+1}$). In such a case, we construct $\tilde{\nu}_{\mathcal{B}}$ hierarchically using more layers, but based on a similar idea. A rigorous proof of Lemma 7, including the proof that appropriate $\alpha, \alpha', \beta, \beta' \in \mathbb{F}$ exist, is presented in §E.1.

Using $\tilde{\nu}_{\mathcal{B}}$, we next construct a σ-network $\tilde{\nu}_{\mathcal{S}}$, for any set $\mathcal{S}$ in $([a, b]_{\mathbb{F}})^d$, whose interval semantics computes that of the scaled indicator function $K\iota_{\mathcal{S}}$.

Lemma 8. *Suppose that for any* $\mathcal{B} \in (\mathbb{I}_{[a,b]})^d$*, there exists a depth-*L σ*-network* $\tilde{\nu}_{\mathcal{B}}$ *without the last affine layer such that* $\tilde{\nu}_{\mathcal{B}}^\sharp = (K\iota_{\mathcal{B}})^\sharp$ *on* $(\mathbb{I}_{[a,b]})^d$*. Then, for any* $\mathcal{S} \subseteq ([a, b]_{\mathbb{F}})^d$*, there exists a depth-*$(L + L_\psi - 1)$ σ*-network* $\tilde{\nu}_{\mathcal{S}} : \overline{\mathbb{F}}^d \to \overline{\mathbb{F}}$ *without the last affine layer such that* $\tilde{\nu}_{\mathcal{S}}^\sharp = (K\iota_{\mathcal{S}})^\sharp$ *on* $(\mathbb{I}_{[a,b]})^d$.

In the proof of Lemma 8, we construct $\tilde{\nu}_{\mathcal{S}}$ using $\tilde{\nu}_{\mathcal{B}}$ and $\psi_{>\eta}$, as follows:

$$\tilde{\nu}_{\mathcal{S}}(\mathbf{x}) := \psi_{>\eta}\left(\left(\bigoplus_{\mathcal{B} \in \mathcal{T}} \alpha'' \otimes \tilde{\nu}_{\mathcal{B}}(\mathbf{x})\right) \oplus \eta\right), \tag{21}$$

where $\mathcal{T}$ denotes the collection of all abstract boxes in $\mathcal{S}$, and $\alpha'' \in \mathbb{F}$ is a constant such that $\eta < \left(\bigoplus_{i=1}^{n} \alpha'' \otimes K\right) \oplus \eta < \infty$ for all $n \geq 1$. We remark that it is possible to make the summation not overflow even for a large n, by cleverly exploiting the rounding errors from $\oplus$. With a proper choice of α'', we can further show that $\tilde{\nu}_{\mathcal{S}}^\sharp = (K\iota_{\mathcal{S}})^\sharp$ on $(\mathbb{I}_{[a,b]})^d$. A formal proof of Lemma 8 is given in §E.2.

Using $\tilde{\nu}_{\mathcal{S}}$, we finally construct a σ-network that coincides, under the interval semantics, with the target function h over $([a, b]_{\mathbb{F}})^d$. This result (Lemma 9) and the above results (Lemmas 7 and 8) directly imply Lemma 3.

Lemma 9. *Assume that for any* $\mathcal{S} \subseteq ([a, b]_{\mathbb{F}})^d$*, there exists a depth-*L' σ*-network* $\tilde{\nu}_{\mathcal{S}}$ *without the last affine layer such that* $\tilde{\nu}_{\mathcal{S}}^\sharp = (K\iota_{\mathcal{S}})^\sharp$ *on* $(\mathbb{I}_{[a,b]})^d$*. Then, for any* $h : \overline{\mathbb{F}}^d \to \overline{\mathbb{F}} \setminus \{\perp\}$*, there exists a* σ*-network* $\nu : \overline{\mathbb{F}}^d \to \overline{\mathbb{F}}$ *such that* $\nu^\sharp = h^\sharp$ *on* $(\mathbb{I}_{[a,b]})^d$.

We now illustrate the main idea of the proof of Lemma 9. For a simpler argument, we assume that h is non-negative; the proof for the general case is similar (see §E.3). Let $0 = z_0 < z_1 < \cdots < z_n = +\infty$ be all non-negative floats (except $\perp$) in increasing order, and let $\mathcal{S}_i \coloneqq \{\mathbf{x} \in ([a,b]_{\mathbb{F}})^d \mid h(\mathbf{x}) \geq z_i\}$ be the level set of h for z_i. Under this setup, we construct ν using $\tilde{\nu}_{\mathcal{S}_i}$, as follows:

$$\nu(\mathbf{x}) \coloneqq \bigoplus_{i=1}^{m} \alpha_i \otimes \tilde{\nu}_{\mathcal{S}_i}(\mathbf{x}), \tag{22}$$

where $m \in \mathbb{N} \cup \{0\}$ and $\alpha_i \in \mathbb{F}$ are chosen so that $z_m = \max\{h(\mathbf{x}) \mid \mathbf{x} \in ([a,b]_{\mathbb{F}})^d\}$ and $\alpha_i \otimes K \approx z_i - z_{i-1}$ for all $i \in [m]$. If $\alpha_i \otimes K$ is close enough to $z_i - z_{i-1}$, then the floating-point summation $\bigoplus_{i=1}^{k} \alpha_i \otimes K$ is exactly equal to the exact summation $\sum_{i=1}^{k} z_i - z_{i-1} = z_k$ for all $k \in [m]$, by the rounding errors of $\oplus$. Using this observation, we can show that $\nu(\mathbf{x}) = h(\mathbf{x})$ for all $\mathbf{x} \in ([a,b]_{\mathbb{F}})^d$, and more importantly, $\nu^\sharp = h^\sharp$ on $(\mathbb{I}_{[a,b]})^d$. The full proof of Lemma 9 is in §E.3.

6 Related Work

Universal approximation. Universal approximation theorems for neural networks are widely studied in the literature, which include results for feedforward networks [12, 27, 28, 55], convolutional networks [71], residual networks [41], and transformers [70]. With the advent of low-precision computing for neural networks (e.g., 8-bit E5M2, 8-bit E4M3 [46, 65]; float16 [45]; bfloat16 [1]), there has been growing interest among researchers in characterizing their expressiveness power in this setting. New UA theorems for "quantized" neural networks, which use finite-precision network parameters with *exact* real arithmetic, have been studied in [15, 21]. These networks differ from the floating-point networks considered in this work, because our networks use *inexact* floating-point arithmetic.

To the best of current knowledge, [30, 53] are the only works that study UA theorems for floating-point neural networks. [53] proves UA theorems for ReLU and step activation functions. Our IUA Theorem 1, by virtue of Eq. (15), is a strict generalization of [53] in two senses: (i) it applies to a much broader class of activations that satisfy Condition 1, which subsumes ReLU and step functions; and (ii) it provides a result for abstract interpretation via interval analysis, of which the pointwise approximation considered in [53] is a special case. Concurrent with this article, [30] generalizes [53] to support a wider range of activation functions and larger input domains. Our Theorem 1 partially subsumes [30] in that it is a result for interval approximation, whereas [30] considers only pointwise approximation. Conversely, a special case of our Theorem 1 for pointwise approximation (i.e., Eq. (15)) is subsumed by [30] in that it applies to smaller classes of activation functions and input domains.

Interval universal approximation. The first work to establish an IUA theorem for neural networks used interval analysis with the ReLU activation [4], which was later extended to the more general class of so-called "squashable" activation functions [66]. Whereas these previous IUA theorems assume the neural network can compute over arbitrary real numbers with infinitely precise real arithmetic, the IUA result (Theorem 1) in this work applies to "machine-implementable" neural networks that use floating-point numbers and operations. To the best our knowledge, no previous work has established an IUA theorem for floating-point neural networks. These different computational models lead to substantial differences in both the proof methods (cf. §3.2 and §5) and the specific technical results—§3.3 gives a detailed discussion of how Theorem 1 differs from previous IUA and robustness results [4, Theorem 1.1]; [66, Theorem 3.7].

Provable robustness. There is an extensive literature on robustness verification and robust training for neural networks, which is surveyed in, e.g., [4, Chapter 1]; [38, 60]. Notable methods among these works are [61, 62], which verify the robustness of a neural network using abstract interpretation with the zonotope and polyhedra domains for a restricted class of activations, and are sound with respect to floating-point arithmetic. Compared to these methods, our contribution is a theoretical result on the inherent expressiveness of provably robust floating-point networks under the interval domain for a broad class of activation functions, rather than new verification algorithms or abstract domains. Indeed, our existence result directly applies to the zonotope and polyhedra domains, as they are more precise than the interval domain. More specific IUA theorems tailored to these domains may yield more compact constructions that witness the existence of a provably robust floating-point neural network. Recently, [32] shows that even if a neural network is provably robust over real arithmetic, it can be non-robust over floating-point arithmetic and remain vulnerable to adversarial attacks. This highlights the importance of establishing robustness in the floating-point setting.

Acknowledgments. G. Hwang and Y. Park were supported by Korea Institute for Advanced Study (KIAS) Individual Grants AP092801 and AP090301, via the Center for AI and Natural Sciences at KIAS. G. Hwang was also supported by National Research Foundation of Korea (NRF) Grants RS-2025-00515264 and RS-2024-00406127, funded by the Korea Ministry of Science and ICT (MSIT); and the Gwangju Institute of Science and Technology (GIST) Global University Project in 2025. Y. Park was also supported by the Sejong University faculty research fund in 2025. S. Park was supported by the Korea Institute of Information & Communications Technology Planning & Evaluation (IITP) Grant RS-2019-II190079, funded by the Korea MSIT; the Information Technology Research Center (IITP-ITRC) Grant IITP-2025-RS-2024-00436857, funded by the Korea MSIT; and the Culture, Sports, and Tourism R&D Program through the Korea Creative Content Agency (KOCCA) Grants RS-2024-00348469 and RS-2024-00345025, funded by the Korea Ministry of Culture, Sports and Tourism (MCST) in 2024. W. Lee and F. Saad were supported by the United States National Science Foundation (NSF) under Grant No. 2311983 and funds from the Computer Science Department at Carnegie Mellon University. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abadi, M., et al.: Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv **1603.04467** (2016). https://doi.org/10.48550/arXiv.1603.04467
2. Albarghouthi, A.: Introduction to Neural Network Verification (2021). https://doi.org/10.48550/arXiv.2109.10317
3. Baader, M.: Expressivity of Certified Neural Networks. Ph.D. thesis, ETH Zurich, Zurich, Switzerland (2024). https://doi.org/10.3929/ETHZ-B-000677199
4. Baader, M., Mirman, M., Vechev, M.T.: Universal approximation with certified networks. In: Proceedings of the International Conference on Learning Representations (2020). https://doi.org/10.48550/arXiv.1909.13846
5. Baader, M., Müller, M.N., Mao, Y., Vechev, M.T.: Expressivity of ReLU-networks under convex relaxations. In: Proceedings of the International Conference on Learning Representations (2024), https://openreview.net/forum?id=awHTL3Hpto
6. Barak, B.: Introduction to theoretical computer science. https://introtcs.org (2023)
7. Beebe, N.H.F.: The Mathematical-Function Computation Handbook: Programming Using the MathCW Portable Software Library. Springer (2017). https://doi.org/10.1007/978-3-319-64110-2
8. Church, A.: A set of postulates for the foundation of logic. Annals of Mathematics **34**(4), 839–864 (1933). https://doi.org/10.2307/1968702
9. Church, A.: An unsolvable problem of elementary number theory. American Journal of Mathematics **58**(2), 345–363 (1936). https://doi.org/10.2307/2371045
10. Cousot, P., Cousot, R.: Abstract interpretation: A unified lattice model for static analysis of programs by construction or approximation of fixpoints. In: Proceedings of the ACM Symposium on Principles of Programming Languages. pp. 238–252 (1977). https://doi.org/10.1145/512950.512973
11. Cousot, P., Cousot, R.: Static determination of dynamic properties of generalized type unions. In: Proceedings of an ACM Conference on Language Design for Reliable Software. pp. 77–94 (1977). https://doi.org/10.1145/800022.808314
12. Cybenko, G.: Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals and Systems **2**(4), 303–314 (1989). https://doi.org/10.1007/BF02551274
13. Daramy-Loirat, C., Defour, D., de Dinechin, F., Gallet, M., Gast, N., Lauter, C., Muller, J.M.: CR-LIBM: A library of correctly rounded elementary functions in double-precision. Research Report ensl-01529804, Laboratoire de l'Informatique du Parallèlisme (Dec 2006), https://ens-lyon.hal.science/ensl-01529804
14. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: QLoRA: Efficient finetuning of quantized LLMs. In: Proceedings of the International Conference on Neural Information Processing Systems (2023), https://dl.acm.org/doi/10.5555/3666122.3666563

15. Ding, Y., Liu, J., Xiong, J., Shi, Y.: On the universal approximability and complexity bounds of quantized ReLU neural networks. In: International Conference on Learning Representations (2019). https://doi.org/10.48550/arXiv.1802.03646
16. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D.: Robust physical-world attacks on deep learning visual classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1625–1634 (2018). https://doi.org/10.1109/CVPR.2018.00175
17. Faissole, F., Geneau de Lamarlière, P., Melquiond, G.: End-to-end formal verification of a fast and accurate floating-point approximation. In: Proceedings of the International Conference on Interactive Theorem Proving. pp. 14:1–14:18 (2024). https://doi.org/10.4230/LIPIcs.ITP.2024.14
18. Finlayson, S.G., Bowers, J.D., Ito, J., Zittrain, J.L., Beam, A.L., Kohane, I.S.: Adversarial attacks on medical machine learning. Science **363**(6433), 1287–1289 (2019). https://doi.org/10.1126/science.aaw4399
19. Gehr, T., Mirman, M., Drachsler-Cohen, D., Tsankov, P., Chaudhuri, S., Vechev, M.T.: AI2: Safety and robustness certification of neural networks with abstract interpretation. In: Proceedings of the IEEE Symposium on Security and Privacy. pp. 3–18 (2018). https://doi.org/10.1109/SP.2018.00058
20. Gödel, K.: On undecidable propositions of formal mathematics systems (Notes by S. C. Kleene and J. B. Rosser). Institute for Advanced Study (1934), https://albert.ias.edu/20.500.12111/7996
21. Gonon, A., Brisebarre, N., Gribonval, R., Riccietti, E.: Approximation speed of quantized versus unquantized ReLU neural networks and beyond. IEEE Transactions on Information Theory **69**(6), 3960–3977 (2023). https://doi.org/10.1109/TIT.2023.3240360
22. Goodfellow, I.J., Bengio, Y., Courville, A.C.: Deep Learning. MIT Press (2016), http://www.deeplearningbook.org/
23. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: Proceedings of the International Conference on Learning Representations (2015). https://doi.org/10.48550/arXiv.1412.6572
24. Gowal, S., Dvijotham, K., Stanforth, R., Bunel, R., Qin, C., Uesato, J., Arandjelovic, R., Mann, T.A., Kohli, P.: Scalable verified training for provably robust image classification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4841–4850 (2019). https://doi.org/10.1109/ICCV.2019.00494
25. Grigore, R.: Java generics are Turing complete. In: Proceedings of the ACM Symposium on Principles of Programming Languages. pp. 73–85 (2017). https://doi.org/10.1145/3093333.3009871
26. Hickey, T.J., Ju, Q., van Emden, M.H.: Interval arithmetic: From principles to implementation. Journal of the ACM **48**(5), 1038–1068 (2001). https://doi.org/10.1145/502102.502106
27. Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Networks **2**(5), 359–366 (1989). https://doi.org/10.1016/0893-6080(89)90020-8

28. Hornik, K., Stinchcombe, M., White, H.: Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. Neural Networks **3**(5), 551–560 (1990). https://doi.org/10.1016/0893-6080(90)90005-6
29. Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R., Bengio, Y.: Quantized neural networks: Training neural networks with low precision weights and activations. Journal of Machine Learning Research **18**(1), 6869–6898 (2017), https://dl.acm.org/doi/10.5555/3122009.3242044
30. Hwang, G., Park, Y., Lee, W., Park, S.: Floating-point neural networks can represent almost all floating-point functions. In: Proceedings of the International Conference on Machine Learning (2025)
31. Institute of Electrical and Electronics Engineers: IEEE Standard for Floating-Point Arithmetic (IEEE Std 754-2019). IEEE, Piscataway, NJ, USA (2019). https://doi.org/10.1109/IEEESTD.2019.8766229
32. Jin, J., Ohrimenko, O., Rubinstein, B.I.P.: Getting a-round guarantees: Floating-point attacks on certified robustness. In: Proceedings of the Workshop on Artificial Intelligence and Security. pp. 53–64 (2024). https://doi.org/10.1145/3689932.3694761
33. Jovanovic, N., Balunovic, M., Baader, M., Vechev, M.T.: On the paradox of certified training. Transactions on Machine Learning Research (2022). https://doi.org/10.48550/arXiv.2102.06700
34. Katz, G., Barrett, C.W., Dill, D.L., Julian, K., Kochenderfer, M.J.: Reluplex: An efficient SMT solver for verifying deep neural networks. In: Proceedings of the International Conference on Computer Aided Verification. pp. 97–117. Springer (2017). https://doi.org/10.1007/978-3-319-63387-9_5
35. Kozen, D.: Automata and computability. Springer (1997). https://doi.org/10.1007/978-1-4612-1844-9
36. Lee, W., Sharma, R., Aiken, A.: On automatically proving the correctness of math.h implementations. Proceedings of the ACM on Programming Languages **2**(POPL), 47:1–47:32 (2018). https://doi.org/10.1145/3158135
37. Li, L., Xie, T., Li, B.: Leaderboard for "SoK: Certified Robustness for Deep Neural Networks". https://sokcertifiedrobustness.github.io/leaderboard/ (2023), retrieved May 19, 2025
38. Li, L., Xie, T., Li, B.: SoK: Certified robustness for deep neural networks. In: Proceedings of the IEEE Symposium on Security and Privacy. pp. 1289–1310. IEEE Press (2023). https://doi.org/10.1109/SP46215.2023.10179303
39. Lim, J.P., Nagarakatte, S.: High performance correctly rounded math libraries for 32-bit floating point representations. In: Proceedings of the ACM Conference on Programming Languages Design and Implementation. pp. 359–374 (2021). https://doi.org/10.1145/3453483.3454049
40. Lim, J.P., Nagarakatte, S.: One polynomial approximation to produce correctly rounded results of an elementary function for multiple representations and rounding modes. Proceedings of the ACM on Programming Languages **6**(POPL), 1–28 (2022). https://doi.org/10.1145/3498664
41. Lin, H., Jegelka, S.: Resnet with one-neuron hidden layers is a universal approximator. In: Proceedings of the International Conference on Neural

Information Processing Systems (2018), https://dl.acm.org/doi/10.5555/3327345.3327515
42. Liu, C., Arnon, T., Lazarus, C., Strong, C.A., Barrett, C.W., Kochenderfer, M.J.: Algorithms for verifying deep neural networks. Foundations and Trends in Optimization **4**(3-4), 244–404 (2021). https://doi.org/10.1561/2400000035
43. Mao, Y., Müller, M.N., Fischer, M., Vechev, M.T.: Understanding certified training with interval bound propagation. In: Proceedings of the International Conference on Learning Representations (2024). https://doi.org/10.48550/arXiv.2306.10426
44. Markstein, P.: IA-64 and Elementary Functions: Speed and Precision. Hewlett-Packard Professional Books, Prentice Hall (2000)
45. Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., Wu, H.: Mixed precision training. In: International Conference on Learning Representations (2018). https://doi.org/10.48550/arXiv.1710.03740
46. Micikevicius, P., Stosic, D., Burgess, N., Cornea, M., Dubey, P., Grisenthwaite, R., Ha, S., Heinecke, A., Judd, P., Kamalu, J., et al.: FP8 formats for deep learning. arXiv **2209.05433** (2022). https://doi.org/10.48550/arXiv.2209.05433
47. Mirman, M., Baader, M., Vechev, M.T.: The fundamental limits of neural networks for interval certified robustness. Transactions on Machine Learning Research (2022), https://openreview.net/forum?id=fsacLLU35V
48. Mirman, M., Gehr, T., Vechev, M.T.: Differentiable abstract interpretation for provably robust neural networks. In: Proceedings of the International Conference on Machine Learning (2018), http://proceedings.mlr.press/v80/mirman18b.html
49. Moore, C., Mertens, S.: The Nature of Computation. Oxford University Press (2011). https://doi.org/10.1093/acprof:oso/9780199233212.001.0001
50. Moore, R.E., Kearfott, R.B., Cloud, M.J.: Introduction to Interval Analysis. Society for Industrial and Applied Mathematics (2009). https://doi.org/10.1137/1.9780898717716
51. Muller, J.: Elementary Functions: Algorithms and Implementation. Birkhäuser Boston, MA, 3rd edn. (2016). https://doi.org/10.1007/978-1-4899-7983-4
52. Muller, J.M., Brisebarre, N., De Dinechin, F., Jeannerod, C.P., Lefevre, V., Melquiond, G., Revol, N., Stehlé, D., Torres, S., et al.: Handbook of Floating-Point Arithmetic. Springer (2018). https://doi.org/10.1007/978-3-319-76526-6
53. Park, Y., Hwang, G., Lee, W., Park, S.: Expressive power of ReLU and step networks under floating-point operations. Neural Networks (2024). https://doi.org/10.1016/j.neunet.2024.106297
54. Pérez, J., Barceló, P., Marinkovic, J.: Attention is turing-complete. Journal of Machine Learning Research **22**, 75:1–75:35 (2021), https://dl.acm.org/doi/10.5555/3546258.3546333
55. Pinkus, A.: Approximation theory of the MLP model in neural networks. Acta Numerica **8**, 143–195 (1999). https://doi.org/10.1017/S0962492900002919

56. Raghunathan, A., Steinhardt, J., Liang, P.: Semidefinite relaxations for certifying robustness to adversarial examples. In: Proceedings of the International Conference on Neural Information Processing Systems (2018), https://dl.acm.org/doi/10.5555/3327546.3327746
57. Rosenberg, I., Shabtai, A., Elovici, Y., Rokach, L.: Adversarial machine learning attacks and defense methods in the cyber security domain. ACM Comput. Surv. **54**(5) (May 2021). https://doi.org/10.1145/3453158
58. Sibidanov, A., Zimmermann, P., Glondu, S.: The CORE-MATH project. In: 29th IEEE Symposium on Computer Arithmetic (ARITH). pp. 26–34 (2022). https://doi.org/10.1109/ARITH54963.2022.00014
59. Siegelmann, H., Sontag, E.: On the computational power of neural nets. Journal of Computer and System Sciences **50**(1), 132–150 (1995). https://doi.org/10.1006/jcss.1995.1013
60. Singh, G.: Building trust and safety in artificial intelligence with abstract interpretation. In: Hermenegildo, M.V., Morales, J.F. (eds.) Static Analysis. pp. 28–38. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-44245-2_3
61. Singh, G., Gehr, T., Mirman, M., Püschel, M., Vechev, M.T.: Fast and effective robustness certification. In: Proceedings of the International Conference on Neural Information Processing Systems (2018), https://dl.acm.org/doi/10.5555/3327546.3327739
62. Singh, G., Gehr, T., Püschel, M., Vechev, M.T.: An abstract domain for certifying neural networks. Proceedings of the ACM on Programming Languages **3**(POPL), 41:1–41:30 (2019). https://doi.org/10.1145/3290354
63. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R.: Intriguing properties of neural networks. In: Proceedings of the International Conference on Learning Representations (2014). https://doi.org/10.48550/arXiv.1312.6199
64. Turing, A.M.: On computable numbers, with an application to the entscheidungsproblem. Proceedings of the London Mathematical Society **s2-42**(1), 230–265 (1937). https://doi.org/10.1112/plms/s2-42.1.230
65. Wang, N., Choi, J., Brand, D., Chen, C.Y., Gopalakrishnan, K.: Training deep neural networks with 8-bit floating point numbers. In: Proceedings of the International Conference on Neural Information Processing Systems (2018), https://dl.acm.org/doi/10.5555/3327757.3327866
66. Wang, Z., Albarghouthi, A., Prakriya, G., Jha, S.: Interval universal approximation for neural networks. Proceedings of the ACM on Programming Languages **6**(POPL), 14.1–14.29 (2022). https://doi.org/10.1145/3498675
67. Wansbrough, K.: Instance declarations are universal. http://www.lochan.org/keith/publications/undec.html (1998)
68. Wong, E., Kolter, J.Z.: Provable defenses against adversarial examples via the convex outer adversarial polytope. In: Proceedings of the International Conference on Machine Learning (2018). https://doi.org/10.48550/arXiv.1711.00851
69. Yarotsky, D.: Optimal approximation of continuous functions by very deep relu networks. In: Proceedings of the Conference on Learning Theory (2018). https://doi.org/10.48550/arXiv.1802.03620

70. Yun, C., Bhojanapalli, S., Rawat, A.S., Reddi, S., Kumar, S.: Are transformers universal approximators of sequence-to-sequence functions? In: Proceedings of the International Conference on Learning Representations (2020). https://doi.org/10.48550/arXiv.1912.10077
71. Zhou, D.X.: Universality of deep convolutional neural networks. Applied and Computational Harmonic Analysis **48**(2), 787–794 (2020). https://doi.org/10.1016/j.acha.2019.06.004
72. Ziv, A., Olshansky, M., Henis, E., Retiman, A.: IBM accurate portable Mathlib (2001), https://github.com/dreal-deps/mathlib

A Preliminaries for the Appendix

A.1 Notation

We introduce additional conventions and notations that are used throughout the appendix. First, we interpret any number in the form of $b = b_0.b_1b_2\ldots$ ($b_i \in \{0,1\}$) as a binary expansion, unless otherwise specified:

$$b = b_0 + b_1 \times 2^{-1} + b_2 \times 2^{-2} + \cdots .$$

Second, we interpret floating-point addition and summation operators ($\oplus$ and $\bigoplus$) in the left-associative way, even when they are mixed together. For instance,

$$z \oplus \bigoplus_{i=1}^{n} x_i := (\cdots((z \oplus x_1) \oplus x_2)\cdots) \oplus x_n.$$

That is, we first expand out all the floating-point addition operations and then perform each operation from left to right. Lastly, in the interval semantics, we abuse notation so that $c \in \mathbb{F}$ denotes the abstract interval $\langle c, c\rangle \in \mathbb{I}$. For instance,

$$\langle a, b\rangle \oplus^\sharp c = \langle a, b\rangle \oplus^\sharp \langle c, c\rangle.$$

A.2 Relaxed Version of Condition 1

In the appendix, we use a relaxed version of Condition 1 to simplify the proof.

Condition 2. *An activation function $\sigma : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ satisfies the following conditions:*

(C1$^+$) *There exist $c_1, c_2 \in \mathbb{F}$ such that $\sigma(c_1) = 0$,*

$$\begin{aligned}\sigma(c_2) \in \mathcal{R} &:= [-\tfrac{5}{4} + 2\varepsilon, -\tfrac{\varepsilon}{2} - 2\varepsilon^2)]_{\mathbb{F}} \cup [\tfrac{\varepsilon}{2} + 2\varepsilon^2, \tfrac{5}{4} - 2\varepsilon]_{\mathbb{F}} \\ &= \left[(1 + 2^{-M+1}) \times 2^{-M-2}, 1 + 2^{-2} - 2^{-M}\right]_{\mathbb{F}} \\ &\quad \cup \left[-(1 + 2^{-2} - 2^{-M}), -(1 + 2^{-M+1}) \times 2^{-M-2}\right]_{\mathbb{F}},\end{aligned}$$

and $\sigma(x)$ lies between $\sigma(c_1)$ and $\sigma(c_2)$ for all x between c_1 and c_2.

(C2$^+$) *There exists $\eta \in \mathbb{F}$ with $\eta \in (-(2^2)^-, (2^2)^-)_{\mathbb{F}}$ such that*

- $5 + \mathfrak{e}_{\min} \le \mathfrak{e}_\eta \le 1$,
- $\mathfrak{e}_{\min} + 5 \le \max\{\mathfrak{e}_{\sigma(\eta)}, \mathfrak{e}_{\sigma(\eta^+)}\} \le \mathfrak{e}_\eta + \mathfrak{e}_{\max} - 5$, *and*
- *for all $x, y \in \mathbb{F}$ with $x \le \eta < \eta^+ \le y$,*

$$\sigma(x) \le \sigma(\eta) < \sigma(\eta^+) \le \sigma(y) \qquad \text{or} \qquad \sigma(x) \ge \sigma(\eta) > \sigma(\eta^+) \ge \sigma(y).$$

(C3$^+$) *There exists $\lambda \in [0, 2^{\mathfrak{e}_{\max}-6} \cdot 2^{\min\{\max\{\mathfrak{e}_{\sigma(\eta)}, \mathfrak{e}_{\sigma(\eta^+)}\}, M+2\}}]$ such that for any $x, y \in \mathbb{F}$ with $x \le \eta < \eta^+ \le y$,*

$$|\sigma(x) - \sigma(\eta)| \le \lambda|x - \eta| \qquad \text{and} \qquad |\sigma(y) - \sigma(\eta^+)| \le \lambda|y - \eta^+|.$$

Lemma 10. *If σ satisfies Condition 1, then σ satisfies Condition 2.*

Proof. We prove each of (C1$^+$)–(C3$^+$) as follows.

(C1$^+$) Because $\varepsilon = 2^{-M-1}$, we have the desired result.
(C2$^+$) Because

$$|\sigma(\eta)|, |\sigma(\eta^+)| \leq 2^{\mathfrak{e}_{\max}-6} \cdot |\eta| \leq 2^{\mathfrak{e}_{\max}-5+\mathfrak{e}_\eta},$$

we have the desired result.
(C3$^+$) Because

$$2^{\mathfrak{e}_{\max}-7} \times |\sigma(\eta)| \leq 2^{\mathfrak{e}_{\max}-7+(\mathfrak{e}_{\sigma(\eta)}+1)} \leq 2^{\mathfrak{e}_{\max}-6+\mathfrak{e}_{\sigma(\eta)}} \leq 2^{\mathfrak{e}_{\max}-6+\max\{\mathfrak{e}_{\sigma(\eta)}, \mathfrak{e}_{\sigma(\eta^+)}\}},$$
$$2^{\mathfrak{e}_{\max}-7} \times 2^{M+3} = 2^{\mathfrak{e}_{\max}-6} \times 2^{M+2},$$

we have the desired result.

□

B Proofs of the Results in §3.1

B.1 Proof of Lemma 1

Proof of (C1′) $\implies$ (C1). Suppose ρ satisfies (C1′). Since $M \geq 3$, we have $\frac{\varepsilon}{2} + 2\varepsilon^2, \frac{5}{4} - 2\varepsilon \in \mathbb{F}$. Hence $\mathrm{rnd}(\rho(c_2')) \in [\frac{\varepsilon}{2} + 2\varepsilon^2, \frac{5}{4} - 2\varepsilon]$. In addition, since $-\frac{\omega}{2} \leq \rho(c_1') \leq \frac{\omega}{2}$, we have $\mathrm{rnd}(\rho(c_1')) = 0$. Since $\mathrm{rnd}(\cdot)$ is order-preserving, $\mathrm{rnd}(\rho(x))$ lies between $\mathrm{rnd}(\rho(c_1'))$ and $\mathrm{rnd}(\rho(c_2'))$ for all x between c_1' and c_2'. Therefore $\mathrm{rnd}(\rho)$ satisfies (C1) of Condition 1 with $c_1 = c_1'$ and $c_2 = c_2'$.

Proof of (C2′) $\implies$ (C2). First suppose

$$\rho(x) \leq \rho(\delta - \tfrac{1}{8}) < \rho(\delta + \tfrac{1}{8}) \leq \rho(y),$$

for all $x, y \in \mathbb{R}$ satisfying $x \leq \delta - \frac{1}{8} < \delta + \frac{1}{8} \leq y$. Since $|\rho(x) - \rho(y)| > \frac{1}{8}|x - y|$ for all $x, y \in [\delta - \frac{1}{8}, \delta + \frac{1}{8}]$, ρ is monotonically increasing on $[\delta - \frac{1}{8}, \delta + \frac{1}{8}]$. Hence $\mathrm{rnd}(\rho)$ is also monotonically increasing on $[\delta - \frac{1}{8}, \delta + \frac{1}{8}] \cap \mathbb{F}$.

We define $\eta \in \mathbb{F}$ as

$$\eta \coloneqq \min\{t \in (\delta - \tfrac{1}{8}, \delta + \tfrac{1}{8}) \cap \mathbb{F} : \mathrm{rnd}(\rho(t)) < \mathrm{rnd}(\rho(t^+))\}.$$

Since the minimum of the distance between the floating-point numbers in $[2^{-2}, 1]$ or $[-1, -2^{-2}]$ is 2^{-M-2} and

$$\rho(\delta + \tfrac{1}{8}) - \rho(\delta - \tfrac{1}{8}) > \frac{1}{8} \times \frac{1}{4} = 2^{-5} \geq 2^{-M-2},$$

there exist $\gamma_1, \gamma_2 \in [\delta - \frac{1}{8}, \delta + \frac{1}{8}] \cap \mathbb{F}$ such that $\mathrm{rnd}(\rho(\gamma_1)) \neq \mathrm{rnd}(\rho(\gamma_2))$. Hence η is well-defined. Note that since $\eta \in (\delta - \frac{1}{8}, \delta + \frac{1}{8})$ we have $|\eta| \in (\frac{1}{4}, 1) \subset [2^{\mathfrak{e}_{\min}+5}, 4 - 8\varepsilon]$. Also note that since $\frac{1}{4}, 1 \in \mathbb{F}$, $|\mathrm{rnd}(\rho(\eta))|, |\mathrm{rnd}(\rho(\eta^+))| \in [\frac{1}{4}, 1]$ leading to

$$\begin{aligned}|\mathrm{rnd}(\rho(\eta))|, |\mathrm{rnd}(\rho(\eta^+))| \in [\tfrac{1}{4}, 1] \subset [2^{-9}, 2^7] &\subset [2^{\mathfrak{e}_{\min}+5}, 2^{\mathfrak{e}_{\max}-8}] \\ &\subset [2^{\mathfrak{e}_{\min}+5}, 2^{\mathfrak{e}_{\max}-6} \cdot \eta].\end{aligned}$$

Let $x, y \in \mathbb{F}$, with $x \leq \eta < \eta^+ \leq y$. Since $\rho(x) \leq \rho(\delta + \frac{1}{8}) \leq \rho(\eta)$, we have

$$\mathrm{rnd}(\rho(x)) \leq \mathrm{rnd}(\rho(\delta - \tfrac{1}{8})) \leq \mathrm{rnd}(\rho(\eta)).$$

If $\eta^+ \leq y \leq \delta + \frac{1}{8}$, since $\mathrm{rnd}(\rho)$ is monotonically increasing on $[\delta - \frac{1}{8}, \delta + \frac{1}{8}] \cap \mathbb{F}$, we have

$$\mathrm{rnd}(\rho(\eta^+)) \leq \mathrm{rnd}(\rho(y)).$$

If $y \geq \delta + \frac{1}{8}$, since $\rho(y) \geq \rho(\delta + \frac{1}{8}) \geq \rho(\eta^+)$ we have

$$\mathrm{rnd}(\rho(\eta^+)) \leq \mathrm{rnd}(\rho(\delta - \tfrac{1}{8})) \leq \mathrm{rnd}(\rho(y)).$$

Symmetrically, if

$$\rho(x) \geq \rho(\delta - \tfrac{1}{8}) < \rho(\delta + \tfrac{1}{8}) \geq \rho(y),$$

we have

$$\mathrm{rnd}(\rho(x)) \geq \mathrm{rnd}(\rho(\eta)) > \mathrm{rnd}(\rho(\eta^+)) \geq \mathrm{rnd}(\rho(y)),$$

for $x \leq \eta < \eta^+ \leq y$.

Proof of (C2′) and (C3′) $\implies$ (C3). Let $\lambda_1 = \max\{\lambda, 2^{\mathfrak{e}_{\min}+1}\}$. Since ρ is λ-Lipschitz, we have

$$|\rho(\alpha) - \rho(\beta)| \leq \lambda|\alpha - \beta| \leq \lambda_1|\alpha - \beta|.$$

for $\alpha, \beta \in \mathbb{R}$.

Let $x, y \in \mathbb{F}$ where $x \leq \eta < \eta^+ \leq y$.

First, suppose $\mathrm{rnd}(\rho(x))$ is normal. Let $C = \max\{|x|, |\eta|\}$. Since $|\mathrm{rnd}(t) - t| \leq |t| \times 2^M$ for normal $t \in \mathbb{F}$, we have

$$\begin{aligned}|\mathrm{rnd}(\rho(x)) - \mathrm{rnd}(\rho(\eta))| &\leq |\mathrm{rnd}(\rho(x)) - \rho(x)| + |\rho(x) - \rho(\eta)| + |\rho(\eta) - \mathrm{rnd}(\rho(\eta))| \\ &\leq \lambda_1(|x| + |\eta|) \times 2^{-M} + \lambda_1|x - \eta| \\ &\leq 2\lambda_1 C \times 2^{-M} + \lambda_1|x - \eta| \leq 5\lambda_1|x - \eta|,\end{aligned}$$

where we use

$$|x - \eta| \geq C \times 2^{-M-1}.$$

Now, suppose $\mathrm{rnd}(\rho(x))$ is subnormal. Since $|\mathrm{rnd}(t) - t| \leq \frac{1}{2} \times \omega$ for subnormal $t \in \mathbb{F}$, we have

$$\begin{aligned}|\mathrm{rnd}(\rho(x)) - \mathrm{rnd}(\rho(\eta))| &\leq \tfrac{1}{2} \times \omega + \lambda_1|\eta| \times 2^{-M} + \lambda_1|x - \eta| \\ &\leq 2\lambda_1 C \times 2^{-M} + \lambda_1|x - \eta| \leq 5\lambda_1|x - \eta|,\end{aligned}$$

where we use

$$2^{\mathfrak{e}_{\min}+1} \leq \lambda_1, \quad C \geq |\eta| \geq 2^{-2}, \quad \tfrac{1}{2}\omega \leq \lambda_1 C \times 2^{-M}.$$

Therefore, we have

$$|\mathrm{rnd}(\rho(x)) - \mathrm{rnd}(\rho(\eta))| \leq 5\lambda_1|x - \eta| \leq \tilde{\lambda}|x - \eta|,$$

where

$$\tilde{\lambda} = 5\lambda_1 \in [0, 2^{\mathfrak{e}_{\max}-9}] = [0, 2^{\mathfrak{e}_{\max}-7} \cdot \min\{|\sigma(\eta)|, 2^{M+3}\}].$$

Similarly, we can show

$$|\mathrm{rnd}(\rho(\eta^+)) - \mathrm{rnd}(\rho(y))| \leq \tilde{\lambda}|\eta^+ - y|.$$

■

B.2 Proof of Corollary 1

If ρ satisfies the conditions of Lemma 1, ρ satisfies Condition 1. Since ReLU, LeakyReLU, GELU, ELU, Mish, softplus, sigmoid, and tanh are increasing on $[\frac{1}{4}, 1]$, we have

$$\rho(x) \leq \rho(\delta - \tfrac{1}{8}) < \rho(\delta + \tfrac{1}{8}) \leq \rho(y),$$

for some $\delta \in [\frac{3}{8}, \frac{7}{8}]$.

To check ρ satisfy the conditions of Lemma 1, we need to check the following requirements.

- $|\rho(c_1')| \leq \frac{\omega}{2}$.
- $|\rho(c_2')| \in [\frac{\varepsilon}{2} + 2\varepsilon^2, \frac{5}{4} - 2\varepsilon]$.
- $\delta \in [\frac{3}{8}, \frac{7}{8}]$.
- $\max\{|c_1|, |c_2|\} \geq 2^{\mathfrak{e}_{\min}+1} = 2^{-2^{E-1}+3} \geq 2^{-13}$.
- $|\rho(x)| \in [\frac{1}{4}, 1]$ for $x \in [\delta - \frac{1}{8}, \delta + \frac{1}{8}]$.
- $\inf_{\frac{1}{4} \leq x \leq 1} |\rho'(x)| > \frac{1}{8}$.
- $\lambda \leq \frac{1}{5} \cdot 2^{\mathfrak{e}_{\max}-9}$.

If $M \geq 3$ and $E \geq 5$, according to Table 2, we have

$$-\Omega \leq -32768,\ [0.0391, 1.125] \subset [\tfrac{\varepsilon}{2} + 2\varepsilon^2, \tfrac{5}{4} - 2\varepsilon],$$
$$\tfrac{\omega}{2} \leq 7.63 \times 10^{-6},\ \tfrac{1}{5} \cdot 2^{\mathfrak{e}_{\max}-9} \geq 12.8.$$

Hence, the above requirements are satisfied by Table 1 under the condition $M \geq 3$, $E \geq 5$ as well as for various floating-point formats presented in Table 2.

To verify $|\rho(-\Omega)| \leq \frac{\omega}{2}$ for sigmoid and softplus, it is sufficient to show

$$\log\left(\rho(-2^{2^{E-1}})\right) \leq (-2^{E-1} - M + 1)\log 2,$$

since ρ is monotonically increasing on $x < 0$ and

$$\rho(-\Omega) = \rho(-(2 - 2^{-M}) \times 2^{2^{E-1}}) \leq \rho(-2^{2^{E-1}}) \leq 2^{-2^{E-1}-M+1} = \tfrac{\omega}{2}.$$

Note that

$$|\mathrm{sigmoid}(x)| = |\frac{1}{1+e^{-x}}| \leq e^x, \quad x < 0,$$
$$|\mathrm{softplus}(x)| = |\log(1+e^x)| \leq e^x, \quad x < -1,$$

and

$$(2^{E-1} + M - 1)\log 2 \leq (2^E - 1)\log 2 \leq 2^{E+1} \leq 2^{2^{E-1}},$$

which is due to $n \leq 2^{n-1} - 1$ for $3 \leq n \in \mathbb{N}$ (Note that $E \geq 5$, $M \leq 2^{E-1}$). Therefore we have

$$\log\left(\rho(-2^{2^{E-1}})\right) \leq -2^{2^{E-1}} \leq (-2^{E-1} - M + 1)\log 2, \tag{23}$$

for $\rho =$ sigmoid or $\rho =$ softplus.

Finally, since ReLU, LeakyReLU, GELU, ELU, Mish and tanh are increasing on $[0, 1]$, softplus and sigmoid are increasing on $[-\infty, 1]$, and $\rho(\cdot)$ is order-preserving, $\rho(x)$ lies between $\rho(c_1')$ and $\rho(c_2')$ for all x between c_1' and c_2'. ■

Table 1: Properties of various activation functions for verifying the conditions. D denotes $[\delta - \frac{1}{8}, \delta + \frac{1}{8}]$ and $\mathrm{Lip}(\rho)$ denotes the Lipschitz constant of ρ. For sigmoid and softplus, we show $|\rho(c_1')| \leq \frac{\omega}{2}$ in Eq. (23). The numbers in the table are represented in decimal form and rounded to two decimal places.

Activation function	c_1'	$\lvert\rho(c_1')\rvert$	c_2'	$\rho(c_2')$	δ	$\inf_{x \in D} \lvert\rho(x)\rvert$	$\sup_{x \in D} \lvert\rho(x)\rvert$	$\inf_{1/4 \leq x \leq 1} \lvert\rho'(x)\rvert$	$\mathrm{Lip}(\rho)$
ReLU	0	0	1	1	0.5	0.37	0.63	1	1
LeakyReLU	0	0	1	1	0.5	0.37	0.63	1	1
GELU	0	0	1	0.84	0.6	0.32	0.57	0.70	1.13
ELU	0	0	1	1	0.5	0.37	0.63	1	1
Mish	0	0	1	0.87	0.5	0.26	0.49	0.75	1.09
softplus	$-\Omega$	Eq. (23)	1	1.31	0.4	0.84	0.99	0.56	1
sigmoid	$-\Omega$	Eq. (23)	1	0.73	0.5	0.59	0.66	0.20	0.25
tanh	0	0	1	0.76	0.5	0.35	0.56	0.42	1

Table 2: Properties of floating-point format for verifying the conditions. The numbers in the table are represented in decimal form and rounded to two decimal places.

Format name	E	M	$-\Omega$	$[\frac{\varepsilon}{2} + 2\varepsilon^2, \frac{5}{4} - 2\varepsilon]$	$\frac{\omega}{2}$	$\frac{1}{5} \cdot 2^{\mathfrak{e}_{\max} - 9}$
9-bit format	5	3	$< -2^{15}$	[0.039,1.125]	7.63×10^{-6}	12.8
bfloat16	8	7	$< -2^{127}$	$[1.93 \times 10^{-3}$,1.24]	9.18×10^{-41}	6.65×10^{34}
float16	5	10	$< -2^{15}$	$[2.54 \times 10^{-4}$ 1.25]	5.96×10^{-8}	12.8
float32	8	23	$< -2^{127}$	$[2.98 \times 10^{-8}$,1.25]	1.40×10^{-45}	6.65×10^{34}
float64	11	52	$< -2^{1023}$	$[5.55 \times 10^{-17}$, 1.25]	5.00×10^{-324}	3.51×10^{304}

C Proofs of the Results in §4

C.1 Proof of Theorem 2

First, consider any $g : \overline{\mathbb{F}}^d \to \overline{\mathbb{F}}^n$ such that for every $\mathbf{x}_0 \in \mathcal{X}$ and $\mathbf{x} \in \mathcal{N}_\delta(\mathbf{x}_0)$,

$$g(\mathbf{x}) = (\underbrace{0, \dots, 0}_{\mathrm{class}(f(\mathbf{x}_0))-1}, 1, 0, \dots, 0). \tag{24}$$

Then, g makes the same prediction as f on $\mathcal{X}$ by Eq. (24), and g is δ-robust on $\mathcal{X}$ since f does so. For each $i \in [n]$, let $g_i : \overline{\mathbb{F}}^d \to \overline{\mathbb{F}}$ be the i-th component of g: $g_i(\mathbf{x}) \coloneqq g(\mathbf{x})_i$. Then, by Theorem 1, there exist σ-neural networks $\nu_1, \dots, \nu_n : \overline{\mathbb{F}}^d \to \overline{\mathbb{F}}$ such that for every $i \in [n]$ and $\mathcal{B} \in \mathbb{I}^d$ in $[-1,1]^d$,

$$\gamma(\nu_i^\sharp(\mathcal{B})) = \big[\min g_i(\gamma(\mathcal{B})), \max g_i(\gamma(\mathcal{B}))\big] \cap \overline{\mathbb{F}}. \tag{25}$$

Next, define $\nu : \overline{\mathbb{F}}^d \to \overline{\mathbb{F}}^n$ by a σ-neural network that stacks up $\nu_1, \dots, \nu_n$ such that for every $\mathcal{B} \in \mathbb{I}^d$,

$$\nu^\sharp(\mathcal{B}) = (\nu_1^\sharp(\mathcal{B}), \dots, \nu_n^\sharp(\mathcal{B})). \tag{26}$$

We can construct such ν because $\nu_1, \dots, \nu_n$ have the same depth by the proof of Theorem 1. Then, ν makes the same prediction as g, and thus as f, by Eqs. (25) and (26). Moreover, we claim that ν is δ-provably robust on $\mathcal{X}$. To prove this, let $\mathbf{x}_0 \in \mathcal{X}$ and $\mathcal{B} \in \mathbb{I}^d$ with $\gamma(\mathcal{B}) = \mathcal{N}_\delta(\mathbf{x}_0)$. Then,

$$\gamma(\nu^\sharp(\mathcal{B})) = \prod_{i=1}^{n} \big[\min g_i(\gamma(\mathcal{B})), \max g_i(\gamma(\mathcal{B}))\big] \cap \overline{\mathbb{F}} \tag{27}$$

$$= \{0\} \times \cdots \times \{0\} \times \{1\} \times \{0\} \times \cdots \times \{0\}, \tag{28}$$

where the first equality is by Eqs. (25) and (26) and the second equality is by Eq. (24). Since $\gamma(\nu^\sharp(\mathcal{B}))$ is a singleton set, $\mathbf{y}, \mathbf{y}' \in \gamma(\nu^\sharp(\mathcal{B}))$ clearly implies $\mathrm{class}(\mathbf{y}) = \mathrm{class}(\mathbf{y}')$, as desired. ■

C.2 Proof of Theorem 3

Theorem 3 is a direct corollary of Lemma 3 and the following lemma. ■

Lemma 11. *Let $\sigma : \mathbb{F} \to \mathbb{F}$ be the identity function, i.e., $\sigma(x) = x$ for all $x \in \mathbb{F}$. Then, σ is $(\mathbb{F}, 1, 1, L_\phi, L_\psi)$-separable for some $L_\phi, L_\psi \in \mathbb{N}$.*

Proof. We define f_0 as

$$f_0(x) \coloneqq \sigma\left(2^{-1} \otimes \sigma\left(x \oplus \omega\right)\right).$$

Then we have

$$f_0^\sharp(\langle 0, 0\rangle) = \langle 0, 0\rangle, \quad f_0^\sharp(\langle \omega, \omega\rangle) = \langle \omega, \omega\rangle,$$
$$f_0^\sharp(\langle -2\omega, 0\rangle) = \langle 0, 0\rangle, \quad f_0^\sharp(\langle 2\omega, 4\omega\rangle) = \langle 2\omega, 2\omega\rangle.$$

Case 1: $x > \omega$ with $x = n \cdot \omega$ for $n \in \mathbb{N}$.

Define $\mathcal{I}_i$ as

$$\mathcal{I}_0 := [2, 2^{M+2} - 1], \tag{29}$$

$$\mathcal{I}_k := [2^{M+1+k}, 2^{M+2+k} - 1], \quad k \in \mathbb{N}. \tag{30}$$

If $n \in \mathcal{I}_0$, we have

$$f_0(x) = \sigma\left(2^{-1} \otimes \sigma\left((n+2)\omega\right)\right) \leq \left(\tfrac{n}{2} + \tfrac{3}{2}\right)\omega.$$

Since the solution of the recurrence relation $a_{i+1} = \frac{1}{2}a_i + \frac{3}{2}$ is $a_i = (a_0 - 3)(\frac{1}{2})^i + 3$, pick $m_1 \in \mathbb{N}$ such that

$$m_1 \geq \left\lceil \frac{\log(2^{M+2} - 4)}{\log(2)} \right\rceil_{\mathbb{Z}}.$$

we have

$$a_m = (a_0 - 3)(\tfrac{1}{2})^{m_1} + 3 \leq (2^{M+2} - 4)(\tfrac{1}{2})^{m_1} + 3 \leq 1 + 3 = 4,$$

which leads to $f^{\circ(m_1)}(x) \leq 4\omega$ and $f^{\circ(m_1+2)}(x) = \omega$ for $2\omega \leq x < 2^{M+2}$.

If $n \in \mathcal{I}_k$, we have

$$f_0(x) = \sigma\left(2^{-1} \otimes \sigma\left(n\omega\right)\right) = \tfrac{n}{2}\omega.$$

Hence $f_0(x) = n_2\omega$ for $n_2 \in \mathcal{I}_{k-1}$.

Therefore we have $f_0^{\circ(\mathfrak{e}_{\max} - \mathfrak{e}_{\min} - 1)}(x) = n_3\omega$ for $n_3 \in \mathcal{I}_0$ which leads to $f_0^{\circ(\mathfrak{e}_{\max} - \mathfrak{e}_{\min} + m_1 + 1)}(x) = \omega$ for $\omega \leq x \leq \Omega$.

Case 2: $x \leq 0$ with $x = -n\omega$ for $n \in \mathbb{N}$.

If $n \in \mathcal{I}_0$, we have

$$f_0(x) = \sigma\left(2^{-1} \otimes \sigma\left((-n)\omega\right)\right) \geq (-\tfrac{n}{2} - \tfrac{1}{2})\omega.$$

Since the solution of the recurrence relation $b_{i+1} = \frac{1}{2}b_i + \frac{1}{2}$ is $b_i = (a_0 - 1)(\frac{1}{2})^i + 1$, pick $m_2 \in \mathbb{N}$ such that

$$m_2 \geq \left\lceil \frac{\log(2^{M+2} - 2)}{\log(2)} \right\rceil_{\mathbb{Z}},$$

we have

$$a_m = (a_0 - 1)\left(\tfrac{1}{2}\right)^{m_1} + 1 \leq (2^{M+2} - 2)\left(\tfrac{1}{2}\right)^{m_1} + 1 \leq 2,$$

which leads to $f^{\circ(m_2)}(x) \geq -2\omega$ and $f^{\circ(m_1+1)}(x) = 0$ for $2^{M+2} < x \leq 0$.

If $n \in \mathcal{I}_k$, we have

$$f_0(x) = \sigma\left(2^{-1} \otimes \sigma\left(-n\omega\right)\right) = -\tfrac{n}{2}\omega.$$

Therefore we have $f_0^{\circ(\mathfrak{e}_{\max}-\mathfrak{e}_{\min}-1)}(x) = -n_4\omega$ for $n_4 \in \mathcal{I}_0$ which leads to $f_0^{\circ(\mathfrak{e}_{\max}-\mathfrak{e}_{\min}+m_2+1)}(x) = \omega$ for $-\Omega \leq x \leq 0$. We define g_0 as $f_0(x) \coloneqq f^{\circ o(\mathfrak{e}_{\max}-\mathfrak{e}_{\min}+\max\{m_1,m_2\}+1)}$, and we have

$$g_0^\sharp(\langle -\Omega, 0\rangle) = \langle 0, 0\rangle, \quad g_0^\sharp(\langle \omega, \Omega\rangle) = \langle \omega, \omega\rangle, \quad g_0^\sharp(\langle -\Omega, \Omega\rangle) = \langle 0, \omega\rangle.$$

For $z \in \mathbb{F}$. we define g_z as

$$g_z(x) \coloneqq \begin{cases} g_0\left(\sigma(x \ominus z)\right) & \text{if} \quad |z| < 2^{\mathfrak{e}_{\max}-M-1}, \\ g_0\left(\sigma(2^{-1} \otimes x \ominus \frac{z}{2}\right) & \text{if} \quad |z| \geq 2^{\mathfrak{e}_{\max}-M-1}, \end{cases}$$

and we have

$$g_z^\sharp(\langle -\Omega, z\rangle) = \langle 0, 0\rangle, \quad g_z^\sharp(\langle z^+, \Omega\rangle) = \langle \omega, \omega\rangle, \quad g_0^\sharp(\langle -\Omega, \Omega\rangle) = \langle 0, \omega\rangle.$$

Finally we define $\iota_{>z}$ as

$$\begin{aligned} \iota_{>z}(x) &\coloneqq \sigma\left(2^{-\mathfrak{e}_{\min}} \otimes \sigma\left(2^{-M} \otimes g_z(x)\right)\right), \\ \iota_{\geq z}(x) &\coloneqq \sigma\left(2^{-\mathfrak{e}_{\min}} \otimes \sigma\left(2^{-M} \otimes g_{z^-}(x)\right)\right), \\ \iota_{<z}(x) &\coloneqq \sigma\left(2^{-\mathfrak{e}_{\min}} \otimes \sigma\left(2^{-M} \otimes g_{-z}(-x)\right)\right), \\ \iota_{\leq z}(x) &\coloneqq \sigma\left(2^{-\mathfrak{e}_{\min}} \otimes \sigma\left(2^{-M} \otimes g_{(-z)^-}(-x)\right)\right). \end{aligned}$$

□

D Proofs of the Results in §5.1

In Lemma 2, we suppose σ satisfies Condition 1. By Lemma 10, we suppose σ satisfies Condition 2, the relaxed version of Condition 1.

D.1 Proof of Lemma 4

Since σ also satisfies Condition 2, it is sufficient to prove the following lemma (Lemma 12). To prove Lemma 12, we need following preliminary technical lemmas: Lemmas 15–17 (presented in §F.2) and Lemma 14 (presented in §F.1). ∎

Lemma 12. *Suppose that $\sigma : \mathbb{F} \to \mathbb{F}$ satisfies Condition 2. Then, there exists a σ-network f without the first and last affine layer such that*

$$f^\sharp\left(\langle -\Omega, \eta\rangle\right) = \langle \sigma(\eta), \sigma(\eta)\rangle, \quad f^\sharp\left(\langle \eta^+, \Omega\rangle\right) = \langle \sigma(\eta^+), \sigma(\eta^+)\rangle. \tag{31}$$

Proof. Let $e_0 \in \mathbb{Z}$ such that $2^{e_0} \le |\sigma(\eta^+) - \sigma(\eta)| < 2^{e_0+1}$. Then note that

$$\max\{\mathfrak{e}_{\sigma(\eta)}, \mathfrak{e}_{\sigma(\eta^+)}\} - M - 1 \le e_0 \le \max\{\mathfrak{e}_{\sigma(\eta)}, \mathfrak{e}_{\sigma(\eta^+)}\} + 1.$$

Define $\mathfrak{e}_\theta, \mathfrak{e}_\zeta \in \mathbb{Z}$, $\tilde{\lambda}$ as

$$\begin{aligned}
\mathfrak{e}_\theta &:= \max\left(\mathfrak{e}_{\min} - M, \mathfrak{e}_{\min} - e_0 - M + 1\right),\\
\mathfrak{e}_\zeta &:= \begin{cases} \mathfrak{e}_\eta - M - 1 & \text{if } \eta > 0 \text{ or } \eta < 0, \eta \ne -2^{\mathfrak{e}_\eta},\\ \mathfrak{e}_\eta - M - 2 & \text{if } \eta < 0, \eta = -2^{\mathfrak{e}_\eta}, \end{cases}\\
\tilde{\lambda} &:= \lambda \times 2^{\mathfrak{e}_\theta + 2}.
\end{aligned}$$

To use Lemma 17, we need to check the assumptions of Lemma 17: $\mathfrak{e}_\theta \le -3$ and $e_0 \le \mathfrak{e}_\eta - M - 3 - \mathfrak{e}_\theta$. Since $\max\{\mathfrak{e}_{\sigma(\eta)}, \mathfrak{e}_{\sigma(\eta^+)}\} \ge \mathfrak{e}_{\min} + 5$, we have

$$-e_0 + \mathfrak{e}_{\min} - M + 1 \le -\max\{\mathfrak{e}_{\sigma(\eta)}, \mathfrak{e}_{\sigma(\eta^+)}\} + \mathfrak{e}_{\min} + 2 \le -3.$$

Therefore we have $\mathfrak{e}_\theta \le -3$. To show $e_0 \le \mathfrak{e}_\eta - M - 3 - \mathfrak{e}_\theta$, first suppose $e_0 \ge 1$. Then we have $\mathfrak{e}_\theta = \mathfrak{e}_{\min} - M$, which leads to

$$e_0 \le \max\{\mathfrak{e}_{\sigma(\eta)}, \mathfrak{e}_{\sigma(\eta^+)}\} + 1 \le \mathfrak{e}_\eta - \mathfrak{e}_{\min} - 3 = \mathfrak{e}_\eta - M - 3 - \mathfrak{e}_\theta.$$

Next suppose $e_0 \le 0$. Then we have $\mathfrak{e}_\theta = \mathfrak{e}_{\min} - e_0 - M + 1$ which leads to

$$\begin{aligned}
\mathfrak{e}_\eta - M - 3 - \mathfrak{e}_\theta &= \mathfrak{e}_\eta - M - 3 - (\mathfrak{e}_{\min} - e_0 - M + 1)\\
&\ge (\mathfrak{e}_{\min} + 5) - \mathfrak{e}_{\min} + e_0 - 4 > e_0.
\end{aligned}$$

We can use Lemma 17 and consider the following cases.

Case 1: $\sigma(\eta) < \sigma(\eta^+)$**.**

By Lemma 17, there exists an affine transformation g such that

$$g^\sharp\left(\langle -\Omega, \Omega\rangle\right) \subset \langle -\Omega, \Omega\rangle, \quad g^\sharp(\mathcal{I}) = \langle \eta, \eta\rangle, \quad g^\sharp(\mathcal{I}^+) = \langle \eta^+, \eta^+\rangle, \tag{32}$$

where

$$\mathcal{I} \coloneqq \langle \sigma(\eta) - 2^{\mathfrak{e}_\varsigma - \mathfrak{e}_\theta}, \sigma(\eta)\rangle, \mathcal{I}^+ \coloneqq \langle \sigma(\eta^+), \sigma(\eta) + 2^{\mathfrak{e}_\varsigma - \mathfrak{e}_\theta}\rangle_{\mathbb{F}}.$$

In addition, if $x - \sigma(\eta) > 2^{\mathfrak{e}_\varsigma - \mathfrak{e}_\theta}$,

$$g(x) - \eta^+ \leq \left(x - \sigma\left(\eta^+\right)\right) \times 2^{\mathfrak{e}_\theta + 2}, \tag{33}$$

and if $x < \sigma(\eta) - 2^{\mathfrak{e}_\varsigma - \mathfrak{e}_\theta}$,

$$\eta - g(x) \leq \left(\sigma\left(\eta\right) - x\right) \times 2^{\mathfrak{e}_\theta + 2}, \tag{34}$$

Since $e_0 \geq \max\{\mathfrak{e}_{\sigma(\eta)}, \mathfrak{e}_{\sigma(\eta^+)}\} - M - 1$, we have

$$\mathfrak{e}_\theta \leq \max\left(\mathfrak{e}_{\min} - M, \mathfrak{e}_{\min} - \max\left\{\mathfrak{e}_{\sigma(\eta)}, \mathfrak{e}_{\sigma(\eta^+)}\right\} + 2\right),$$

which leads to

$$\begin{aligned}
\tilde{\lambda} &= \lambda \times 2^{\mathfrak{e}_\theta + 2} \leq \lambda \times 2^{\mathfrak{e}_{\min} + 4} \cdot 2^{\max\{-M-2, -\max\{\mathfrak{e}_{\sigma(\eta)}, \mathfrak{e}_{\sigma(\eta^+)}\}\}} \\
&\leq 2^{-1} \cdot 2^{\min\{\max\{\mathfrak{e}_{\sigma(\eta)}, \mathfrak{e}_{\sigma(\eta^+)}\}, M+2\}} \cdot 2^{\max\{-M-2, -\max\{\mathfrak{e}_{\sigma(\eta)}, \mathfrak{e}_{\sigma(\eta^+)}\}\}} \\
&\leq 1/2.
\end{aligned}$$

Define g_1 as

$$g_1(x) \coloneqq \sigma(g(x)).$$

Then, if $x \geq \sigma\left(\eta^+\right)$, by Eq. (33),

$$\begin{aligned}
g_1(x) - g_1\left(\sigma\left(\eta^+\right)\right) &= \sigma\left(g(x)\right) - \sigma\left(\eta^+\right) \leq \lambda(g(x) - \eta^+) \\
&\leq \lambda\left(x - \sigma\left(\eta^+\right)\right) \times 2^{\mathfrak{e}_\theta + 2} \leq \tilde{\lambda}\left(x - \sigma\left(\eta^+\right)\right).
\end{aligned}$$

Similarly, if $x \leq \sigma(\eta)$, by Eq. (34),

$$g_1\left(\sigma(\eta)\right) - g_1(x) \leq \tilde{\lambda}\left(\sigma(\eta) - x\right).$$

Therefore, we define $n_1 \in \mathbb{Z}_{\geq 0}$ such that

$$n_1 \coloneqq \max\left\{\left\lceil \log_{\tilde{\lambda}^{-1}}\left(\frac{\Omega - \sigma(\eta^+)}{\sigma(\eta) - \sigma(\eta^+) + 2^{\mathfrak{e}_\varsigma - \mathfrak{e}_\theta}}\right)\right\rceil_{\mathbb{Z}}, \left\lceil \log_{\tilde{\lambda}^{-1}}\left(\frac{\sigma(\eta) + \Omega}{2^{\mathfrak{e}_\varsigma - \mathfrak{e}_\sigma}}\right)\right\rceil_{\mathbb{Z}}\right\}.$$

Note that since $\sigma(\eta) - \sigma(\eta^+) + 2^{\mathfrak{e}_\varsigma - \mathfrak{e}_\theta} \geq -2^{e_0} + 2^{\mathfrak{e}_\varsigma - \mathfrak{e}_\theta} > 0$, n is well-defined.

We define $h_1(x)$ as $h_1(x) \coloneqq g_1^{\circ n_1}(x)$. Then we have

$$\begin{aligned}
h_1(x) - h_1\left(\sigma(\eta^+)\right) &\leq \tilde{\lambda}^n\left(x - \sigma(\eta^+)\right) \leq \sigma(\eta) - \sigma(\eta^+) + 2^{\mathfrak{e}_\varsigma - \mathfrak{e}_\theta} && \text{if } \sigma(\eta^+) \leq x \leq \Omega, \\
h_1\left(\sigma(\eta)\right) - h_1(x) &\leq \tilde{\lambda}^n\left(\sigma(\eta) - x\right) \leq 2^{\mathfrak{e}_\varsigma - \mathfrak{e}_\sigma} && \text{if } -\Omega \leq x \leq \sigma(\eta),
\end{aligned}$$

leading to

$$h_1(x) \le \sigma(\eta) + 2^{\mathfrak{e}_\varsigma - \mathfrak{e}_\theta} \quad \text{if} \quad \sigma(\eta^+) \le x \le \Omega,$$
$$h_1(x) \ge \sigma(\eta) - 2^{\mathfrak{e}_\varsigma - \mathfrak{e}_\sigma} \quad \text{if} \quad -\Omega \le x \le \sigma(\eta).$$

Finally, we define $f_1(x)$ as $f_1(x) \coloneqq g_1 \circ h_1 \circ \sigma(x) = g_1^{\circ(n+1)} \circ \sigma(x)$. Together with Eq. (32), we have

$$f_1^\sharp(\langle -\Omega, \eta\rangle) = \langle \sigma(\eta), \sigma(\eta)\rangle, \quad f_1^\sharp\left(\langle \eta^+, \Omega\rangle\right) = \langle \sigma(\eta^+), \sigma(\eta^+)\rangle.$$

Case 2: $\sigma(\eta) > \sigma(\eta^+)$.

By Lemma 17, there exists an affine transformation g such that

$$g^\sharp(\langle -\Omega, \Omega\rangle) \subset \langle -\Omega, \Omega\rangle, \quad g^\sharp(\mathcal{I}) = \langle \eta, \eta\rangle, \quad g^\sharp(\mathcal{I}^+) = \langle \eta^+, \eta^+\rangle. \tag{35}$$

where

$$\theta \coloneqq -2^{\mathfrak{e}_\theta}, \mathcal{I}^+ \coloneqq \langle \sigma(\eta) - 2^{\mathfrak{e}_\varsigma - \mathfrak{e}_\theta}, \sigma\left(\eta^+\right)\rangle, \quad \mathcal{I} \coloneqq \langle \sigma(\eta), \sigma(\eta) + 2^{\mathfrak{e}_\varsigma - \mathfrak{e}_\theta}\rangle,$$

In addition,

$$\eta - g(x) \le (x - \sigma(\eta)) \times 2^{\mathfrak{e}_\theta + 2} \quad \text{for} \quad \sigma(\eta) + 2^{\mathfrak{e}_\varsigma - \mathfrak{e}_\theta} \le x \in \mathbb{F}, \tag{36}$$
$$g(x) - \eta^+ \le \left(\sigma\left(\eta^+\right) - x\right) \times 2^{\mathfrak{e}_\theta + 2} \quad \text{for} \quad \sigma(\eta) - 2^{\mathfrak{e}_\varsigma - \mathfrak{e}_\theta} \ge x \in \mathbb{F}, \tag{37}$$

Define g_2 as

$$g_2(x) \coloneqq \sigma(g(x)).$$

Then, if $\sigma(\eta) \le x \in \mathbb{F}$, by Eq. (36) we have

$$g_2(\sigma(\eta)) - g_2(x) \le \tilde{\lambda}(x - \sigma(\eta)) \le \tilde{\lambda}\left(x - \sigma(\eta^+)\right).$$

If $\sigma(\eta^+) \ge x \in \mathbb{F}$, by Eq. (37) we have

$$g_2(x) - g_2\left(\sigma\left(\eta^+\right)\right) \le \tilde{\lambda}\left(\sigma\left(\eta^+\right) - x\right) \le \tilde{\lambda}(\sigma(\eta) - x).$$

We define h as $h \coloneqq g_2 \circ g_2(x)$. Then we have

$$h_2^\sharp(\mathcal{I}_2) = \langle \sigma(\eta^+), \sigma(\eta^+)\rangle, \quad h_2^\sharp\left(\mathcal{I}_2^+\right) = \langle \sigma(\eta), \sigma(\eta)\rangle,$$

where

$$\mathcal{I}_2 = \langle \sigma(\eta) - \tilde{\lambda}^{-2}\left(2^{\mathfrak{e}_\varsigma - \mathfrak{e}_\theta}\right), \sigma(\eta^+)\rangle, \ \mathcal{I}_2^+ = \langle \sigma(\eta), \sigma(\eta) + \tilde{\lambda}^{-2}\left(2^{\mathfrak{e}_\varsigma - \mathfrak{e}_\theta}\right)\rangle.$$

Hence Therefore, we define $n_2 \in \mathbb{Z}_{\ge 0}$ such that

$$n_2 \coloneqq \max\left\{\left\lceil \log_{\tilde{\lambda}^{-2}}\left(\frac{\Omega - \sigma(\eta^+)}{2^{\mathfrak{e}_\varsigma - \mathfrak{e}_\theta}}\right)\right\rceil_{\mathbb{Z}}, \left\lceil \log_{\tilde{\lambda}^{-2}}\left(\frac{\sigma(\eta) + \Omega}{2^{\mathfrak{e}_\varsigma - \mathfrak{e}_\sigma}}\right)\right\rceil_{\mathbb{Z}}\right\},$$

and define f_2 as $f_2(x) \coloneqq g_2 \circ h_2^{\circ(n_2)} \circ \sigma(x) = g_2^{\circ(2n_2+1)} \circ \sigma(x)$. Together with Eq. (35), we have

$$f_2^\sharp(\langle -\Omega, \eta\rangle) = \langle \sigma(\eta), \sigma(\eta)\rangle, \quad f_2^\sharp\left(\langle \eta^+, \Omega\rangle\right) = \langle \sigma(\eta^+), \sigma(\eta^+)\rangle,$$

and this completes the proof. □

D.2 Proof of Lemma 5

Since σ also satisfies Condition 2, we have $|\sigma(\mathfrak{e}_\eta)|| \in [(1+2^{-M+1} \times 2^{-M-2}, 1+2^{-2}-2^{-M}]$ leading to the fact $\sigma(\eta)$ is normal. In addition, we have $\max\{\mathfrak{e}_\theta, \mathfrak{e}_{\theta'}\} \geq \mathfrak{e}_{\min}+1$ since $\max\{|\theta|, |\theta'|\} \geq \frac{\omega}{\varepsilon}$.

By the lemma below (Lemma 13), we have the desired result. To prove Lemma 13, we need following preliminary technical lemmas: Lemmas 18–21 (presented in §F.3) and Lemma 14 (presented in §F.1). ■

Lemma 13. *Suppose that $\sigma : \mathbb{F} \to \mathbb{F}$ satisfies Condition 2. Let $\gamma_1, \gamma_2, \kappa_1, \kappa_2 \in \mathbb{F}$ with $|\kappa_1| < |\kappa_2|$ with $\mathfrak{e}_{\kappa_2} \geq \mathfrak{e}_{\min}+1$. Suppose there exist $c_2 \in \mathbb{F}$ such that $K \coloneqq \sigma(c_2)$ is normal. Then, there exist $n \in \mathbb{N}$, $w_1, \alpha_i, z_i, b \in \mathbb{F}$ such that*

$$(w_1 \otimes \sigma(\gamma_1)) \oplus \bigoplus_{i=1}^{n} (\alpha_i \otimes \sigma(z_i)) \oplus b = \kappa_1,$$

$$(w_1 \otimes \sigma(\gamma_2)) \oplus \bigoplus_{i=1}^{n} (\alpha_i \otimes \sigma(z_i)) \oplus b = \kappa_2.$$

Proof. By Lemma 14, either $\mathfrak{s}_K^{\parallel} \in (2^{-1}, 1]_{\mathbb{F}}$ or $\mathfrak{s}_K^{\dagger} \in [2^{-1}, 1)_{\mathbb{F}}$ exists such that

$$\mathfrak{s}_K^{\dagger} \otimes \mathfrak{s}_K = 1^- = 1 - 2^{-M-1} = 0.\underbrace{1\ldots1}_{M+1 \text{ times}},$$

$$\mathfrak{s}_K^{\parallel} \otimes \mathfrak{s}_K = 1.$$

Let define $\tilde{\kappa} \in \mathbb{F}$ such that

$$\tilde{\kappa} \oplus \kappa_1 = \kappa_2.$$

Since $\mathfrak{e}_{\kappa_2} \geq 1+\mathfrak{e}_{\min}$, we have $\mathfrak{e}_{\tilde{\kappa}} \geq \mathfrak{e}_{\min}$. Let $\sigma(\bar{\gamma}) = \max\{|\sigma(\gamma_1)|, |\sigma(\gamma_2)|\}$.

By Lemma 18, there exists $w_1 \in \mathbb{F}$ such that

$$w_1 \otimes (\sigma(\gamma_2) \ominus \sigma(\gamma_1)) + \epsilon_1 = \tilde{\kappa}, \quad \mathfrak{e}_{w_1} \leq \mathfrak{e}_{\tilde{\kappa}} - e_0,$$

for some $\epsilon_1 = 0$ or $\pm 2^{-M+\mathfrak{e}_{\tilde{\kappa}}}$. Then by Lemma 19,

$$w_1 \otimes \sigma(\gamma_2) = w_1 \otimes (\sigma(\gamma_2) \ominus \sigma(\gamma_1)) + (w_1 \otimes \sigma(\gamma_1)) + (C \times 2^{-M-1}),$$

where

$$\begin{aligned} |C| &\leq |w_1| \left((2+2^{-M-1})|\sigma(\eta^+) + \sigma(\eta)| + |\sigma(\gamma_2)| + |\sigma(\gamma_1)|\right) \\ &\leq (6+2^{-M})|w_1||\sigma(\bar{\gamma})| \leq 7 \times 2^{\mathfrak{e}_{\tilde{\kappa}}+1}. \end{aligned}$$

Hence we have

$$w_1 \otimes \sigma(\gamma_2) = w_1 \otimes (\sigma(\gamma_2) \ominus \sigma(\gamma_1)) + (w_1 \otimes \sigma(\gamma_1)) + \epsilon_2,$$

for some $|\epsilon_2| \leq 7 \times 2^{-M+\mathfrak{e}_{\tilde{\kappa}}}$. Therefore we have

$$w_1 \otimes \sigma(\gamma_2) = \tilde{\kappa} + (w_1 \otimes \sigma(\gamma_1)) + \epsilon_3,$$

where $|\epsilon_3| \leq 8 \times 2^{-M+\tilde{\kappa}}$. Since $M \geq 3$, the exponents of $w_1 \otimes \sigma(\gamma_1)$ and $w_1 \otimes \sigma(\gamma_2)$ are at most $\mathfrak{e}_{\tilde{\kappa}} + 2$.

We consider the following cases.

Case 1: $\mathfrak{s}_K^{\|}$ exists.

Let $\zeta = w_1 \otimes \sigma(\gamma_1)$. For $i = 0, \dots, M$, we define α_i, z_i, b as

$$(\alpha_i, z_i) = \begin{cases} (0, c_2) & \text{if } \mathfrak{s}_{\zeta,1} = 0 \\ (-\mathrm{sign}(\zeta) \times \mathfrak{s}_K^{\|} \times 2^{-i+\mathfrak{e}_\zeta}, c_2) & \text{if } \mathfrak{s}_{\zeta,1} = 1 \end{cases}, \qquad b = \kappa_1.$$

Then since $\bigoplus_{i=0}^{M} (\alpha_i \otimes \sigma(z_i)) = \zeta$, we have

$$(w_1 \otimes \sigma(\gamma_1)) \oplus \bigoplus_{i=0}^{M} (\alpha_i \otimes \sigma(z_i)) = 0.$$

Since the exponents of $w_1 \otimes \sigma(\gamma_1)$ and $w_1 \otimes \sigma(\gamma_2)$ are at most $\mathfrak{e}_{\tilde{\kappa}} + 2$, we have

$$(w_1 \otimes \sigma(\gamma_2)) \oplus \bigoplus_{i=0}^{M} (\alpha_i \otimes \sigma(z_i)) = (w_1 \otimes \sigma(\gamma_2)) + \zeta + \epsilon_4, \quad |\epsilon_4| \le 2^{\mathfrak{e}_{\tilde{\kappa}}+2},$$

$$(w_1 \otimes \sigma(\gamma_2)) \oplus \bigoplus_{i=0}^{M} (\alpha_i \otimes \sigma(z_i)) \oplus b = \tilde{\kappa} + \epsilon_5, \quad |\epsilon_5| \le 12 \times 2^{-M+\tilde{\kappa}}.$$

By Lemma 20, there exist $\delta_1, \dots, \delta_{36} \in \mathbb{F}$ such that

$$\bigoplus_{i=1}^{36} \delta_i = 0,$$

$$\tilde{\kappa} + \epsilon_4 \oplus \bigoplus_{i=1}^{36} \delta_i = \tilde{\kappa},$$

where $\delta_i = \pm 2^{\mathfrak{e}_{\delta_i}}$. Hence, let $(\alpha_{i+M}, z_{i+M}) = (\mathrm{sign}(\delta_i) \times \mathfrak{s}_K^{\|} \times 2^{\mathfrak{e}_{\delta_i}}, c_2)$. Then,

$$(w_1 \otimes \sigma(\gamma_1)) \oplus \bigoplus_{i=0}^{M} (\alpha_i \otimes \sigma(z_i)) \oplus \bigoplus_{i=M+1}^{M+37} (\alpha_i \otimes \sigma(z_i)) \oplus b = \kappa_1,$$

$$(w_1 \otimes \sigma(\gamma_2)) \oplus \bigoplus_{i=0}^{M} (\alpha_i \otimes \sigma(z_i)) \oplus \bigoplus_{i=M+1}^{M+37} \oplus b = (\tilde{\kappa} + \epsilon_2) \oplus \bigoplus_{i=M+1}^{M+37} \delta_i \oplus \kappa_1$$

$$= \tilde{\kappa} \oplus \kappa_1 = \kappa_2.$$

Case 2: $\mathfrak{s}_K^{\dagger}$ exists.

Let $\zeta_0 = w_1 \otimes \sigma(\gamma_1)$. We recursively define ζ_i, α_i, z_i as

$$(\alpha_i, z_i) \coloneqq (\mathfrak{s}_K^{\dagger} \times 2^{-i+\mathfrak{e}_{\zeta_{i-1}}}, c_2), \quad \zeta_i \coloneqq \zeta_{i-1} \ominus 0.\underbrace{1\dots1}_{M+1 \text{ times}} \times 2^{\mathfrak{e}_{\zeta_{i-1}}}, \ b = \kappa_1.$$

Then we have

$$\alpha_i \otimes \sigma(z_i) = 0.\underbrace{1\dots1}_{M+1 \text{ times}} \times 2^{\mathfrak{e}_{\zeta_{i-1}}}.$$

Let $n_2 = \min\{n \in \mathbb{N} : \mathfrak{e}_{\zeta_i} \leq -4 - 2M + \mathfrak{e}_{\tilde{\kappa}}\}$.
Note that $\bigoplus_{i=1}^{n_2} (\alpha_i \otimes \sigma(z_i)) = \zeta \pm 2^{\mathfrak{e}_\zeta - M}$. Then we have

$$|\,(w_1 \otimes \sigma(\gamma_1)) \oplus \bigoplus_{i=1}^{n_2} (\alpha_i \otimes \sigma(z_i))\,| \leq (2 - 2^{-M}) \times 2^{-4-2M+\mathfrak{e}_{\tilde{\kappa}}}.$$

Since the exponents of $w_1 \otimes \sigma(\gamma_1)$ and $w_1 \otimes \sigma(\gamma_2)$ are at most $\mathfrak{e}_{\tilde{\kappa}} + 2$, we have

$$(w_1 \otimes \sigma(\gamma_2)) \oplus \bigoplus_{i=1}^{n_2} (\alpha_i \otimes \sigma(z_i)) = (w_1 \otimes \sigma(\gamma_2)) + \zeta + \epsilon_6, \quad |\epsilon_6| \leq 2 \times 2^{\mathfrak{e}_{\tilde{\kappa}}+2},$$

$$(w_1 \otimes \sigma(\gamma_2)) \oplus \bigoplus_{i=1}^{n_2} (\alpha_i \otimes \sigma(z_i)) \oplus b = \tilde{\kappa} + \epsilon_7, \quad |\epsilon_7| \leq 16 \times 2^{-M+\tilde{\kappa}}.$$

By Lemma 21, there exist $\delta_1, \ldots, \delta_{48} \in \mathbb{F}$ such that

$$|\,(w_1 \otimes \sigma(\gamma_1)) \oplus \bigoplus_{i=1}^{n_2} (\alpha_i \otimes \sigma(z_i))\,| \oplus \bigoplus_{i=1}^{48} \delta_i = 0,$$

$$\tilde{\kappa} + \epsilon_7 \oplus \bigoplus_{i=1}^{48} \delta_i = \tilde{\kappa},$$

where $\delta_i = \pm 2^{\mathfrak{e}_{\delta_i}}$. Hence let $(\alpha_{i+M}, z_{i+M}) = (\mathrm{sign}(\delta_i) \times \mathfrak{s}_K^{\dagger} \times 2^{\mathfrak{e}_{\delta_i}}, c_2)$. Then,

$$\begin{aligned}
&(w_1 \otimes \sigma(\gamma_1)) \oplus \bigoplus_{i=1}^{n_2} (\alpha_i \otimes \sigma(z_i)) \oplus \bigoplus_{i=n_2+1}^{n_2+49} (\alpha_i \otimes \sigma(z_i)) \oplus b = \kappa_1, \\
&(w_1 \otimes \sigma(\gamma_2)) \oplus \bigoplus_{i=1}^{n_2} (\alpha_i \otimes \sigma(z_i)) \oplus \bigoplus_{i=n_2+1}^{n_2+49} \oplus b = (\tilde{\kappa} + \epsilon_2) \oplus \bigoplus_{i=n_2+1}^{n_2+49} \delta_i \oplus \kappa_1 \\
&= \tilde{\kappa} \oplus \kappa_1 = \kappa_2.
\end{aligned}$$

□

D.3 Proof of Lemma 6

To prove Lemma 6, we need the following preliminary technical lemma: Lemma 23 (presented in §F.4) and Lemma 14 (presented in §F.1)

Define $\mu_z(\cdot)$ as $\mu_z(x) = (w \otimes x) \oplus b$. Since $\eta \in [-4 + 8\varepsilon, 4 - 8\varepsilon]_{\mathbb{F}}, |\eta| \geq \frac{\omega}{\varepsilon}$ we have $\mathfrak{e}_{\eta^-}, \mathfrak{e}_\eta, \mathfrak{e}_{\eta^+} \leq 1$.

We represent η, z as $\eta = \mathfrak{s}_\eta \times 2^{\mathfrak{e}_\eta}$, $z = \mathfrak{s}_z \times 2^{\mathfrak{e}_z}$ with $1 + \mathfrak{e}_{\min} \leq e_\eta \leq 1$.

Let $c_z = \max\{\mathfrak{e}_{\min} - \mathfrak{e}_z, 0\} \geq 0$. Note that $\mathfrak{e}_z + c_z \geq \mathfrak{e}_{\min}$. By Lemma 14, at least one of $\mathfrak{s}_z^{\|}$ or $\mathfrak{s}_z^{\ddagger}$ exists. We consider the following cases.

Case 1: $\eta > 0\ z \geq 0$**.**

In this case, (1) in Lemma 6 holds if

$$w > 0, \quad \begin{cases} \mu_z(-1) & \geq -\Omega, \\ \mu_z(z) & = \eta, \\ \mu_z(z^+) & = \eta^+, \\ \mu_z(1) & \leq \Omega. \end{cases}$$

Note that $z^+ = (\mathfrak{s}_z + 2^{-M+c_z}) \times 2^{\mathfrak{e}_z}$. Let

$$w = 2^{-\mathfrak{e}_z + \mathfrak{e}_\eta - c_z}, \quad b = (\mathfrak{s}_\eta - \mathfrak{s}_z \times 2^{-c_z}) \times 2^{\mathfrak{e}_\eta}.$$

Since $1 + \mathfrak{e}_{\min} \leq \mathfrak{e}_\eta \leq 1$, we have $\mathfrak{e}_{\min} + \mathfrak{e}_\eta \leq 1 - \mathfrak{e}_{\min} = \mathfrak{e}_{\max}$ which leads to $w, b \in \mathbb{F}$. Then we have

$$\begin{aligned} \mu_z(z) &= (w \otimes z) \oplus b = \mathfrak{s}_z \times 2^{\mathfrak{e}_\eta - c_x} \oplus b \\ &= \mathrm{rnd}(\mathfrak{s}_z \times 2^{\mathfrak{e}_\eta - c_z} + (\mathfrak{s}_\eta - \mathfrak{s}_z \times 2^{-c_z}) \times 2^{\mathfrak{e}_\eta}) = \eta, \\ \mu_z(z^+) &= (w \otimes z^+) \oplus b = (\mathfrak{s}_z \times 2^{\mathfrak{e}_\eta - c_z} + 2^{-M+\mathfrak{e}_\eta}) \oplus b \\ &= \mathrm{rnd}(\eta + 2^{-M+\mathfrak{e}_\eta}) = \eta^+. \end{aligned}$$

In addition, we have

$$\begin{aligned} \mu_z(-1) = -w \oplus b \geq -2^{\mathfrak{e}_{\max}} > -\Omega, \\ \mu_z(1) = w \oplus b \leq 2^{\mathfrak{e}_{\max}} \leq \Omega. \end{aligned}$$

Case 2: $\eta < 0,\ z < 0$**.**

In this case, (1) in Lemma 6 holds if

$$w > 0, \quad \begin{cases} \mu_z(-1) & \geq -\Omega, \\ \mu_z(z) & = \eta, \\ \mu_z(z^+) & = \eta^+, \\ \mu_z(1) & \leq \Omega. \end{cases}$$

Note that we have $z = -\mathfrak{s}_z \times 2^{\mathfrak{e}_z}$ and $\eta = -\mathfrak{s}_\eta \times 2^{\mathfrak{e}_\eta}$.

Case 2-1: $\eta \neq -2^{\mathfrak{e}_\eta}$**.**

In this case, we have $\eta^+ = -(\mathfrak{s}_\eta - 2^{-M}) \times 2^{\mathfrak{e}_\eta}$.

Case 2-1-1: $-2^{1+\mathfrak{e}_{\min}} \leq z < 0$**, or** $z < -2^{1+\mathfrak{e}_{\min}},\ z \neq -2^{\mathfrak{e}_z}$**.**

In this case, we have $z^+ = -(\mathfrak{s}_z - 2^{-M+c_z}) \times 2^{\mathfrak{e}_z}$. Let

$$w = 2^{-\mathfrak{e}_z + \mathfrak{e}_\eta - c_z}, \quad b = (-\mathfrak{s}_\eta + \mathfrak{s}_z \times 2^{-c_z}) \times 2^{\mathfrak{e}_\eta}.$$

Then we have

$$\mu_z(z) = (w \otimes z) \oplus b = -\mathfrak{s}_z \times 2^{\mathfrak{e}_\eta - c_z} \oplus b$$

$$
\begin{aligned}
&= \mathrm{rnd}(-\mathfrak{s}_z \times 2^{\mathfrak{e}_\eta - c_z} + (-\mathfrak{s}_\eta + \mathfrak{s}_z \times 2^{-c_z}) \times 2^{\mathfrak{e}_\eta}) = \eta, \\
\mu_z(z^+) &= (w \otimes z^+) \oplus b \\
&= (-\mathfrak{s}_z \times 2^{\mathfrak{e}_\eta - c_z} + 2^{-M+\mathfrak{e}_\eta}) \oplus b = \mathrm{rnd}(-\eta + 2^{-M+\mathfrak{e}_\eta}) = \eta^+, \\
\mu_z(-1) &= -w \oplus b \geq -2^{\mathfrak{e}_{\max}} > -\Omega, \\
\mu_z(1) &= w \oplus b \leq 2^{\mathfrak{e}_{\max}} < \Omega.
\end{aligned}
$$

Case 2-1-2: $z < -2^{1+\mathfrak{e}_{\min}},\ z = -2^{\mathfrak{e}_z}$**.**

In this case, we have $1+\mathfrak{e}_{\min} \leq \mathfrak{e}_z \leq 1$, $c_z = 0$ and $z^+ = -(2-2^{-M}) \times 2^{-1+\mathfrak{e}_z}$. By Lemma 23, we have $(1+2^{-M}) \otimes z = -(1+2^{-M}) \times 2^{-\mathfrak{e}_z}$ and $(1+2^{-M}) \otimes (z^+) = -2^{\mathfrak{e}_z}$. Let

$$
w = (1 + 2^{-M}) \times 2^{-\mathfrak{e}_z + \mathfrak{e}_\eta}, \quad b = (-\mathfrak{s}_\eta + \mathfrak{s}_z + 2^{-M}) \times 2^{\mathfrak{e}_\eta}.
$$

Then we have

$$
\begin{aligned}
\mu_z(z) &= (w \otimes z) \oplus b = -(1 + 2^{-M}) \times 2^{\mathfrak{e}_\eta} \oplus b \\
&= \mathrm{rnd}(-(1 + 2^{-M}) \times 2^{\mathfrak{e}_\eta} + (-\mathfrak{s}_\eta + 1 + 2^{-M}) \times 2^{\mathfrak{e}_\eta}) = \eta, \\
\mu_z(z^+) &= (w \otimes z^+) \oplus b = -2^{\mathfrak{e}_\eta} \oplus b = \mathrm{rnd}(-\eta + 2^{-M+\mathfrak{e}_\eta}) = \eta^+, \\
\mu_z(-1) &= -w \oplus b \geq -2^{\mathfrak{e}_{\max}} > -\Omega, \\
\mu_z(1) &= w \oplus b \leq 2^{\mathfrak{e}_{\max}} < \Omega.
\end{aligned}
$$

Case 2-2: $\eta = -2^{\mathfrak{e}_\eta}$**.**

In this case, we have $\eta^+ = -(2 - 2^{-M}) \times 2^{-1+\mathfrak{e}_\eta}$.

Case 2-2-1: $-2^{1+\mathfrak{e}_{\min}} \leq z < 0$**, or** $z < -2^{1+\mathfrak{e}_{\min}},\ z \neq -2^{\mathfrak{e}_z}$**.**

In this case, we have $z^+ = -(\mathfrak{s}_z - 2^{-M+c_z}) \times 2^{\mathfrak{e}_z}$ and $(-1 + \mathfrak{s}_z \times 2^{-1-c_z})$ is exact. Let

$$
w = 2^{-1-\mathfrak{e}_z+\mathfrak{e}_\eta-c_z}, \quad b = (-1 + \mathfrak{s}_z \times 2^{-1-c_z}) \times 2^{\mathfrak{e}_\eta}.
$$

Then we have

$$
\begin{aligned}
\mu_z(z) &= (w \otimes z) \oplus b = -\mathfrak{s}_z \times 2^{-1+\mathfrak{e}_\eta - c_z} \oplus b \\
&= \mathrm{rnd}(-\mathfrak{s}_z \times 2^{-1+\mathfrak{e}_\eta - c_z} + (-1 + \mathfrak{s}_z \times 2^{-1-c_z}) \times 2^{\mathfrak{e}_\eta}) = \eta, \\
\mu_z(z^+) &= (w \otimes z^+) \oplus b = (-\mathfrak{s}_z \times 2^{-1+\mathfrak{e}_\eta - c_z} + 2^{-1-M+\mathfrak{e}_\eta}) \oplus b \\
&= \mathrm{rnd}(-\eta + 2^{-1-M+\mathfrak{e}_\eta}) = \eta^+, \\
\mu_z(-1) &= -w \oplus b \geq -2^{\mathfrak{e}_{\max}} > -\Omega, \\
\mu_z(1) &= w \oplus b \leq 2^{\mathfrak{e}_{\max}} < \Omega.
\end{aligned}
$$

Case 2-2-2: $z < -2^{1+\mathfrak{e}_{\min}},\ z = -2^{\mathfrak{e}_z}$**.**

In this case, we have $c_z = 0$ and $z^+ = -(2 - 2^{-M}) \times 2^{-1+\mathfrak{e}_z}$. Let

$$
w = 2^{-\mathfrak{e}_z+\mathfrak{e}_\eta-c_z}, \quad b = (-1 + 2^{-c_z}) \times 2^{\mathfrak{e}_\eta}.
$$

Then we have

$$
\begin{aligned}
\mu_z(z) &= (w \otimes z) \oplus b = -2^{-\mathfrak{e}_\eta - c_z} \oplus b = \mathrm{rnd}(-2^{\mathfrak{e}_\eta - c_z} + (-1 + 2^{-c_z}) \times 2^{\mathfrak{e}_\eta}) = \eta, \\
\mu_z(z^+) &= (w \otimes z^+) \oplus b = (-2^{\mathfrak{e}_\eta - c_z} + 2^{-1-M+\mathfrak{e}_\eta}) \oplus b = \mathrm{rnd}(-\eta + 2^{-1-M+\mathfrak{e}_\eta}) = \eta^+, \\
\mu_z(-1) &= -w \oplus b \geq -2^{\mathfrak{e}_{\max}} > -\Omega, \\
\mu_z(1) &= w \oplus b \leq 2^{\mathfrak{e}_{\max}} < \Omega.
\end{aligned}
$$

Case 3: $\eta > 0,\ z < 0$.

In this case, (2) in Lemma 6 holds if

$$
w > 0, \quad \begin{cases} \mu_z(-1) & \leq \Omega, \\ \mu_z(z) & = \eta^+, \\ \mu_z(z^+) & = \eta, \\ \mu_z(1) & \geq -\Omega. \end{cases}
$$

In this case, let $\nu := -(\eta^+)$. Using **Case 2**, there exist $w_\nu, b_\nu \in \mathbb{F}$ with $w_\nu > 0$ such that

$$
\begin{aligned}
(w_\nu \otimes z) \oplus b_\nu &= \nu = -(\eta^+), \\
(w_\nu \otimes z^+) \oplus b_\nu &= \nu^+ = -\eta.
\end{aligned}
$$

Let $w = -w_\nu, b = -b_\nu$. Then we have

$$
\begin{aligned}
\mu_z(z) &= (w \otimes z) \oplus b = \eta^+, \\
\mu_z(z^+) &= (w \otimes z^+) \oplus b = \eta, \\
\mu_z(-1) &= -w \oplus b = w_\nu \oplus (-b_\nu) \leq 2^{\mathfrak{e}_{\max}} < \Omega, \\
\mu_z(1) &= w \oplus b = (-w_\nu) \oplus (-b_\nu) \geq -2^{\mathfrak{e}_{\max}} > -\Omega.
\end{aligned}
$$

Case 4: $\eta < 0,\ z \geq 0$.

Let $\nu := -(\eta^+)$. Using **Case 1**, there exist w_ν, b_ν with $w_n u > 0$ such that

$$
\begin{aligned}
(w_\nu \otimes z) \oplus b_\nu &= \nu = -(\eta^+), \\
(w_\nu \otimes z^+) \oplus b_\nu &= \nu^+ = -\eta.
\end{aligned}
$$

Let $w = -w_\nu, b = -b_\nu$. Then we have

$$
\begin{aligned}
\mu_z(z) &= (w \otimes z) \oplus b = \eta^+, \\
\mu_z(z^+) &= (w \otimes z^+) \oplus b = \eta, \\
\mu_z(-1) &= -w \oplus b = w_\nu \oplus (-b_\nu) \leq 2^{\mathfrak{e}_{\max}} < \Omega, \\
\mu_z(1) &= w \oplus b = (-w_\nu) \oplus (-b_\nu) \geq -2^{\mathfrak{e}_{\max}} > -\Omega.
\end{aligned}
$$

■

E Proofs of the Results in §5.2

Throughout this section, we use $\mathcal{R} = \{x \in \mathbb{F} : |x| \in [\frac{\varepsilon}{2} + 2\varepsilon^2, \frac{5}{4} - 2\varepsilon]_{\mathbb{F}}\}$ and $K = \sigma(c_2)$, i.e., $K \in \mathcal{R}$ by Condition 1.

We present preliminary technical lemmas for the proofs in this section: Lemmas 24–26 (presented in §F.5).

E.1 Proof of Lemma 7

We prove Lemma 7 using the mathematical induction on $\lceil \log_{2^M} d \rceil$. Consider the base case where $\lceil \log_{2^M} d \rceil = 1$, i.e., $d \in [2^M]$. Let $\mathcal{B} = (\langle a_1, b_1 \rangle, \dots, \langle a_d, b_d \rangle$ and $\mu_{K,\eta,n} : \mathbb{F}^n \to \mathbb{F}$ be a network without the first affine layer such that for $x_1, \dots, x_n \in \{0, K\}$,

$$\begin{aligned} &\mu_{K,\eta,n}(x_1, \dots, x_d) > \eta \quad \text{if } x_i = K \text{ for all } i \in [d], \\ &\mu_{K,\eta,n}(x_1, \dots, x_d) \le \eta \quad \text{otherwise.} \end{aligned}$$

We note that such $\mu_{K,\eta,n}$ always exists for all $n \in [2^M]$ by Lemma 26. We then construct $\tilde{\nu}_{\mathcal{B}}$ as follows:

$$\tilde{\nu}_{\mathcal{B}}(\mathbf{x}) = \psi_{>\eta}\Big(\mu_{K,\eta,d}\big(\tilde{\nu}_1(x_1), \dots, \tilde{\nu}_d(x_d)\big)\Big),$$

where

$$\tilde{\nu}_i(x_i) = \psi_{>\eta}\left(\mu_{K,\eta,2}\Big(\phi_{\ge a_i}(x_i), \phi_{\le b_i}(x_i)\Big)\right) \quad \text{for all } i \in [d].$$

Then, one can observe that $\tilde{\nu}_{\mathcal{B}}$ has depth $L_\phi + 2L_\psi - 2$ and does not have the last affine layer.

Let $\mathcal{C} = (\langle s_1, t_1 \rangle, \dots, \langle s_d, t_d \rangle)$ be an abstract box in $[a, b]^d_{\mathbb{F}}$, and let $\bar{\mathcal{B}} = \gamma(\mathcal{B})$ and $\bar{\mathcal{C}} = \gamma(\mathcal{C})$. We now show that $\tilde{\nu}^\sharp_{\mathcal{B}}(\mathcal{C}) = (K\iota_{\mathcal{B}})^\sharp(\mathcal{C})$ by considering the following three cases: (1) $\bar{\mathcal{B}} \cap \bar{\mathcal{C}} = \emptyset$, (2) $\bar{\mathcal{B}} \cap \bar{\mathcal{C}} \ne \emptyset$ and $\bar{\mathcal{C}} \not\subset \bar{\mathcal{B}}$, and (3) $\bar{\mathcal{C}} \subset \bar{\mathcal{B}}$.

Case 1: $\bar{\mathcal{B}} \cap \bar{\mathcal{C}} = \emptyset$.

In this case, there exists $i^* \in [d]$ such that $[s_{i^*}, t_{i^*}]_{\mathbb{F}} \cap [a_{i^*}, b_{i^*}]_{\mathbb{F}} = \emptyset$; otherwise, $\bar{\mathcal{B}} \cap \bar{\mathcal{C}} \ne \emptyset$. This implies that

$$\tilde{\nu}^\sharp_{i^*}(\langle s_{i^*}, t_{i^*} \rangle) = \langle 0, 0 \rangle.$$

Since $\gamma(g^\sharp_i(\langle c_i, d_i \rangle) \subset [0, K]$ by the definition of $\psi_{>\eta}$, we have

$$\mu^\sharp_{K,\eta,2}\Big(\tilde{\nu}^\sharp_1(\langle s_1, t_1 \rangle), \dots, \tilde{\nu}^\sharp_d(\langle s_d, t_d \rangle)\Big) = \langle u, v \rangle,$$

for some $u, v \in \mathbb{F}$ such that $u \le v < \eta$. Hence,

$$\tilde{\nu}^\sharp_{\mathcal{B}}(\mathcal{C}) = \psi_{>\eta}\big(\langle u, v \rangle\big) = \langle 0, 0 \rangle.$$

Case 2: $\bar{\mathcal{B}} \cap \bar{\mathcal{C}} \neq \emptyset$ and $\bar{\mathcal{C}} \not\subset \bar{\mathcal{B}}$.

Since $\bar{\mathcal{B}} \cap \bar{\mathcal{C}} \neq \emptyset$, $[a_i, b_i]_{\mathbb{F}} \cap [s_i, t_i]_{\mathbb{F}} \neq \emptyset$ for all $i \in [d]$. This implies that for each $i \in [d]$,

$$\tilde{\nu}_i^\sharp(\langle s_i, t_i \rangle) = \langle w_i, 1 \rangle, \tag{38}$$

for some $w_i \in \{0, 1\}$. Furthermore, since $\bar{\mathcal{B}} \not\subset \bar{\mathcal{C}}$, there exists $i^* \in [n]$ such that $[s_{i^*}, t_{i^*}]_{\mathbb{F}} \not\subset [a_{i^*}, b_{i^*}]_{\mathbb{F}}$, i.e., $w_{i^*} = 0$. Combining Eq. (38) and $u_{i^*} = 0$ implies that

$$\mu^\sharp_{K,\eta,2}\Big(\tilde{\nu}_1^\sharp(\langle s_1, t_1 \rangle), \ldots, \tilde{\nu}_d^\sharp(\langle s_d, t_d \rangle)\Big) = \langle u, v \rangle,$$

for some $u < \eta < v$. Hence, $\tilde{\nu}^\sharp_{\mathcal{B}}(\mathcal{C}) = \psi_{>\eta}\big(\langle u, v \rangle\big) = \langle 0, K \rangle$.

Case 3: $\bar{\mathcal{C}} \subset \bar{\mathcal{B}}$.

Since $\mathcal{C} \subset \mathcal{B}$, $[s_i, t_i]_{\mathbb{F}} \subset [a_i, b_i]_{\mathbb{F}}$ for all $i \in [n]$. This implies that

$$\tilde{\nu}_i^\sharp(\langle s_i, t_i \rangle) = \langle 1, 1 \rangle,$$

for all $i \in [n]$. Thus, it holds that $\tilde{\nu}^\sharp_{\mathcal{B}}(\mathcal{C}) = \langle K, K \rangle$. By considering all three cases, one can conclude that when $\lceil \log_{2^M} d \rceil = 1$, then a depth-$(L_\phi + 2L_\psi - 2)$ σ-network $\tilde{\nu}^\sharp_{\mathcal{B}} = (K\iota_{\mathcal{B}})^\sharp$ on $[a, b]^d_{\mathbb{F}}$.

Now, suppose that $\lceil \log_{2^M} d \rceil = n > 1$. Let $k \in [2^M - 1]$ and $r \in [2^{(n-1)M} - 1] \cup \{0\}$ such that $d = k2^{(n-1)M} + r$, and let

$$\mathcal{B}_j = \big(\langle a_{(j-1)2^{(n-1)M}+1}, b_{(j-1)2^{(n-1)M}+1} \rangle, \ldots, \langle a_{j2^{(n-1)M}}, b_{j2^{(n-1)M}} \rangle\big) \quad \text{for all } j \in [k],$$
$$\mathcal{B}_{k+1} = \big(\langle a_{k2^{(n-1)M}+1}, b_{k2^{(n-1)M}+1} \rangle, \ldots, \langle a_{k2^{(n-1)M}+r}, b_{k2^{(n-1)M}+r} \rangle\big).$$

Then, by the inductive hypothesis, there exist depth-$(L_\phi + nL_\phi - n)$ σ-networks $\tilde{\nu}_{\mathcal{B}_1}, \ldots, \tilde{\nu}_{\mathcal{B}_k}$ and a depth-$(L_\phi + (\lceil \log_{2^M} r \rceil + 1)L_\phi - (\lceil \log_{2^M} r \rceil + 1))$ σ-network $\tilde{\nu}_{\mathcal{B}_{k+1}}$ such that $\tilde{\nu}^\sharp_{\mathcal{B}_j} = (K\iota_{\mathcal{B}_j})^\sharp$ on $[a, b]^d_{\mathbb{F}}$ for all $j \in [k+1]$. Let $\hat{\nu}_{\mathcal{B}_{k+1}}$ be the composition of $\tilde{\nu}_{\mathcal{B}_{k+1}}$ and $(n - 1 - \lceil \log_{2^M} r \rceil)$ times composition of $\psi_{>\eta} \circ \mu_{K,\eta,1}$, i.e., $\hat{\nu}_{\mathcal{B}_{k+1}}$ is a depth-$(L_\phi + nL_\phi - n)$ σ-network satisfying $\hat{\nu}^\sharp_{\mathcal{B}_{k+1}} = (K\iota_{\mathcal{B}_{k+1}})^\sharp$. Here, we note that $\tilde{\nu}_{\mathcal{B}_1}, \ldots, \tilde{\nu}_{\mathcal{B}_k}, \tilde{\nu}_{\mathcal{B}_1}, \ldots, \hat{\nu}_{\mathcal{B}_{k+1}}$ have the same depth and they do not have the last affine layers. Under this observation, we construct $\tilde{\nu}_{\mathcal{B}}$ as

$$\tilde{\nu}^\sharp_{\mathcal{B}} = \psi^\sharp_{>\eta}\Big(\mu^\sharp_{K,\eta,k+1}\big(\tilde{\nu}^\sharp_{\mathcal{B}_1}, \ldots, \tilde{\nu}^\sharp_{\mathcal{B}_k}, \hat{\nu}^\sharp_{\mathcal{B}_{k+1}}\big)\Big).$$

Then, $\tilde{\nu}^\sharp_{\mathcal{B}} = (K\iota_{\mathcal{B}})^\sharp$ and $\tilde{\nu}_{\mathcal{B}}$ has depth $L_\phi + (n+1)L_\psi - n - 1$. This proves the inductive step and completes the proof. ■

E.2 Proof of Lemma 8

Let $\mathcal{T}$ be the collection of all abstract boxes in $\mathcal{S}$. We construct $\tilde{\nu}_{\mathcal{S}}(\mathbf{x})$ as

$$\tilde{\nu}_{\mathcal{S}}(\mathbf{x}) = \psi_{>\eta}\big(g(\mathbf{x})\big), \quad g(\mathbf{x}) = \left(\bigoplus_{\mathcal{B} \in \mathcal{T}} w \otimes \tilde{\nu}_{\mathcal{B}}(\mathbf{x})\right) \oplus \eta,$$

where $w \in \mathbb{F}$ is chosen so that $w \otimes K \in (2^{\mathfrak{e}_\eta - M-1}, (1+2^{-1}) \times 2^{\mathfrak{e}_\eta - M})_{\mathbb{F}}$, i.e., $(w \otimes K) \oplus \eta \geq \eta^+$. Such w always exists since $K \in \mathcal{R}$ and by Lemma 24. Furthermore, from our choice of w, we note that $\eta \leq g(\mathbf{x}) \leq 2^{\mathfrak{e}_\eta+1}$ for all $\mathbf{x} \in \mathbb{F}$, i.e., overflow does not occur in the evaluation of g under $2^{E-1} \geq M \geq 3$. Here, note that $\tilde{\nu}_{\mathcal{S}}(\mathbf{x})$ has depth $L + L_\phi - 1$.

We now show that $\tilde{\nu}_{\mathcal{S}}^\sharp = (K\iota_{\mathcal{S}})^\sharp$ on $[a,b]_{\mathbb{F}}^d$. Let $\mathcal{C} = (\langle s_1, t_1\rangle, \dots, \langle s_d, t_d\rangle)$ be an abstract box in $[a,b]_{\mathbb{F}}^d$, and let $\bar{\mathcal{C}} = \gamma(\mathcal{C})$ and $\bar{\mathcal{B}} = \gamma(\mathcal{B})$ for all $\mathcal{B} \in \mathcal{T}$. Here, if $\bar{\mathcal{C}} \cap \mathcal{S} = \emptyset$, then $\bar{\mathcal{B}} \cap \bar{\mathcal{C}} = \emptyset$ for all $\mathcal{B} \in \mathcal{T}$. This implies that

$$g^\sharp(\mathcal{C}) = \langle \eta, \eta\rangle \quad \text{and} \quad \tilde{\nu}_{\mathcal{S}}^\sharp(\mathcal{B}) = \langle 0, 0\rangle,$$

by the definition of $\tilde{\nu}_{\mathcal{B}}$ (see Lemma 7). In addition, if $\bar{\mathcal{C}} \cap \mathcal{S} \neq \emptyset$ and $\bar{\mathcal{C}} \not\subset \mathcal{S}$, then $\bar{\mathcal{C}} \not\subset \bar{\mathcal{B}}$ for all $\mathcal{B} \in \mathcal{T}$ and there exists $\mathcal{B}^* \in \mathcal{T}$ such that $\bar{\mathcal{B}}^* \cap \mathcal{C} \neq \emptyset$. This implies that $\tilde{\nu}_{\mathcal{B}}(\mathcal{C}) = \langle 0, u_{\mathcal{B}}\rangle$ for some $u_{\mathcal{B}} \in \{0, K\}$ for all $\mathcal{B} \in \mathcal{T}$ and $u_{\mathcal{B}^*} = K$. This implies that for some $v \geq \eta^+$, we have

$$g^\sharp(\mathcal{C}) = \langle \eta, v\rangle \quad \text{and} \quad \tilde{\nu}_{\mathcal{B}}^\sharp(\mathcal{C}) = \langle 0, K\rangle.$$

Lastly, suppose that $\bar{\mathcal{C}} \subset \mathcal{S}$. Since $\mathcal{C}$ is a box in $\mathcal{S}$, this implies $\mathcal{C} \in \mathcal{T}$ and $\tilde{\nu}_{\mathcal{C}}^\sharp(\mathcal{C}) = \langle K, K\rangle$. Thus, for some $\eta^+ \leq u \leq v$, we have

$$g^\sharp(\mathcal{C}) = \langle u, v\rangle \quad \text{and} \quad \tilde{\nu}_{\mathcal{S}}^\sharp(\mathcal{C}) = \langle K, K\rangle.$$

This completes the proof. ■

E.3 Proof of Lemma 9

Let $\mathcal{D} = [a,b]_{\mathbb{F}}^d$ and define

$$h_+(\mathbf{x}) = \max\{0, h(\mathbf{x})\} \quad \text{and} \quad h_-(\mathbf{x}) \max\{0, -h(\mathbf{x})\},$$

i.e., $h^\sharp = h_+^\sharp \ominus^\sharp h_-^\sharp$. Let $k_+, k_- \in \mathbb{N} \cup \{0\}$ be the numbers such that $\max_{\mathbf{x} \in \mathcal{D}} h_+(\mathbf{x})$ is the $(k_+ + 1)$-th smallest non-negative number in $\mathbb{F}$ and $\max_{\mathbf{x}\in\mathcal{D}} h_-(\mathbf{x})$ is the $(k_- + 1)$-th smallest non-negative number in $\mathbb{F}$. Let $0 = z_0 < z_1 < \cdots < z_{(|\mathbb{F}|-1)/2} = \Omega < z_{(|\mathbb{F}|+1)/2} = \infty$ be all non-negative numbers in $\mathbb{F} \cup \{\infty\}$; here, we have $z_{k_+} = \max_{\mathbf{x}\in\mathcal{D}} h_+(\mathbf{x})$ and $z_{k_-} = \max_{\mathbf{x}\in\mathcal{D}} h_-(\mathbf{x})$.

Let $\mathcal{S}_{\tau,i} = \{\mathbf{x} \in \mathcal{D} : h_\tau(\mathbf{x}) \geq z_i\}$ for $\tau \in \{-,+\}$ for all $i \in [(|\mathbb{F}|+1)/2] \cup \{0\}$ and define $f : \mathcal{D} \to \mathbb{F}$ as

$$\nu(\mathbf{x}) = \left(\bigoplus_{i=1}^{k_+} \left(w_i \otimes \tilde{\nu}_{\mathcal{S}_{+,i}}(\mathbf{x}) \right) \right) \oplus \bigoplus_{i=1}^{k_-} \left((-w_i) \otimes \tilde{\nu}_{\mathcal{S}_{-,i}}(\mathbf{x}) \right),$$

where $w_i \in \mathbb{F}$ is chosen so that $w_i \otimes K \in (2^{\mathfrak{e}_{z_i} - M-1}, (1+2^{-1}) \times 2^{\mathfrak{e}_{z_i} - M})_{\mathbb{F}}$. Such w_i exists for all i by $K \in \mathcal{R}$ and Lemma 24. Here, we note that ν has depth L.

Let $\mathcal{B}$ be a box in $\mathcal{D}$ and choose $i_{\tau,\max}, i_{\tau,\min}$ such that $z_{i_{\tau,\max}} = \max_{\mathbf{x}\in\gamma(\mathcal{B})} h_\tau(\mathbf{x})$ and $z_{i_{\tau,\min}} = \min_{\mathbf{x}\in\gamma(\mathcal{B})} h_\tau(\mathbf{x})$ for $\tau \in \{-,+\}$. Then, one can observe that for $\tau \in \{-,+\}$ and $i \in [(|\mathbb{F}|+1)/2] \cup \{0\}$,

$$\tilde{\nu}^\sharp_{\mathcal{S}_{\tau,i}}(\mathcal{B}) = \begin{cases} \langle K, K\rangle & \text{if } i \le i_{\tau,\min} \\ \langle 0, K\rangle & \text{if } i_{\min} < i \le i_{\tau,\max} \\ \langle 0, 0\rangle & \text{if } i_{\tau,\max} < i \end{cases}.$$

By the definition of w_i and Lemma 25, this implies that

$$\sum_{i=1}^{k_+}{}^\sharp \left(w_i \otimes^\sharp \tilde{\nu}^\sharp_{\mathcal{S}_{+,i}}(\mathcal{B})\right) = \langle z_{i_{+,\min}}, z_{i_{+,\max}}\rangle.$$

Here, if $z_{i_{+,\min}} > 0$, then by the definition h_- and $\mathcal{S}_{\tau,i}$, we have $\tilde{\nu}_{\mathcal{S}_{-,i}}(\mathcal{B}) = \langle 0, 0\rangle$ for all i. This implies that

$$\nu^\sharp(\mathcal{B}) = \sum_{i=1}^{k_+}{}^\sharp \left(w_i \otimes^\sharp \tilde{\nu}^\sharp_{\mathcal{S}_{+,i}}(\mathcal{B})\right) = \langle z_{i_{+,\min}}, z_{i_{+,\max}}\rangle = h^\sharp(\mathcal{B}).$$

If $z_{i_{+,\min}} = 0$, then

$$(-w_i) \otimes^\sharp \tilde{\nu}_{\mathcal{S}^\sharp_{-,i}(\mathcal{B})} = \langle -u_i, 0\rangle,$$

for some $u_i \in (2^{\mathfrak{e}_{z_i}-M-1}, (1+2^{-1}) \times 2^{\mathfrak{e}_{z_i}-M})_{\mathbb{F}}$. Hence, by Lemma 25 and the definition of $\oplus^\sharp$, we have $\nu^\sharp(\mathcal{B}) = \langle -z_{i_{-,\max}}, z_{i_{+,\max}}\rangle = h^\sharp(\mathcal{B})$. ∎

F Technical Lemmas

F.1 Common Technical Lemma

We present the technical lemma used in multiple proofs.

Lemma 14. *Suppose $M \geq 3, E \geq 2$. For any $x \in [1,2)_{\mathbb{F}}$, at least one of the followings holds:*

- *There exists $y \in (2^{-1}, 1]_{\mathbb{F}}$ such that $x \otimes y = 1$. In this case, we denote y as $x^{\parallel} := y$.*
- *There exists $y \in [2^{-1}, 1)_{\mathbb{F}}$ such that $x \otimes y_1 = 1^- = 1 - 2^{-1-M}$. In this case, we denote y as $x^{\dagger} := y_1$.*

Proof. If $x = 1$, then $x^{\parallel} = 1$, $x^{\dagger} = 1 - 2^{-1-M}$.
Now suppose $1 < x < 2$. Note that $x = 1.\underbrace{m_{x,1}...m_{x,M}}_{M \text{ times}} \times 2^0$. Let $n_x = x \times 2^M \in \mathbb{N}$.
Note that $2^M < n_x < 2^{M+1}$. By dividing 2^{2M} by n_x, we have the following

$$2^{2M} = n_x n_u + r \quad (0 < r < n_x). \tag{39}$$

Note that since $2^{M-1} \leq n_u < 2^M$, $n_u \times 2^{1-M}$ has the form of $1.\underbrace{m_1...m_M}_{M \text{ times}} \times 2^{-1}$ whose significand has $\leq M+1$ binary digits. We consider the following cases.

- **Case (1)** $0 < r \leq 2^{-2+M}$.
  In this case, $x^{\parallel} = n_u \times 2^{-M}$ exists since

$$1 - 2^{-2-M} \leq x \times (n_u \times 2^{-M}) = (n_x \times n_u) \times 2^{-2M} = 1 - r \times 2^{-2M} < 1.$$

- **Case (2)** $2^{-2+M} < r < n_x - 2^{-1+M}$.
  In this case, note that $n_x - 2^{-1+M} < 2^{-1+M} < 3 \times 2^{-2+M}$. Then $x^{\dagger} = n_u \times 2^{-M}$ exists since

$$1 - 3 \times 2^{-2-M} < x \times (n_u \times 2^{-M}) = (n_x \times n_u) \times 2^{-2M} = 1 - r \times 2^{-2M} < 1 - 2^{-2-M}.$$

- **Case (3)** $n_x - 2^{-1+M} \leq r < n_x$.
  In this case, $x^{\parallel} = (n_u + 1) \times 2^{-M}$ exists since

$$1 < x \times ((n_u + 1) \times 2^{-M}) = 1 + (n_x - r) \times 2^{-2M} \leq 1 + 2^{-1-M}.$$

Now, we show $x^{\parallel} > 2^{-1}$. If $n_x \leq 2^{1+M} - 4$, we have $n_u \geq 2^{-1+M} + 1$. If $n_x \geq 2^{1+M} - 3$, we have $n_u = 2^{-1+M}, r \geq 2^{-1+M}$ which does not belong to **Case (1)**. Therefore we have $x^{\parallel} > 2^{-1}$. This completes the proof. $\square$

F.2 Technical Lemmas for Lemma 4

This subsection presents technical lemmas for the proof of Lemma 4 (§D.1).

Lemma 15. *Suppose* $\eta \in (-2^{1+\mathfrak{e}_{\min}}, 2^{1+\mathfrak{e}_{\min}})_{\mathbb{F}}$*. For any normal* $x \in (-1-2^{-1}, 1+2^{-1})_{\mathbb{F}}$*, there exist* $y_1, y_2 \in \mathbb{F}$ *such that* $(y_1 \otimes x) \oplus (y_2 \otimes x) = \eta$*.*

Proof. Without loss of generality, we assume $\eta, x > 0$. Since $\eta \in (-2^{1+\mathfrak{e}_{\min}}, 2^{1+\mathfrak{e}_{\min}})_{\mathbb{F}}$, we write η as $\eta = n_\eta \times 2^{-M+\mathfrak{e}_{\min}}$ for $1 \le n_\eta < 2^{1+M}, n_\eta \in \mathbb{N}$. First define y_* as

$$y_* := \begin{cases} 2^{-1-M+\mathfrak{e}_{\min}-\mathfrak{e}_x} & \text{if} \quad x < 1, \mathfrak{s}_x = 1, \\ 2^{-M+\mathfrak{e}_{\min}-\mathfrak{e}_x} & \text{if} \quad x < 1, \mathfrak{s}_x > 1, \\ 2^{-M+\mathfrak{e}_{\min}} & \text{if} \quad 1 \le x < 1+2^{-1}. \end{cases}$$

If $n_\eta = 1$, let $y_1 = y_*$.

Then we have $y_1 \otimes x = \begin{cases} \mathrm{rnd}(2^{-M+\mathfrak{e}_{\min}}) = 2^{-M+\mathfrak{e}_{\min}} & \text{if} \quad x < 1, \mathfrak{s}_x = 1, \\ \mathrm{rnd}(\mathfrak{s}_x \times 2^{-1-M+\mathfrak{e}_{\min}}) = 2^{-M+\mathfrak{e}_{\min}} & \text{if} \quad x < 1, \mathfrak{s}_x > 1, \\ \mathrm{rnd}(\mathfrak{s}_x \times 2^{-M+\mathfrak{e}_{\min}}) = 2^{-M+\mathfrak{e}_{\min}} & \text{if} \quad 1 \le x < 1+2^{-1}. \end{cases}$

If $n_\eta > 1$, we have $2 \le n_\eta \le 2^{1+M} - 1$. Since x is normal, we can write x as $x = n_x \times 2^{-M+\mathfrak{e}_x}$ for some $2^M \le n_x < 2^{1+M}, n_x \in \mathbb{N}$.

By dividing $2^M \times n_\eta$ by n_x, we have

$$2^M \times n_\eta = m \times n_x + r \quad (0 \le r < n_x),$$

for some $1 \le m \le 2^{1+M} - 1$.

Let

$$(n_{y_1}, y_2) = \begin{cases} (m, 0) & \text{if} \quad r < 2^{M-1} \text{ or } r = 2^{M-1}, n_\eta \equiv 0 \pmod 2 \\ (m, y_*) & \text{if} \quad 2^{M-1} < r < 3 \times 2^{M-1} \text{ or } r \in \{2^{M-1}, 3 \times 2^{M-1}\}, n_\eta \equiv 1 \pmod 2 \\ (m+1, 0) & \text{if} \quad r > 3 \times 2^{M-1} \text{ or } r = 3 \times 2^{M-1}, n_\eta \equiv 0 \pmod 2 \end{cases}$$

and $y_1 = n_{y_1} \times 2^{-M+\mathfrak{e}_{\min}-\mathfrak{e}_x} \in \mathbb{F}$. We consider the following cases.

Case 1: $r < 2^{M-1}$ **or** $r = 2^{M-1}, n_\eta \equiv 0 \pmod 2$**.**

In this case, we have

$$y_1 \otimes x = \mathrm{rnd}((n_{y_1} \times n_x \times 2^{-M}) \times 2^{-M+\mathfrak{e}_{\min}}) = \mathrm{rnd}((n_\eta - r \times 2^{-M}) \times 2^{-M+\mathfrak{e}_{\min}}) = \eta.$$

Case 2: $2^{M-1} < r < 3 \times 2^{M-1}$ **or** $r \in \{2^{M-1}, 3 \times 2^{M-1}\}, n_\eta \equiv 1 \pmod 2$**.**

In this case, we have

$$\begin{aligned} (y_1 \otimes x) \oplus (y_2 \otimes x) &= \mathrm{rnd}((n_\eta - r \times 2^{-M}) \times 2^{-M+\mathfrak{e}_{\min}}) \oplus 2^{-M+\mathfrak{e}_{\min}} \\ &= (n_\eta - 1) \times 2^{-M+\mathfrak{e}_{\min}} \oplus 2^{-M+\mathfrak{e}_{\min}} = \eta. \end{aligned}$$

Case 3: $r > 3 \times 2^{M-1}$ or $r = 3 \times 2^{M-1}, n_\eta \equiv 0 \pmod 2$.

In this case, since $3 \times 2^{M-1} \le r, n_x < 2^{M+1}$, we have $0 \le n_x - r < 2^{M-1}$. Therefore,

$$(y_1 \otimes x) = \mathrm{rnd}((n_\eta + (n_x - r) \times 2^{-M}) \times 2^{-M+\mathfrak{e}_{\min}}) = \eta.$$

□

Lemma 16. *Let $K \in [(1 + 2^{-M+1}) \times 2^{-M-2}, 1 + 2^{-2} - 2^{-M}]_{\mathbb{F}}$. Consider $\mathfrak{e}_\zeta \in \mathbb{Z}$ such that $\mathfrak{e}_{\min} + 1 \le \mathfrak{e}_\zeta \le \mathfrak{e}_{\max} - M - 1$. For $n \in \mathbb{N}$, $i \in [n]$, and $0 < \alpha_i \in \mathbb{F}$, define $f : \mathbb{F} \to \mathbb{F}$ as*

$$f(x) \coloneqq x \oplus \bigoplus_{i=1}^{n} (\alpha_i \otimes K)\,.$$

Then, there exists n and α_is such that one of the following statements holds:

$$f^\sharp(\langle -1.1 \times 2^{\mathfrak{e}_\zeta}, 0\rangle) \subset \langle -(2^{\mathfrak{e}_\zeta})^-, (2^{\mathfrak{e}_\zeta})^-\rangle,\ f^\sharp(\langle \omega, 1.1 \times 2^{\mathfrak{e}_\zeta}\rangle) \subset \langle 2^{\mathfrak{e}_\zeta}, (1.1\times 2^{\mathfrak{e}_\zeta+1})^-\rangle, \tag{40}$$

or

$$f^\sharp(\langle -1.1 \times 2^{\mathfrak{e}_\zeta}, 0\rangle) \subset \langle -(2^{\mathfrak{e}_\zeta})^-, 2^{\mathfrak{e}_\zeta}\rangle,\ f^\sharp(\langle \omega, 1.1 \times 2^{\mathfrak{e}_\zeta}\rangle) \subset \langle (2^{\mathfrak{e}_\zeta})^+, (1.1\times 2^{\mathfrak{e}_\zeta+1})^-\rangle. \tag{41}$$

If $\mathfrak{e}_\zeta \le \mathfrak{e}_{\min}$, for each statement of Eq. (40) or Eq. (41), there exist n and α_is such that satisfy each equation, respectively.
Additionally, if $2^{\mathfrak{e}_\zeta} \le x \in \mathbb{F}$, then we have

$$f(x) \le x \oplus (2^{\mathfrak{e}_\zeta})^+\,. \tag{42}$$

Proof. Since f is a sum of increasing functions, we only need to consider the endpoints of $\langle -1.1 \times 2^{\mathfrak{e}_\zeta}, 0\rangle$ and $\langle \omega, 1.1 \times 2^{\mathfrak{e}_\zeta}\rangle$, namely $x = 0, -1.1 \times 2^{\mathfrak{e}_\zeta}, 1.1 \times 2^{\mathfrak{e}_\zeta}$, and ω. In other words, we suffice to show

$$\begin{cases} f(-1.1 \times 2^{\mathfrak{e}_\zeta}) > -2^{\mathfrak{e}_\zeta} \\ f(0) < 2^{\mathfrak{e}_\zeta} \\ f(\omega) \ge 2^{\mathfrak{e}_\zeta} \\ f(1.1 \times 2^{\mathfrak{e}_\zeta}) < 1.1 \times 2^{\mathfrak{e}_\zeta+1}. \end{cases} \tag{43}$$

or

$$\begin{cases} f(-1.1 \times 2^{\mathfrak{e}_\zeta}) > -2^{\mathfrak{e}_\zeta} \\ f(0) \le 2^{\mathfrak{e}_\zeta} \\ f(\omega) > 2^{\mathfrak{e}_\zeta} \\ f(1.1 \times 2^{\mathfrak{e}_\zeta}) < 1.1 \times 2^{\mathfrak{e}_\zeta+1}. \end{cases} \tag{44}$$

Let K be represented as

$$K = \mathfrak{s}_K \times 2^{\mathfrak{e}_K}, \quad -M-2 \le \mathfrak{e}_K \le 0.$$

By Lemma 14, at least one of $\mathfrak{s}_K^{\|} \in (2^{-1}, 1]_{\mathbb{F}}$ or $\mathfrak{s}_K^{\dagger} \in [2^{-1}, 1)_{\mathbb{F}}$ exists such that

$$\mathfrak{s}_K^{\dagger} \otimes \mathfrak{s}_K = 1^- = 1 - 2^{-M-1} = 0.\underbrace{1\dots1}_{M+1 \text{ times}},$$
$$\mathfrak{s}_K^{\|} \otimes \mathfrak{s}_K = 1.$$

We consider the following cases.

Case 1: $\mathfrak{s}_K^{\dagger}$ exists.

In this case, we will show Eq. (43). First, define $m_1 \in \mathbb{N}_{\geq 0}$ as

$$m_1 := \left\lceil \frac{-1 + \mathfrak{e}_\zeta - \mathfrak{e}_{\min}}{M+2} \right\rceil_{\mathbb{Z}} \in \mathbb{N}_{\geq 0}.$$

Then m_1 is the unique non-negative natural number satisfying

$$\mathfrak{e}_\zeta - (M+2)(m_1+1) \leq \mathfrak{e}_{\min} - M - 1 < \mathfrak{e}_\zeta - (M+2)\,m_1,$$

or equivalently,

$$\mathfrak{e}_{\min} - M \leq \mathfrak{e}_\zeta - (M+2)\,m_1 \leq \mathfrak{e}_{\min} + 1, \tag{45}$$

where the first inequality is due to $\mathfrak{e}_{\min} - M - 1 < \mathfrak{e}_\zeta - (M+2)\,m_1$ and $\mathfrak{e}_\zeta - (M+2)\,m_1 \in \mathbb{Z}$.

For $i \in [m_1]$, we define β_i as

$$\beta_i := (2 - 2^{-M}) \times 2^{\mathfrak{e}_\zeta - (M+2)(m_1-i)-1} = 0.\underbrace{1\dots1}_{M+1 \text{ times}} \times 2^{\mathfrak{e}_\zeta - (M+2)(m_1-i)}.$$

As

$$\begin{aligned}\mathfrak{e}_{\min} \leq \mathfrak{e}_{\beta_i} &= \mathfrak{e}_\zeta - (M+2)(m_1-1) - 1 = \mathfrak{e}_\zeta - (M+2)\,m_1 + M + 1\\ &\leq \mathfrak{e}_\zeta - (M+2)(m_1-i) - 1 \leq \mathfrak{e}_\zeta - 1 \leq \mathfrak{e}_{\max} - M - 2,\end{aligned}$$

we have $\beta_i \in \mathbb{F}$. Define α_i' for $i \in [m_1]$ as

$$\alpha_i' := \mathfrak{s}_K^{\dagger} \times 2^{\mathfrak{e}_\zeta - (M+2)(m_1-i) - \mathfrak{e}_K}.$$

Since $\mathfrak{s}_K^{\dagger} \in \left[\frac{1}{2}, 1\right)_{\mathbb{F}}$, we have $\mathfrak{s}_{\alpha_i'} = 2\mathfrak{s}_K^{\dagger}$ and $\mathfrak{e}_{\alpha_i'} = \mathfrak{e}_\zeta - (M+2)(m_1-i) - 1$. Since $0 \leq -\mathfrak{e}_K \leq M+2$, we have

$$\mathfrak{e}_{\min} \leq \mathfrak{e}_{\alpha_i'} = \mathfrak{e}_\zeta - (M+2)(m_1-i) - \mathfrak{e}_K - 1 \leq \mathfrak{e}_{\max}.$$

Therefore, we have $\alpha_i' \in \mathbb{F}$, and

$$\alpha_i' \otimes K = \beta_i. \tag{46}$$

Consider $2^{\mathfrak{e}_\varsigma-(M+2)m_1}-\omega$. By Eq. (45), $2^{\mathfrak{e}_\varsigma-(M+2)m_1}$ is subnormal and

$$2^{\mathfrak{e}_\varsigma-(M+2)m_1}-\omega = 1.\underbrace{1\dots1}_{\mathfrak{e}_\varsigma-(M+2)m_1-1-\mathfrak{e}_{\min}+M\text{ times}}\times 2^{\mathfrak{e}_\varsigma-(M+2)m_1-1}\in\mathbb{F}.$$

Since $2^{\mathfrak{e}_\varsigma-(M+2)m_1}-\omega\in(-2^{1+\mathfrak{e}_{\min}},2^{1+\mathfrak{e}_{\min}})_{\mathbb{F}}$, by Lemma 15, there exist $\widetilde{\beta}_1,\widetilde{\beta}_2\in\mathbb{F}$ such that

$$2^{\mathfrak{e}_\varsigma-(M+2)m_1}-\omega=\left(\widetilde{\beta}_1\otimes K\right)\oplus\left(\widetilde{\beta}_2\otimes K\right). \tag{47}$$

We define $f_1(x)$ as

$$f_1(x)\coloneqq\begin{cases}x\oplus\left(\widetilde{\beta}_1\otimes K\right)\oplus\left(\widetilde{\beta}_2\otimes K\right) & \text{if}\quad m_1=0\\ x\oplus\left(\widetilde{\beta}_1\otimes K\right)\oplus\left(\widetilde{\beta}_2\otimes K\right)\oplus\bigoplus_{i=1}^{m_1}\alpha_i'\otimes K & \text{if}\quad m_1\ge 1\end{cases}$$

By Eq. (46) we have

$$f_1(x)=\begin{cases}x\oplus\left(\widetilde{\beta}_1\otimes K\right)\oplus\left(\widetilde{\beta}_2\otimes K\right) & \text{if}\quad m_1=0\\ x\oplus\left(\widetilde{\beta}_1\otimes K\right)\oplus\left(\widetilde{\beta}_2\otimes K\right)\oplus\bigoplus_{i=1}^{m_1}\beta_i & \text{if}\quad m_1\ge 1\end{cases}$$

Next, we present the following claim.

<u>Claim 1-1:</u> For any $k\in[m_1]$, we have

$$\left(\widetilde{\beta}_1\otimes K\right)\oplus\left(\widetilde{\beta}_2\otimes K\right)\oplus\bigoplus_{i=1}^{k}\beta_i=\beta_k.$$

We show the claim using the induction on k.

Base step ($k=1$):

By Eq. (47), we have

$$\left(\widetilde{\beta}_1\otimes K\right)\oplus\left(\widetilde{\beta}_2\otimes K\right)\oplus\beta_1=\left(2^{\mathfrak{e}_\varsigma-(M+2)m_1}-\omega\right)\oplus 0.\underbrace{1\dots1}_{M+1\text{ times}}\times 2^{\mathfrak{e}_\varsigma-(M+2)(m_1-1)}$$

$$=\mathrm{rnd}(0.\underbrace{1\dots1}_{M+1\text{ times}}\;0\;\underbrace{1\dots1}_{\mathfrak{e}_\varsigma-(M+2)m_1-1-\mathfrak{e}_{\min}+M\text{ times}}\times 2^{\mathfrak{e}_\varsigma-(M+2)(m_1-1)})$$

$$=0.\underbrace{1\dots1}_{M+1\text{ times}}\times 2^{\mathfrak{e}_\varsigma-(M+2)(m_1-1)}=\beta_1.$$

Induction step:

Assume that the induction hypothesis is satisfied for k. Then we have

$$\bigoplus_{i=1}^{k+1}\beta_i=\beta_k\oplus\beta_{k+1}=\mathrm{rnd}(0.\underbrace{1\dots1}_{M+1\text{ times}}\;0\;\underbrace{1\dots1}_{M+1\text{ times}}\times 2^{\mathfrak{e}_\varsigma-(M+2)(m_1-k-1)})$$

$$=0.\underbrace{1\dots1}_{M+1\text{ times}}\times 2^{\mathfrak{e}_\varsigma-(M+2)(m_1-k-1)}=\beta_{k+1}.$$

Therefore, we prove the claim for any $k \in [m_1]$.

Thus, we have

$$f_1(0) = \begin{cases} \left(\widetilde{\beta}_1 \otimes K\right) \oplus \left(\widetilde{\beta}_2 \otimes K\right) = 2^{\mathfrak{e}_\zeta} - \omega < 2^{\mathfrak{e}_\zeta} & \text{if} \quad m_1 = 0, \\ \beta_{m_1} = 0.\underbrace{1\ldots1}_{M+1 \text{ times}} \times 2^{\mathfrak{e}_\zeta} < 2^{\mathfrak{e}_\zeta} & \text{if} \quad m_1 \geq 1. \end{cases} \tag{48}$$

Now, we present the following claim.

<u>**Claim 1-2:**</u> For any $k \in [m_1]$, we have

$$\omega \oplus \left(\widetilde{\beta}_1 \otimes K\right) \oplus \left(\widetilde{\beta}_2 \otimes K\right) \oplus \bigoplus_{i=1}^{k} \beta_i \geq 2^{\mathfrak{e}_\zeta - (M+2)(m_1 - k)}. \tag{49}$$

We show the claim using the induction on k.

Base step ($k = 1$):
As $\widetilde{\beta}_1 \otimes K < 2^{\mathfrak{e}_{\min}+1}$, the summation $\omega \oplus (K \otimes \widetilde{\beta}_1)$ is exact. Thus,

$$\omega \oplus \left(\widetilde{\beta}_1 \otimes K\right) \oplus \left(\widetilde{\beta}_2 \otimes K\right) = \left(\omega + K \otimes \widetilde{\beta}_1\right) \oplus \left(K \otimes \widetilde{\beta}_2\right) = \mathrm{rnd}(\omega + K \otimes \widetilde{\beta}_1 + K \otimes \widetilde{\beta}_2)_{\mathbb{F}}$$
$$= \mathrm{rnd}(\omega + \left(K \otimes \widetilde{\beta}_1 \oplus K \otimes \widetilde{\beta}_2\right))_{\mathbb{F}} = 2^{\mathfrak{e}_\zeta - (M+2)m_1}.$$

Therefore, we have

$$\omega \oplus \left(\widetilde{\beta}_1 \otimes K\right) \oplus \left(\widetilde{\beta}_2 \otimes K\right) \oplus \beta_1 = \mathrm{rnd}(0.\underbrace{1\ldots1}_{M+2 \text{ times}} \times 2^{\mathfrak{e}_\zeta - (M+2)(m_1-1)})$$
$$= 2^{\mathfrak{e}_\zeta - (M+2)(m_1-1)}.$$

Induction step:
Assume that the induction hypothesis is satisfied for k and consider the case of $k+1$. By the induction hypothesis,

$$\omega \oplus \left(\widetilde{\beta}_1 \otimes K\right) \oplus \left(\widetilde{\beta}_2 \otimes K\right) \oplus \bigoplus_{i=1}^{k+1} \beta_i \geq 2^{\mathfrak{e}_\zeta - (M+2)(m_1-k)} \oplus \beta_{k+1}$$
$$= \mathrm{rnd}(0.\underbrace{1\ldots1}_{M+2 \text{ times}} \times 2^{\mathfrak{e}_\zeta - (M+2)(m_1-k-1)}) = 2^{\mathfrak{e}_\zeta - (M+2)(m_1-k-1)}.$$

Thus, the induction hypothesis holds for any $k \in [m_1]$, which proves the claim. If $k = m_1$ in Eq. (49), we have

$$f_1(\omega) = \begin{cases} \omega \oplus \left(\widetilde{\beta}_1 \otimes K\right) \oplus \left(\widetilde{\beta}_2 \otimes K\right) = 2^{\mathfrak{e}_\zeta} \geq 2^{\mathfrak{e}_\zeta} & \text{if} \quad m_1 = 0, \\ \omega \oplus \left(\widetilde{\beta}_1 \otimes K\right) \oplus \left(\widetilde{\beta}_2 \otimes K\right) \oplus \bigoplus_{i=1}^{m_1} \beta_i \geq 2^{\mathfrak{e}_\zeta} & \text{if} \quad m_1 \geq 1. \end{cases} \tag{50}$$

For $x = 1.1 \times 2^{\mathfrak{e}_\zeta}$, we consider three cases with respect to $\mathfrak{e}_\zeta$: $\mathfrak{e}_\zeta \geq \mathfrak{e}_{\min} + 2$, $\mathfrak{e}_\zeta = \mathfrak{e}_{\min} + 1$, and $\mathfrak{e}_\zeta \leq \mathfrak{e}_{\min}$.

If $\mathfrak{e}_\zeta \geq \mathfrak{e}_{\min}+2$, we have $m_1 \geq 1$. Since $\widetilde{\beta}_1 \otimes K, \widetilde{\beta}_2 \otimes K, \beta_i < 2^{\mathfrak{e}_\zeta} \times 2^{-M-1}$ for $i \in [m_1-1]$, we have

$$\begin{aligned} f_1\left(1.1 \times 2^{\mathfrak{e}_\zeta}\right) &= 1.1 \times 2^{\mathfrak{e}_\zeta} \oplus \left(\widetilde{\beta}_1 \otimes K\right) \oplus \left(\widetilde{\beta}_2 \otimes K\right) \oplus \bigoplus_{i=1}^{m_1} \beta_i = 1.1 \times 2^{\mathfrak{e}_\zeta} \oplus \beta_{m_1} \\ &= 1.1 \times 2^{\mathfrak{e}_\zeta} \oplus 0.\underbrace{1\ldots1}_{M+1 \text{ times}} \times 2^{\mathfrak{e}_\zeta} \leq 1.01 \times 2^{\mathfrak{e}_\zeta+1} < 1.1 \times 2^{\mathfrak{e}_\zeta+1}. \end{aligned} \tag{51}$$

If $\mathfrak{e}_\zeta = \mathfrak{e}_{\min}+1$, we have $m_1 = 0$, which leads to $\left(\widetilde{\beta}_1 \otimes K\right) \oplus \left(\widetilde{\beta}_2 \otimes K\right) = 2^{\mathfrak{e}_\zeta} - \omega$ by Eq. (47). Therefore we have,

$$\begin{aligned} f_1\left(1.1 \times 2^{\mathfrak{e}_\zeta}\right) &= 1.1 \times 2^{\mathfrak{e}_\zeta} \oplus \left(\widetilde{\beta}_1 \otimes K\right) \oplus \left(\widetilde{\beta}_2 \otimes K\right) \\ &\leq \operatorname{rnd}\left(1.1 \times 2^{\mathfrak{e}_\zeta} + \left(\widetilde{\beta}_1 \otimes K\right) + \left(\widetilde{\beta}_2 \otimes K\right) + 2\omega\right)_{\mathbb{F}} \\ &\leq \operatorname{rnd}(1.01 \times 2^{\mathfrak{e}_\zeta+1} + 2\omega)_{\mathbb{F}} < 1.1 \times 2^{\mathfrak{e}_\zeta+1}. \end{aligned} \tag{52}$$

If $\mathfrak{e}_\zeta \leq \mathfrak{e}_{\min}$, we also have $m_1 = 0$ and $\left(\widetilde{\beta}_1 \otimes K\right) \oplus \left(\widetilde{\beta}_2 \otimes K\right) = 2^{\mathfrak{e}_\zeta} - \omega$. Since the summation is exact, we have

$$f_1\left(1.1 \times 2^{\mathfrak{e}_\zeta}\right) \leq 1.01 \times 2^{\mathfrak{e}_\zeta+1} < 1.1 \times 2^{\mathfrak{e}_\zeta+1}. \tag{53}$$

For $x = -1.1 \times 2^{\mathfrak{e}_\zeta}$, we consider three cases with respect to $\mathfrak{e}_\zeta$, that is, $\mathfrak{e}_\zeta \geq \mathfrak{e}_{\min}+2$, $\mathfrak{e}_\zeta = \mathfrak{e}_{\min}+1$, and $\mathfrak{e}_\zeta \leq \mathfrak{e}_{\min}$.

If $\mathfrak{e}_\zeta \geq \mathfrak{e}_{\min}+2$, we have $m_1 \geq 1$. Hence

$$\begin{aligned} f_1\left(-1.1 \times 2^{\mathfrak{e}_\zeta}\right) &\geq -1.1 \times 2^{\mathfrak{e}_\zeta} \oplus \left(\widetilde{\beta}_1 \otimes K\right) \oplus \left(\widetilde{\beta}_2 \otimes K\right) \oplus \bigoplus_{i=1}^{m_1} \beta_i \\ &\geq -1.1 \times 2^{\mathfrak{e}_\zeta} \oplus \beta_m = -1.1 \times 2^{\mathfrak{e}_\zeta} \oplus 0.\underbrace{1\ldots1}_{M+1 \text{ times}} \times 2^{\mathfrak{e}_\zeta} = -2^{\mathfrak{e}_\zeta-1}. \end{aligned} \tag{54}$$

If $\mathfrak{e}_\zeta = \mathfrak{e}_{\min}+1$, we have $m_1 = 0$. Therefore,

$$\begin{aligned} f_1\left(-1.1 \times 2^{\mathfrak{e}_\zeta}\right) &\geq -1.1 \times 2^{\mathfrak{e}_\zeta} \oplus \left(\widetilde{\beta}_1 \otimes K\right) \oplus \left(\widetilde{\beta}_2 \otimes K\right) \\ &\geq \operatorname{rnd}\left(-1.1 \times 2^{\mathfrak{e}_\zeta} + \left(\widetilde{\beta}_1 \otimes K\right) + \left(\widetilde{\beta}_2 \otimes K\right) - 2\omega\right)_{\mathbb{F}} > -2^{\mathfrak{e}_\zeta}. \end{aligned} \tag{55}$$

If $\mathfrak{e}_\zeta \leq \mathfrak{e}_{\min}$, we also have $m_1 = 0$. Since the summation is exact, we have

$$f_1\left(-1.1 \times 2^{\mathfrak{e}_\zeta}\right) = -2^{\mathfrak{e}_\zeta-1} - \omega > -2^{\mathfrak{e}_\zeta-1}. \tag{56}$$

By Eqs. (48) and (50)–(56), we show Eq. (43). Therefore we conclude that

$$f^\sharp\left(\langle -1.1 \times 2^{\mathfrak{e}_\zeta}, 0\rangle\right) \subset \langle -(2^{\mathfrak{e}_\zeta})^-, (2^{\mathfrak{e}_\zeta})^-\rangle,\; f^\sharp\left(\langle \omega, 1.1 \times 2^{\mathfrak{e}_\zeta}\rangle\right) \subset \langle 2^{\mathfrak{e}_\zeta}, (1.1{\times}2^{\mathfrak{e}_\zeta+1})^-\rangle,$$

To show Eq. (42), we can apply by similar argument to $x = 1.1 \times 2^{\mathfrak{e}_\zeta}$. If $\mathfrak{e}_\zeta \geq \mathfrak{e}_{\min} + 2$, we have

$$f_1(x) = x \oplus \beta_m \leq x \oplus 2^{\mathfrak{e}_\zeta} \leq x \oplus (2^{\mathfrak{e}_\zeta})^+.$$

If $\mathfrak{e}_\zeta \leq \mathfrak{e}_{\min} + 1$, we have

$$f_1(x) \leq \operatorname{rnd}(x \oplus \beta_m + 2\omega) \leq x \oplus (2^{\mathfrak{e}_\zeta})^+.$$

Case 2: $\mathfrak{s}_K^{\|}$ exists.

In this case, we show Eq. (44). Define $m_2 \in \mathbb{N}_{\geq 0}$ as

$$m_2 := \left\lfloor \frac{-\mathfrak{e}_{\min} + M + \mathfrak{e}_\zeta}{M+1} \right\rfloor_{\mathbb{Z}} \in \mathbb{N}_{\geq 1}.$$

Then m_2 is the unique non-negative natural number satisfying

$$\mathfrak{e}_\zeta - (M+1)\, m_2 < \mathfrak{e}_{\min} + 1 \leq \mathfrak{e}_\zeta - (M+1)\,(m_2 - 1),$$

or equivalently,

$$\mathfrak{e}_{\min} - M \leq \mathfrak{e}_\zeta - (M+1)\, m_2 < \mathfrak{e}_{\min} + 1. \tag{57}$$

For $i \in [m_2]$, define β_i as

$$\beta_i := 2^{\mathfrak{e}_\zeta - (M+1)(m_2 - i)}.$$

Then, as

$$\mathfrak{e}_{\min} + 1 \leq \mathfrak{e}_\zeta - (M+1)\,(m-1) \leq \mathfrak{e}_{\beta_i} = \mathfrak{e}_\zeta - (M+1)\,(m-i) \leq \mathfrak{e}_\zeta \leq \mathfrak{e}_{\max} - M - 1,$$

we have $\beta_i \in \mathbb{F}$. For $i \in [n]$, define α_i as

$$\alpha_i' := \begin{cases} \mathfrak{s}_K^{\|} \times 2^{\mathfrak{e}_\zeta - (M+1)(m_2 - i) - \mathfrak{e}_K} & \text{if} \quad 2^{-1} \leq \mathfrak{s}_K^{\|} < 1 \\ 2^{\mathfrak{e}_\zeta - (M+1)(m_2 - i) - \mathfrak{e}_K - 1} & \text{if} \quad \mathfrak{s}_K^{\|} = 1 \end{cases}$$

As $0 \leq -\mathfrak{e}_K \leq M + 2$, we have

$$\mathfrak{e}_{\min} \leq \mathfrak{e}_{\alpha_i'} = \mathfrak{e}_\zeta - (M+1)\,(m_2 - i) - \mathfrak{e}_K - 1 \leq \mathfrak{e}_{\max}.$$

Hence we have $\alpha_i' \in \mathbb{F}$.

Consider $2^{\mathfrak{e}_\zeta - (M+1)m_2}$. Since $\mathfrak{e}_{\min} - M \leq \mathfrak{e}_\zeta - (M+1)\, m_2 \leq \mathfrak{e}_{\min}$, we have $2^{\mathfrak{e}_\zeta - (M+1)m_2} \in \mathbb{F}$ and $2^{\mathfrak{e}_\zeta - (M+1)m_2} \in (-2^{1+\mathfrak{e}_{\min}}, 2^{1+\mathfrak{e}_{\min}})_{\mathbb{F}}$. By Lemma 15, there exist $\widetilde{\beta}_1, \widetilde{\beta}_2 \in \mathbb{F}$ such that

$$2^{\mathfrak{e}_\zeta - (M+1)m_2} = \left(\widetilde{\beta}_1 \otimes K\right) \oplus \left(\widetilde{\beta}_2 \otimes K\right).$$

Define $f_2(x)$ as

$$\begin{aligned} f_2(x) &:= x \oplus \left(\widetilde{\beta}_1 \otimes K\right) \oplus \left(\widetilde{\beta}_2 \otimes K\right) \oplus \bigoplus_{i=1}^{m_2} \alpha_i' \otimes K \\ &= x \oplus \left(\widetilde{\beta}_1 \otimes K\right) \oplus \left(\widetilde{\beta}_2 \otimes K\right) \oplus \bigoplus_{i=1}^{m_2} \beta_i. \end{aligned}$$

We present the following claim.

Claim 2-1: For any $k \in [m_2]$, we have

$$\left(\widetilde{\beta}_1 \otimes K\right) \oplus \left(\widetilde{\beta}_2 \otimes K\right) \oplus \bigoplus_{i=1}^{k} \beta_i = \beta_k.$$

We show the claim using the induction on k.

Base step ($k = 1$):

$$\begin{aligned}\left(\widetilde{\beta}_1 \otimes K\right) \oplus \left(\widetilde{\beta}_2 \otimes K\right) \oplus \beta_1 &= 2^{\mathfrak{e}_\zeta - (M+1)m_2} \oplus 2^{\mathfrak{e}_\zeta - (M+1)(m_2-1)} \\ &= 2^{\mathfrak{e}_\zeta - (M+1)(m_2-1)} = \beta_1.\end{aligned}$$

Induction step:

Assume that the induction hypothesis is satisfied for k. Then we have

$$\begin{aligned}&\left(\widetilde{\beta}_1 \otimes K\right) \oplus \left(\widetilde{\beta}_2 \otimes K\right) \oplus \bigoplus_{i=1}^{k+1} \beta_i = \beta_k \oplus \beta_{k+1} \\ &= \mathrm{rnd}(1.\underbrace{0\ldots0}_{M\text{ times}}1 \times 2^{\mathfrak{e}_\zeta - (M+1)(m_2-k-1)}) = 2^{\mathfrak{e}_\zeta - (M+1)(m_2-k-1)} = \beta_{k+1}.\end{aligned}$$

Therefore the induction hypothesis is satisfied for any $k \le m$, and we prove the claim. Thus,

$$f_2(0) = \beta_{m_2} = 2^{\mathfrak{e}_\zeta}. \tag{58}$$

Now we will show that for $f_2(\omega) > 2^{\mathfrak{e}_\zeta}$. We present the following claim.

Claim 2-2: For any $k \in [m_2]$, we have

$$\omega \oplus \left(\widetilde{\beta}_1 \otimes K\right) \oplus \left(\widetilde{\beta}_2 \otimes K\right) \oplus \bigoplus_{i=1}^{k} \beta_i \ge \beta_k^+.$$

We show the claim using the induction on k.

Base step ($k = 1$):

Since $\omega \oplus \left(\widetilde{\beta}_1 \otimes K\right)$ is exact, we have

$$\begin{aligned}\omega \oplus \left(\widetilde{\beta}_1 \otimes K\right) \oplus \left(\widetilde{\beta}_2 \otimes K\right) &= \left(\omega + K \otimes \widetilde{\beta}_1\right) \oplus \left(K \otimes \widetilde{\beta}_2\right) \\ =&\mathrm{rnd}(\omega + K \otimes \widetilde{\beta}_1 + K \otimes \widetilde{\beta}_2)_{\mathbb{F}} = 2^{\mathfrak{e}_\zeta - (M+1)m_2} + \omega > 2^{\mathfrak{e}_\zeta - (M+1)m_2}.\end{aligned}$$

Thus,

$$\begin{aligned}&\omega \oplus \left(\widetilde{\beta}_1 \otimes K\right) \oplus \left(\widetilde{\beta}_2 \otimes K\right) \oplus \beta_1 \ge \left(2^{\mathfrak{e}_\zeta - (M+1)m_2} + \omega\right) \oplus 2^{\mathfrak{e}_\zeta - (M+1)(m_2-1)} \\ &= \left(2^{\mathfrak{e}_\zeta - (M+1)(m_2-1)}\right)^+ = \beta_1^+.\end{aligned}$$

Induction step: Assume that the induction hypothesis is satisfied for k and consider the case of $k+1$.

$$\begin{aligned}
&\omega \oplus \left(\widetilde{\beta}_1 \otimes K\right) \oplus \left(\widetilde{\beta}_2 \otimes K\right) \oplus \bigoplus_{i=1}^{k+1} \beta_i \geq \beta_k^+ \oplus \beta_{k+1} \\
&= 1.\underbrace{0\ldots0}_{M-1 \text{ times}}1 \times 2^{\mathfrak{e}_\zeta-(M+1)(m-k)} \oplus 2^{\mathfrak{e}_\zeta-(M+1)(m-k-1)} \\
&= \mathrm{rnd}(1.\underbrace{0\ldots0}_{M \text{ times}}1\underbrace{0\ldots0}_{M-1 \text{ times}}1 \times 2^{\mathfrak{e}_\zeta-(M+1)(m-k-1)}) \\
&= (1+2^{-M}) \times 2^{\mathfrak{e}_\zeta-(M+1)(m-k-1)} = \beta_{k+1}^+.
\end{aligned}$$

Thus, the induction hypothesis is satisfied for any $k \in [m_2]$ and we prove the claim. Therefore, we have

$$f_2(\omega) = \omega \oplus \left(\widetilde{\beta}_1 \otimes K\right) \oplus \left(\widetilde{\beta}_2 \otimes K\right) \oplus \bigoplus_{i=1}^{m} \beta_i \geq \beta_m^+ > \beta_m = 2^{\mathfrak{e}_\zeta}. \quad (59)$$

Now, we show Eq. (44) for $x = 1.1 \times 2^{\mathfrak{e}_\zeta}$ and $x = -1.1 \times 2^{\mathfrak{e}_\zeta}$. We consider two cases with respect to $\mathfrak{e}_\zeta$: $\mathfrak{e}_\zeta \leq \mathfrak{e}_{\min}$ and $\mathfrak{e}_\zeta \geq \mathfrak{e}_{\min}+1$. If $\mathfrak{e}_\zeta \leq \mathfrak{e}_{\min}$, since the summation exact, we have the desired results.
If $\mathfrak{e}_\zeta \geq \mathfrak{e}_{\min}+1$ and $x = 1.1 \times 2^{\mathfrak{e}_\zeta}$, since $\beta_i < 2^{\mathfrak{e}_\zeta} \times 2^{-M-1}$ for $i \in [m_2-2]$, we have

$$\begin{aligned}
f_2\left(1.1 \times 2^{\mathfrak{e}_\zeta}\right) &= \left(1.1 \times 2^{\mathfrak{e}_\zeta}\right) \oplus \beta_{n-1} \oplus \beta_n = 1.1 \times 2^{\mathfrak{e}_\zeta} \oplus 2^{\mathfrak{e}_\zeta-M-1} \oplus 2^{\mathfrak{e}_\zeta} \\
&= \left(1.1 \times 2^{\mathfrak{e}_\zeta}\right)^+ \oplus 2^{\mathfrak{e}_\zeta} < 1.1 \times 2^{\mathfrak{e}_\zeta+1}. \quad (60)
\end{aligned}$$

If $\mathfrak{e}_\zeta \geq \mathfrak{e}_{\min}+1$ and $x = -1.1 \times 2^{\mathfrak{e}_\zeta}$, we have

$$f_2\left(-1.1 \times 2^{\mathfrak{e}_\zeta}\right) \geq -1.1 \times 2^{\mathfrak{e}_\zeta} \oplus \bigoplus_{i=1}^{m_2} \beta_i \geq -1.1 \times 2^{\mathfrak{e}_\zeta} \oplus \beta_m = -2^{\mathfrak{e}_\zeta-1} > -2^{\mathfrak{e}_\zeta}. \quad (61)$$

Due to Eqs. (58)–(61), we show Eq. (44). Therefore we conclude that

$$f^\sharp\left(\langle -1.1 \times 2^{\mathfrak{e}_\zeta}, 0\rangle\right) \subset \langle -(2^{\mathfrak{e}_\zeta})^-, 2^{\mathfrak{e}_\zeta}\rangle,\ f^\sharp\left(\langle \omega, 1.1 \times 2^{\mathfrak{e}_\zeta}\rangle\right) \subset \langle (2^{\mathfrak{e}_\zeta})^+, (1.1{\times}2^{\mathfrak{e}_\zeta+1})^-\rangle.$$

To show Eq. (42), we consider two cases: $2^{\mathfrak{e}_\zeta} \leq x < 2^{\mathfrak{e}_\zeta+1}$ and $x \geq 2^{\mathfrak{e}_\zeta+1}$.
If $2^{\mathfrak{e}_\zeta} \leq x < 2^{\mathfrak{e}_\zeta+1}$, we have

$$f_2(x) = x \oplus \beta_{m_2-1} \oplus \beta_{m_2} \leq x^+ \oplus 2^{\mathfrak{e}_\zeta} = x \oplus \left(2^{\mathfrak{e}_\zeta}\right)^+.$$

If $x \geq 2^{\mathfrak{e}_\zeta+1}$, we have

$$f_2(x) = x \oplus \beta_{m_2} \leq x \oplus 2^{\mathfrak{e}_\zeta} \leq x \oplus \left(2^{\mathfrak{e}_\zeta}\right)^+.$$

Hence we show Eq. (42).

Finally, note that in both cases, if $\mathfrak{e}_\zeta \le \mathfrak{e}_{\min}$, then, $m_1 = 0$ and $m_2 = 0$, which implies that we do not need the existence of $\mathfrak{s}_K^{\dagger}$ and $\mathfrak{s}_K^{\|}$ in the definition of $f_1(x)$ and $f_2(x)$. Therefore, if $\mathfrak{e}_\zeta \le \mathfrak{e}_{\min}$, for both statements Eq. (40) and Eq. (41), there exists n and α_i such that satisfying the statements.

This completes the proof. □

Lemma 17. *Suppose that $\sigma : \mathbb{F} \to \mathbb{F}$ satisfies Condition 2. Assume that there exists an integer $e_0 \in \mathbb{Z}$ such that $2^{e_0} \le |\sigma(\eta^+) - \sigma(\eta)| < 2^{e_0+1}$ and define $\mathfrak{e}_\theta$ as*

$$\mathfrak{e}_\theta \coloneqq \max\left(\mathfrak{e}_{\min} - M, -e_0 + \mathfrak{e}_{\min} - M + 1\right).$$

Suppose that $\mathfrak{e}_\theta \le -3$, $e_0 \le \mathfrak{e}_\eta - M - 3 - \mathfrak{e}_\theta$.
If $\sigma(\eta) < \sigma(\eta^+)$, define $\theta, \mathcal{I}, \mathcal{I}^+$ as

$$\theta \coloneqq 2^{\mathfrak{e}_\theta}, \mathcal{I} \coloneqq \langle \sigma(\eta) - 2^{\mathfrak{e}_\zeta - \mathfrak{e}_\theta}, \sigma(\eta)\rangle, \text{ and } \mathcal{I}^+ \coloneqq \langle \sigma(\eta^+), \sigma(\eta) + 2^{\mathfrak{e}_\zeta - \mathfrak{e}_\theta}\rangle,$$

and if $\sigma(\eta) > \sigma(\eta^+)$, define $\theta, \mathcal{I}, \mathcal{I}^+$ as

$$\theta \coloneqq -2^{\mathfrak{e}_\theta}, \mathcal{I}^+ \coloneqq \langle \sigma(\eta) - 2^{\mathfrak{e}_\zeta - \mathfrak{e}_\theta}, \sigma\left(\eta^+\right)\rangle, \text{ and } \mathcal{I} \coloneqq \langle \sigma(\eta), \sigma(\eta) + 2^{\mathfrak{e}_\zeta - \mathfrak{e}_\theta}\rangle,$$

where

$$\mathfrak{e}_\zeta \coloneqq \begin{cases} \mathfrak{e}_\eta - M - 1 & \text{if } \eta > 0 \text{ or } \eta < 0, \eta \ne -2^{\mathfrak{e}_\eta}, \\ \mathfrak{e}_\eta - M - 2 & \text{if } \eta < 0, \eta = -2^{\mathfrak{e}_\eta}. \end{cases}$$

Then, there exists $k \in \mathbb{N}$ and $\alpha_1, \ldots, \alpha_k, z_1, \ldots, z_k \in \mathbb{F}$ such that for

$$f(x) = (\theta \otimes x) \oplus \bigoplus_{i=1}^{k} (\alpha_i \otimes \sigma(z_i)) \oplus \eta,$$

the followings hold:

$$f^\sharp(\langle -\Omega, \Omega\rangle) \subset \langle -\Omega, \Omega\rangle, \quad f^\sharp(\mathcal{I}) = \langle \eta, \eta\rangle, \quad f^\sharp(\mathcal{I}^+) = \langle \eta^+, \eta^+\rangle. \tag{62}$$

In addition, if $\sigma(\eta) < \sigma(\eta^+)$

$$f(x) - \eta^+ \le \left(x - \sigma\left(\eta^+\right)\right) \times 2^{\mathfrak{e}_\theta + 2} \quad \text{for} \quad \sigma(\eta) + 2^{\mathfrak{e}_\zeta - \mathfrak{e}_\theta} \le x \in \mathbb{F}, \tag{63}$$

$$\eta - f(x) \le \left(\sigma\left(\eta\right) - x\right) \times 2^{\mathfrak{e}_\theta + 2} \quad \text{for} \quad \sigma(\eta) - 2^{\mathfrak{e}_\zeta - \mathfrak{e}_\theta} \ge x \in \mathbb{F}, \tag{64}$$

and if $\sigma(\eta) > \sigma(\eta^+)$, we have

$$\eta - f(x) \le \left(x - \sigma\left(\eta\right)\right) \times 2^{\mathfrak{e}_\theta + 2} \quad \text{for} \quad \sigma(\eta) + 2^{\mathfrak{e}_\zeta - \mathfrak{e}_\theta} \le x \in \mathbb{F}, \tag{65}$$

$$f(x) - \eta^+ \le \left(\sigma\left(\eta^+\right) - x\right) \times 2^{\mathfrak{e}_\theta + 2} \quad \text{for} \quad \sigma(\eta) - 2^{\mathfrak{e}_\zeta - \mathfrak{e}_\theta} \ge x \in \mathbb{F}, \tag{66}$$

Proof. First, $\mathcal{I}^+$ is not empty because $|\sigma(\eta^+) - \sigma(\eta)| < 2^{e_0+1} \le 2^{\mathfrak{e}_\eta - M - 2 - \mathfrak{e}_\theta} \le 2^{\mathfrak{e}_\zeta - \mathfrak{e}_\theta}$. Since f is monotone, we only need to consider the endpoints of $\mathcal{I}$ and $\mathcal{I}^+$.

Note that $2^{\mathfrak{e}_\zeta}, 2^{\mathfrak{e}_\zeta + 1} \in \mathbb{F}$ (since $\mathfrak{e}_\eta \ge 5 + \mathfrak{e}_{\min}$) and the following hold:

$$2^{\mathfrak{e}_\zeta + 1} = \eta^+ - \eta, \ \left(1.1 \times 2^{\mathfrak{e}_\zeta + 1}\right)^- \oplus \eta = \eta^+. \tag{67}$$

Now, represent η as

$$\eta = \mathfrak{s}_\eta \times 2^{\mathfrak{e}_\eta} = 1.\mathfrak{s}_{\eta,1} \dots \mathfrak{s}_{\eta,M} \times 2^{\mathfrak{e}_\eta}.$$

Then we have if $\mathfrak{s}_{\eta,M} = 0$, we have

$$\eta \oplus 2^{\mathfrak{e}_\zeta} = \eta,\ \eta \oplus (-2^{\mathfrak{e}_\zeta}) = \eta,\ \eta \oplus (2^{\mathfrak{e}_\zeta})^+ = \eta^+,\ \eta \oplus \left(-(2^{\mathfrak{e}_\zeta})^+\right) = \eta^-, \quad (68)$$

and if $\mathfrak{s}_{\eta,M} = 1$, we have

$$\eta \oplus (2^{\mathfrak{e}_\zeta})^- = \eta,\ \eta \oplus \left(-(2^{\mathfrak{e}_\zeta})^-\right) = \eta,\ \eta \oplus 2^{\mathfrak{e}_\zeta} = \eta^+,\ \eta \oplus (-2^{\mathfrak{e}_\zeta}) = \eta^-. \quad (69)$$

By Lemma 16, there exists $n \in \mathbb{N}$, $i \in [n]$, $\widetilde{\alpha_i} \in \mathbb{F}$ such that for $g : \mathbb{F} \to \mathbb{F}$ defined as

$$g(x) \coloneqq x \oplus \bigoplus_{i=1}^{n} (\widetilde{\alpha_i} \otimes K_\sigma),$$

one of the following statements holds:

$$g^\sharp(\langle -1.1 \times 2^{\mathfrak{e}_\zeta}, 0\rangle) \subset \langle -(2^{\mathfrak{e}_\zeta})^-, (2^{\mathfrak{e}_\zeta})^-\rangle,\ g^\sharp(\langle \omega, 1.1 \times 2^{\mathfrak{e}_\zeta}\rangle) \subset \langle 2^{\mathfrak{e}_\zeta}, (1.1\times 2^{\mathfrak{e}_\zeta+1})^-\rangle, \quad (70)$$

or

$$g^\sharp(\langle -1.1 \times 2^{\mathfrak{e}_\zeta}, 0\rangle) \subset \langle -(2^{\mathfrak{e}_\zeta})^-, 2^{\mathfrak{e}_\zeta}\rangle,\ g^\sharp(\langle \omega, 1.1 \times 2^{\mathfrak{e}_\zeta}\rangle) \subset \langle (2^{\mathfrak{e}_\zeta})^+, (1.1\times 2^{\mathfrak{e}_\zeta+1})^-\rangle. \quad (71)$$

with

$$g(x) \le x \oplus (2^{\mathfrak{e}_\zeta})^+ \quad \text{if} \quad 2^{\mathfrak{e}_\zeta} \le x \in \mathbb{F}. \quad (72)$$

Furthermore, if $\mathfrak{e}_\zeta \le \mathfrak{e}_{\min}$, there exist g_1 and g_2 such that they satisfy Eq. (70) and Eq. (71), respectively. Hence we pick g as $g = g_2$ if $\mathfrak{s}_{\eta,M} = 0$ and $g = g_1$ if $\mathfrak{s}_{\eta,M} = 1$.

If $\mathfrak{e}_\zeta \ge \mathfrak{e}_{\min}$, by Lemma 24, there exists $\beta \in \mathbb{F}$ such that the following inequality holds:

$$\frac{1}{2} \times 2^{\mathfrak{e}_\zeta - M} < \beta \otimes K_\sigma < \frac{5}{4} \times 2^{\mathfrak{e}_\zeta - M}.$$

Therefore we have

$$2^{\mathfrak{e}_\zeta} \le 2^{\mathfrak{e}_\zeta} \oplus (\beta \otimes K_\sigma) \le (2^{\mathfrak{e}_\zeta})^+,\ -\left((2^{\mathfrak{e}_\zeta})^+\right) \le -2^{\mathfrak{e}_\zeta} \oplus (\beta \otimes K_\sigma) \le -2^{\mathfrak{e}_\zeta}. \quad (73)$$

Define $\tilde{\beta} \in \mathbb{F}$ as

$$\tilde{\beta} \coloneqq \begin{cases} 0 & \text{if } \mathfrak{e}_\zeta \le \mathfrak{e}_{\min}. \\ \beta & \text{if } \mathfrak{e}_\zeta \ge \mathfrak{e}_{\min} + 1, \mathfrak{s}_{\eta,M} = 0 \text{ and Eq. (70) holds,} \\ 0 & \text{if } \mathfrak{e}_\zeta \ge \mathfrak{e}_{\min} + 1, \mathfrak{s}_{\eta,M} = 1 \text{ and Eq. (70) holds,} \\ 0 & \text{if } \mathfrak{e}_\zeta \ge \mathfrak{e}_{\min} + 1, \mathfrak{s}_{\eta,M} = 0 \text{ and Eq. (71) holds,} \\ -\beta & \text{if } \mathfrak{e}_\zeta \ge \mathfrak{e}_{\min} + 1, \mathfrak{s}_{\eta,M} = 1 \text{ and Eq. (71) holds.} \end{cases} \quad (74)$$

Define $f(x) : \mathbb{F} \to \mathbb{F}$ as

$$f(x) \coloneqq (\theta \otimes x) \oplus (-\theta \otimes \sigma(\eta)) \oplus \bigoplus_{i=1}^{n} (\widetilde{\alpha}_i \otimes K_\sigma) \oplus \left(\tilde{\beta} \otimes K_\sigma\right) \oplus \eta.$$

Now we analyze the abstract interval arithmetic of f. First, consider the function $\tilde{g}$ defined as

$$\tilde{g}(x) \coloneqq g(x) \oplus \left(\tilde{\beta} \otimes K_\sigma\right) \oplus \eta = x \oplus \bigoplus_{i=1}^{n} (\widetilde{\alpha}_i \otimes K_\sigma) \oplus \left(\tilde{\beta} \otimes K_\sigma\right) \oplus \eta. \tag{75}$$

Together with Eqs. (68), (69), (73) and (74), we have

$$\tilde{g}^\sharp\langle -1.1 \times 2^{\mathfrak{e}_\varsigma}, 0\rangle \subset$$
$$\begin{cases} \langle -(2^{\mathfrak{e}_\varsigma})^-, 2^{\mathfrak{e}_\varsigma}\rangle \oplus^\sharp \eta & \subset \langle \eta, \eta\rangle \text{ if } \mathfrak{e}_\zeta \le \mathfrak{e}_{\min} \text{ and } \mathfrak{s}_{\eta,M} = 0, \\ \langle -(2^{\mathfrak{e}_\varsigma})^-, (2^{\mathfrak{e}_\varsigma})^-\rangle \oplus^\sharp \eta & \subset \langle \eta, \eta\rangle \text{ if } \mathfrak{e}_\zeta \le \mathfrak{e}_{\min} \text{ and } \mathfrak{s}_{\eta,M} = 1, \\ \langle -(2^{\mathfrak{e}_\varsigma})^-, (2^{\mathfrak{e}_\varsigma})^-\rangle \oplus^\sharp \left(\tilde{\beta} \otimes K_\sigma\right) \oplus^\sharp \eta & \subset \langle \eta, \eta\rangle \text{ if } \mathfrak{e}_\zeta \ge \mathfrak{e}_{\min} + 1, \mathfrak{s}_{\eta,M} = 0 \text{ and (70) holds}, \\ \langle -(2^{\mathfrak{e}_\varsigma})^-, (2^{\mathfrak{e}_\varsigma})^-\rangle \oplus^\sharp \eta & \subset \langle \eta, \eta\rangle \text{ if } \mathfrak{e}_\zeta \ge \mathfrak{e}_{\min} + 1, \mathfrak{s}_{\eta,M} = 1 \text{ and (70) holds}, \\ \langle -(2^{\mathfrak{e}_\varsigma})^-, 2^{\mathfrak{e}_\varsigma}\rangle \oplus^\sharp \eta & \subset \langle \eta, \eta\rangle \text{ if } \mathfrak{e}_\zeta \ge \mathfrak{e}_{\min} + 1, \mathfrak{s}_{\eta,M} = 0 \text{ and (71) holds}, \\ \langle -(2^{\mathfrak{e}_\varsigma})^-, 2^{\mathfrak{e}_\varsigma}\rangle \oplus^\sharp \left(-\tilde{\beta} \otimes K_\sigma\right) \oplus^\sharp \eta & \subset \langle \eta, \eta\rangle \text{ if } \mathfrak{e}_\zeta \ge \mathfrak{e}_{\min} + 1, \mathfrak{s}_{\eta,M} = 1 \text{ and (71) holds}, \end{cases}$$

and thus, we have $g'\langle -1.1 \times 2^{\mathfrak{e}_\varsigma}, 0\rangle = \langle \eta, \eta\rangle$. Similarly, together with Eqs. (67)–(69), (73) and (74), we have $\tilde{g}^\sharp\left(\langle \omega, 1.1 \times 2^{\mathfrak{e}_\varsigma}\rangle\right) = \langle \eta^+, \eta^+\rangle$ by the following argument:

$$\tilde{g}^\sharp\left(\langle \omega, 1.1 \times 2^{\mathfrak{e}_\varsigma}\rangle\right)$$
$$\subset \begin{cases} \langle (2^{\mathfrak{e}_\varsigma})^+, (1.1 \times 2^{\mathfrak{e}_\varsigma+1})^-\rangle \oplus^\sharp \eta & \subset \langle \eta^+, \eta^+\rangle \text{ if } \mathfrak{e}_\zeta \le \mathfrak{e}_{\min} \text{ and } \mathfrak{s}_{\eta,M} = 0, \\ \langle 2^{\mathfrak{e}_\varsigma}, (1.1 \times 2^{\mathfrak{e}_\varsigma+1})^-\rangle \oplus^\sharp \eta & \subset \langle \eta^+, \eta^+\rangle \text{ if } \mathfrak{e}_\zeta \le \mathfrak{e}_{\min} \text{ and } \mathfrak{s}_{\eta,M} = 1, \\ \langle 2^{\mathfrak{e}_\varsigma}, (1.1 \times 2^{\mathfrak{e}_\varsigma+1})^-\rangle \oplus^\sharp \left(\tilde{\beta} \otimes K_\sigma\right) \oplus^\sharp \eta & \subset \langle \eta^+, \eta^+\rangle \text{ if } \mathfrak{e}_\zeta \ge \mathfrak{e}_{\min} + 1, \mathfrak{s}_{\eta,M} = 0 \text{ and (70) holds}, \\ \langle 2^{\mathfrak{e}_\varsigma}, (1.1 \times 2^{\mathfrak{e}_\varsigma+1})^-\rangle \oplus^\sharp \eta & \subset \langle \eta^+, \eta^+\rangle \text{ if } \mathfrak{e}_\zeta \ge \mathfrak{e}_{\min} + 1, \mathfrak{s}_{\eta,M} = 1 \text{ and (70) holds}, \\ \langle (2^{\mathfrak{e}_\varsigma})^+, (1.1 \times 2^{\mathfrak{e}_\varsigma+1})^-\rangle \oplus^\sharp \eta & \subset \langle \eta^+, \eta^+\rangle \text{ if } \mathfrak{e}_\zeta \ge \mathfrak{e}_{\min} + 1, \mathfrak{s}_{\eta,M} = 0 \text{ and (71) holds}, \\ \langle (2^{\mathfrak{e}_\varsigma})^+, (1.1 \times 2^{\mathfrak{e}_\varsigma+1})^-\rangle \oplus^\sharp \left(-\tilde{\beta} \otimes K_\sigma\right) \oplus^\sharp \eta & \subset \langle \eta^+, \eta^+\rangle \text{ if } \mathfrak{e}_\zeta \ge \mathfrak{e}_{\min} + 1, \mathfrak{s}_{\eta,M} = 1 \text{ and (71) holds}. \end{cases}$$

Now we define $f(x)$ and $h(x)$ as

$$h(x) \coloneqq \begin{cases} (2^{\mathfrak{e}_\theta} \otimes x) \oplus (-2^{\mathfrak{e}_\theta} \otimes \sigma(\eta)) & \text{if } \sigma(\eta) < \sigma(\eta^+), \\ (-2^{\mathfrak{e}_\theta} \otimes x) \oplus (2^{\mathfrak{e}_\theta} \otimes \sigma(\eta)) & \text{if } \sigma(\eta) > \sigma(\eta^+), \end{cases}$$

$$f(x) \coloneqq h(x) \oplus \bigoplus_{i=1}^{n} (\widetilde{\alpha}_i \otimes K_\sigma) \oplus \left(\tilde{\beta} \otimes K_\sigma\right) \oplus \eta,$$

We need to show the followings to show Eq. (62).

$$h^\sharp(\mathcal{I}) \subset \langle -1.1 \times 2^{\mathfrak{e}_\varsigma}, 0\rangle, \tag{76}$$
$$h^\sharp\left(\mathcal{I}^+\right) \subset \langle \omega, 1.1 \times 2^{\mathfrak{e}_\varsigma}\rangle. \tag{77}$$

To show this, we consider the following cases.

Case 1: $\sigma(\eta) < \sigma(\eta^+)$**.**

In this case, we need to show

$$\begin{cases} h\left(\sigma(\eta) - 2^{\mathfrak{e}_\varsigma - \mathfrak{e}_\theta}\right) \geq -1.1 \times 2^{\mathfrak{e}_\varsigma} \\ h\left(\sigma(\eta)\right) \leq 0 \\ h\left(\sigma(\eta^+)\right) \geq \omega \\ h\left(\sigma(\eta^+) + 2^{\mathfrak{e}_\varsigma - \mathfrak{e}_\theta}\right) \leq 1.1 \times 2^{\mathfrak{e}_\varsigma}. \end{cases} \quad . \tag{78}$$

First, note that for any $\gamma \in \mathbb{F}$, unless $2^{\mathfrak{e}_\theta} \otimes \gamma$ is subnormal, $2^{\mathfrak{e}_\theta} \otimes \gamma$ is exact. Hence we have

$$|2^{\mathfrak{e}_\theta} \otimes \gamma - 2^{\mathfrak{e}_\theta} \times \gamma| \leq \frac{1}{2}\omega.$$

For $x = \sigma(\eta) - 2^{\mathfrak{e}_\varsigma - \mathfrak{e}_\theta}$, we have

$$\begin{aligned} &|\mathrm{rnd}(2^{\mathfrak{e}_\theta} \otimes \sigma(\eta) - 2^{\mathfrak{e}_\theta} \otimes x)_{\mathbb{F}}| \leq |\mathrm{rnd}(2^{\mathfrak{e}_\theta} \times \sigma(\eta) - 2^{\mathfrak{e}_\theta} \times x + \omega)_{\mathbb{F}}| \\ &\leq |\mathrm{rnd}((x - \sigma(\eta)) \times 2^{\mathfrak{e}_\theta} + \omega)_{\mathbb{F}}| \leq \mathrm{rnd}(2^{\mathfrak{e}_\varsigma} + \omega)_{\mathbb{F}} \leq 1.1 \times 2^{\mathfrak{e}_\varsigma}. \end{aligned}$$

Therefore,

$$h(x) = (2^{\mathfrak{e}_\theta} \otimes x) \oplus (-2^{\mathfrak{e}_\theta} \otimes \sigma(\eta)) \geq -1.1 \times 2^{\mathfrak{e}_\varsigma}.$$

For $x = \sigma(\eta)$, we have

$$h\left(\sigma(\eta)\right) = (2^{\mathfrak{e}_\theta} \otimes \sigma(\eta)) \oplus (-2^{\mathfrak{e}_\theta} \otimes \sigma(\eta)) = 0.$$

For $x = \sigma(\eta^+)$, by Lemma 22, since $|\sigma(\eta^+) - \sigma(\eta)| \geq 2^{e_0}$ and $-e_0 - M + \mathfrak{e}_{\min} + 1 = \mathfrak{e}_\theta \in \mathbb{F}$, we have

$$h\left(\sigma(\eta^+)\right) = \left(2^{\mathfrak{e}_\theta} \otimes \sigma\left(\eta^+\right)\right) \oplus (2^{\mathfrak{e}_\theta} \otimes -\sigma(\eta)) \geq \omega.$$

For $x = \sigma(\eta^+) + 2^{\mathfrak{e}_\varsigma - \mathfrak{e}_\theta}$, similar to the case of $x = \sigma(\eta) - 2^{\mathfrak{e}_\varsigma - \mathfrak{e}_\theta}$, we have

$$h(x) = (2^{\mathfrak{e}_\theta} \otimes x) \oplus (-2^{\mathfrak{e}_\theta} \otimes \sigma(\eta)) \leq 1.1 \times 2^{\mathfrak{e}_\varsigma}.$$

Therefore, due to Eq. (78), Eqs. (76) and (77) hold.

Additionally, we need to show Eqs. (63) and (64).
For $x > \sigma(\eta) + 2^{\mathfrak{e}_\varsigma - \mathfrak{e}_\theta}$, we have

$$\begin{aligned} (\theta \otimes x) \oplus (-2^{\mathfrak{e}_\theta} \otimes \sigma(\eta)) &= \mathrm{rnd}(\theta \otimes x - 2^{\mathfrak{e}_\theta} \otimes \sigma(\eta))_{\mathbb{F}} \geq \mathrm{rnd}(x \times 2^{\mathfrak{e}_\theta} - \sigma(\eta) \times 2^{\mathfrak{e}_\theta} - \omega)_{\mathbb{F}} \\ &= \mathrm{rnd}((x - \sigma(\eta)) \times 2^{\mathfrak{e}_\theta} - \omega)_{\mathbb{F}} \\ &\geq \mathrm{rnd}\Big(\left(\left(\sigma(\eta) + 2^{\mathfrak{e}_\varsigma - \mathfrak{e}_\theta}\right)^+ - \sigma(\eta)\right) \times 2^{\mathfrak{e}_\theta} - \omega\Big)_{\mathbb{F}} \geq 2^{\mathfrak{e}_\varsigma}. \end{aligned}$$

Hence, by Eq. (72), we have

$$g\left((2^{\mathfrak{e}_\theta} \otimes x) \oplus (2^{\mathfrak{e}_\theta} \otimes \sigma(\eta))\right) \leq \mathrm{rnd}((x - \sigma(\eta)) \times 2^{\mathfrak{e}_\theta} + \omega)_{\mathbb{F}} \oplus (2^{\mathfrak{e}_\varsigma})^+.$$

Therefore,

$$
\begin{aligned}
f(x) &\le \operatorname{rnd}((x-\sigma(\eta))\times 2^{\mathfrak{e}_\theta}+\omega)_{\mathbb{F}} \oplus (2^{\mathfrak{e}_\varsigma})^+ \oplus (\beta\otimes K_\sigma)\oplus \eta \\
&\le \operatorname{rnd}((x-\sigma(\eta))\times 2^{\mathfrak{e}_\theta}+\omega)_{\mathbb{F}} \oplus (2^{\mathfrak{e}_\varsigma})^+ \oplus \left(\frac{5}{4}\times 2^{\mathfrak{e}_\varsigma - M}\right)\oplus \eta \\
&\le \operatorname{rnd}(\left(x-\sigma\left(\eta^+\right)\right)\times 2^{\mathfrak{e}_\theta}+2^{\mathfrak{e}_\theta+e_0+1}+\omega)_{\mathbb{F}} \oplus (2^{\mathfrak{e}_\varsigma})^+ \oplus \left(\frac{5}{4}\times 2^{\mathfrak{e}_\varsigma - M}\right)\oplus \eta \\
&\le \operatorname{rnd}(\left(x-\sigma\left(\eta^+\right)\right)\times 2^{\mathfrak{e}_\theta}+2^{\mathfrak{e}_\varsigma})_{\mathbb{F}} \oplus (2^{\mathfrak{e}_\varsigma})^+ \oplus \left(\frac{5}{4}\times 2^{\mathfrak{e}_\varsigma - M}\right)\oplus \eta.
\end{aligned}
$$

where we use $2^{\mathfrak{e}_\varsigma}\ge 8\omega$.

If $(x-\sigma(\eta^+))\times 2^{\mathfrak{e}_\theta}\le 2^{\mathfrak{e}_\varsigma-1}$, since

$$
\operatorname{rnd}(\left(x-\sigma\left(\eta^+\right)\right)\times 2^{\mathfrak{e}_\theta}+2^{\mathfrak{e}_\varsigma}+\omega)_{\mathbb{F}} \oplus (2^{\mathfrak{e}_\varsigma})^+ \oplus \left(\frac{5}{4}\times 2^{\mathfrak{e}_\varsigma - M}\right) < 1.1\times 2^{\mathfrak{e}_\varsigma+1},
$$

we have $f(x)\le \eta^+$.

If $(x-\sigma(\eta^+))\times 2^{\mathfrak{e}_\theta} > 2^{\mathfrak{e}_\varsigma-1}$, there exist $k\in\mathbb{N}$ such that $k\times 2^{\mathfrak{e}_\varsigma-1}\le (x-\sigma(\eta^+))\times 2^{\mathfrak{e}_\theta} < (k+1)\times 2^{\mathfrak{e}_\varsigma-1}$. Then,

$$
\begin{aligned}
f(x) &\le \operatorname{rnd}(\left(x-\sigma\left(\eta^+\right)\right)\times 2^{\mathfrak{e}_\theta}+2^{\mathfrak{e}_\varsigma})_{\mathbb{F}} \oplus (2^{\mathfrak{e}_\varsigma})^+ \oplus \left(\frac{5}{4}\times 2^{\mathfrak{e}_\varsigma - M}\right)\oplus \eta \\
&\le \operatorname{rnd}(\left(x-\sigma\left(\eta^+\right)\right)\times 2^{\mathfrak{e}_\theta}+2^{\mathfrak{e}_{\varsigma+1}}+2^{\mathfrak{e}_\varsigma - M})_{\mathbb{F}} \oplus \left(\frac{5}{4}\times 2^{\mathfrak{e}_\varsigma - M}\right)\oplus \eta \\
&\le \eta^+ + \left\lceil \frac{k+1}{4}\right\rceil_{\mathbb{Z}} \times 2^{\mathfrak{e}_\varsigma+1} \le \eta^+ + 4\times \left(x-\sigma\left(\eta^+\right)\right)\times 2^{\mathfrak{e}_\theta}.
\end{aligned}
$$

Similarly, for $x<\sigma(\eta)-2^{\mathfrak{e}_\varsigma-\mathfrak{e}_\theta}$ we have

$$
\eta - f(x) \le (x-\sigma(\eta))\times 4\theta.
$$

Case 2: $\sigma(\eta)>\sigma(\eta^+)$**.**

In this case, we need to show

$$
\begin{cases}
h\left(\sigma(\eta)-2^{\mathfrak{e}_\varsigma-\mathfrak{e}_\theta}\right)\le 1.1\times 2^{\mathfrak{e}_\varsigma} \\
h\left(\sigma(\eta^+)\right)\ge \omega \\
h\left(\sigma(\eta)\right)\le 0 \\
h\left(\sigma(\eta^+)+2^{\mathfrak{e}_\varsigma-\mathfrak{e}_\theta}\right)\ge -1.1\times 2^{\mathfrak{e}_\varsigma}.
\end{cases}
\quad . \tag{79}
$$

By similar arguments to **Case 1**, we can show Eq. (79) and Eqs. (65) and (66). □

F.3 Technical Lemmas for Lemma 5

This subsection presents technical lemmas for the proof of Lemma 5 (§D.2).

Lemma 18. *Let $x \in \mathbb{F}$ be normal. Then for any $c \in \mathbb{F}$ such that $\mathfrak{e}_c \geq \mathfrak{e}_{\min}$ and $|\mathfrak{e}_x - \mathfrak{e}_c| \leq \mathfrak{e}_{\max}$, there exists $w \in \mathbb{F}$ such that*

$$w \otimes x = c \pm 2^{-M+\mathfrak{e}_c}.$$

Proof. Without loss of generality, we assume $x, c > 0$ and $\mathfrak{e}_x = 1$. For any normal $a = 1.\mathfrak{s}_{a,1} \ldots \mathfrak{s}_{a,M}$, we have

$$(a^+ \otimes x) - (a \otimes x) = \mathrm{rnd}(a^+ \otimes x) - \mathrm{rnd}(a \otimes x) = 2^{-M} \text{ or } 2^{1-M}.$$

since $(a^+ \times x) - (a \times x) = 2^{-M} \times x < 2^{1-M}$. Since the gap is 2^{-M} or 2^{1-M}, we can pick $w \in \mathbb{F}$ such that $w \otimes x = c \pm 2^{-M+\mathfrak{e}_c}$. □

Lemma 19 (Floating-point distributive law). *Let $a, b, c \in \mathbb{F}$ be normal. Then we have*

$$a \otimes (b \oplus c) = (a \otimes b) + (a \otimes c) + (C \times 2^{-M-1}),$$

where $|C| \leq |a| \left((2 + 2^{-M-1})|b + c| + |b| + |c|\right)$.

Proof. Since

$$b \oplus c = (b + c)(1 + \delta_1), \quad |\delta_1| \leq \varepsilon = 2^{-M-1}.$$

We have

$$\begin{aligned} a \otimes (b \oplus c) &= \mathrm{rnd}(a \times \mathrm{rnd}(b + c)) = (a \times (b + c))\,(1 + \delta_1)(1 + \delta_2) \\ &= (ab + ac)\,(1 + \delta_1)(1 + \delta_2), \quad |\delta_1|, |\delta_2| \leq \varepsilon. \end{aligned}$$

Since

$$(a \otimes b) + (a \otimes c) = ab(1 + \delta_3) + ac(1 + \delta_4) = a(b + c) + a(b\delta_3 + c\delta_4), \quad |\delta_3|, |\delta_4| \leq \varepsilon$$

the difference between $a \otimes (b \oplus c)$ and $(a \otimes b) + (a \otimes c)$ is

$$a \otimes (b \oplus c) - ((a \otimes b) + (a \otimes c)) = (ab + ac)\,(\delta_1 + \delta_2 + \delta_1\delta_2) - a(b\delta_3 + c\delta_4)$$

Therefore,

$$|a \otimes (b \oplus c) - ((a \otimes b) + (a \otimes c))\,| \leq |a| \left(|b + c|(2\varepsilon + \varepsilon^2) + (|b| + |c|)\varepsilon\right).$$

□

Lemma 20. *Let $x, y \in \mathbb{F}$. Suppose $x = 0$ or $\mathfrak{e}_x \leq -3 - 2M + \mathfrak{e}_y$ where $\mathfrak{e}_y \geq 1 + \mathfrak{e}_{\min}$. Then for $a \in \{-2^{-M+\mathfrak{e}_y}, 2^{-M+\mathfrak{e}_y}\}$, there exist $\alpha_1, \ldots, \alpha_3 \in \mathbb{F}$ such that*

$$x \oplus \bigoplus_{i=1}^{3} \alpha_i = 0,$$

$$y \oplus \bigoplus_{i=1}^{3} \alpha_i = y + a.$$

Proof. Without loss of generality, we assume $y > 0$.
Case (1) $\mathfrak{s}_{y,M} = 0$.
If $a = -2^{-M+\mathfrak{e}_y}$, let $\alpha_1 = \alpha_2 = 2^{-1-M+\mathfrak{e}_y}$, $\alpha_3 = -2^{-M+\mathfrak{e}_y}$.
If $a = 2^{-M+\mathfrak{e}_y}$, let $\alpha_1 = \alpha_2 = -2^{-1-M+\mathfrak{e}_y}$, $\alpha_3 = 2^{-M+\mathfrak{e}_y}$.
Case (2) $\mathfrak{s}_{y,M} = 1$.
If $a = -2^{-M+\mathfrak{e}_y}$, let $\alpha_1 = -2^{-M+\mathfrak{e}_y}$, $\alpha_2 = \alpha_3 = 2^{-1-M+\mathfrak{e}_y}$.
If $a = 2^{-M+\mathfrak{e}_y}$, let $\alpha_1 = 2^{-M+\mathfrak{e}_y}$, $\alpha_2 = \alpha_3 = -2^{-1-M+\mathfrak{e}_y}$.

□

Lemma 21. *Let $x, y \in \mathbb{F}$. Suppose $x = 0$ or $\mathfrak{e}_x \leq -4 - 2M + \mathfrak{e}_y$ where $\mathfrak{e}_y \geq 1 + \mathfrak{e}_{\min}$. Then for $a \in \{-2^{-M+\mathfrak{e}_y}, 2^{-M+\mathfrak{e}_y}\}$, there exist $\alpha_1, \dots, \alpha_5 \in \mathbb{F}$ of the form $\alpha_i = 1.\underbrace{1\dots1}_{M\ times} \times 2^{\mathfrak{e}_{\alpha_i}}$ such that*

$$x \oplus \bigoplus_{i=1}^{5} \alpha_i = 0,$$

$$y \oplus \bigoplus_{i=1}^{5} \alpha_i = y + a.$$

Proof. Without loss of generality, we assume $y > 0$.
First, note that

$$1.\underbrace{1\dots1}_{M \text{ times}} \oplus 1.\underbrace{1\dots1}_{M \text{ times}} = 11.\underbrace{1\dots1}_{M-1 \text{ times}},$$

$$11.\underbrace{1\dots1}_{M-1 \text{ times}} \oplus 1.\underbrace{1\dots1}_{M \text{ times}} = \mathrm{rnd}(101.\underbrace{1\dots1}_{M-2 \text{ times}}01) = 101.\underbrace{1\dots1}_{M-2 \text{ times}},$$

$$101.\underbrace{1\dots1}_{M-2 \text{ times}} \oplus 1.\underbrace{1\dots1}_{M \text{ times}} = \mathrm{rnd}(111.\underbrace{1\dots1}_{M-3 \text{ times}}011) = 111.\underbrace{1\dots1}_{M-2 \text{ times}}.$$

Hence $\bigoplus_{i=1}^{4} 1.\underbrace{1\dots1}_{M \text{ times}} = 1.\underbrace{1\dots1}_{M \text{ times}} \times 2^2$.
Therefore, if $a = \pm 2^{-M+\mathfrak{e}_y}$, let

$$\alpha_1 = \alpha_2 = \mp 1.\underbrace{1\dots1}_{M \text{ times}} \times 2^{-2-M+\mathfrak{e}_y}, \alpha_3 = \pm 1.\underbrace{1\dots1}_{M \text{ times}} \times 2^{-1-M+\mathfrak{e}_y}.$$

Then we have

$$x \oplus \bigoplus_{i=1}^{3} \alpha_i = 0, \quad y \oplus \bigoplus_{i=1}^{3} \alpha_i = y + a.$$

□

Lemma 22. *Consider normal floating-point numbers $\gamma_1, \gamma_2 \in \mathbb{F}$ with $\mathfrak{e}_{\gamma_1} \geq \mathfrak{e}_{\gamma_2} \geq \mathfrak{e}_{\min} + 1$. Suppose an integer $e_0 \in \mathbb{Z}$ satisfies the following:*

$$2^{e_0} \leq |\gamma_1 - \gamma_2| < 2^{e_0+1}.$$

Define $\mathfrak{e}_\theta$ *as*

$$\mathfrak{e}_\theta \coloneqq \max\left(\mathfrak{e}_{\min} - M, -e_0 + \mathfrak{e}_{\min} - M + 1\right).$$

Then, we have

$$(2^{\mathfrak{e}_\theta} \otimes \gamma_1) \oplus (-2^{\mathfrak{e}_\theta} \otimes \gamma_2) \neq 0.$$

Proof. For any $\gamma \in \mathbb{F}$, unless $2^{\mathfrak{e}_\theta} \otimes \gamma$ is subnormal , $2^{\mathfrak{e}_\theta} \otimes \gamma$ is exact. Hence,

$$|2^{\mathfrak{e}_\theta} \otimes \gamma_1 - 2^{\mathfrak{e}_\theta} \times \gamma_1| \leq \frac{1}{2}\omega,$$
$$|2^{\mathfrak{e}_\theta} \otimes \gamma_2 - 2^{\mathfrak{e}_\theta} \times \gamma_2| \leq \frac{1}{2}\omega.$$

Since $e_0 \geq \mathfrak{e}_{\gamma_1} - M - 1$ and $e_0 \geq \mathfrak{e}_{\gamma_2} - M - 1$, we have

$$\mathfrak{e}_\theta + \mathfrak{e}_{\gamma_1} = -e_0 + \mathfrak{e}_{\min} - M + 1 + \mathfrak{e}_{\gamma_1} \leq \mathfrak{e}_{\min}$$

Therefore following inequalities hold:

$$\begin{aligned}
&|2^{\mathfrak{e}_\theta} \otimes \gamma_1 - 2^{\mathfrak{e}_\theta} \otimes \gamma_2| \\
&\geq |2^{\mathfrak{e}_\theta} \times \gamma_1 - 2^{\mathfrak{e}_\theta} \times \gamma_2| - |2^{\mathfrak{e}_\theta} \otimes \gamma_1 - 2^{\mathfrak{e}_\theta} \times \gamma_1| - |2^{\mathfrak{e}_\theta} \otimes \gamma_2 - 2^{\mathfrak{e}_\theta} \times \gamma_2| \\
&\geq 2^{\mathfrak{e}_{\min} - M + 1} - \omega = \omega.
\end{aligned}$$

Since $|2^{\mathfrak{e}_\theta} \otimes \gamma_1|, |2^{\mathfrak{e}_\theta} \otimes \gamma_2| \leq 2^{1+\mathfrak{e}_{\min}}$, their gap to adjacent number is ω. Therefore they are distinct. □

F.4 Technical Lemma for Lemma 6

This subsection presents technical lemma for the proof of Lemma 6 (§D.3).

Lemma 23. *For* $1 \leq x < 1 + 2^{-1}$*, we have* $(1 + 2^{-M}) \otimes x = x^+$*. For* $1 + 2^{-1} \leq x \leq 2 - 2^{-1-M}$*, we have* $(1 + 2^{-M}) \otimes x = x^{++}$*. For* $x = 2 - 2^{-M}$ *we have* $(1 + 2^{-M}) \otimes x = x^+$*.*

Proof. $(2^{-1} + 2^{-1-M}) \otimes x = \mathrm{rnd}(1 + 2^{-1-M} - 2^{-1-2M}) = 1.$ □

F.5 Technical Lemmas for §E

This subsection presents technical lemma for the proofs in §E.

Lemma 24. *Let* $K \in \mathbb{F}$ *with* $|K| \in [(1 + 2^{-M+1}) \times 2^{-M-2}, 1 + 2^{-2} - 2^{-M}]_{\mathbb{F}}$*. Consider* $\mathfrak{e}_\zeta \in \mathbb{Z}$ *such that* $\mathfrak{e}_{\min} - M \leq \mathfrak{e}_\zeta \leq \mathfrak{e}_{\max} - M$*. Then, there exists* $\gamma \in \mathbb{F}$ *such that the following inequality holds:*

$$\frac{1}{2} \times 2^{\mathfrak{e}_\zeta} < \gamma \otimes K \leq \frac{5}{4} \times 2^{\mathfrak{e}_\zeta}.$$

Proof. Let K be represented as

$$K = \mathfrak{s}_K \times 2^{\mathfrak{e}_K}, \quad -M - 2 \leq \mathfrak{e}_K \leq 0.$$

We consider the following cases.

Case 1: $\mathfrak{e}_{\max} - M - 1 \le \mathfrak{e}_\zeta \le \mathfrak{e}_{\max} - M$**.**

If $\mathfrak{s}_K \in [1, 1+2^{-M}]_{\mathbb{F}}$, we have $-M-1 \le \mathfrak{e}_K \le 0$. We define γ as

$$\gamma \coloneqq 2^{\mathfrak{e}_\zeta - \mathfrak{e}_K - 1}.$$

Since $\mathfrak{e}_\zeta - \mathfrak{e}_K - 1 \le \mathfrak{e}_{\max} - M - (M-1) - 1 = \mathfrak{e}_{\max}$, we have $\gamma \in \mathbb{F}$. Then,

$$2^{\mathfrak{e}_\zeta - 1} \le \gamma \times K \le (1 + 2^{-M}) \times 2^{\mathfrak{e}_\zeta - 1}.$$

If $\mathfrak{s}_K \in [1 + 2^{-M+1}, 2)_{\mathbb{F}}$, we define γ as

$$\gamma \coloneqq (2 - 2^{-M}) \times 2^{\mathfrak{e}_\zeta - \mathfrak{e}_K - 2}.$$

Since $\mathfrak{e}_\zeta - \mathfrak{e}_K - 2 \le \mathfrak{e}_{\max} - M - (-M-2) - 2 = \mathfrak{e}_{\max}$, we have $\gamma \in \mathbb{F}$. Because

$$\begin{aligned}(2 - 2^{-M}) \times (1 + 2^{-M+1}) = 2 + 3 \times 2^{-M} - 2^{-2M+1} > 2,\\ (2 - 2^{-M}) \times (2 - 2^{-M}) = 2^2 - 2^{2-M} + 2^{-2M} < 4,\end{aligned}$$

we have

$$\frac{1}{2} \times 2^{\mathfrak{e}_\zeta} < \gamma \otimes K < 2^{\mathfrak{e}_\zeta}.$$

Case 2: $\mathfrak{e}_{\min} - M + 2 \le \mathfrak{e}_\zeta \le \mathfrak{e}_{\max} - M - 2$**.**

If $\mathfrak{s}_K \in \left[1, 1 + 2^{-2}\right]_{\mathbb{F}}$, define γ as

$$\gamma \coloneqq 2^{\mathfrak{e}_\zeta - \mathfrak{e}_K}.$$

As $\mathfrak{e}_{\min} - M + 2 \le \mathfrak{e}_\zeta - \mathfrak{e}_K \le \mathfrak{e}_{\max}$, we have $\gamma \in \mathbb{F}$. Then, we have

$$2^{\mathfrak{e}_\zeta} \le \gamma \otimes K = \mathrm{rnd}(\mathfrak{s}_K \times 2^{\mathfrak{e}_\zeta})_{\mathbb{F}} \le \mathrm{rnd}\Big(\frac{5}{4} \times 2^{\mathfrak{e}_\zeta}\Big)_{\mathbb{F}} = \frac{5}{4} \times 2^{\mathfrak{e}_\zeta},$$

where the last equality is followed by $\mathfrak{e}_\zeta \ge \mathfrak{e}_{\min} - M + 2$.
If $\mathfrak{s}_K \in \left(1 + 2^{-2}, 2\right)_{\mathbb{F}}$, define γ as

$$\gamma \coloneqq 2^{\mathfrak{e}_\zeta - \mathfrak{e}_K - 1}.$$

As $\mathfrak{e}_\zeta - \mathfrak{e}_K - 1 \ge \mathfrak{e}_{\min} - M + 1$, we have $\gamma \in \mathbb{F}$. Then, we have

$$\frac{1}{2} \times 2^{\mathfrak{e}_\zeta} < \gamma \otimes K = \mathrm{rnd}(\mathfrak{s}_K \times 2^{\mathfrak{e}_\zeta - 1})_{\mathbb{F}} \le 2^{\mathfrak{e}_\zeta},$$

where the first inequality is followed from $\mathfrak{e}_\zeta - 1 \ge \mathfrak{e}_{\min} - M + 1$ and $\mathfrak{s}_K > 1 + 2^{-2}$.

Case 3: $\mathfrak{e}_\zeta = \mathfrak{e}_{\min} - M + 1$**.**

If $\mathfrak{s}_K \in [1, \frac{5}{4})$, define γ as

$$\gamma := 2^{\mathfrak{e}_\zeta - \mathfrak{e}_K}.$$

Since $\mathfrak{e}_\zeta - \mathfrak{e}_K \geq \mathfrak{e}_{\min} - M + 1$, we have $\gamma \in \mathbb{F}$. Then,

$$\gamma \otimes K = \mathrm{rnd}(\mathfrak{s}_K \times 2^{\mathfrak{e}_{\min} - M + 1})_{\mathbb{F}} = 2\omega = 2^{\mathfrak{e}_\zeta},$$

since $2\omega \leq \mathfrak{s}_K \times 2^{\mathfrak{e}_{\min} - M + 1} < \frac{5}{2}\omega$.
If $\mathfrak{s}_K \in \left[\frac{5}{4}, \frac{5}{3}\right)$, define γ as

$$\gamma := (1 + 2^{-1}) \times 2^{\mathfrak{e}_\zeta - \mathfrak{e}_K - 1}.$$

As $\mathfrak{e}_K \leq -1$, $\mathfrak{e}_\zeta - \mathfrak{e}_K - 1 \geq \mathfrak{e}_{\min} - M + 1$, and we have $\gamma \in \mathbb{F}$. Then,

$$\gamma \otimes K = \mathrm{rnd}(\mathfrak{s}_K \times (1 + 2^{-1}) \times 2^{\mathfrak{e}_{\min} - M})_{\mathbb{F}} = \mathrm{rnd}(\mathfrak{s}_K \times (1 + 2^{-1}) \times \omega)_{\mathbb{F}} = 2\omega = 2^{\mathfrak{e}_\zeta},$$

since $\frac{3}{2}\omega < \frac{15}{8}\omega \leq \mathfrak{s}_K \times (1 + 2^{-1}) \times \omega < \frac{5}{2}\omega$.

If $\mathfrak{s}_K \in \left(\frac{5}{3}, 2\right)$, define γ as

$$\gamma := 2^{\mathfrak{e}_\zeta - \mathfrak{e}_K - 1}.$$

Since $\mathfrak{e}_\zeta - \mathfrak{e}_K \geq \mathfrak{e}_{\min} - M$, we have $\gamma \in \mathbb{F}$. Then,

$$\gamma \otimes K = \mathrm{rnd}(\mathfrak{s}_K \times 2^{\mathfrak{e}_{\min} - M})_{\mathbb{F}} = \mathrm{rnd}(\mathfrak{s}_K \times \omega)_{\mathbb{F}} = 2\omega = 2^{\mathfrak{e}_\zeta},$$

since $\frac{3}{2}\omega < \frac{5}{3}\omega \leq \mathfrak{s}_K \times \omega < 2\omega$.

Case 4: $\mathfrak{e}_\zeta = \mathfrak{e}_{\min} - M$**.**

If $\mathfrak{s}_K \in [1, \frac{3}{2})$, define γ as

$$\gamma := 2^{\mathfrak{e}_\zeta - \mathfrak{e}_K}.$$

Since $\mathfrak{e}_\zeta - \mathfrak{e}_K \geq \mathfrak{e}_{\min} - M$, we have $\gamma \in \mathbb{F}$. Then,

$$\gamma \otimes K = \mathrm{rnd}(\mathfrak{s}_K \times 2^{\mathfrak{e}_{\min} - M})_{\mathbb{F}} = \mathrm{rnd}(\mathfrak{s}_K \times \omega)_{\mathbb{F}} = \omega = 2^{\mathfrak{e}_\zeta},$$

since $\omega \leq \mathfrak{s}_K \times \omega < \frac{3}{2}\omega$.
If $\mathfrak{s}_K \in (\frac{3}{2}, 2)$, we have $\mathfrak{e}_K \leq -1$ by the assumption. Define γ as

$$\gamma := 2^{\mathfrak{e}_\zeta - \mathfrak{e}_K - 1}.$$

Since $\mathfrak{e}_\zeta - \mathfrak{e}_K - 1 \geq \mathfrak{e}_{\min} - M$, we have $\gamma \in \mathbb{F}$. Then,

$$\gamma \otimes K = \mathrm{rnd}(\mathfrak{s}_K \times 2^{\mathfrak{e}_{\min} - M - 1})_{\mathbb{F}} = \mathrm{rnd}\Big(\frac{1}{2} \times \mathfrak{s}_K \times \omega\Big)_{\mathbb{F}} = \omega = 2^{\mathfrak{e}_\zeta},$$

since $\frac{1}{2}\omega < \frac{3}{4}\omega \leq \frac{1}{2} \times \mathfrak{s}_K \times \omega < \omega$. This completes the proof. $\square$

Lemma 25. *Let* $0 = z_0 < z_1 < \cdots < z_{(|\mathbb{F}|-1)/2} = \Omega < z_{(|\mathbb{F}|+1)/2} = \infty$ *be all non-negative floats in* $\mathbb{F}$. *Then, for any* $j \in [(|\mathbb{F}|+1)/2] \cup \{0\}$ *and* $x_i \in (2^{-1} \times (z_i - z_{i-1}), (1+2^{-1}) \times (z_i - z_{i-1}))_{\mathbb{F}}$, *it holds that*

$$\bigoplus_{i=1}^{j} x_i = z_j.$$

Proof. Since it is obvious for $j = 0$ we consider $j \geq 1$, and we use the mathematical induction on j.

Base step: $j = 1$.

For the base step, note that $z_1 = \omega$. Then we have

$$x_1 \in (2^{-1} \times (z_1 - z_0), (1+2^{-1}) \times (z_1 - z_0))_{\mathbb{F}} = (\frac{1}{2} \times \omega, \frac{3}{2} \times \omega)_{\mathbb{F}} = \{\omega\} = \{z_1\}.$$

Induction step.

Assume that $\bigoplus_{i=1}^{j} x_i = z_j$ (inductive hypothesis). We write z_j as $\mathfrak{s}_{z_j} \times 2^{\mathfrak{e}_{z_j}}$. We consider the following cases.

Case 1: $2^{\mathfrak{e}_{z_j}} \leq \mathfrak{e}_{\min} + 1$.

In this case, $z_{j+1} - z_j = \omega$. Consider the case of $j+1$ as follows:

$$\bigoplus_{i=1}^{j+1} x_i = \left(\bigoplus_{i=1}^{j} x_i\right) \oplus x_{j+1} = z_j \oplus x_{j+1} = z_{j+1},$$

where the last equality follows from

$$x_{j+1} \in \big(2^{-1} \times (z_{j+1} - z_j), (1+2^{-1}) \times (z_{j+1} - z_j)\big)_{\mathbb{F}} = (\frac{1}{2} \times \omega, \frac{3}{2} \times \omega)_{\mathbb{F}} = \{\omega\}.$$

Case 2: $2^{\mathfrak{e}_{z_j}} \geq \mathfrak{e}_{\min} + 2$.

In this case, $z_{j+1} - z_j = 2^{-M+\mathfrak{e}_{z_j}}$ Consider the case of $j+1$ as follows:

$$\bigoplus_{i=1}^{j+1} x_i = \left(\bigoplus_{i=1}^{j} x_i\right) \oplus x_{j+1} = z_j \oplus x_{j+1} = z_{j+1},$$

where the last equality follows from

$$\begin{aligned} x_{j+1} &\in \big(2^{-1} \times (z_{j+1} - z_j), (1+2^{-1}) \times (z_{j+1} - z_j)\big)_{\mathbb{F}} \\ &= \left(2^{-M-1+\mathfrak{e}_{x_j}}, (1+2^{-1}) \times 2^{-M-1+\mathfrak{e}_{x_j}}\right)_{\mathbb{F}}. \end{aligned}$$

This completes the proof. □

Lemma 26. *For any* $\eta \in [-2^3, 2^3)_{\mathbb{F}}$, $x \in [(1+2^{-p+1}) \times 2^{-p-2}, 1+2^{-1}-2^{-p}]_{\mathbb{F}}$, *and* $n \in [2^p]$, *there exist* $k \in \mathbb{N}$ *and* $\alpha, \beta, z_1, \ldots, z_k, \theta_1, \ldots, \theta_k \in \mathbb{F}$ *such that*

$$\left(\bigoplus_{i=1}^{n} \alpha \otimes x\right) \oplus (\theta_1 \otimes \sigma(z_1)) \oplus \cdots \oplus (\theta_k \otimes \sigma(z_k)) \oplus \beta \in [\eta^+, \infty)_{\mathbb{F}},$$

$$\left(\bigoplus_{i=1}^{n-1} \alpha \otimes x\right) \oplus (\theta_1 \otimes \sigma(z_1)) \oplus \cdots \oplus (\theta_k \otimes \sigma(z_k)) \oplus \beta \in (-\infty, \eta]_{\mathbb{F}}.$$

Proof. If η is normal, by Lemma 24, there exists $\alpha \in \mathbb{F}$ such that the following inequality holds: if $\eta > 0$,

$$\frac{1}{2} \times 2^{\mathfrak{e}_\eta - M} \leq \alpha \otimes x < \frac{5}{4} \times 2^{\mathfrak{e}_\eta - M}.$$

if $\eta < 0$,

$$\frac{1}{2} \times 2^{\mathfrak{e}_\eta - M - 1} \leq \alpha \otimes x < \frac{5}{4} \times 2^{\mathfrak{e}_\eta - M - 1}.$$

Then, for $n = 1$,

$$\eta \oplus \left(\sum_{i=1}^{n} \alpha \otimes x\right) = \eta \oplus (\alpha \otimes x) = \eta^+. \tag{80}$$

If η is subnormal, consider $\alpha \in \mathbb{F}$ such that $\alpha \otimes x = \omega$. Then, for $n = 1$,

$$\eta \oplus \left(\sum_{i=1}^{n} \alpha \otimes x\right) = \eta \oplus \omega = \eta^+. \tag{81}$$

This completes the proof. □