Floating-Point Neural Networks Can Represent Almost All Floating-Point Functions

Geonho Hwang [1] **Yeachan Park** [2] **Wonyeol Lee** [3] **Sejun Park** [4]

Abstract

Existing works on the expressive power of neural networks typically assume real-valued parameters and exact mathematical operations during the evaluation of networks. However, neural networks run on actual computers can take parameters only from a small subset of the reals and perform inexact mathematical operations with round-off errors and overflows. In this work, we study the expressive power of *floating-point* neural networks, i.e., networks with floating-point parameters and operations. We first observe that for floating-point neural networks to represent *all* functions from floating-point vectors to floating-point vectors, it is necessary to *distinguish* different inputs: the first layer of a network should be able to generate different outputs for different inputs. We also prove that such distinguishability is *sufficient*, along with mild conditions on activation functions. Our result shows that with practical activation functions, floating-point neural networks can represent floating-point functions from a wide domain to all finite or infinite floats. For example, the domain is all finite floats for $\mathrm{Sigmoid}$ and $\tanh$, and it is all finite floats of magnitude less than $1/8$ times the largest float for ReLU, ELU, SeLU, GELU, Swish, Mish and $\sin$.

1. Introduction

Deep neural networks have achieved remarkable success in various fields of science and engineering (LeCun et al., 2015). One of the theoretical foundations of neural networks is the *universal approximation theorem*, which states that neural networks can approximate a large class of target functions. Classical results show that two-layer fully-connected networks using a non-polynomial activation function can approximate any continuous function on a compact domain within an arbitrary accuracy (Cybenko, 1989; Hornik et al., 1989; Pinkus, 1999; Leshno et al., 1993). Such results have been extended to width-bounded setups (Lu et al., 2017; Kidger & Lyons, 2020; Park et al., 2021) and more practical networks (Zhou, 2020; Tabuada & Gharesifard, 2021; Yun et al., 2020; Ramanujan et al., 2020; Yuan & Agaian, 2023).

These universal approximation results assume real-valued network parameters and/or exact mathematical operations (e.g., addition and multiplication) during the evaluation of networks. In practice, however, neural networks are executed on actual computers in which network parameters can take values only from a finite subset of the reals and mathematical operations can be inexact due to round-off errors. This presents a clear gap between theoretical assumptions and practical environments for neural networks, which becomes more significant under low-precision setups: network parameters can take only a tiny number of values, and mathematical operations can have huge round-off errors.

Several works have studied the expressive power of neural networks that take parameter values from a fixed *finite* set. For example, Ding et al. (2019) investigate the universal approximation of ReLU networks with quantized parameters, and show that such "quantized" ReLU networks with at least two distinct parameters can approximate Sobolev functions $\mathcal{W}^{n,\infty}$ ($n \geq 1$). Similarly, Gonon et al. (2023) examine the impact of quantization on the approximation ability of ReLU networks, and show that a uniformly quantized network can approximate a network with real parameters. However, all these works assume *exact* mathematical operations, which limits our understanding of neural networks under finite-precision operations with round-off errors.

Some recent studies have investigated more realistic setups, where parameters can take values from a finite set and mathematical operations can be *inexact*. For example, Hwang et al. (2024) study the expressive power of neural networks under *fixed-point* arithmetic, and show that such "fixed-point" networks with finite-precision weights and infinite-precision biases can approximate any continuous function when they

[1]Department of Mathematical Sciences, GIST [2]Department of Mathematics and Statistics, Sejong University [3]Department of Computer Science and Engineering, POSTECH [4]Department of Artificial Intelligence, Korea University. Correspondence to: Sejun Park <sejun.park000@gmail.com>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

use popular activation functions such as Sigmoid, ReLU, ELU, SoftPlus, SiLU, Mish, and GELU. Likewise, Park et al. (2024) analyze the expressive power of networks under *floating-point* arithmetic. They show that such "floating-point" networks can represent all floating-point functions over a unit cube when they use ReLU or Step (i.e., the step function) as activation functions.

Despite all these advances, no existing work has studied the expressive power of neural networks under the following arguably most practical setup:

- Parameters are represented as floating-point numbers.
- Operations are performed using floating-point arithmetic.
- A broad class of activation functions is supported.

For instance, Hwang et al. (2024) considers fixed-point arithmetic, not floating-point arithmetic; and Park et al. (2024) considers only two activation functions (ReLU and Step), leaving out many widely-used activation functions such as GELU (Hendrycks & Gimpel, 2016), SeLU (Klambauer et al., 2017), Swish (Ramachandran et al., 2017), Mish (Bochkovskiy et al., 2020), and sin (Sitzmann et al., 2020). We remark that the proof technique of Park et al. (2024) relies heavily on the piecewise linearity of activation functions, so extending their proof to non-piecewise-linear activation functions is highly non-trivial, if possible at all. To summarize, the following question has remained open:

> What is the expressivity of *floating-point* neural networks using general *practical* activation functions?

1.1. Contribution

In this work, we investigate the class of floating-point functions that can be represented by feed-forward neural networks using general activation functions under floating-point arithmetic. To this end, we first observe that to represent *all functions* from some domain $\mathcal{X}$ of floating-point numbers to floating-point numbers, it is *necessary* for networks to *distinguish* any pair of distinct inputs in the domain in the first layer (Lemma 3.2). Namely, for any $x, x' \in \mathcal{X}$ with $x \neq x'$, there should exist a floating-point affine transformation ϕ such that $\sigma(\phi(x)) \neq \sigma(\phi(x'))$, where σ denotes the activation function. Using this property, we show that networks using the correctly rounded version of the cos activation function cannot represent some floating-point functions (Lemma 3.3): e.g., some functions over $[-2^8, 2^8]$ for the 16-bit half-precision floating-point format.

We then show that such distinguishability is *sufficient* for networks to represent all floating-point functions under mild conditions on activation functions σ (Theorem 3.4): σ can output zero and two moderate values (e.g., values around 1) (Condition 1). We also provide easily verifiable conditions on activation functions that ensure distinguishability (Lemmas 3.6, 3.7, and 3.10). Using these conditions, we show that floating-point neural networks can represent a large class of floating-point functions for various activation functions (e.g., Identity, Sigmoid, tanh, ReLU, ELU, SeLU, GELU, Swish), under various floating-point formats from low-precision ones (e.g., float8, bfloat16) to high-precision ones (e.g., float32, float64) (Corollaries 3.8 and 3.11).

We note that a concurrent work by Hwang et al. (2025) studies universal approximation under floating-point arithmetic. Their work considers *interval* universal approximation, which generalizes the classical pointwise approximation. However, a special case of their result for pointwise approximation is *subsumed* by our result (Theorem 3.4): their condition is strictly stronger, and their network construction requires more depth than ours. In addition, they do not study the *necessity* of their condition unlike our work.

2. Preliminaries

In this section, we introduce the basic notations and concepts used throughout the paper, including floating-point arithmetic and floating-point neural networks.

We begin with notations. We use $\mathbb{N}$, $\mathbb{Z}$, and $\mathbb{R}$ to denote the sets of natural numbers, integers, and real numbers, respectively. For $a, b \in \mathbb{R}$, we define $[a, b] \coloneqq \{x \in \mathbb{R} : a \leq x \leq b\}$ and $(a, b) \coloneqq \{x \in \mathbb{R} : a < x < b\}$; we define $[a, b)$ and $(a, b]$ analogously. For $\mathcal{S} \subset \mathbb{R}$, we define $[a, b]_{\mathcal{S}} \coloneqq [a, b] \cap \mathcal{S}$, with $(a, b)_{\mathcal{S}}$, $[a, b)_{\mathcal{S}}$, and $(a, b]_{\mathcal{S}}$ defined analogously. For $n \in \mathbb{N}$, we write $[n] \coloneqq [1, n]_{\mathbb{N}}$. For $d \in \mathbb{N}$, a set $\mathcal{S}$, and $x \in \mathcal{S}^d$, we define x_i as the i-th coordinate of x. In this paper, all fractional numbers with a radix point are assumed to be in binary representation: e.g., $1.101 = 2^0 + 2^{-1} + 2^{-3} = 13/8$.

2.1. Floating-Point Arithmetic

Floating-point numbers. For $p, q \in \mathbb{N}$, we define $\mathbb{F}_{p,q}$ as the set of *finite* floating-point numbers:

$$\begin{aligned}\mathbb{F}_{p,q} \coloneqq \big\{ & s \times (1.m_1 \cdots m_p) \times 2^e : s \in \{-1, 1\}, \\ & m_1, \ldots, m_p \in \{0, 1\}, e \in [\mathfrak{e}_{\min}, \mathfrak{e}_{\max}]_{\mathbb{Z}} \big\} \\ \cup \big\{ & s \times (0.m_1 \cdots m_p) \times 2^{\mathfrak{e}_{\min}} : \\ & s \in \{-1, 1\}, m_1, \ldots, m_p \in \{0, 1\} \big\}, \end{aligned} \tag{1}$$

where $\mathfrak{e}_{\min}$ and $\mathfrak{e}_{\max}$ are defined as $\mathfrak{e}_{\min} \coloneqq -2^{q-1} + 2$ and $\mathfrak{e}_{\max} \coloneqq 2^{q-1} - 1$. Here, s, $m_1 \ldots m_p$, and e are called the *sign*, *mantissa*, and *exponent* of a floating-point number, respectively. Note that $p + q + 1$ bits suffice to represent all numbers in $\mathbb{F}_{p,q}$: p bits and q bits for representing the mantissa and exponent, respectively, and one additional bit for the sign. For simplicity, we omit the subscript and write $\mathbb{F}$ for $\mathbb{F}_{p,q}$ when p and q are clear from the context.

We use ∞ and $-\infty$ to denote positive and negative *infinities*, and assume the usual order: $-\infty < x < \infty$ for any $x \in \mathbb{R}$.

Format	(p, q)
16-bit half precision (IEEE, 2019)	$(10, 5)$
32-bit single precision (IEEE, 2019)	$(23, 8)$
64-bit double precision (IEEE, 2019)	$(52, 11)$
8-bit E5M2 (Micikevicius et al., 2022)	$(2, 5)$
8-bit E4M3 (Micikevicius et al., 2022)	$(3, 4)$
bfloat16 (Google; Abadi et al., 2016)	$(7, 8)$

Table 1: List of frequently-used floating-point formats.

We use NaN to denote *not-a-number*, which can be produced, e.g., when ∞ is added to $-\infty$ under floating-point arithmetic. We assume any operation including NaN produces NaN; this does hold for floating-point addition, subtraction, and multiplication. We use $\overline{\mathbb{F}}$ to denote the set of *all* floating-point numbers (or floats) $\overline{\mathbb{F}} \coloneqq \mathbb{F} \cup \{-\infty, \infty, \mathrm{NaN}\}$. We note that $\overline{\mathbb{F}}_{p,q}$ can also be represented using $p+q+1$ bits since we are not using the whole 2^q representations for the exponent of floats in $\mathbb{F}_{p,q}$. For $x \in \mathbb{F}$, x^+ and x^- denote the smallest and largest floats greater than and less than x, respectively. For $x \in \mathbb{F}$, we use $\mathfrak{e}_x \coloneqq \max\{\mathfrak{e}_{\min}, \lfloor \log_2 |x| \rfloor\}$. The smallest and largest finite positive floats are denoted by $\omega \coloneqq 2^{\mathfrak{e}_{\min}-p}$ and $\Omega \coloneqq (2-2^{-p}) \times 2^{\mathfrak{e}_{\max}}$, respectively.

The IEEE-754 standard (IEEE, 2019) defines (p, q) for widely used floating-point formats: e.g., $(10, 5)$ for the 16-bit half precision (float16), $(23, 8)$ for the 32-bit single precision (float32), and $(52, 11)$ for the 64-bit double precision (float64). In this paper, we assume that (p, q) satisfies $2 \le p \le 2^{q-1} - 3$. Many popular floating-point formats satisfy this condition, as illustrated in Table 1.

Floating-point operations. The rounding function $\lceil \cdot \rfloor_{\mathbb{F}} : \mathbb{R} \cup \{-\infty, \infty, \mathrm{NaN}\} \to \overline{\mathbb{F}}$ is defined as

$$\lceil x \rfloor_{\mathbb{F}} \coloneqq \begin{cases} \arg\min_{y \in \mathbb{F}} |x - y| & \text{if } |x| < \Omega\left(1 + 2^{-p-1}\right), \\ \infty & \text{if } x \ge \Omega\left(1 + 2^{-p-1}\right), \\ -\infty & \text{if } x \le -\Omega\left(1 + 2^{-p-1}\right), \\ \mathrm{NaN} & \text{if } x = \mathrm{NaN}. \end{cases}$$

There can be two floats equidistant from a real number. In such a case, we break the tie using the tie-to-even rule: $\lceil x \rfloor_{\mathbb{F}}$ is defined by the (unique) float whose last mantissa bit m_p (see Eq. (1)) is zero. If $\mathbb{F}$ is clear from the text, we omit the subscript $\mathbb{F}$ in $\lceil \cdot \rfloor_{\mathbb{F}}$.

For $\rho : \mathbb{R} \to \mathbb{R}$, we define the *correctly rounded* function $\lceil \rho \rfloor : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ of ρ as follows:

$$\lceil \rho \rfloor(x) \coloneqq \begin{cases} \lceil \rho(x) \rfloor & \text{if } x \in \mathbb{F}, \\ \lceil l \rfloor & \text{if } x = -\infty \wedge \exists \lim_{x \to -\infty} \rho(x), \\ \lceil r \rfloor & \text{if } x = \infty \wedge \exists \lim_{x \to \infty} \rho(x), \\ \mathrm{NaN} & \text{otherwise}, \end{cases}$$

where $l = \lim_{x \to -\infty} \rho(x)$ and $r = \lim_{x \to \infty} \rho(x)$. Here, the existence of l, r includes the case $l, r \in \{-\infty, \infty\}$.

For $x, y \in \overline{\mathbb{F}}$, we define the floating-point operations $\oplus, \ominus$, and $\otimes$ as $x \oplus y \coloneqq \lceil x + y \rfloor$, $x \ominus y \coloneqq \lceil x - y \rfloor$, and $x \otimes y \coloneqq \lceil x \times y \rfloor$. Note that the addition and multiplication are not associative: e.g., $(x \oplus y) \oplus z \neq x \oplus (y \oplus z)$ in general. Therefore, we must be very careful about the ordering of the operations. To concisely represent the addition of floating-point numbers $x_1, \ldots, x_n$, we use $\bigoplus$ as follows:

$$\bigoplus_{i=1}^{n} x_i \coloneqq x_1 \oplus \cdots \oplus x_n,$$

where addition is performed from left to right. Likewise, we use $\oplus$ and $\bigoplus$ in a left-associative manner. For example,

$$\bigoplus_{i=1}^{n} \bigoplus_{j=1}^{m} x_{i,j} \coloneqq x_{1,1} \oplus \cdots \oplus x_{1,m} \oplus x_{2,1} \oplus \cdots \oplus x_{2,m} \oplus \cdots \oplus x_{n,1} \oplus \cdots \oplus x_{n,m},$$

where addition is performed from left to right. These definitions ensure that the addition is always performed sequentially, starting from the first element. When performing summation on the product of finite sets of integers, the summation is always carried out in lexicographic order. E.g.,

$$\bigoplus_{x,y \in \{1,2\} \times \{3,4\}} (x \oplus y) \coloneqq (1 \oplus 3) \oplus (1 \oplus 4) \oplus (2 \oplus 3) \oplus (2 \oplus 4).$$

For $\mathcal{X} \subset \mathbb{F}^d$, we use $\mathbb{1}_{\mathcal{X}} : \mathbb{F}^d \to \mathbb{F}$ to denote the indicator function of $\mathcal{X}$: $\mathbb{1}_{\mathcal{X}}(x)$ is one if $x \in \mathcal{X}$ and zero otherwise. If $\mathcal{X} = \{x_0\}$ is a singleton set, we use $\mathbb{1}_{x_0}$ to denote $\mathbb{1}_{\mathcal{X}}$.

2.2. Floating-Point Neural Networks

To define floating-point neural networks, we first define a floating-point affine transformation. For $d_1, d_2 \in \mathbb{N}$, let $w_i = (w_{i,1}, \ldots, w_{i,d_1}) \in \mathbb{F}^{d_1}$ and $b_i \in \mathbb{F}$ for all $i \in [d_2]$. Then, for $I = (w_1, \ldots, w_{d_2}, b_1, \ldots, b_{d_2})$, we define the (floating-point) affine transformation $\mathrm{aff}_I : \overline{\mathbb{F}}^{d_1} \to \overline{\mathbb{F}}^{d_2}$ as

$$\mathrm{aff}_I(x_1, \ldots, x_{d_1})_i \coloneqq \left(\bigoplus_{j=1}^{d_1} (w_{i,j} \otimes x_j) \right) \oplus b_i$$

for all $i \in [d_2]$. For a floating-point activation function $\sigma : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$, we will slightly abuse notation so that σ also performs a vectorized operation: for $d \in \mathbb{N}$ and $x \in \overline{\mathbb{F}}^d$,

$$\sigma(x) \coloneqq (\sigma(x_1), \ldots, \sigma(x_d)).$$

Based on the above definitions, we define a σ neural network as an alternating composition of floating-point affine transformations and activation functions. Concretely, let $l \in \mathbb{N}$,

$d_0, \dots, d_l \in \mathbb{N}$, and let $\mathrm{aff}_{I_i} : \overline{\mathbb{F}}^{d_{i-1}} \to \overline{\mathbb{F}}^{d_i}$ be affine transformations for all $i \in [l]$. Then, for $\mathcal{I} = (I_1, \dots, I_l)$, a function $\mathcal{N}_\mathcal{I} : \mathbb{F}^{d_1} \to \overline{\mathbb{F}}^{d_l}$ defined by

$$\mathcal{N}_\mathcal{I} = \mathrm{aff}_{I_l} \circ \sigma \circ \cdots \circ \mathrm{aff}_{I_2} \circ \sigma \circ \mathrm{aff}_{I_1} \tag{2}$$

is called a (floating-point) σ *neural network*. We define the number of layers of a network $\mathcal{N}$ as the number of affine transformations in $\mathcal{N}$: e.g., $\mathcal{N}_\mathcal{I}$ has l layers. For technical purposes, we also define a neural network *ending with activation functions*: for $\mathcal{I}' = (I_1, \dots, I_{l-1})$, the function

$$\mathcal{M}_{\mathcal{I}'} = \sigma \circ \mathrm{aff}_{I_{l-1}} \circ \sigma \circ \cdots \circ \mathrm{aff}_{I_2} \circ \sigma \circ \mathrm{aff}_{I_1},$$

which omits the last affine transformation, is called a neural network ending with activation functions. Here, $\mathcal{M}_{\mathcal{I}'}$ has $l-1$ layers as it contains $l-1$ affine transformations.

In this paper, we investigate the representation power of floating-point neural networks. For $d_1, d_2 \in \mathbb{N}$, $\mathcal{X} \subset \mathbb{F}^{d_1}$, and $f : \mathcal{X} \to (\mathbb{F} \cup \{-\infty, \infty\})^{d_2}$, we say that "$f$ can be represented by a (floating-point) neural network" if there exists a floating-point neural network $\mathcal{N}$ such that

$$f(x) = \mathcal{N}(x) \text{ for all } x \in \mathcal{X}.$$

Note that f can output ∞ and $-\infty$.

3. Main Results

In this section, we formally present our main results. We first introduce a necessary condition on activation functions for floating-point networks to represent floating-point functions (Section 3.1). We then introduce a sufficient condition and compare it with our necessary condition (Section 3.2). We lastly provide easily verifiable conditions on activation functions that imply the sufficient condition, and show that networks using practical activation functions can represent all floating-point functions on a wide domain (Section 3.3).

3.1. Necessary Condition on Activation Functions

Given a floating-point activation function $\sigma : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ and a domain $\mathcal{X} \subset \mathbb{F}^d$, we are interested in identifying whether σ networks can represent all functions from $\mathcal{X}$ to $\mathbb{F} \cup \{-\infty, \infty\}$. A natural observation is that such universal representation is *impossible* if there exist $x, x' \in \mathcal{X}$ such that $\sigma(\phi(x)) = \sigma(\phi(x'))$ for all floating-point affine transformations ϕ. That is, for any σ network, the outputs of the first layer at x and x' are identical. By the definition of neural networks (Eq. (2)), this implies that the final outputs of the network at x and x' must also be identical for all σ networks; and thus, σ networks cannot represent any function $f : \mathcal{X} \to \mathbb{F} \cup \{-\infty, \infty\}$ such that $f(x) \neq f(x')$.

To formally describe the above idea, we define the *distinguishability* of an input domain.

Definition 3.1 (Distinguishability)**.** Let $\sigma : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$, $d \in \mathbb{N}$, $\mathcal{X} \subset \mathbb{F}^d$, and $\mathcal{Y} \subset \overline{\mathbb{F}}$. We say that "$\mathcal{X}$ is σ-distinguishable with range $\mathcal{Y}$" if for every $x, x' \in \mathcal{X}$ with $x \neq x'$, there exists an affine transformation $\phi : \mathbb{F}^d \to \overline{\mathbb{F}}$ such that

$$\sigma(\phi(x)) \neq \sigma(\phi(x')) \quad \text{and} \quad \sigma(\phi(\mathcal{X})) \subset \mathcal{Y}. \tag{3}$$

To represent all functions from $\mathcal{X} \subset \mathbb{F}^d$ to $\mathbb{F} \cup \{-\infty, \infty\}$, one can observe that $\mathcal{X}$ should be σ-distinguishable with range $\mathbb{F} \cup \{-\infty, \infty\}$ as stated in the following lemma. See Appendix D.1 for the formal proof.

Lemma 3.2. *Let $d \in \mathbb{N}$, $\mathcal{X} \subset \mathbb{F}^d$, and $\sigma : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$. If $\mathcal{X}$ is not σ-distinguishable with range $\mathbb{F} \cup \{-\infty, \infty\}$, then there exists $f : \mathcal{X} \to \mathbb{F} \cup \{-\infty, \infty\}$ such that $f \neq g$ on $\mathcal{X}$ for all σ networks g.*

We note that if a one-dimensional subset $\mathcal{X} \subset \mathbb{F}$ is σ-distinguishable with some range, then for any $d \in \mathbb{N}$, $\mathcal{X}^d$ is also σ-distinguishable with the same range.

Using Lemma 3.2, we show in Lemma 3.3 that networks using the $\lceil \cos \rfloor$ activation function cannot represent all functions from $\left[-2^{\lfloor (p+7/2) \rfloor}, 2^{\lfloor (p+7/2) \rfloor}\right]_\mathbb{F}$ to $\mathbb{F} \cup \{-\infty, \infty\}$ (e.g., $p = 10$ for float16 and $p = 23$ for float32). The proof of Lemma 3.3 is in Appendix D.2.

Lemma 3.3. *Any $f : [-2^{\lfloor (p+7/2) \rfloor}, 2^{\lfloor (p+7/2) \rfloor}]_\mathbb{F} \to \mathbb{F}$ with $f(0) \neq f(\omega)$ cannot be represented by a $\lceil \cos \rfloor$ network.*

Nevertheless, for most practical activation functions, floating-point neural networks can represent all functions over a wide domain $(-2^{\mathfrak{e}_{\max}-2}, 2^{\mathfrak{e}_{\max}-2})_\mathbb{F}$. We will see this in the next subsection.

3.2. Sufficient Condition on Activation Functions

While the distinguishability of $\mathcal{X} \subset \mathbb{F}^d$ with range $\mathbb{F} \cup \{-\infty, \infty\}$ is necessary for representing all floating-point functions from $\mathcal{X}$ to $\mathbb{F} \cup \{-\infty, \infty\}$, the distinguishability is also sufficient under mild assumptions on its range and the activation function (Theorem 3.4).

We first introduce the following condition to explain our sufficient condition.

Condition 1. For an activation function $\sigma : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$, there exist $C_0, C_1, C_2 \in \mathbb{F}$ such that $|C_i|, |C_i - C_j| \leq 2^{\mathfrak{e}_{\max}}$ for all $0 \leq i, j \leq 2$, and

$$\sigma(C_0) = 0, \;\; 2^{\mathfrak{e}_{\min}} \leq |\sigma(C_1)| < \frac{5}{4}, \;\; |\sigma(C_2)| > (2^{-p-2})^+.$$

Condition 1 requires the existence of three points C_0, C_1, C_2 for an activation function $\sigma : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$. This condition can be easily satisfied for the correctly rounded versions of popular activation functions. For example, $\sigma(C_0) = 0$ can be satisfied for $C_0 = 0$ (e.g., ReLU, GELU, sin,

tanh) or for C_0 of large magnitude such as $-2^{\mathfrak{e}_{\max}}$ (e.g., Sigmoid). Furthermore, the conditions on C_1, C_2 can be easily satisfied by choosing some $C_1 = C_2$ so that $\sigma(C_1 = C_2) \in ((2^{-p-2})^+, 5/4)$.

Under Condition 1, we show that the distinguishability of $\mathcal{X}$ with range $[-2^{\mathfrak{e}_{\max}}, 2^{\mathfrak{e}_{\max}}]$ suffices for representing all functions from $\mathcal{X}$ to $(\mathbb{F} \cup \{-\infty, \infty\})^{d_{\text{out}}}$.

Theorem 3.4. *Let $\sigma : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$, $d_{\text{in}}, d_{\text{out}} \in \mathbb{N}$, $\mathcal{X} \subset \mathbb{F}^{d_{\text{in}}}$, and $f : \mathcal{X} \to (\mathbb{F} \cup \{-\infty, \infty\})^{d_{\text{out}}}$. Suppose that σ satisfies Condition 1 and $\mathcal{X}$ is σ-distinguishable with range $[-2^{\mathfrak{e}_{\max}}, 2^{\mathfrak{e}_{\max}}]_{\mathbb{F}}$. Then, there exists a four-layer σ network g such that $f = g$ on $\mathcal{X}$.*

There are two notable differences between our necessary condition (Lemma 3.2) and sufficient condition (Theorem 3.4): Theorem 3.4 requires Condition 1, and considers a smaller range $[-2^{\mathfrak{e}_{\max}}, 2^{\mathfrak{e}_{\max}}]_{\mathbb{F}}$ for distinguishability. First, we use Condition 1 for networks to generate all possible values in $\mathbb{F} \cup \{-\infty, \infty\}$ (see Lemma 4.2). If an activation function σ can only output too large values (e.g., $[2^{\mathfrak{e}_{\max}}, \infty]$) or too small values (e.g., in $[-\omega, \omega]$), then a σ network may not be able to generate some values in $\mathbb{F} \cup \{-\infty, \infty\}$. Second, the smaller range is due to technical reasons in our proof, which is used to avoid overflow during the evaluation of networks. If a network can generate values with large absolute values (e.g., close to Ω) while distinguishing inputs, then multiplying/adding constants to those values may incur overflow and the network may output NaN. However, if σ has a well-bounded range (i.e., $\sigma(\mathbb{F} \cup \{-\infty, \infty\}) \subset [2^{-\mathfrak{e}_{\max}}, 2^{\mathfrak{e}_{\max}}]$) as in the correctly rounded versions of Sigmoid and tanh, then this range condition is automatically satisfied.

3.3. Sufficient Conditions for Distinguishability

In this subsection, we provide easily verifiable conditions on floating-point activation functions (Lemmas 3.6 and 3.7) and real activation functions (Lemma 3.10) that imply distinguishability. Using these conditions, we show that networks with popular activation functions are universal function representers (Corollaries 3.8 and 3.11).

To describe our conditions, we first define the *separating points* of a floating-point activation function σ.

Definition 3.5 (Separating Point)**.** For $\sigma : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$, we say that $\eta \in \mathbb{F}$ is a "separating point of σ" if

$$\sigma(\eta^-) \notin \{\sigma(\eta), \sigma(\eta^+)\} \quad \text{or} \quad \sigma(\eta^+) \notin \{\sigma(\eta), \sigma(\eta^-)\}.$$

Here, η^- (or η^+) denotes the largest (or smallest) float that is smaller (or larger) than η (see Section 2.1).

We design our sufficient conditions using separating points. Specifically, for each distinct pair (x, x') of inputs in a domain $\mathcal{X}$, we aim to find a floating-point affine transformation $\phi_{x,x'}$ and a separating point $\eta_{x,x'}$ of σ such that $\phi_{x,x'}(x) = \eta^-_{x,x'}$ (or $\eta^+_{x,x'}$) and $\phi_{x,x'}(x') \in \{\eta_{x,x'}, \eta^+_{x,x'}\}$ (or $\{\eta^-_{x,x'}, \eta_{x,x'}\}$). If we can find such an affine transformation and a separating point for all distinct pairs in the domain, then the domain is distinguishable since $\sigma(\phi_{x,x'}(x)) \neq \sigma(\phi_{x,x'}(x'))$ (see Eq. (3)).

Based on the above idea, we propose our sufficient condition for distinguishability in the following lemma. The proof of Lemma 3.6 is in Appendix D.3.

Lemma 3.6. *Let $\sigma : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$, $n \in \mathbb{N}$, and $\eta_1, \dots, \eta_n \in \mathbb{F}$ be separating points of σ with $|\eta_1| \leq \cdots \leq |\eta_n|$. Suppose that integers $e_1, e_2 \in [\mathfrak{e}_{\min} + 1, \mathfrak{e}_{\max}]$ satisfy*

$$[\mathfrak{e}_{\min}, e_2]_{\mathbb{Z}} \subset \bigcup_{i=1}^{n} [\mathfrak{e}_{\eta_i} - e_1, \mathfrak{e}_{\eta_i} + \mathfrak{e}_{\max} - 2]_{\mathbb{Z}}.$$

Then, $(-2^{e_2+1}, 2^{e_2+1})_{\mathbb{F}}$ is σ-distinguishable with range

$$\mathcal{R}_{e_1,e_2} := \sigma\big([-(2^{e_1+e_2+1} \oplus |\eta_n|^+), 2^{e_1+e_2+1} \oplus |\eta_n|^+]_{\mathbb{F}}\big).$$

In Lemma 3.6, one can observe that having two separating points—one with small magnitude and one with moderate-to-large magnitude—suffices to distinguish a large domain. In particular, if σ has two separating points η_1, η_2 with

$$|\eta_1| < 2^{\mathfrak{e}_{\min}+1} \text{ and } 4 \leq |\eta_2| < 2^{\mathfrak{e}_{\max}-p-1},$$

then $(-2^{e_2+1}, 2^{e_2+1})_{\mathbb{F}}$ is σ-distinguishable with range $\mathcal{R}_{0,e_2}$ for all $e_2 \in [\mathfrak{e}_{\min} + 1, \mathfrak{e}_{\max}]_{\mathbb{Z}}$. By choosing the largest e_2^* such that $\mathcal{R}_{0,e_2^*} \subset [-2^{\mathfrak{e}_{\max}}, 2^{\mathfrak{e}_{\max}}]$ and by using Theorem 3.4, we can show that σ networks can represent all functions from $(-2^{e_2^*+1}, 2^{e_2^*+1})_{\mathbb{F}}$ to $\mathbb{F} \cup \{-\infty, \infty\}$. For example, $e_2^* = \mathfrak{e}_{\max} - 1$ when $\sigma = \text{ReLU}$.

We note that having a separating point with small absolute value (e.g., $\approx 2^{\mathfrak{e}_{\min}}$) is critical for distinguishing a large domain using Lemma 3.6, while avoiding overflow. This is because, to distinguish two small numbers (e.g., 0 and ω), we find $w, b \in \mathbb{F}$ and a separating point $\eta \in \mathbb{F}$ such that $w \otimes 0 \oplus b = \eta^-$ (or η^+) and $w \otimes \omega \oplus b \in \{\eta, \eta^+\}$ (or $\{\eta, \eta^-\}$). If η is large in magnitude, then w and b must also be large. As a result, overflow may occur when the domain contains large numbers. We also note that a separating point with moderate-to-large absolute value (e.g., ≥ 1) is necessary to distinguish a large domain using Lemma 3.6.

Using Lemma 3.6, we provide a sufficient condition for the distinguishability of floating-point activation functions that have finite values at all inputs including $\pm\infty$, such as $\lceil \text{Sigmoid} \rfloor$. The proof of Lemma 3.7 is in Appendix D.4.

Lemma 3.7. *Let $\sigma : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ such that $\sigma(\mathbb{F} \cup \{-\infty, \infty\}) \subset [-2^{\mathfrak{e}_{\max}}, 2^{\mathfrak{e}_{\max}}]_{\mathbb{F}}$. Suppose that σ has two separating points $|\eta_1| < 2$ and $|\eta_2| \geq 4$. Then, $\mathbb{F}$ is σ-distinguishable with range $[-2^{\mathfrak{e}_{\max}}, 2^{\mathfrak{e}_{\max}}]_{\mathbb{F}}$.*

Lemma 3.7 states that if $|\sigma(\mathbb{F} \cup \{\infty, -\infty\})|$ is bounded by $2^{\mathfrak{e}_{\max}}$ and there exist two separating points of moderate size,

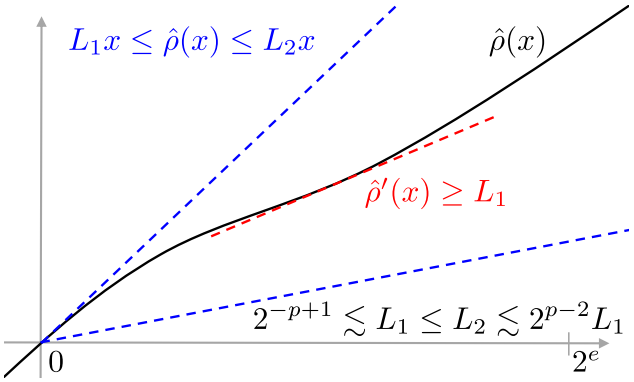

Figure 1: Visualization of the conditions in Lemma 3.10.

then $\mathbb{F}$ is σ-distinguishable with range $[-2^{\mathfrak{e}_{\max}}, 2^{\mathfrak{e}_{\max}}]_{\mathbb{F}}$. By Theorem 3.4, this implies that σ networks can represent all functions from $\mathbb{F}^d$ to $\mathbb{F} \cup \{-\infty, \infty\}$, leading to the following corollary for $\lceil \mathrm{Sigmoid} \rfloor$ and $\lceil \tanh \rfloor$.

Corollary 3.8. *Let σ be one of $\lceil \mathrm{Sigmoid} \rfloor$ and $\lceil \tanh \rfloor$. Then, for any $d \in \mathbb{N}$, σ networks can represent all functions from $\mathbb{F}^d$ to $\mathbb{F} \cup \{-\infty, \infty\}$.*

We now consider a more realistic scenario: σ is the correctly rounded version of a real activation function ρ. For this case, we first introduce a sufficient condition for having a separating point. The proof of Lemma 3.9 is in Appendix D.5.

Lemma 3.9. *Let $\rho : \mathbb{R} \to \mathbb{R}$, $a, b \in \mathbb{F}$, and $e \in [\mathfrak{e}_{\min}, \mathfrak{e}_{\max}]_{\mathbb{Z}}$ with $|\lceil \rho \rfloor (a)| \le 2^{e+1}(1 - 2^{-p})$. Suppose that there exists $L > 0$ such that $L(b - a) \ge 2^{e-p}$ and either*

- *$\lceil \rho \rfloor (a) \ge 0$ and $\rho'(x) \ge L$ for all $x \in [a, b]$, or*
- *$\lceil \rho \rfloor (a) \le 0$ and $\rho'(x) \le -L$ for all $x \in [a, b]$.*

Then, there exists a separating point $\eta \in [a, b]_{\mathbb{F}}$ of $\lceil \rho \rfloor$.

Lemma 3.9 states that if ρ is sufficiently increasing or decreasing on a long enough interval, then that interval contains a separating point of $\lceil \rho \rfloor$. Using this lemma, we now present a sufficient condition on a real activation function ρ that guarantees the distinguishability of its correctly rounded version $\lceil \rho \rfloor$. The proof of Lemma 3.10 is in Appendix D.6.

Lemma 3.10. *Let $\rho, \zeta : \mathbb{R} \to \mathbb{R}$ with $\zeta(x) = -x$. Suppose that there exist $\hat{\rho} \in \{\rho, \zeta \circ \rho, \rho \circ \zeta, \zeta \circ \rho \circ \zeta\}$, $L_1, L_2 > 0$, and $e \in \mathbb{Z}$ such that the following hold:*

- *$L_1 x \le \hat{\rho}(x) \le L_2 x$ for all $x \in [0, 2^e)$.*
- *$\hat{\rho}'(x) \ge L_1$ for all $x \in (0, 2^e)$.*
- *$-p \le l_1 \le l_2 \le l_1 + p$, where $l_1 = \lfloor \log_2(L_1/2) \rfloor$ and $l_2 = \lfloor \log_2(2L_2) \rfloor$.*

Then, $(-2^{e'}, 2^{e'})_{\mathbb{F}}$ is $\lceil \rho \rfloor$-distinguishable with range $[-2^{\mathfrak{e}_{\max}}, 2^{\mathfrak{e}_{\max}}]_{\mathbb{F}}$, where $e' = \mathfrak{e}_{\max} + \min\{e - 2, -l_2\}$.

Lemma 3.10 roughly states that for some moderate-size $L_1, L_2 > 0$, if a real activation function ρ is bounded between $L_1 x$ and $L_2 x$, and ρ' is lower bounded by L_1 for all inputs between 0 and 2^e, then the domain $(-2^{e'}, 2^{e'})_{\mathbb{F}}$ is $\lceil \rho \rfloor$-distinguishable with range $[-2^{\mathfrak{e}_{\max}}, 2^{\mathfrak{e}_{\max}}]_{\mathbb{F}}$, where $e' \approx \mathfrak{e}_{\max} + e - 2$. By Theorem 3.4, this implies that $\lceil \rho \rfloor$ networks can represent all functions from $(-2^{e'}, 2^{e'})_{\mathbb{F}}$ to $\mathbb{F} \cup \{-\infty, \infty\}$. Since Lemma 3.10 covers various practical activation functions ρ with $\rho(0) = 0$, the following corollary holds for the correctly rounded version of such ρ. The proof of Corollary 3.11 is in Appendix D.7.

Corollary 3.11. *Let σ be one of $\lceil \mathrm{Identity} \rfloor$, $\lceil \mathrm{ReLU} \rfloor$, $\lceil \mathrm{ELU} \rfloor$, $\lceil \mathrm{SeLU} \rfloor$, $\lceil \mathrm{GELU} \rfloor$, $\lceil \mathrm{Swish} \rfloor$, $\lceil \mathrm{Mish} \rfloor$, and $\lceil \sin \rfloor$. Then, for any $d \in \mathbb{N}$, σ networks can represent all functions from $(-2^{\mathfrak{e}_{\max}-2}, 2^{\mathfrak{e}_{\max}-2})_{\mathbb{F}}^d$ to $\mathbb{F} \cup \{-\infty, \infty\}$.*

4. Proof of Theorem 3.4

We prove Theorem 3.4 by explicitly constructing the target four-layer network g under Condition 1. Specifically, we construct g as a linear combination of indicator functions of points (Lemma 4.1). Given z in the domain, our construction of the indicator function of z consists of three parts. In the first part, we create an injective layer consisting of an affine transformation followed by an activation function based on the distinguishability. Due to the injectivity, all inputs in the domain have distinct output vectors after passing this part. Here, we use $y = (y_1, \ldots, y_k)$ to denote the output of the first part of z. In the second part, we construct $2n$ binary step functions $f_{1,1}, f_{1,2}, \ldots, f_{n,1}, f_{n,2}$ where $f_{i,1}(x) = (C_1 - C_0)1[x \ge y_i] + C_0$ and $f_{i,2}(-x) = (C_1 - C_0)1[x \le y_i] + C_0$ where C_0, C_1 are from Condition 1. Namely, all $f_{i,1}$ (or $f_{i,2}$) is C_1 if and only if the input to the network is z (or $-z$). To implement such $f_{i,j}$, we exploit the round-off error in floating-point operations and show that we can increase the gap between arbitrary two numbers by sequentially adding the same floats to those two numbers (see Lemma 4.6). In the third part, we construct a function that outputs C_1 or C_2 if all $f_{i,j}$ are C_1 and outputs C_0 otherwise. We then apply the activation function after this function so that our indicator function is either zero (i.e., $\sigma(C_0)$) if the input is z or some non-zero value (i.e., $\sigma(C_1)$ or $\sigma(C_2)$) otherwise.

We now formally prove Theorem 3.4.

Lemma 4.1. *Let $\sigma : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$, $d \in \mathbb{N}$, and $\mathcal{X} \subset \mathbb{F}^d$. Suppose that σ satisfies Condition 1 and $\mathcal{X}$ is σ-distinguishable with range $[-2^{\mathfrak{e}_{\max}}, 2^{\mathfrak{e}_{\max}}]_{\mathbb{F}}$. Then, for any $z \in \mathbb{F}^d$ and $c \in \{C_1, C_2\}$, there exists a three-layer σ network $f : \mathcal{X} \to \mathbb{F}$ ending with the activation function such that*

$$f(x) = \sigma(c)\, \mathbb{1}_z(x).$$

We present the proof of Lemma 4.1 in Section 4.2. Lemma 4.1 states that if an activation function σ satisfies Condition 1 and the domain $\mathcal{X}$ is σ-distinguishable,

then we can construct an indicator function of $z \in \mathbb{F}$ with a coefficient of $\sigma(C_1)$ or $\sigma(C_2)$. With Condition 1, this implies that we can construct indicator functions with coefficients $\sigma(C_1) \in [2^{\mathfrak{e}_{\min}}, 5/4]$ and $\sigma(C_2) > (2^{-p-2})^+$. The following lemma states that applying affine transformations to these coefficients suffice for generating all floats in $\mathbb{F} \cup \{-\infty, \infty\}$.

Lemma 4.2. *Suppose that $\sigma : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ satisfies Condition 1. Then, for any $x \in \mathbb{F} \cup \{-\infty, \infty\}$, there exist $n \in \mathbb{N}$, $w_1, \dots, w_n \in \mathbb{F}$, and $z_1, \dots, z_n \in \{C_1, C_2\}$ such that*

$$x = (w_1 \otimes \sigma(z_1)) \oplus \cdots \oplus (w_n \otimes \sigma(z_n)).$$

To prove Lemma 4.2, we show that for each $x \in \mathbb{F}$, there exist $w \in \mathbb{F}$ and $z \in \{C_1, C_2\}$ such that $x \oplus (w \otimes \sigma(z)) = x^+$. Here, we note that $w \otimes \sigma(z)$ may not be exactly equal to $x^+ - x$ due to the round-off error. For a more formal argument, see the full proof in Appendix D.8.

We now prove Theorem 3.4 using Lemmas 4.1 and 4.2. Without loss of generality, we assume $d_2 = 1$, i.e., the target function f is scalar-valued. To represent $f' : \mathcal{X} \to (\mathbb{F} \cup \{-\infty, \infty\})^{d_2}$ with $d_2 > 1$, we can construct d_2 networks that represent coordinatewise functions $x \mapsto f'(x)_i$ and concatenate them.

For any $f : \mathcal{X} \to \mathbb{F} \cup \{-\infty, \infty\}$, we can represent f as follows:

$$f(x) = \sum_{y \in \mathcal{X}} f(y) \mathbb{1}_y(x).$$

Let $c \coloneqq f(y)$. Then, by Condition 1 and Lemma 4.2, for any $c \in \mathbb{F} \cup \{-\infty, \infty\}$, there exist $n_c \in \mathbb{N}$, $z_{c,1}, \dots, z_{c,n_c} \in \{C_1, C_2\}$, and $w_{c,1}, \dots, w_{c,n_c} \in \mathbb{F}$ such that

$$c = \bigoplus_{i=1}^{n_c} (w_{c,i} \otimes \sigma(z_{c,i})).$$

By Lemma 4.1, for each $f(y) = c \in \mathbb{F}$ and $i \in [n_c]$, there exists a three-layer σ-network $h_{c,i} : \mathcal{X} \to \mathbb{F}$ ending with the activation function such that $h_{c,i}(x) = \sigma(z_{c,i}) \mathbb{1}_y(x)$. We construct the target four-layer σ network g as follows:

$$g = \bigoplus_{y \in \mathcal{X}} \bigoplus_{i=1}^{n_{f(y)}} w_{f(y),i} \otimes h_{f(y),i}.$$

Since $\mathbb{1}_y(x) = 0$ if $y \neq x$, it holds that for each $x \in \mathcal{X}$,

$$g(x) = \bigoplus_{i=1}^{n_{f(x)}} (w_{f(x),i} \otimes \sigma(z_{f(x),i})) = f(x).$$

This completes the proof of Theorem 3.4.

4.1. Sequential Addition and Transferability

To describe our proof of Lemma 4.1, we introduce the *sequential addition* defined as follows.

$$f_1 \circ f_2(x) = f_2(x) \oplus z_1 \oplus \cdots \oplus z_n \equiv f_2(x) \oplus (w_1 \otimes \sigma(c_1)) \oplus \cdots \oplus (w_n \otimes \sigma(c_n))$$

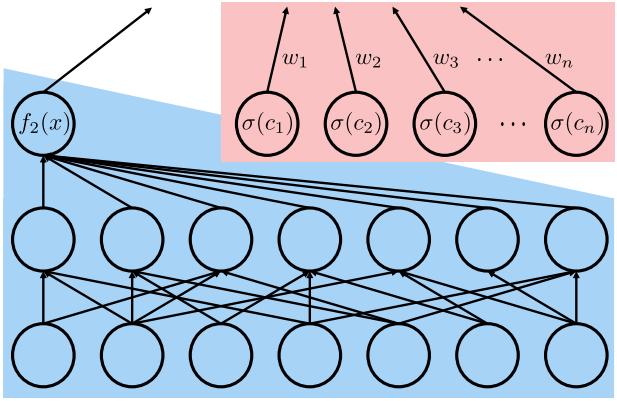

Figure 2: A composition of a network f_2 ending with an activation function (blue) and a sequential addition f_1 (red). Each z_i in the sequential addition can be represented by $w_i \otimes \sigma(c_i)$ and $f_1 \circ f_2$ can be represented by a network.

Definition 4.3 (Sequential addition)**.** Let $\sigma : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ and $\Sigma_\sigma \coloneqq \{w \otimes \sigma(c) : w, c \in \mathbb{F} \text{ with } w \otimes \sigma(c) \in \mathbb{F}\}$. We say a function $f : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ is a "sequential addition using σ" if $f(\mathbb{F}) \subset \mathbb{F}$ and there exist $n \in \mathbb{N}$ and $z_1, \dots, z_n \in \Sigma_\sigma$ such that for each $x \in \mathbb{F}$,

$$f(x) = x \oplus z_1 \oplus \cdots \oplus z_n. \tag{4}$$

We often drop σ and use Σ to denote Σ_σ if it is clear from the context.

We often compose a sequential addition with a network ending with an activation function: for a sequential addition f_1 using σ and a σ network $f_2 : \mathbb{F}^d \to \overline{\mathbb{F}}$, $f_1 \circ f_2$ is also a σ network (see Fig. 2 for an illustration). Here, we use additional activation functions and biases to represent z_i in Eq. (4). However, since the floating-point addition is not associative, the sequential addition in Eq. (4) can be different from adding a single float (e.g., a bias) to x, i.e., $x \mapsto x \oplus (z_1 \oplus \cdots \oplus z_n)$. We note that for sequential additions g_1, g_2 using σ, their composition $g_1 \circ g_2$ is also a sequential addition.

By Definition 4.3, for any sequential addition f, we have $f(\mathbb{F}) \subset \mathbb{F}$, i.e., overflow does not occur for finite inputs. Furthermore, by the monotonicity of the floating-point addition, we have $f(x_1) \leq f(x_2)$ for all finite floats $x_1 \leq x_2$, i.e., a sequential addition preserves the order of inputs. These properties also imply the following lemma. The proof of Lemma 4.4 is presented in Appendix D.9.

Lemma 4.4. *For any $\sigma : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ and a sequential addition f using σ, $f([-2^{\mathfrak{e}_{\max}}, 2^{\mathfrak{e}_{\max}}]_{\mathbb{F}}) \subset [-2^{\mathfrak{e}_{\max}}, 2^{\mathfrak{e}_{\max}}]_{\mathbb{F}}$.*

To concisely describe what the sequential addition can do,

we define the *transferability* as follows.

Definition 4.5 (Transferability)**.** Let $n \in \mathbb{N}$ and $(x_1, \dots, x_n), (y_1, \dots, y_n) \in \mathbb{F}^n$. We say "$(x_1, \dots, x_n)$ is transferable to $(y_1, \dots, y_n)$ using σ" or write "$(x_1, \dots, x_n) \overset{\sigma}{\Longrightarrow} (y_1, \dots, y_n)$" if there exists a sequential addition $f : \mathbb{F} \to \mathbb{F}$ using σ such that $f(x_i) = y_i$ for all $i \in [n]$.

By Definitions 4.3 and 4.5, one can observe that

$$(x_1, x_2) \overset{\sigma}{\Longrightarrow} (x_1 \oplus z_1 \oplus \cdots \oplus z_n, x_2 \oplus z_1 \oplus \cdots \oplus z_n)$$

for all $z_1, \dots, z_n \in \Sigma_\sigma$. We now describe what sequential addition can do in the following lemma.

Lemma 4.6. *Let $\sigma : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ and suppose that σ satisfies Condition 1. Then, for any $y \in [-2^{\mathfrak{e}_{\max}}, 2^{\mathfrak{e}_{\max}})_{\mathbb{F}}$ and $x_1, x_2 \in [-2^{\mathfrak{e}_{\max}}, 2^{\mathfrak{e}_{\max}}]_{\mathbb{F}}$ such that $x_2 - x_1 \in (0, 2^{\mathfrak{e}_{\max}}]_{\mathbb{F}}$, it holds that*

$$(-2^{\mathfrak{e}_{\max}}, y, y^+, 2^{\mathfrak{e}_{\max}}) \overset{\sigma}{\Longrightarrow} (x_1, x_1, x_2, x_2).$$

Due to the order-preserving property of sequential additions, for the sequential addition (say f) in Lemma 4.6, it holds that $f([-2^{\mathfrak{e}_{\max}}, y]_{\mathbb{F}}) = \{x_1\}$ and $f([y^+, 2^{\mathfrak{e}_{\max}}]_{\mathbb{F}}) = \{x_2\}$. Namely, we can *split* the set $[-2^{\mathfrak{e}_{\max}}, 2^{\mathfrak{e}_{\max}}]_{\mathbb{F}}$ into two parts with respect to the threshold y using f. We note that such a sequential addition f cannot be constructed by adding a single bias term; we use the round-off error and non-associativity of floating-point addition to prove Lemma 4.6. For more details, see the proof of Lemma 4.6 in Section 4.3.

4.2. Proof of Lemma 4.1

We are now ready to prove Lemma 4.1. By the definition of the distinguishability, there exist $n \in \mathbb{N}$ and affine transformations $\phi_1, \dots, \phi_n : \overline{\mathbb{F}}^d \to \overline{\mathbb{F}}$ satisfying the following two properties: (i) for any $y \in \mathcal{X}$, there exists $j_y \in [n]$ such that $\sigma(\phi_{j_y}(z)) \neq \sigma(\phi_{j_y}(y))$ and (ii) $\sigma(\phi_j(\mathcal{X})) \subset [-2^{\mathfrak{e}_{\max}}, 2^{\mathfrak{e}_{\max}}]_{\mathbb{F}}$ for all $j \in [n]$.

Consider $C_0, C_1 \in \mathbb{F}$ in Condition 1. Without loss of generality, we assume $C_0 < C_1$, $\sigma(C_1) > 0$, and $C_0 < c$; the proof for the remaining cases can be done similarly.

By Lemma 4.6, there exist sequential additions $f_{j,1}, f_{j,2}$ for all $j \in [n]$ such that

$$f_{j,1}(x) = \begin{cases} C_1 & \text{if } \sigma(\phi_j(z)) \le x \le 2^{\mathfrak{e}_{\max}}, \\ C_0 & \text{if } -2^{\mathfrak{e}_{\max}} \le x < \sigma(\phi_j(z)), \end{cases}$$

$$f_{j,2}(-x) = \begin{cases} C_0 & \text{if } \sigma(\phi_j(z)) < x \le 2^{\mathfrak{e}_{\max}}, \\ C_1 & \text{if } -2^{\mathfrak{e}_{\max}} \le x \le \sigma(\phi_j(z)). \end{cases}$$

Define $g_{j,k} : \mathcal{X} \to \mathbb{F}$ as

$$g_{j,k}(x) \coloneqq \sigma\left(f_{j,k}\left((-1)^{k-1} \times \sigma(\phi_j(x))\right)\right).$$

Then, one can observe that $g_{j,k}$ is a two-layer network ending with the activation function. Furthermore, for any $x \in \mathcal{X}$, $g_{j,k}(x) = \sigma(C_1)$ for all $j \in [n]$ and $k \in \{1, 2\}$ if and only if $x = z$; otherwise, there exists some j, k such that $g_{j,k}(x) = \sigma(C_0) = 0$.

We now construct the target three-layer network f ending with the activation function σ as

$$f(x) \coloneqq \sigma(h(g_{1,1}(x), g_{1,2}(x), \dots, g_{n,1}(x), g_{n,2}(x)))$$
$$h(y_1, \dots, y_{2n}) \coloneqq h_{2n-1}\left(\cdots h_2\left(h_1(y_1 \oplus y_2) \oplus y_3\right) \cdots \oplus y_{2n}\right)$$

for some sequential additions $h_1, \dots, h_{2n-1}$. Here, we choose $h_1, h_2, \dots, h_{2n-1}$ such that for $i \in [2n-2]$,

$$h_i : (0, \sigma(C_1), 2\sigma(C_1)) \overset{\sigma}{\Longrightarrow} (0, 0, \sigma(C_1)),$$
$$h_{2n-1} : (0, \sigma(C_1), 2\sigma(C_1)) \overset{\sigma}{\Longrightarrow} (C_0, C_0, c).$$

Note that such sequential additions exist by Lemma 4.6. Under these choices of $h_1, \dots, h_{2n}$, one can observe that for any $x \in \mathcal{X}$, $f(x) = 0$ if $x \neq z$ and $f(x) = \sigma(c)$ if $x = z$. This completes the proof of Lemma 4.1 which leads to the proof of Theorem 3.4

4.3. Proof of Lemma 4.6

To describe our proof of Section 4.3, we introduce the following lemmas. The proofs of Lemmas 4.7–4.9 are presented in Appendices D.10–D.12, respectively.

Lemma 4.7. *Let $\sigma : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ and suppose that σ satisfies Condition 1. Then, for any $x_1, x_2 \in [-2^{\mathfrak{e}_{\max}}, 2^{\mathfrak{e}_{\max}}]_{\mathbb{F}}$ with $x_1 < x_2$, there exists a constant $y \in (0, 2^{\mathfrak{e}_{\max}}]_{\mathbb{F}}$ such that*

$$(x_1, x_2) \overset{\sigma}{\Longrightarrow} (0, y).$$

Lemma 4.8. *Let $\sigma : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ and suppose that σ satisfies Condition 1. Then, for any $x_1, x_2 \in [-2^{\mathfrak{e}_{\max}}, 2^{\mathfrak{e}_{\max}}]_{\mathbb{F}}$ with $x_1 x_2 > 0$,*

$$(0, x_1) \overset{\sigma}{\Longrightarrow} (0, x_2).$$

Lemma 4.9. *Let $\sigma : \overline{\mathbb{F}} \to \overline{\mathbb{F}}$ and suppose that σ satisfies Condition 1. Then, for any $x_1, x_2 \in [-2^{\mathfrak{e}_{\max}}, 2^{\mathfrak{e}_{\max}}]_{\mathbb{F}}$ with $x_2 - x_1 \in (0, 2^{\mathfrak{e}_{\max}}]$, there exists $x \in (0, 2^{\mathfrak{e}_{\max}}]_{\mathbb{F}}$ such that*

$$(0, x) \overset{\sigma}{\Longrightarrow} (x_1, x_2).$$

We now prove Lemma 4.6. By Lemma 4.7, there exists $c \in (0, 2^{\mathfrak{e}_{\max}}]_{\mathbb{F}}$ and a sequential addition $f_1 : (y, y^+) \overset{\sigma}{\Longrightarrow} (0, c)$. Then, by the order-preserving property of sequential additions, it holds that

$$f_1(x) \in [-2^{\mathfrak{e}_{\max}}, 0]_{\mathbb{F}}, \quad f_1(z) \in [c, 2^{\mathfrak{e}_{\max}}]_{\mathbb{F}},$$

for all $x \in [-2^{\mathfrak{e}_{\max}}, y]_{\mathbb{F}}$ and $z \in [y^+, 2^{\mathfrak{e}_{\max}}]_{\mathbb{F}}$. Furthermore, by Lemma 4.8, there exists a sequential addition

$f_2 : (0, c) \overset{\sigma}{\Longrightarrow} (0, 2^{\mathfrak{e}_{\max}})$. Again, by the order-preserving property of sequential additions, we have

$$f_2 \circ f_1(x) \ominus \omega \in [-2^{\mathfrak{e}_{\max}}, -\omega]_{\mathbb{F}}, \;\; f_2 \circ f_1(z) \ominus \omega = 2^{\mathfrak{e}_{\max}}$$

for all $x \in [-2^{\mathfrak{e}_{\max}}, y]_{\mathbb{F}}$ and $z \in [y^+, 2^{\mathfrak{e}_{\max}}]_{\mathbb{F}}$. Here, note that $f_3 \coloneqq f_2 \circ f_1 \ominus \omega$ is also a sequential addition. We also choose a sequential addition $f_4 : (-\omega, 0) \overset{\sigma}{\Longrightarrow} (-2^{\mathfrak{e}_{\max}}, 0)$ which exists by Lemma 4.8. Then, it holds that

$$f_4 \circ f_3(x) = -2^{\mathfrak{e}_{\max}}, \;\; f_4 \circ f_3(z) = f_4(2^{\mathfrak{e}_{\max}}) \in [0, 2^{\mathfrak{e}_{\max}}]_{\mathbb{F}}$$

for all $x \in [-2^{\mathfrak{e}_{\max}}, y]_{\mathbb{F}}$ and $z \in [y^+, 2^{\mathfrak{e}_{\max}}]_{\mathbb{F}}$.

By Lemma 4.9, there exist $x^* \in [-2^{\mathfrak{e}_{\max}}, 2^{\mathfrak{e}_{\max}}]_{\mathbb{F}}$ and a sequential addition $f_5 : (0, x^*) \overset{\sigma}{\Longrightarrow} (x_1, x_2)$. In addition, by Lemmas 4.7 and 4.8, there exists a sequential addition $f_6 : (-2^{\mathfrak{e}_{\max}}, f_4(2^{\mathfrak{e}_{\max}})) \overset{\sigma}{\Longrightarrow} (0, x^*)$. Then, we have

$$f_5 \circ f_6 \circ f_4 \circ f_3(x) = x_1, \;\; f_5 \circ f_6 \circ f_4 \circ f_3(z) = x_2$$

for all $x \in [-2^{\mathfrak{e}_{\max}}, y]_{\mathbb{F}}$ and $z \in [y^+, 2^{\mathfrak{e}_{\max}}]_{\mathbb{F}}$. This implies that $(-2^{\mathfrak{e}_{\max}}, y, y^+, 2^{\mathfrak{e}_{\max}}) \overset{\sigma}{\Longrightarrow} (x_1, x_1, x_2, x_2)$ and completes the proof of Lemma 4.6.

5. Quantitative Analysis

Our network construction uses width $O(nd_{\text{in}}2^{(p+q)(d_{\text{in}}+1)})$, where d_{in} is the input dimension, p and q are the numbers of bits for the mantissa and exponent, respectively, and n is a constant depending on the activation function. Specifically, we use width $O(nd_{\text{in}}2^{p+q})$ for each indicator function, and the entire network consists of $O(2^{(p+q)d_{\text{in}}})$ such indicator functions—one for each input. This results in a total width of $O(nd_{\text{in}}2^{(p+q)(d_{\text{in}}+1)})$.

We note that a width of $2^{\Theta((p+q)d_{\text{in}})}$ is *necessary* for floating-point networks of constant depth to represent all functions from a floating-point domain (e.g., $[-1, 1]_{\mathbb{F}}^{d_{\text{in}}}$) to $\mathbb{F} \cup \{-\infty, \infty\}$. Consider a floating-point network with width W and constant depth. Such a network has at most $O(W^2)$ parameters. Since each parameter can take at most $O(2^{p+q})$ distinct values, the network can represent at most $2^{O((p+q)W^2)}$ distinct functions. In contrast, the number of all functions from $[-1, 1]_{\mathbb{F}}^{d_{\text{in}}}$ to $\mathbb{F} \cup \{-\infty, \infty\}$ is $2^{(p+q)2^{\Theta((p+q)d_{\text{in}})}}$, as there are $2^{\Theta((p+q)d_{\text{in}})}$ inputs and $\Theta(2^{p+q})$ outputs. Therefore, to represent all such functions, the network must have width at least $2^{\Theta((p+q)d_{\text{in}})}$.

Such exponential growth in d_{in} also arises in the real-valued setting: under constant depth, networks constructed in prior works (e.g., Yarotsky (2017; 2018); Park et al. (2021); Zhang et al. (2024)) require width $\varepsilon^{-\Theta(d)}$, where ε denotes the target approximation error. Moreover, this exponential growth has been shown to be *necessary* for ReLU networks (Yarotsky, 2018). Our results in the floating-point setting closely mirror these findings in the real-valued setting.

Nevertheless, our results do not establish (i) the minimum depth and width or (ii) depth-width trade-offs for floating-point universal approximation. This contrasts with existing results for real-valued universal approximation:

(i) It is well-known that depth 2 is necessary and sufficient for universal approximation with non-polynomial activation functions (Pinkus, 1999). Recently, it has been shown that width $\max\{d_{\text{in}}, d_{\text{out}}, 2\}$ is necessary and sufficient to approximate any continuous function from $[0, 1]^{d_{\text{in}}}$ to $\mathbb{R}^{d_{\text{out}}}$ in the L^p distance ($1 \leq p < \infty$), for activation functions that can approximate the identity function and the binary step function (Shin et al., 2025). The minimum width in the L^∞ distance is known only for specific pairs of d_{in} and d_{out} (Kim et al., 2024).

(ii) For ReLU networks, the number of parameters required for universal approximation decreases from $\Theta(\varepsilon^{-d_{\text{in}}})$ to $\Theta(\varepsilon^{-d_{\text{in}}/2})$ as the network depth increases (Yarotsky, 2018). This benefit of depth has also been observed in the approximation of specific function classes (Telgarsky, 2016; Safran & Shamir, 2017; Chatziafratis et al., 2020).

We believe identifying the minimum depth/width and understanding depth-width trade-offs for floating-point universal approximation are important directions for future research.

6. Conclusion

In this work, we propose necessary and sufficient conditions on activation functions for floating-point neural networks to represent a large class of floating-point functions. Specifically, we demonstrate that the distinguishability of an activation function is crucial for determining the representability of neural networks. Furthermore, our results show that networks using correctly rounded practical activation functions can represent all floating-point functions on a wide domain. We believe that our research contributes to the theoretical understanding of practical floating-point neural networks and will provide a solid foundation for future research.

Acknowledgements

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

and AP090301); National Research Foundation of Korea grant funded by MSIT (RS-2025-00515264 and RS-2024-00406127); the Global University Project grant funded by GIST in 2025; and the Sejong University faculty research fund in 2025.

Impact Statement

This paper investigates the representability of floating-point neural networks. We could not find notable potential societal consequences of our work.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

Chatziafratis, V., Nagarajan, S. G., Panageas, I., and Wang, X. Depth-width trade-offs for ReLU networks via Sharkovsky's theorem. In *International Conference on Learning Representations (ICLR)*, 2020.

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

Ding, Y., Liu, J., Xiong, J., and Shi, Y. On the universal approximability and complexity bounds of quantized ReLU neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.

Gonon, A., Brisebarre, N., Gribonval, R., and Riccietti, E. Approximation speed of quantized versus unquantized ReLU neural networks and beyond. *IEEE Transactions on Information Theory*, 69(6):3960–3977, 2023.

Google. Improve your model's performance with bfloat16. https://cloud.google.com/tpu/docs/bfloat16.

Hendrycks, D. and Gimpel, K. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

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

Hwang, G., Park, Y., and Park, S. On expressive power of quantized neural networks under fixed-point arithmetic. *arXiv preprint arXiv:2409.00297*, 2024.

Hwang, G., Lee, W., Park, Y., Park, S., and Saad, F. Floating-point neural networks are provably robust universal approximators. In *International Conference on Computer Aided Verification (CAV)*, 2025.

IEEE. *IEEE Standard for Floating-Point Arithmetic (IEEE Std 754-2019)*. IEEE, Piscataway, NJ, USA, 2019. doi: 10.1109/IEEESTD.2019.8766229.

Kidger, P. and Lyons, T. Universal approximation with deep narrow networks. In *Conference on Learning Theory (COLT)*, 2020.

Kim, N., Min, C., and Park, S. Minimum width for universal approximation using ReLU networks on compact domain. In *International Conference on Learning Representations (ICLR)*, 2024.

Klambauer, G., Unterthiner, T., Mayr, A., and Hochreiter, S. Self-normalizing neural networks. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature*, 521(7553):436–444, 2015.

Leshno, M., Lin, V. Y., Pinkus, A., and Schocken, S. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 1993.

Lu, Z., Pu, H., Wang, F., Hu, Z., and Wang, L. The expressive power of neural networks: A view from the width. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

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 preprint arXiv:2209.05433*, 2022.

Park, S., Yun, C., Lee, J., and Shin, J. Minimum width for universal approximation. In *International Conference on Learning Representations (ICLR)*, 2021.

Park, Y., Hwang, G., Lee, W., and Park, S. Expressive power of ReLU and step networks under floating-point operations. *Neural Networks*, 175:106297, 2024.

Pinkus, A. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8:143 – 195, 1999.

Ramachandran, P., Zoph, B., and Le, Q. V. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

Ramanujan, V., Wortsman, M., Kembhavi, A., Farhadi, A., and Rastegari, M. What's hidden in a randomly weighted neural network? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Safran, I. and Shamir, O. Depth-width tradeoffs in approximating natural functions with neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

Shin, J., Kim, N., Hwang, G., and Park, S. Minimum width for universal approximation using squashable activation functions. In *International Conference on Machine Learning (ICML)*, 2025.

Sitzmann, V., Martel, J., Bergman, A., Lindell, D., and Wetzstein, G. Implicit neural representations with periodic activation functions. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Tabuada, P. and Gharesifard, B. Universal approximation power of deep residual neural networks via nonlinear control theory. In *International Conference on Learning Representations (ICLR)*, 2021.

Telgarsky, M. Benefits of depth in neural networks. In *Conference on Learning Theory (COLT)*, 2016.

Yarotsky, D. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017.

Yarotsky, D. Optimal approximation of continuous functions by very deep ReLU networks. In *Conference on Learning Theory (COLT)*, 2018.

Yuan, C. and Agaian, S. S. A comprehensive review of binary neural network. *Artificial Intelligence Review*, 56 (11):12949–13013, 2023.

Yun, C., Bhojanapalli, S., Rawat, A. S., Reddi, S., and Kumar, S. Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations (ICLR)*, 2020.

Zhang, S., Lu, J., and Zhao, H. Deep network approximation: Beyond ReLU to diverse activation functions. *Journal of Machine Learning Research (JMLR)*, 25(35): 1–39, 2024.

Zhou, D.-X. Universality of deep convolutional neural networks. *Applied and Computational Harmonic Analysis*, 48(2):787–794, 2020.

A. Notation Table

Table 2: Notation table

Symbol	Description	Definition	Reference
$\mathbb{F}_{p,q} = \mathbb{F}$	set of all finite floating-point numbers.		Section 2.1
$\overline{\mathbb{F}}_{p,q} = \overline{\mathbb{F}}$	set of all floating-point numbers	$\mathbb{F} \cup \{-\infty, \infty, \text{NaN}\}$	Section 2.1
q	bit for exponent.		Section 2.1
p	bit for mantissa.		Section 2.1
$\mathfrak{e}_{\min}$	the smallest exponent of $\mathbb{F}$	$-2^{q-1}+2$.	Section 2.1
$\mathfrak{e}_{\max}$	the largest exponent of $\mathbb{F}$.	$2^{q-1}-1$	Section 2.1
ω	the smallest positive float.	$2^{\mathfrak{e}_{\min}-p}$	Section 2.1
Ω	the largest positive float.	$2^{\mathfrak{e}_{\max}} \times (2-2^{-p})$	Section 2.1
∞ , $-\infty$	positive, negative infinity.		Section 2.1
NaN	Not a number.		Section 2.1
$\mathfrak{e}_x$	the exponent of $x \in \mathbb{F}$.		Section 2.1
$\mathfrak{m}_x$	the significand of $x \in \mathbb{F}$.		Section 2.1
$\widetilde{\mathfrak{e}}_x$	normalized exponent of $x \in \mathbb{F}$.	Eq. (7)	Appendix B
$\widetilde{\mathfrak{m}}_x$	normalized mantissa of $x \in \mathbb{F}$.	Eq. (7)	Appendix B
x^-	the predecessor of $x \in \mathbb{F}$.		Section 2.1
x^+	the successor of $x \in \mathbb{F}$.		Section 2.1
$\lceil \cdot \rfloor_{\mathbb{F}}$	rounding operation.	round to nearest (tie-to-even rule)	Section 2.1
ρ	real-valued activation function.		Section 2.1
$\lceil \rho \rfloor$	corrected rounded function of ρ.		Section 2.1
σ	floating activation function.		Section 2.1
$\oplus, \ominus, \otimes$	floating-point operations.		Section 2.1
$\bigoplus$	addition of multiple floats.		Section 2.1
$\mathbb{1}$	indicator function.		Section 2.1
σ-distinguishable with range $\mathcal{R}$		Definition 3.1	Section 3.1
C_0, C_1, C_2	constants for Condition 1.	Condition 1	Section 3.2
$\eta_1, \eta_2, \ldots, \eta_n$	separating points of σ.	Definition 3.5	Section 3.3
e_1, e_2	constants in Lemma 3.6.		Section 3.3
$\Sigma_\sigma = \Sigma$	set of output after activation.	$\mathbb{F} \cap \{w \otimes \sigma(c) : w, c \in \mathbb{F}\}$.	Section 4.1
sequential addition using σ.		Definition 4.3	Section 4.1
$\overset{\sigma}{\Longmapsto}$	transferrable using σ.	Definition 4.5	Section 4.1
$\lceil x \rceil_{\mathbb{Z}}$	ceiling function	$\min\{m \in \mathbb{Z} : m \geq x\}$	Appendix B
$\lfloor x \rfloor_{\mathbb{Z}}$	floor function	$\max\{m \in \mathbb{Z} : m \leq x\}$	Appendix B

B. Additional notations

The set of floating-point numbers consists of two disjoint subsets,

$$\{0\} \cup \left\{ s \times (1.m_1 \cdots m_p) \times 2^e : s \in \{-1,1\},\ m_1, \ldots, m_p \in \{0,1\},\ e \in [\mathfrak{e}_{\min}, \mathfrak{e}_{\max}]_{\mathbb{Z}} \right\} \tag{5}$$

and

$$\left\{ s \times (0.m_1 \cdots m_p) \times 2^e : s \in \{-1,1\},\ m_1, \ldots, m_p \in \{0,1\},\ e = \mathfrak{e}_{\min} \right\} \setminus \{0\}. \tag{6}$$

An element of subset (5) is called a *normal floating-point number*. And an element of subset (6) is denoted as a *subnormal floating-point number*. For each $x \in \mathbb{F}$, x can be represented by products of three parts, the *sign* s, *significand* $m_0.m_1 \ldots m_p$, and the *exponent* e:

$$x = s \times m_0.m_1 \ldots m_p \times 2^e.$$

To clearly denote the components of a floating-point number $x \in \mathbb{F}$, we denote its entire significand $m_0.m_1 \dots m_p$ of x as $\mathfrak{m}_x$ with each binary digit m_i is denoted as $\mathfrak{m}_{x,i}$. In other words, the significand can be expressed as $\mathfrak{m}_x = \mathfrak{m}_{x,0}.\mathfrak{m}_{x,1} \dots \mathfrak{m}_{x,p}$. Note that for a normal floating-point number x, $\mathfrak{m}_{x,0} = 1$ whereas for a subnormal floating-point number, $\mathfrak{m}_{x,0} = 0$. The sign s of x is denoted as s_x, the exponent e of x is denoted as $\mathfrak{e}_x$. By the definition, we have $\mathfrak{e}_x \in [\mathfrak{e}_{\min}, \mathfrak{e}_{\max}]_{\mathbb{Z}}$ for any $x \in \mathbb{F}$ and $\mathfrak{e}_x = \mathfrak{e}_{\min}$ for a subnormal floating-point number x. In short, x can be represented as

$$x = s_x \times \mathfrak{m}_x \times 2^{\mathfrak{e}_x}.$$

As the significant of a normal floating-point number is always in $[1, 2)$, it is often convenient to express a subnormal floating-point in the same form. For $x \in \mathbb{F}$, $\widetilde{\mathfrak{m}}_x$ and $\widetilde{\mathfrak{e}}_x$ are defined as the unique numbers satisfying $s_x \in \{-1, 1\}$, $\widetilde{\mathfrak{m}}_x \in [1, 2)_{\mathbb{F}}$, $\widetilde{\mathfrak{e}}_x \in \mathbb{Z}$ and

$$x = s_x \times \widetilde{\mathfrak{m}}_x \times 2^{\widetilde{\mathfrak{e}}_x}. \tag{7}$$

Note that if x is normal, we have $\widetilde{\mathfrak{m}}_x = \mathfrak{m}_x$ and $\widetilde{\mathfrak{e}}_x = \mathfrak{e}_x$. If x is subnormal, we have $\mathfrak{e}_{\min} - p \le \widetilde{\mathfrak{e}}_x \le \mathfrak{e}_{\min} - 1$.

Additionally, we define the ceiling function $\lceil x \rceil_{\mathbb{Z}}$ and the floor function $\lfloor x \rfloor_{\mathbb{Z}}$ as follows.

$$\lceil x \rceil_{\mathbb{Z}} \coloneqq \min\{m \in \mathbb{Z} : m \ge x\},$$
$$\lfloor x \rfloor_{\mathbb{Z}} \coloneqq \max\{m \in \mathbb{Z} : m \le x\}.$$

C. Technical lemmas

Lemma C.1. *Assume that σ satisfies Condition 1. Consider $x_1, x_2 \in \mathbb{F}$ such that $\mathfrak{e}_{x_1} \le \mathfrak{e}_{x_2} - 2$. Then,*

$$(x_1, x_2) \overset{\sigma}{\Longmapsto} (0, x_2).$$

This lemma implies that if we have two floating-point numbers whose exponents differ by two or more, performing sequential addition can effectively eliminate the smaller number. The lemma is used to simplify calculations in the proofs of other lemmas.
Proof technique: If there is a discrepancy between the exponents of two floating-point numbers, one can select a floating-point number such that its addition is ignored by the larger number and modifies the smaller number.

Proof of Lemma C.1. Fix x_2 and use mathematical induction on the absolute value of x_1 to prove the statement. If $x_1 = 0$, there is nothing to prove. Assume that for $x' \in \mathbb{F}$ such that $|x'| < |x_1|$, the induction hypothesis is satisfied. By Lemma C.6, there exists γ such that $\gamma \in \left(2^{\mathfrak{e}_{x_2}-p-3}, 2^{\mathfrak{e}_{x_2}-p-2}\right]_{\mathbb{F}} \cap \Sigma$. Then, $x_2 \oplus (\pm\gamma) = x_2$. For $x_1 > 0$, $|x_1 \oplus (-\gamma)| < x_1$, and for $x_1 < 0$, $|x_1 \oplus \gamma| < x_1$. Therefore, there exists x' such that $|x'| < |x_1|$, and

$$(x_1, x_2) \overset{\sigma}{\Longmapsto} (x', x_2) \overset{\sigma}{\Longmapsto} (0, x_2),$$

where the last relation is by the induction hypothesis. This completes the proof. □

The following lemma states that if the exponent of the floating-point number x is not too close to $\mathfrak{e}_{\min}$, then, $(0, x)$ is transferable to $(0, x^-)$ and $(0, x^+)$.

Lemma C.2. *Assume that σ satisfies Condition 1. Consider a floating-point number $0 < x \in \mathbb{F}$ such that $\mathfrak{e}_{\min} + 2 \le \mathfrak{e}_x < \mathfrak{e}_{\max}$. Then,*

$$(0, x) \overset{\sigma}{\Longmapsto} (0, x^+).$$

And if $2^{\mathfrak{e}_{\min}+2} \le x \le 2^{\mathfrak{e}_{\max}}$,

$$(0, x) \overset{\sigma}{\Longmapsto} (0, x^-).$$

Proof of Lemma C.2. By Lemma C.6, there exists γ such that $\gamma \in \left(2^{\mathfrak{e}_x-p-1}, 2^{\mathfrak{e}_x-p}\right]_{\mathbb{F}} \cap \Sigma$. Therefore,

$$(0, x) \overset{\sigma}{\Longmapsto} (\gamma, x \oplus \gamma) = (\gamma, x^+) \overset{\sigma}{\Longmapsto} (0, x^+),$$

where the last relation is by Lemma C.1. Similarly, if $x \neq 2^{\mathfrak{e}_x}$,

$$(0, x) \overset{\sigma}{\Longmapsto} (-\gamma, x \oplus (-\gamma)) = (-\gamma, x^-) \overset{\sigma}{\Longmapsto} (0, x^-).$$

If $x = 2^{\mathfrak{e}_x}$, there exists γ such that $\gamma \in \left(2^{\mathfrak{e}_x - p - 2}, 2^{\mathfrak{e}_x - p - 1}\right]_{\mathbb{F}} \cap \Sigma$, and similar arguments hold. This completes the proof. □

The following lemma states that if the exponent of a floating-point number x is close to $\mathfrak{e}_{\min}$, then, $(0, x)$ is transferable to $(0, x^-)$. Together with Lemma C.2, for any x such that $0 < x \leq 2^{\mathfrak{e}_{\max}}$, $(0, x)$ can be transferable to $(0, x^-)$.

Lemma C.3. *Assume that σ satisfies Condition 1. For $x \in \mathbb{F}$ such that $0 < x < 2^{\mathfrak{e}_{\min}+2}$, the following relation holds:*

$$(0, x) \overset{\sigma}{\Longmapsto} (0, x^-).$$

Proof of Lemma C.3. By Lemma C.6, $\omega, 2\omega \in \Sigma$. If $2^{\mathfrak{e}_{\min}+1} < x < 2^{\mathfrak{e}_{\min}+2}$ and $\mathfrak{m}_{x,p} = 0$, then,

$$(0, x) \overset{\sigma}{\Longmapsto} (\omega, x \oplus \omega) = (\omega, x) \overset{\sigma}{\Longmapsto} (2\omega, x) \overset{\sigma}{\Longmapsto} (0, x \oplus (-2\omega)) = \left(0, x^-\right).$$

If $\mathfrak{e}_x = \mathfrak{e}_{\min} + 1$ and $\mathfrak{m}_{x,p} = 1$, then, $\mathfrak{e}_{x^-} = \mathfrak{e}_{\min} + 1$ or $\mathfrak{e}_{x^+} = \mathfrak{e}_{\min} + 1$. If $\mathfrak{e}_{x^-} = \mathfrak{e}_{\min} + 1$, then, $\mathfrak{m}_{x,p} = 1$ and thus,

$$(0, x) \overset{\sigma}{\Longmapsto} (-2\omega, x \oplus (-2\omega)) = (-2\omega, x^-) \overset{\sigma}{\Longmapsto} \left(-\omega, x^- \oplus \omega\right) = \left(-\omega, x^-\right) \overset{\sigma}{\Longmapsto} \left(0, x^-\right).$$

Symmetric arguments hold for $\mathfrak{e}_{x^+} = \mathfrak{e}_{\min} + 1$ case.

If $x \leq 2^{\mathfrak{e}_{\min}+1}$, then $x = N\omega$ where $N \in \left[2^{p+1}\right]$.

$$\begin{aligned}
(0, N\omega) &\overset{\sigma}{\Longmapsto} (\omega, (N+1)\omega) \overset{\sigma}{\Longmapsto} \cdots \overset{\sigma}{\Longmapsto} \left(\left(2^{p+1} - N\right)\omega, 2^{p+1}\omega\right) \\
&\overset{\sigma}{\Longmapsto} \left(\left(2^{p+1} - N\right)\omega \oplus \omega, 2^{p+1}\omega \oplus \omega\right) = \left(\left(2^{p+1} - N + 1\right)\omega, 2^{p+1}\omega\right) \\
&\overset{\sigma}{\Longmapsto} \left(\left(2^{p+1} - N + 1\right)\omega \oplus (-\omega), 2^{p+1} \oplus (-\omega)\right) = \left(\left(2^{p+1} - N\right)\omega, \left(2^{p+1} - 1\right)\omega\right) \\
&\overset{\sigma}{\Longmapsto} \left(\left(2^{p+1} - N\right)\omega \oplus (-\omega), \left(2^{p+1} - 1\right)\omega \oplus (-\omega)\right) = \left(\left(2^{p+1} - N - 1\right)\omega, \left(2^{p+1} - 2\right)\omega\right) \\
&\overset{\sigma}{\Longmapsto} \cdots \overset{\sigma}{\Longmapsto} (0, (N-1)\omega) = \left(0, x^-\right).
\end{aligned}$$

□

The following lemma is the corresponding version of Lemma C.3 for x^+; that is, if the exponent of a positive floating-point number x is not too big, then, $(0, x)$ is transferable to $(0, y)$ for some larger floating-point numbers y. Note that, unlike the previous lemma, it only claims the existence of larger y, not x^+. However, together with the combination of Lemmas C.2 and C.3, one can easily check that $(0, x)$ is transferable to $(0, x^+)$.

Lemma C.4. *Assume that σ satisfies Condition 1. Consider $0 \neq x \in \mathbb{F}$ such that $\mathfrak{e}_x \leq \mathfrak{e}_{\max} - 2p - 2$. Then, there exists $y \in \mathbb{F}$ such that $xy > 0$, $2^{\mathfrak{e}_{\max}} \geq |y| > |x|$, and*

$$(0, x) \overset{\sigma}{\Longmapsto} (0, y).$$

Proof of Lemma C.4. As the floating-point number is symmetric with respect to zero, we only need to consider the case of $x > 0$. By Condition 1, there exists $K \in \Sigma$ such that $\mathfrak{e}_K = \widetilde{\mathfrak{e}}_x + p + 1$. If $\mathfrak{m}_{K,p} = 1$ or $x > 2^{\widetilde{\mathfrak{e}}_x}$,

$$\begin{aligned}
(0, x) &\overset{\sigma}{\Longmapsto} \left(\mathfrak{m}_K \times 2^{\widetilde{\mathfrak{e}}_x + p + 1}, \mathfrak{m}_K \times 2^{\widetilde{\mathfrak{e}}_x + p + 1} \oplus x\right) = \left(\mathfrak{m}_K \times 2^{\widetilde{\mathfrak{e}}_x + p + 1}, \left(\mathfrak{m}_K \times 2^{\widetilde{\mathfrak{e}}_x + p + 1}\right)^+\right) \\
&\overset{\sigma}{\Longmapsto} \left(0, \left(\mathfrak{m}_K \times 2^{\widetilde{\mathfrak{e}}_x + p + 1}\right)^+ \ominus \mathfrak{m}_K \times 2^{\widetilde{\mathfrak{e}}_x + p + 1}\right) = \left(0, 2^{\widetilde{\mathfrak{e}}_x + 1}\right) \text{ or } \left(0, 2^{\widetilde{\mathfrak{e}}_x + 2}\right).
\end{aligned}$$

If $\mathfrak{m}_{K,p} = 0$, $x = 2^{\widetilde{\mathfrak{e}}_x}$ and x is normal, by Lemma C.6, there exists $\gamma \in \Sigma$ such that $2^{\mathfrak{e}_x - p - 1} < \gamma \leq 2^{\mathfrak{e}_x - p}$. Then,

$$(0, x) \overset{\sigma}{\Longmapsto} (\gamma, x \oplus \gamma) = \left(\gamma, x^+\right) \overset{\sigma}{\Longmapsto} \left(\mathfrak{m}_K \times 2^{\widetilde{\mathfrak{e}}_x + p + 1} \oplus \gamma, \mathfrak{m}_K \times 2^{\widetilde{\mathfrak{e}}_x + p + 1} \oplus x^+\right)$$

$$= \left(\mathfrak{m}_K \times 2^{\widetilde{\mathfrak{e}}_x+p+1}, \left(\mathfrak{m}_K \times 2^{\widetilde{\mathfrak{e}}_x+p+1}\right)^+\right) \overset{\sigma}{\Longmapsto} \left(0, \left(\mathfrak{m}_K \times 2^{\widetilde{\mathfrak{e}}_x+p+1}\right)^+ \ominus \mathfrak{m}_K \times 2^{\widetilde{\mathfrak{e}}_x+p+1}\right)$$
$$= \left(0, 2^{\widetilde{\mathfrak{e}}_x+1}\right) \text{ or } \left(0, 2^{\widetilde{\mathfrak{e}}_x+2}\right).$$

If $\mathfrak{m}_{K,p} = 0$, $x = 2^{\widetilde{\mathfrak{e}}_x}$ and x is subnormal, by Lemma C.6, $\omega \in \Sigma$. Therefore,

$$(0, x) \overset{\sigma}{\Longmapsto} (\omega, \omega \oplus x) = (\omega, x^+) \overset{\sigma}{\Longmapsto} \left(\mathfrak{m}_K \times 2^{\widetilde{\mathfrak{e}}_x+p+1}, \left(\mathfrak{m}_K \times 2^{\widetilde{\mathfrak{e}}_x+p+1}\right)^+\right) \overset{\sigma}{\Longmapsto} \left(0, 2^{\widetilde{\mathfrak{e}}_x+1}\right) \text{ or } \left(0, 2^{\widetilde{\mathfrak{e}}_x+2}\right).$$

This completes the proof. □

Lemma C.5. *Consider a floating-point $\eta \in \mathbb{F}$. Then, for any $x_1, x_2 \in \mathbb{F}$ such that $x_1 \neq x_2$, $|x_1| \leq |x_2|$, and*

$$\mathfrak{e}_{\min} + 1 \leq \mathfrak{e}_\eta - \mathfrak{e}_{x_2} \leq \mathfrak{e}_{\max},$$

there exist $w, b \in \mathbb{F}$ such that

$$\{w \otimes x_1 \oplus b, w \otimes x_2 \oplus b\} = \{\eta, \eta^+\} \text{ or } \{\eta^-, \eta^+\}.$$

Furthermore,

$$|w| \leq \left(1 + 2^{-p}\right) \times 2^{\mathfrak{e}_\eta - \mathfrak{e}_{x_2}},$$

and

$$|b| \leq \eta^+.$$

Proof of Lemma C.5.
Case 1: $0 \leq x_1 < x_2$
Consider the case

$$0 \leq x_1 < 2^{\widetilde{\mathfrak{e}}_{x_2}} < x_2,$$

define w as $w \coloneqq 2^{\mathfrak{e}_\eta - \widetilde{\mathfrak{e}}_{x_2} - p - 1}$ and b as $b \coloneqq \eta$. Then,

$$w \otimes x_2 \oplus \eta > 2^{\mathfrak{e}_\eta - p - 1} \oplus \eta = \eta^+,$$

and as $w \otimes x_1 \leq 2^{\mathfrak{e}_\eta - p - 1}$,

$$\eta \leq w \otimes x_1 \oplus \eta \leq \eta.$$

Consider the case

$$0 \leq x_1 < 2^{\widetilde{\mathfrak{e}}_{x_2}} = x_2,$$

If $\mathfrak{m}_{\eta,p} = 1$, define w as $w \coloneqq 2^{\mathfrak{e}_\eta - \widetilde{\mathfrak{e}}_{x_2} - p - 1}$ and b as $b \coloneqq \eta$. Then, similar to $x_2 > 2^{\mathfrak{e}_{x_2}}$ case, $w \otimes x_2 \oplus b = \eta$ and $w \otimes x_1 \oplus b = \eta$. If $\mathfrak{m}_{\eta,p} = 0$, define w as $w \coloneqq -2^{\mathfrak{e}_\eta - \widetilde{\mathfrak{e}}_{x_2} - p - 1}$ and b as $b \coloneqq \eta^+$. Then, $w \otimes x_2 \oplus b = \eta$ and $w \otimes x_1 \oplus b = \eta^+$.

Consider the case x_1, x_2 are subnormal or

$$2^{\mathfrak{e}_{x_2}} \leq x_1 < x_2.$$

Then, there exists $i \in \mathbb{N}_0$ such that

$$\mathfrak{m}_{x_1,j} = \mathfrak{m}_{x_2,j} \text{ for } j \in [i],$$

and

$$0 = \mathfrak{m}_{x_1,i+1} \neq \mathfrak{m}_{x_2,i+1} = 1.$$

If $\mathfrak{m}_{\eta,p} = 1$ define $w \in \mathbb{F}$ as

$$w \coloneqq 2^{\mathfrak{e}_\eta + i - \mathfrak{e}_{x_2} - p}.$$

Define b as

$$b \coloneqq \eta - 1.\mathfrak{m}_{x_2,1} \ldots \mathfrak{m}_{x_2,i} \underbrace{0 \ldots 0}_{p-i} \times 2^{\mathfrak{e}_\eta - p + i}.$$

As

$$1.\mathfrak{m}_{x_2,1}\dots\mathfrak{m}_{x_2,i}\underbrace{0\dots0}_{p-i}\times w = 1\mathfrak{m}_{x_2,1}\dots\mathfrak{m}_{x_2,i}\times 2^{\mathfrak{e}_\eta - p},$$

the operation is exact: $\eta - 1.\mathfrak{m}_{x_2,1}\dots\mathfrak{m}_{x_2,i}\underbrace{0\dots0}_{p-i}\times 2^{\mathfrak{e}_\eta-p+i}\in\mathbb{F}$. Then,

$$w\otimes x_1\oplus b = 1.\mathfrak{m}_{x_2,1}\dots\mathfrak{m}_{x_2,i}0\mathfrak{m}_{x_1,i+2}\dots\mathfrak{m}_{x_1,p}\times 2^{\mathfrak{e}_\eta-p+i}\oplus b = \left\lceil \eta + 0.\mathfrak{m}_{x_1,i+2}\dots\mathfrak{m}_{x_1,p}\times 2^{\mathfrak{e}_\eta-p-1}\right\rfloor = \eta,$$

and

$$w\otimes x_2\oplus b = 1.\mathfrak{m}_{x_2,1}\dots\mathfrak{m}_{x_2,i}1\mathfrak{m}_{x_2,i+2}\dots\mathfrak{m}_{x_2,p}\times 2^{\mathfrak{e}_\eta-p+i}\oplus b = \left\lceil \eta + 1.\mathfrak{m}_{x_2,i+2}\dots\mathfrak{m}_{x_2,p}\times 2^{\mathfrak{e}_\eta-p-1}\right\rfloor = \eta^+.$$

If $\mathfrak{m}_{\eta,p}=0$, define b as

$$b \coloneqq \eta^+ - 1.\mathfrak{m}_{x_2,1}\dots\mathfrak{m}_{x_2,i}\underbrace{0\dots0}_{p-i}\times 2^{\mathfrak{e}_\eta-p+i}.$$

and w as $\coloneqq -2^{\mathfrak{e}_\eta+i-\mathfrak{e}_{x_2}-p}$. Then, similarly,

$$w\otimes x_1\oplus b = 1.\mathfrak{m}_{x_2,1}\dots\mathfrak{m}_{x_2,i}0\mathfrak{m}_{x_1,i+2}\dots\mathfrak{m}_{x_1,p}\times 2^{\mathfrak{e}_\eta-p+i}\oplus b = \left\lceil \eta^+ - 0.\mathfrak{m}_{x_1,i+2}\dots\mathfrak{m}_{x_1,p}\times 2^{\mathfrak{e}_\eta-p-1}\right\rfloor = \eta^+,$$

and

$$w\otimes x_2\oplus b = 1.\mathfrak{m}_{x_2,1}\dots\mathfrak{m}_{x_2,i}1\mathfrak{m}_{x_2,i+2}\dots\mathfrak{m}_{x_2,p}\times 2^{\mathfrak{e}_\eta-p+i}\oplus +b = \left\lceil \eta^+ - 1.\mathfrak{m}_{x_2,i+2}\dots\mathfrak{m}_{x_2,p}\times 2^{\mathfrak{e}_\eta-p-1}\right\rfloor = \eta.$$

Case 2: $x_1<0<x_2$ Define b as $b\coloneqq\eta$. If $\widetilde{\mathfrak{m}}_{x_2}<1.1$, define w as

$$w \coloneqq -2^{\mathfrak{e}_\eta-\widetilde{\mathfrak{e}}_{x_2}-p}.$$

Then,

$$w\otimes x_2 = \widetilde{\mathfrak{m}}_{x_2}\times 2^{\mathfrak{e}_\eta-p},$$

and

$$w\otimes x_2\oplus b = \eta^-.$$

And as

$$|w\otimes x_2|\le|w\otimes x_1|,$$
$$\eta\le w\otimes x_2\oplus b\le\eta^+.$$

If $\widetilde{\mathfrak{m}}_{x_2}\ge 1.1$, define w as

$$w\coloneqq -2^{\mathfrak{e}_\eta-\widetilde{\mathfrak{e}}_{x_2}-p-1},$$

and we get the same conclusion. This completes the proof. □

Lemma C.6. *Suppose that $\sigma:\overline{\mathbb{F}}\to\overline{\mathbb{F}}$ satisfies Condition 1 with constants C_1 and C_2 and let $e\in[\mathfrak{e}_{\min}-p,\mathfrak{e}_{\max}-p]_{\mathbb{Z}}$. Then, there exists $\gamma\in\mathbb{F}$ and $i\in[2]$ such that*

$$2^{e-1}<\gamma\otimes\sigma(C_i)\le 2^e.$$

Proof of Lemma C.6. Let C_1 and C_2 from Condition 1 be represented as

$$\sigma(C_1)=\mathfrak{m}_{\sigma(C_1)}\times 2^{\mathfrak{e}_{\sigma(C_1)}}\quad\text{ and }\sigma(C_2)=\mathfrak{m}_{\sigma(C_2)}\times 2^{\mathfrak{e}_{\sigma(C_2)}}.$$

Consider the case $\mathfrak{e}_{\min}-p-2\le e\le 2$. If $\mathfrak{m}_{\sigma(C_1)}\in\left[1,\frac{5}{4}\right]_{\mathbb{F}}$, define γ as

$$\gamma\coloneqq 1.1\times 2^{e-\mathfrak{e}_{\sigma(C_1)}-1}.$$

As $\mathfrak{e}_{\sigma(C_1)} \geq \mathfrak{e}_{\min} = -2^{q-1}+2$, $e - \mathfrak{e}_{\sigma(C_1)} - 1 \leq 2^{q-1} - 1 = \mathfrak{e}_{\max}$, and $\gamma \in \mathbb{F}$. Then,

$$\frac{3}{4} \times 2^e \leq \gamma \times \sigma(C_1) = \mathfrak{m}_{\sigma(C_1)} \times 0.11 \times 2^e \leq \frac{15}{16} \times 2^e,$$

and thus

$$\frac{3}{4} \times 2^e \leq \gamma \otimes \sigma(C_1) = \left\lceil \mathfrak{m}_{\sigma(C_1)} \times 1.1 \times 2^e \right\rfloor_{\mathbb{F}} \leq 2^e,$$

where the first inequality is satisfied as $e \geq \mathfrak{e}_{\min} - p + 2$. If $\mathfrak{m}_{\sigma(C_1)} \in \left(\frac{5}{4}, 2\right)_{\mathbb{F}}$, define γ as

$$\gamma \coloneqq 2^{e - \mathfrak{e}_{\sigma(C_1)} - 1}.$$

Similar to the above case, $\gamma \in \mathbb{F}$. Then,

$$\frac{1}{2} \times 2^e < \gamma \otimes \sigma(C_1) = \left\lceil \mathfrak{m}_{\sigma(C_1)} \times 2^{e-1} \right\rfloor_{\mathbb{F}} \leq 2^e,$$

where the first inequality is satisfied as $e - 1 \geq \mathfrak{e}_{\min} - p + 1$.

Consider the case $2 < e < \mathfrak{e}_{\max} - p$. If $\mathfrak{m}_{\sigma(C_2)} \in \left[1, \frac{5}{4}\right]_{\mathbb{F}}$, define γ as

$$\gamma \coloneqq 1.1 \times 2^{e - \mathfrak{e}_{\sigma(C_2)} - 1}.$$

As $e - \mathfrak{e}_{\sigma(C_2)} - 1 \leq \mathfrak{e}_{\max} - p - 1 - \mathfrak{e}_{\sigma(C_2)} - 1 \leq \mathfrak{e}_{\max}$, $\gamma \in \mathbb{F}$. Then, similar to the above case,

$$\frac{3}{4} \times 2^e \leq \lceil \gamma \times \sigma(C_2) \rfloor \leq 2^e.$$

If $\mathfrak{m}_{\sigma(C_2)} \in \left(\frac{5}{4}, 2\right)_{\mathbb{F}}$, define γ as

$$\gamma \coloneqq 2^{e - \mathfrak{e}_{\sigma(C_2)} - 1},$$

and similar arguments holds.

Consider the case $e = \mathfrak{e}_{\max} - p$. If $\mathfrak{m}_{\sigma(C_2)} \geq 1^{++}$, define γ as

$$\gamma \coloneqq \left(2 - 2^{-p}\right) \times 2^{\mathfrak{e}_{\max} - p - 2 - \mathfrak{e}_{\sigma(C_2)}}.$$

Then, $\gamma \in \mathbb{F}$, and we have

$$2^{\mathfrak{e}_{\max} - p - 1} < \gamma \otimes \sigma(C_2) = \left(\mathfrak{m}_{\sigma(C_2)} \otimes \left(2 - 2^{-p}\right)\right) \times 2^{\mathfrak{e}_{\max} - p - 2} \leq 2^{\mathfrak{e}_{\max} - p}.$$

If $\mathfrak{m}_{\sigma(C_2)} \leq 1^{+}$, then, $\mathfrak{e}_{\sigma(C_2)} \geq -p - 1$. If $\mathfrak{m}_{\sigma(C_2)} \in \left[1, \frac{5}{4}\right]_{\mathbb{F}}$, define γ as

$$\gamma \coloneqq 1.1 \times 2^{\mathfrak{e}_{\max} - p - 1 - \mathfrak{e}_{\sigma(C_2)}},$$

and if $\mathfrak{m}_{\sigma(C_2)} \in \left(\frac{5}{4}, 2\right)_{\mathbb{F}}$, define γ as

$$\gamma \coloneqq 2^{\mathfrak{e}_{\max} - p - 1 - \mathfrak{e}_{\sigma(C_2)}}.$$

Then, $\gamma \in \mathbb{F}$, and similar arguments hold as in $2 < e < \mathfrak{e}_{\max} - p$ case.

Consider the case $e = \mathfrak{e}_{\min} - p + 1$. If $1 \leq \mathfrak{m}_{\sigma(C_1)} < \frac{5}{4}$, define γ as

$$\gamma \coloneqq 2^{e - \mathfrak{e}_{\sigma(C_1)}}.$$

Then,

$$\gamma \otimes \sigma(C_1) = \left\lceil \mathfrak{m}_{\sigma(C_1)} \times 2^e \right\rfloor_{\mathbb{F}} = \left\lceil \mathfrak{m}_{\sigma(C_1)} \times 2\omega \right\rfloor_{\mathbb{F}} = 2\omega = 2^e.$$

If $\frac{5}{4} \leq \mathfrak{m}_{\sigma(C_1)} < \frac{5}{3}$, define γ as

$$\gamma \coloneqq 1.1 \times 2^{e - \mathfrak{e}_{\sigma(C_1)} - 1}.$$

As $\mathfrak{e}_{\sigma(C_1)} \leq -1$, $\gamma \in \mathbb{F}$. Then,

$$\gamma \otimes \sigma(C_1) = \left\lceil \mathfrak{m}_{\sigma(C_1)} \times 1.1 \times 2^{e-1} \right\rfloor_{\mathbb{F}} = \left\lceil \mathfrak{m}_{\sigma(C_1)} \times 1.1 \times \omega \right\rfloor_{\mathbb{F}} = 2\omega = 2^e.$$

If $\frac{5}{3} \leq \mathfrak{m}_{\sigma(C_1)} < 2$, define γ as

$$\gamma \coloneqq 2^{e-\mathfrak{e}_{\sigma(C_1)}-1}.$$

Then,

$$\gamma \otimes \sigma(C_1) = \left\lceil \mathfrak{m}_{\sigma(C_1)} \times 2^{e-1} \right\rfloor_{\mathbb{F}} = \left\lceil \mathfrak{m}_{\sigma(C_1)} \times \omega \right\rfloor_{\mathbb{F}} = 2\omega = 2^e.$$

Consider the case $e = \mathfrak{e}_{\min} - p$. If $1 \leq \mathfrak{m}_{\sigma(C_1)} < 1.1$, define γ as

$$\gamma \coloneqq 2^{e-\mathfrak{e}_{\sigma(C_1)}}.$$

Then,

$$\gamma \otimes \sigma(C_1) = \left\lceil \mathfrak{m}_{\sigma(C_1)} \times 2^{e} \right\rfloor_{\mathbb{F}} = \left\lceil \mathfrak{m}_{\sigma(C_1)} \times \omega \right\rfloor_{\mathbb{F}} = \omega = 2^e.$$

If $1.1 \leq \mathfrak{m}_{\sigma(C_1)} < 2$, define γ as

$$\gamma \coloneqq 2^{e-\mathfrak{e}_{\sigma(C_1)}-1}.$$

As $\mathfrak{e}_{\sigma(C_1)} \leq -1$, $\gamma \in \mathbb{F}$. Then,

$$\gamma \otimes \sigma(C_1) = \left\lceil \mathfrak{m}_{\sigma(C_1)} \times 2^{e} - 1 \right\rfloor_{\mathbb{F}} = \left\lceil \mathfrak{m}_{\sigma(C_1)} \times \frac{1}{2}\omega \right\rfloor_{\mathbb{F}} = \omega = 2^e.$$

This completes the proof. □

D. Proof of lemmas

D.1. Proof of Lemma 3.2

Proof. Suppose that $\mathcal{X}$ is not σ-distinguishable with range $\mathbb{F}$. Then, there exist $x_1, x_2 \in \mathcal{X}$ such that for any $d_2 \in \mathbb{N}$ and affine transformations $\phi_1, \phi_2 : \overline{\mathbb{F}}^d \to \overline{\mathbb{F}}^{d_2}$,

$$\sigma(\phi_1(x_1)) = \sigma(\phi_1(x_2)).$$

This implies that for any σ network $g : \mathcal{X} \to \overline{\mathbb{F}}$, $g(x_1) = g(x_2)$. In other words, σ networks cannot represent a function $f : \mathcal{X} \to \mathbb{F}$ with $f(x_1) \neq f(x_2)$. This completes the proof. □

D.2. Proof of Lemma 3.3

Proof. By Lemma 3.2, it is sufficient to prove that $\lceil \cos \rfloor$ can not distinguish the domain with range $\mathbb{F} \cup \{-\infty, \infty\}$. Let $\sigma(x) \coloneqq \lceil \cos(x) \rfloor$. As $\sigma(0) = 1$ and $\cos(x) \geq 1 - \frac{x^2}{2}$, for $x \in \mathbb{F}$ such that $0 < |x| \leq 2^{\frac{-p-1}{2}}$,

$$\cos(x) > 1 - \frac{x^2}{2} \geq 1 - 2^{-p-2}$$

Therefore, for $|x| \leq 2^{\frac{-p-1}{2}}$,

$$\sigma(x) = \lceil \cos(x) \rfloor = 1.$$

Consider $w, b \in \mathbb{F}$ such that $\sigma(w \otimes 0 \oplus b) \neq \sigma(w \otimes \omega \oplus b)$. Then, w should satisfy

$$|w \otimes \omega| > 2^{-\left\lfloor \frac{p+1}{2} \right\rfloor_{\mathbb{Z}} - p - 1}.$$

Thus,

$$|w| > 2^{-\mathfrak{e}_{\min} - 1 - \left\lfloor \frac{p+1}{2} \right\rfloor_{\mathbb{Z}}}.$$

And one of $|w \otimes x \oplus b|$, $|w \otimes (-x) \oplus b|$ is greater than or equal to $w \otimes x$. Therefore, for $|x| \geq 2^{3 + \left\lfloor \frac{p+1}{2} \right\rfloor_{\mathbb{Z}}}$,

$$w \otimes x \oplus b = \infty \text{ or } -\infty,$$

which means that

$$\sigma(w \otimes x \oplus b) = \text{NaN}.$$

This completes the proof. □

D.3. Proof of Lemma 3.6

Proof. Consider arbitrary $x_1, x_2 \in \left(-2^{e_2+1}, 2^{e_2+1}\right)_{\mathbb{F}}$ such that $x_1 \neq x_2$ and $|x_1| \leq |x_2|$. By the condition of the lemma, there exists $i \in [n]$ such that

$$\mathfrak{e}_{x_2} \in [\mathfrak{e}_{\eta_i} - e_1, \mathfrak{e}_{\eta_i} + \mathfrak{e}_{\max} - 2]_{\mathbb{Z}}.$$

By Lemma C.5, there exist w and b such that

$$\sigma(w \otimes x_1 \oplus b) \neq \sigma(w \otimes x_2 \oplus b),$$
$$|w| \leq \left(1 + 2^{-p}\right) \times 2^{e_1}, \text{ and } |b| \leq |\eta_i|^+.$$

Then,

$$\sigma\left(\left(-2^{e_2+1}, 2^{e_2+1}\right)_{\mathbb{F}} \otimes w \oplus b\right) \subset \sigma\left(\left[-2^{e_2+e_1+1}, 2^{e_2+e_1+1}\right]_{\mathbb{F}} \oplus b\right) \subset \sigma\left(\left[-\left(2^{e_2+e_1+1} \oplus |\eta_n|^+\right), 2^{e_2+e_1+1} \oplus |\eta_n|^+\right]_{\mathbb{F}}\right).$$

This completes the proof. □

D.4. Proof of Lemma 3.7

Proof. This is a direct consequence of Lemma 3.6. By defining $e_1 = e_2 = \mathfrak{e}_{\max}$, as $\mathfrak{e}_{\eta_1} \leq 1$ and $\mathfrak{e}_{\eta_2} \geq 2$, we have

$$[\mathfrak{e}_{\min}, \mathfrak{e}_{\max}]_{\mathbb{Z}} \subset \bigcup_{i=1}^{2} [\mathfrak{e}_{\eta_i} - \mathfrak{e}_{\max}, \mathfrak{e}_{\eta_i} + \mathfrak{e}_{\max} - 2]_{\mathbb{Z}}.$$

Thus, $\mathbb{F} = \left(-2^{\mathfrak{e}_{\max}+1}, 2^{\mathfrak{e}_{\max}+1}\right)_{\mathbb{F}}$ is σ-distinguishable with range

$$\sigma\left(\left[-\left(2^{e_1+e_2+1} \oplus |\eta_2|^+\right), 2^{e_1+e_2+1} \oplus |\eta_2|^+\right]_{\mathbb{F}}\right) \subset \sigma\left(\mathbb{F} \cup \{\infty, -\infty\}\right) \subset [-2^{\mathfrak{e}_{\max}}, 2^{\mathfrak{e}_{\max}}]_{\mathbb{F}}.$$

This completes the proof. □

D.5. Proof of Lemma 3.9

Proof. Without loss of generality, consider the case: $\rho'(x) \geq L$ for all $x \in [a, b]$ and $\lceil \rho \rfloor (a) \geq 0$. As

$$\rho(b) - \rho(a) \geq L(b - a) \geq 2^{e-p},$$

$\sigma(b) \neq \sigma(a)$. Therefore, there exists $\eta \in [a, b]_{\mathbb{F}}$ such that $\sigma(\eta^-) < \sigma(\eta) \leq \sigma(\eta^+)$. This completes the proof. □

D.6. Proof of Lemma 3.10

Proof. Define σ as $\sigma \coloneqq \lceil \hat{\rho} \rfloor$. We will prove that for any $e_1 \in [\mathfrak{e}_{\min}, e-1]_{\mathbb{Z}}$, there exists a distinguishing point η such that $\mathfrak{e}_\eta = e$. For any floating-point number $x \in \mathbb{F}$, if $\sigma(x)$ is normal, as

$$|\sigma(x)| = |\lceil \hat{\rho}(x) \rfloor| \leq L_2 x \times \left(1 + 2^{-p}\right),$$

$\mathfrak{e}_{\sigma(x)} \leq \mathfrak{e}_x + \lceil \log_2\left(L_2 \times (1 + 2^{-p})\right) \rceil_{\mathbb{Z}}$. Note that for any $x \geq 2^{\mathfrak{e}_{\min}}$,

$$\sigma(x) \geq \sigma\left(2^{\mathfrak{e}_{\min}}\right) \geq 2^{l_1} 2^{\mathfrak{e}_{\min}} \geq \omega,$$

Consider Lemma 3.9 with $a = 2^{e_1}$, $b = \left(2^{e_1+1}\right)^-$, $L = L_1$, and $e = \mathfrak{e}_{\sigma(x)}$. As

$$L_1\left(\left(2^{e_1+1}\right)^- - 2^{e_1}\right) = L_1\left(1 - 2^{-p}\right) 2^{e_1} \geq L_1\left(1 - 2^{-p}\right) 2^{\mathfrak{e}_{\sigma(x)} - \lceil \log_2\left(L_2 \times \left(1 + 2^{-p}\right)\right) \rceil_{\mathbb{Z}}} \geq 2^{\mathfrak{e}_{\sigma(x)} - p},$$

there exists a distinguishing point $\eta \in [2^{e_1}, \left(2^{e_1+1}\right)^-]$ such that $\mathfrak{e}_\eta = e_1$. Then, by Lemma 3.6 with distinguishing points with exponents $\mathfrak{e}_{\min}, \mathfrak{e}_{\min}+1, \ldots, e-1$, $e_1 = 0$, and $e_2 = \min(\mathfrak{e}_{\max} - l_2 - 1, \mathfrak{e}_{\max} + e - 3)$, the interval $\left(-2^{e_2+1}, 2^{e_2+1}\right)_{\mathbb{F}}$ is σ-distinguishable with range

$$\begin{aligned}\sigma\left(\left[-\left(2^{e_1+e_2+1} \oplus |\eta_n|^+\right), 2^{e_1+e_2+1} \oplus |\eta_n|^+\right]\right) &\subset L_2\left(1 + 2^{-p}\right)\left[-\left(2^{e_1+e_2+1} \oplus |\eta_n|^+\right), 2^{e_1+e_2+1} \oplus |\eta_n|^+\right] \\ &\subset 2^{l_2}\left[-\left(2^{\mathfrak{e}_{\max}-l_2}\right), 2^{\mathfrak{e}_{\max}-l_2}\right] \subset \left[-2^{\mathfrak{e}_{\max}}, 2^{\mathfrak{e}_{\max}}\right].\end{aligned}$$

This completes the proof. □

D.7. Proof of Corollary 3.11

Proof. Let $\hat{\rho}(x) = \rho(x)$. The result is straightforward by Table 3. □

$\rho(x)$	e	L_1	L_2	l_1	l_2	e'
ReLU	$\mathfrak{e}_{\max}$	1	1	-1	1	$\mathfrak{e}_{\max} - 1$
ELU	$\mathfrak{e}_{\max}$	1	1	-1	1	$\mathfrak{e}_{\max} - 1$
GELU	$\mathfrak{e}_{\max}$	0.5	1	-2	1	$\mathfrak{e}_{\max} - 1$
SeLU	$\mathfrak{e}_{\max}$	1.05	2	-1	2	$\mathfrak{e}_{\max} - 1$
Swish	$\mathfrak{e}_{\max}$	0.5	1	-2	1	$\mathfrak{e}_{\max} - 1$
Mish	$\mathfrak{e}_{\max}$	0.6	1	-2	1	$\mathfrak{e}_{\min} - 1$
sin	0	0.540	1	-2	1	$\mathfrak{e}_{\min} - 2$

Table 3: Properties of floating-point format for verifying the conditions. Numbers in the table are rounded to the second decimal place.

D.8. Proof of Lemma 4.2

Proof. By Lemma C.6, for any $e \in [\mathfrak{e}_{\min} - p, \mathfrak{e}_{\max} - p]$, there exists γ and $i \in [2]$ such that

$$2^{e-1} < \gamma \otimes \sigma(C_i) \leq 2^e.$$

Therefore, for any $x \in \mathbb{F}$, there exists γ and $i \in [2]$ such that $x \oplus \gamma \otimes \sigma(C_i) = x^+$. This completes the proof. □

D.9. Proof of Lemma 4.4

Proof. Represent f as

$$f(x) = x \oplus \bigoplus_{i=1}^{n} y_i.$$

where $n \in \mathbb{N}$ and $y_i \in \Sigma$ for $i \in [n]$. We first note that for any y_i, if $|y_i| < 2^{\mathfrak{e}_{\max}-p-1}$, then, $\Omega \oplus y_i = \Omega$ and $-\Omega \oplus y_i = -\Omega$. And if $|y_i| \geq 2^{\mathfrak{e}_{\max}-p-1}$, then $\Omega \oplus y_i = \infty$ or $-\Omega \oplus y_i = -\infty$. As f is a function from $\mathbb{F}$ to $\mathbb{F}$, $|y_i| < 2^{\mathfrak{e}_{\max}-p-1}$ for any $i \in [n]$, and thus, $\Omega \oplus \bigoplus_{i=1}^{j} y_i = \Omega$ and $-\Omega \oplus \bigoplus_{i=1}^{j} y_i = -\Omega$ for any $j \in [n]$.

Now, we will use proof by contradiction. Without loss of generality, assume that $f(2^{\mathfrak{e}_{\max}}) > 2^{\mathfrak{e}_{\max}}$. Then, there exists $i_0 \in [n]$ such that $2^{\mathfrak{e}_{\max}} \oplus \bigoplus_{i=1}^{i_0-1} y_i \leq 2^{\mathfrak{e}_{\max}}$ and $2^{\mathfrak{e}_{\max}} \oplus \bigoplus_{i=1}^{i_0} y_i > 2^{\mathfrak{e}_{\max}}$. Then, $y_{i_0} > 2^{\mathfrak{e}_{\max}-p-1}$ and it gives the contradiction. This completes the proof. □

D.10. Proof of Lemma 4.7

Proof. Without loss of generality, assume that $x_1 < x_2$. Use mathematical induction on $|x_1|$. If $x_1 = 0$, there is nothing to prove. Assume that if $|x'| < x_1$ for any $x_2' \neq x'$, then there exists y' such that $(x', x_2') \overset{\sigma}{\Longmapsto} (0, y')$.

If $0 < x_1 < x_2$, by Lemma C.6, there exists $\gamma \in \Sigma$ such that $\gamma < x_1$ and $x_1 \oplus (-\gamma) < x_1$ and $x_2 \oplus (-\gamma) \neq x_1 \oplus (-\gamma)$. Therefore,

$$(x_1, x_2) \overset{\sigma}{\Longmapsto} (x_1 \oplus (-\gamma), x_2 \oplus (-\gamma)) \overset{\sigma}{\Longmapsto} (0, y'),$$

by the induction hypothesis.

If $x_1 < 0 < x_2$, by Lemma C.6, there exists $\gamma \in \Sigma$ such that $x_1 \oplus \gamma > x_1$. Therefore,

$$(x_1, x_2) \overset{\sigma}{\Longmapsto} (x_1 \oplus \gamma, x_2 \oplus \gamma) \overset{\sigma}{\Longmapsto} (0, y'),$$

by the induction hypothesis. This completes the proof. □

D.11. Proof of Lemma 4.8

We only prove the case $x_1, x_2 > 0$ and the remaining case is by symmetry. We first prove that for any $0 < x \in \mathbb{F}$ such that $x \in (0, 2^{\mathfrak{e}_{\max}})_{\mathbb{F}}$,

$$(0, x) \overset{\sigma}{\Longmapsto} (0, x^+).$$

If $\mathfrak{e}_x \geq \mathfrak{e}_{\min} + 2$, it is by Lemma C.2. If $\mathfrak{e}_x \leq \mathfrak{e}_{\min} + 1$, then by Lemma C.4, there exists $y \in \mathbb{F}$ such that $(0, x) \overset{\sigma}{\Longmapsto} (0, y)$. Use mathematical induction on decreasing order to prove that for any $x < x' \leq y$, $(0, x) \overset{\sigma}{\Longmapsto} (0, x')$. Assume that if $x' < x''$, then $(0, x) \overset{\sigma}{\Longmapsto} (0, x'')$. By Lemmas C.2 and C.3, $(0, x'^+) \overset{\sigma}{\Longmapsto} (0, x')$. Therefore,

$$(0, x) \overset{\sigma}{\Longmapsto} (0, x'^+) \overset{\sigma}{\Longmapsto} (0, x').$$

Therefore, the induction hypothesis holds for any x', which leads to $(0, x^+)$.

Use mathematical induction on the difference $|x_2 - x_1|$. If $x_1 = x_2$, there is nothing to prove. Assume that the lemma holds for any x_1', x_2' such that $|x_2' - x_1'| < |x_2 - x_1|$. If $x_1 < x_2$, by Lemmas C.2 and C.4 and the induction hypothesis (note that $|x_1^+ - x_2| < |x_2 - x_1|$.)

$$(0, x_1) \overset{\sigma}{\Longmapsto} (0, x_1^+) \overset{\sigma}{\Longmapsto} (0, x_2).$$

Similarly, if $x_1 > x_2$, by Lemmas C.2 and C.3 and the induction hypothesis,

$$(0, x_1) \overset{\sigma}{\Longmapsto} (0, x_1^-) \overset{\sigma}{\Longmapsto} (0, x_2).$$

This completes the proof.

D.12. Proof of Lemma 4.9

Proof. By the symmetry, we only need to consider the case $|x_2| \geq |x_1|$. Use mathematical induction on the absolute value of x_1. Assume that there exists $x \in \mathbb{F}$ such that

$$(0, x) \overset{\sigma}{\Longmapsto} (x_1', x_2),$$

for any $|x_1'| < |x_1|$. First consider the case $x_1 \geq 0$. If $x_1 = 0$, there is nothing to prove.
If $x_1 \leq 2^{\mathfrak{e}_{\min}+1}$, by Lemma C.6, we have $\omega \in \Sigma$. Since there exists $x_1', x_2' \in \mathbb{F}$ such that $x_1' \oplus \omega = x_1$, $x_2' \oplus \omega = x_2$, and $x_1' < x_2'$, we have

$$(x_1', x_2') \overset{\sigma}{\Longmapsto} (x_1' \oplus \omega, x_2' \oplus \omega) = (x_1, x_2).$$

As $|x_1'| < |x_1|$, the induction hypothesis is satisfied.

If $x_1 > 2^{\mathfrak{e}_{\min}+1}$, define e as $e \coloneqq \widetilde{\mathfrak{e}}_{x_1^-}$. By Lemma C.6, there exists $\gamma \in \Sigma$ such that

$$2^{e-p-1} < \gamma \leq 2^{e-p}.$$

Then, there exists $x_2' = x_2$ or x_2^- such that

$$x_2' \oplus \gamma = x_2.$$

Then,

$$\left(x_1^-, x_2'\right) \overset{\sigma}{\Longmapsto} \left(x_1^- \oplus \gamma, x_2' \oplus \gamma\right) = \left(x_1, x_2\right).$$

As $\left|x_1^-\right| < x_1$ and $x_1 \leq 2^{\mathfrak{e}_{\max}}$, the induction hypothesis is satisfied.

Similarly, if $x_1 < 0$, there exists γ such that

$$(x_1', x_2') \overset{\sigma}{\Longmapsto} (x_1' \ominus \gamma, x_2' \ominus \gamma) = (x_1, x_2). \tag{8}$$

Therefore, the induction hypothesis is satisfied. This completes the proof. □