

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

Geonho Hwang
with Yeachan Park, Wonyeol Lee, and Sejun Park

Department of Mathematical Sciences,
Gwangju Institute of Science and Technology

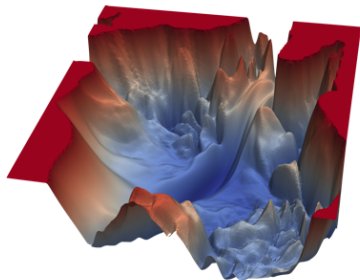
May 28, 2025

Motivation

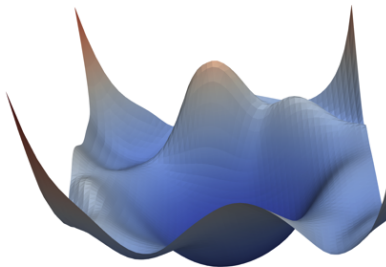
- Many foundational theories in deep learning are formulated over the field of real numbers, relying on the assumptions of infinite precision and continuity.
- For example, results such as the universal approximation theorem are typically stated in terms of real-valued functions, weights, and activations, existing within the realm of pure mathematics.

Motivation

No residual connections



With residual connections

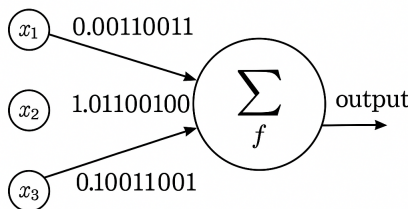


Same general network architecture

Figure: Landscape of Real Neural Network

Motivation

But actual neural networks operate on computer systems using finite-precision arithmetic.



Motivation

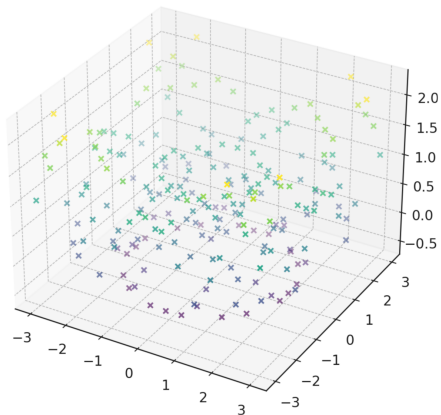


Figure: Floating-point landscape?

Motivation

- Modern LLM models actively adopt low-precision formats to reduce weight storage and computation.
- OpenAI uses mixed-precision training in models like GPT-3.5 and GPT-4—primarily employing FP16 to accelerate training and reduce GPU memory usage.
- This further widens the gap between theoretical analysis over the reals and practical implementation.

Motivation

For each theorem formulated over the reals, we need a corresponding version that applies to finite-precision computation.

- Universal Approximation Theorem \rightarrow Floating-point Universal Approximation Theorem
- Optimization Theory \rightarrow Optimization under Floating-Point Arithmetic
- Generalization Error Analysis \rightarrow Generalization Error Analysis under Floating-Point Arithmetic

Universal Approximation Theorem for Floating-point Setting

In this presentation, we focus on the universal approximation theorem in the context of floating-point neural networks.

Traditional universal approximation theorem: for every $f : \mathbb{R}^d \supset K \rightarrow \mathbb{R}$, there exists a neural network $g : K \rightarrow \mathbb{R}$ such that

$$\|f - g\| < \epsilon. \quad (1)$$

Floating-point universal approximation theorem: for every $f : \mathbb{F}^d \supset \mathcal{M} \rightarrow \mathbb{F}$, there exists a floating-point neural network $g : \mathcal{M} \rightarrow \mathbb{F}$ such that

$$f = g. \quad (2)$$

Floating-point Numbers

In modern computers, most numbers are saved in the format of floating-point numbers.

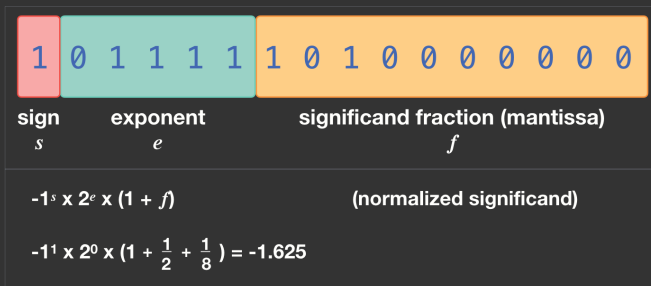
Definition

Let $p, q \in \mathbb{N}$ denote the number of mantissa bits and exponent bits, respectively. We define $\mathbb{F}_{p,q}$ to be the set of finite floating-point numbers:

$$\begin{aligned} \mathbb{F}_{p,q} := \\ \{s \times (1.m_1 \cdots m_p) \times 2^e : s \in \{-1, 1\}, m_1, \dots, m_p \in \{0, 1\}, e \in [\epsilon_{\min}, \epsilon_{\max}] \mathbb{Z}\} \\ \cup \{s \times (0.m_1 \cdots m_p) \times 2^{\epsilon_{\min}} : s \in \{-1, 1\}, m_1, \dots, m_p \in \{0, 1\}\} \quad (3) \end{aligned}$$

Floating-point Numbers

Traditional floating point (IEEE 754 style)



Format	(p, q)
8-bit E5M2 [MSB ⁺ 22]	(2, 5)
8-bit E4M3 [MSB ⁺ 22]	(3, 4)
16-bit half-precision float (float16)	(10, 5)
32-bit single precision float (float32)	(23, 8)
64-bit double precision float (float64)	(52, 11)
bfloat16 [Goo, AAB ⁺ 16]	(7, 8)

Table: List of frequently used floating-point formats.

The IEEE-754 standard [IEE19] defines p and q for widely used floating-point formats:

Floating-point Numbers

Exponents range from $\epsilon_{\min} := -2^{q-1} + 2$ to $\epsilon_{\max} := 2^{q-1} - 1$.

For example, in FP32, where $p = 23$ and $q = 8$, exponents are from -126 to 127 .

The smallest and the largest finite positive floating-point numbers are

$$\omega = 2^{\epsilon_{\min} - p} = 2^{-126 - 23} = 2^{-149}, \quad (4)$$

and

$$\Omega = (2 - 2^{-p}) \times 2^{\epsilon_{\max}} = 1.1 \dots 1 \times 2^{127}, \quad (5)$$

respectively. We define $\mathbb{F} = \mathbb{F}_{p,q}$. And

$$\overline{\mathbb{F}} = \mathbb{F} \cup \{\infty, -\infty, \text{NaN}\}. \quad (6)$$

Rounding Operation

The rounding operation $\lceil \cdot \rceil_{\mathbb{F}} : \mathbb{R} \cup \{-\infty, \infty, \text{NaN}\} \rightarrow \overline{\mathbb{F}}$ is defined as

$$\lceil x \rceil_{\mathbb{F}} = \begin{cases} \operatorname{argmin}_{y \in \mathbb{F}} |x - y| & \text{if } |x| < \Omega(1 + 2^{-p-1}), \\ \infty & \text{if } x \geq \Omega(1 + 2^{-p-1}), \\ -\infty & \text{if } x \leq -\Omega(1 + 2^{-p-1}), \\ \text{NaN} & \text{if } x = \text{NaN}. \end{cases}$$

Rounding Operation

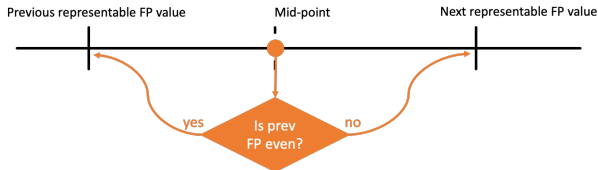


Figure: Round-to-even rounding rule

There may be two floating-point numbers equidistant from a real number. In such a case, we break the tie using the tie-to-even rule: $\lceil x \rceil_{\mathbb{F}}$ is y , where y is the unique floating-point number whose last mantissa bit m_p (see Equation (3)) is zero.

Floating-point Operation

For $x, y \in \overline{\mathbb{F}}$, we define the floating-point operations \oplus , \ominus , and \otimes as

$$x \oplus y := \lceil x + y \rceil, \quad (7)$$

$$x \ominus y := \lceil x - y \rceil, \quad (8)$$

and

$$x \otimes y := \lceil x \times y \rceil. \quad (9)$$

Pathological Behavior of Floating-point Numbers

Floating-point addition is not associative:

$$(x \oplus y) \oplus z \neq x \oplus (y \oplus z). \quad (10)$$

For example,

$$1 = (1 \oplus 2^{-p-1}) \oplus 2^{-p-1} = 1 \oplus (2^{-p-1} \oplus 2^{-p-1}) = 1 + 2^{-p}. \quad (11)$$

Correctly Rounded Activation Function

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

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

Floating-point Neural Networks

The affine transformation $\text{aff}_I : \mathbb{F}^{d_1} \rightarrow \mathbb{F}^{d_2}$ is defined as

$$\text{aff}_I(x_1, \dots, x_{d_1})_i = \left(\bigoplus_{j=1}^{d_1} (w_{i,j} \otimes x_j) \right) \oplus b_i \text{ for } i \in [d_2].$$

We must be very careful about the ordering of the operations. For multiple floating-point numbers x_1, \dots, x_n , we define \bigoplus as

$$\bigoplus_{i=1}^n x_i := x_1 \oplus \dots \oplus x_n = (\dots (x_1 \oplus x_2) \oplus x_3) \oplus \dots) \oplus x_n.$$

Then, for $\mathcal{I} = \{l_1, \dots, l_l\}$, a σ network $\mathcal{N}_{\mathcal{I}} : \mathbb{F}^{d_1} \rightarrow \mathbb{F}^{d_l}$ is defined as the composition:

$$\mathcal{N}_{\mathcal{I}} = \text{aff}_{l_l} \circ \sigma \circ \dots \circ \text{aff}_{l_2} \circ \sigma \circ \text{aff}_{l_1}.$$

Universal Approximation of Floating-point Neural Networks

Consider the case where we aim to approximate an arbitrary function $f : [-2, 2]_{\mathbb{F}} \rightarrow \mathbb{F}$. Suppose we define an activation function $\sigma : \mathbb{F} \rightarrow \mathbb{F}$ as follows:

$$\sigma(x) = \begin{cases} 0 & \text{if } |x| \leq 2^p, \\ 1 & \text{if } \infty > |x| > 2^p, \\ \text{NaN} & \text{if } |x| = \infty. \end{cases} \quad (12)$$

To achieve the universal approximation, we must be able to distinguish between 0 and ω , which means that there exists $w, b \in \mathbb{F}$ such that

$$\sigma(w \otimes 0 \oplus b) \neq \sigma(w \otimes \omega \oplus b). \quad (13)$$

If w is small,

$$\sigma(w \otimes 0 \oplus b) = \sigma(w \otimes \omega \oplus b). \quad (14)$$

If w is large, then,

$$\sigma(w \otimes 2 \oplus b) = \text{NaN}. \quad (15)$$

Distinguishability

Definition (Distinguishability)

Let $\sigma : \overline{\mathbb{F}} \rightarrow \overline{\mathbb{F}}$, $d \in \mathbb{N}$, $\mathcal{M} \subset \mathbb{F}^d$, and $\mathcal{R} \subset \overline{\mathbb{F}}$. We say that “ \mathcal{M} is σ -distinguishable with range \mathcal{R} ” if for any $x \in \mathcal{M}$, there exist $n \in \mathbb{N}$ and affine transformations $\phi_1, \dots, \phi_n : \mathbb{F}^d \rightarrow \overline{\mathbb{F}}$ satisfying the following: for each $y \in \mathcal{M}$, there exists $i_y \in [n]$ such that

$$\sigma(\phi_{i_y}(y)) \neq \sigma(\phi_{i_y}(x))$$

and $\sigma(\phi_i(\mathcal{M})) \subset \mathcal{R}$ for all $i \in [n]$.

Distinguishability

To represent all functions from $\mathcal{M} \subset \mathbb{F}^d$ to \mathbb{F} , one can observe that \mathcal{M} should be σ -distinguishable with range $\mathbb{F} \cup \{-\infty, \infty\}$.

Lemma

Let $d \in \mathbb{N}$, $\mathcal{M} \subset \mathbb{F}^d$, and $\sigma : \overline{\mathbb{F}} \rightarrow \overline{\mathbb{F}}$. If \mathcal{M} is not σ -distinguishable with range $\mathbb{F} \cup \{-\infty, \infty\}$, then there exists $f : \mathcal{M} \rightarrow \mathbb{F} \cup \{-\infty, \infty\}$ such that for any σ network g , there exists $x \in \mathcal{M}$ such that $g(x) \neq f(x)$.

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

Non-distinguishable Activation Function

Lemma

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

Proof Sketch.

$\cos(x) \approx 1 - \frac{x^2}{2}$ near zero, and if $\cos(x) > 1 - 2^{-p-1}$,

$$\lceil \cos \rceil (x) = 1. \quad (16)$$

Therefore, for $|x| \lesssim 2^{(-p-1)/2}$,

$$\lceil \cos \rceil (x) = 1. \quad (17)$$



Universal Approximation of Floating-point Neural Networks

Theorem

Let $\sigma : \overline{\mathbb{F}} \rightarrow \overline{\mathbb{F}}$, $d_1, d_2 \in \mathbb{N}$, $\mathcal{M} \subset \mathbb{F}^{d_1}$, and $f : \mathcal{M} \rightarrow (\mathbb{F} \cup \{-\infty, \infty\})^{d_2}$. Suppose that σ satisfies the following condition and \mathcal{M} is σ -distinguishable with range $[-2^{\epsilon_{\max}}, 2^{\epsilon_{\max}}]_{\mathbb{F}}$. Then, there exists a four-layer σ network g such that $g = f$ on \mathcal{M} .

Condition

For an activation function $\sigma : \overline{\mathbb{F}} \rightarrow \overline{\mathbb{F}}$, there exist $C_0, C_1, C_2 \in \mathbb{F}$ with $|C_i|, |C_j - C_i| \leq 2^{\epsilon_{\max}}$ for all $0 \leq i, j \leq 2$ such that

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

Sufficient Condition for Distinguishability

Lemma

Let $\rho : \mathbb{R} \rightarrow \mathbb{R}$ and $\hat{\rho}(x)$ be one of the following: $\rho(x), \rho(-x), -\rho(x), -\rho(-x)$. Suppose that there exist $e \in \mathbb{Z}$, $L_1, L_2 > 0$, and $\hat{\rho}(x)$ is one of $\rho(x), \rho(-x), -\rho(x), -\rho(-x)$ satisfying the following:

- $\hat{\rho}(x) \leq L_2 x$ for all $x \in [0, 2^e]$,
- $L_1 \leq \hat{\rho}'(x)$ for $0 < x < 2^e$,
- for $l_1 := \lfloor -1 + \log_2 L_1 \rfloor$ and $l_2 := \lfloor 1 + \log_2 L_2 \rfloor$, it holds that $l_1 \geq -p$ and $p \geq l_2 - l_1$.

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

Sufficient Condition for Distinguishability

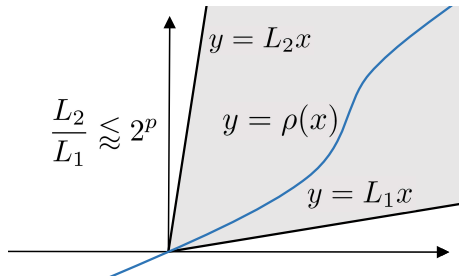


Figure: Visualization of the conditions in the lemma. $\rho : \mathbb{R} \rightarrow \mathbb{R}$ is located between L_1x and L_2x . The ratio between L_2 and L_1 must be at most approximately 2^p .

Distinguishable Activation Functions

Corollary

Let σ be a correctly rounded version of any of the following: identity, ReLU, ELU, SeLU, GELU, Swish, Mish and sin. For any $d \in \mathbb{N}$, σ networks can represent any functions from $(-2^{e_{\max}-2}, 2^{e_{\max}-2})_{\mathbb{F}}^d$ to $\mathbb{F} \cup \{-\infty, \infty\}$.

Distinguishable Activation Functions

Lemma

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

Corollary

For any $d \in \mathbb{N}$, networks using [Sigmoid] or [tanh] can represent any functions from \mathbb{F}^d to $\overline{\mathbb{F}}$.

Indicator Function

Lemma

Let $\sigma : \overline{\mathbb{F}} \rightarrow \overline{\mathbb{F}}$, $d \in \mathbb{N}$, and $\mathcal{M} \subset \mathbb{F}^d$. Suppose that σ satisfies Condition 5.1 and \mathcal{M} is σ -distinguishable with range $[-2^{\epsilon_{\max}}, 2^{\epsilon_{\max}}]_{\mathbb{F}}$. Then, for any $z \in \mathbb{F}^d$ and $c \in \{C_1, C_2\}$, there exists a three-layer σ network $f : \mathcal{M} \rightarrow \mathbb{F}$ ending with the activation function such that

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

Sequential Addition

Definition (Sequential addition)

Let $\sigma : \overline{\mathbb{F}} \rightarrow \overline{\mathbb{F}}$. We say a function $f : \overline{\mathbb{F}} \rightarrow \overline{\mathbb{F}}$ 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 $x \in \mathbb{F}$,

$$f(x) = x \oplus z_1 \oplus \dots \oplus z_n \quad (18)$$

where $\Sigma_\sigma := \{w \otimes \sigma(c) : w, c \in \mathbb{F} \text{ with } w \otimes \sigma(c) \in \mathbb{F}\}$. We often drop σ and use Σ to denote Σ_σ if it is clear from the context.

Sequential Addition

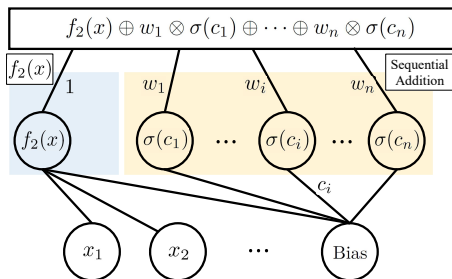


Figure: Visualization of the sequential addition embodied by an affine transformation. If f_2 is the output of some σ network ending with the activation function, then, for a sequential addition f_1 using σ , $f_1 \circ f_2$ can be represented by σ networks.

Power of Sequential Addition

Definition (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 “ $(x_1, \dots, x_n) \xrightarrow{\sigma} (y_1, \dots, y_n)$ ” if there exists a sequential addition $f : \mathbb{F} \rightarrow \mathbb{F}$ using σ such that $f(x_i) = y_i$ for all $i \in [n]$.

Lemma

Let $\sigma : \overline{\mathbb{F}} \rightarrow \overline{\mathbb{F}}$ and suppose that σ satisfies Condition 5.1. Then, for any $y \in [-2^{\epsilon_{\max}}, 2^{\epsilon_{\max}}]_{\mathbb{F}}$ and $x_1, x_2 \in [-2^{\epsilon_{\max}}, 2^{\epsilon_{\max}}]_{\mathbb{F}}$ such that $x_2 - x_1 \in (0, 2^{\epsilon_{\max}}]_{\mathbb{F}}$, it holds that

$$(-2^{\epsilon_{\max}}, y, y^+, 2^{\epsilon_{\max}}) \xrightarrow{\sigma} (x_1, x_1, x_2, x_2).$$

Entire Proof Sketch

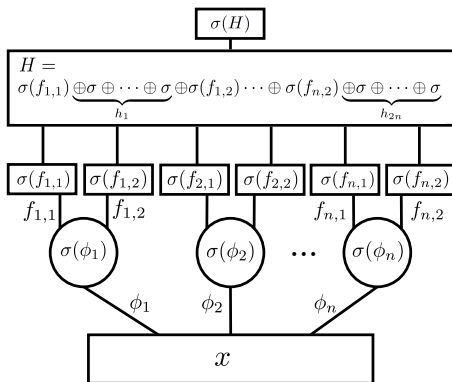


Figure: Structure of the indicator constructed by the σ network in Theorem 10. Affine transformations are represented as lines, and the output of activation functions are represented as circles and rectangles.

Conclusion

- In this work, we propose the necessary and sufficient conditions for activation functions in neural networks to represent arbitrary floating-point functions.
- Specifically, we demonstrate that the distinguishability of an activation function is crucial for determining the representability of neural networks.
- Our results cover almost all practical activation functions.

Future Work

- How does backpropagation work under floating-point arithmetic?
- Does it converge to a minimum?

References I



Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al., *Tensorflow: Large-scale machine learning on heterogeneous distributed systems*, arXiv preprint arXiv:1603.04467 (2016).



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



IEEE, *IEEE standard for floating-point arithmetic (IEEE Std 754-2019)*, IEEE, Piscataway, NJ, USA, 2019.



Paulius Micikevicius, Dusan Stosic, Neil Burgess, Marius Cornea, Pradeep Dubey, Richard Grisenthwaite, Sangwon Ha, Alexander Heinecke, Patrick Judd, John Kamalu, et al., *Fp8 formats for deep learning*, arXiv preprint arXiv:2209.05433 (2022).