CONSTRUCTIVE MATHEMATICAL ANALYSIS
xx (2011), No. 1, pp. 00-00
http://dergipark.org.tr/en/pub/cma
ISSN 2651 - 2939

Research Article

Elementary proof of Funahashi's theorem

MITSUO IZUKI, TAKAHIRO NOI, YOSHIHIRO SAWANO*, AND HIROKAZU TANAKA

ABSTRACT. Funahashi established that the space of two-layer feedforward neural networks is dense in the space of all continuous functions defined over compact sets in n-dimensional Euclidean space. The purpose of this short survey is to reexamine the proof of Theorem 1 in Funahashi [3]. The Tietze extension theorem, whose proof is contained in the appendix, will be used. This paper is based on harmonic analysis, real analysis, and Fourier analysis. However, the audience in this paper is supposed to be researchers who do not specialize in these fields of mathematics. Some fundamental facts that are used in this paper without proofs will be collected after we present some notation in this paper.

Keywords: neural network, activation function, Funahashi's theorem, Fourier analysis, uniform approximation

2020 Mathematics Subject Classification: 42B35, 47B33, 46E30.

1. INTRODUCTION

The goal of this survey is to prove the following theorem due to Funahashi using theorems on uniform convergence in harmonic analysis and real analysis:

Theorem 1.1 (Theorem 1 in Funahashi [3]). *Let $\phi(t)$ be a nonconstant, bounded, increasing, and continuous function on $\mathbb{R}$, and let $K \subset \mathbb{R}^n$ a compact set. Let $\varepsilon > 0$ and $f(x)$ be a continuous real-valued function on K. Then there exist a natural number N_1 and real constants c_k, θ_k, w_{kj} ($1 \le k \le N_1$, $1 \le j \le n$) such that*

$$\max_{x \in K} \left| f(x) - \tilde{f}(x) \right| < \varepsilon \tag{1.1}$$

holds, where

$$\tilde{f}(x) = \sum_{k=1}^{N_1} c_k \phi \left(\sum_{j=1}^{n} w_{kj} x_j - \theta_k \right) \quad (x = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n).$$

Mathematically, Theorem 1.1 can be understood as a theorem on uniform approximation. Uniform approximation is important when we consider the change of the limit and integration over compact sets. It is also important in the field of numerical analysis.

We say that $\tilde{f}(x)$ belongs to the space of two-layer feedforward neural networks generated by $\phi(t)$. In the branch of the neural network, $\phi(t)$ is called (0-)sigmoidal.

The field of artificial neural networks (or neural networks in short) began in 1943 when McCulloch and Pitts demonstrated that a combination of neuron-like computational units could

Received: xx.xx.xxxx; Accepted: xx.xx.xxxx; Published Online: xx.xx.xxxx
*Corresponding author: Yoshihiro Sawano; yoshihiro-sawano@celery.ocn.ne.jp
DOI: 10.33205/cma.xxxxxx

perform any logical operations [8]. Following this seminal work, in 1958, Rosenblatt formulated a single-layer neural network called a perceptron inspired by information processing in the central nervous system [11]. As a neuron emits an action potential when the sum of synaptic inputs exceeds the threshold, a perceptron performs a classification task by computing its activation according to a weighted sum of multiple inputs. Two notable theoretical analyses of the perceptron included the convergence theorem and the counting theorem; the former guarantees that a perceptron can learn a decision boundary when a training set is linearly separable [10], and the latter estimates the number of training points that a perceptron can learn [2]. Despite these conceptual and theoretical developments, interest in neural networks waned in the 1970s after Minksy and Pepert suggested that a perceptron cannot perform nonlinear operations as simple as exclusive or (XOR) [9]. A multilayer neural network could realize such nonlinear functions, but no learning algorithms were known to train a multilayer neural network.

The field of neural networks was revived in the early 1980s when the backpropagation algorithm was invented to train multilayer neural networks [13]. Errors in the output units propagate backward to hidden units, and the weights connected to hidden units are updated according to the backpropagated errors. The backpropagation algorithm allows a multilayer network to learn from any training set of non-linear relations. Introducing hidden units in a multilayer network resulted in two significant consequences. First, the multilayer neural network can find latent representations in hidden layers related to, but not the same as, network inputs and outputs. Such latent representations allow for abstraction and dimensional reduction of network input. Second, a multilayer network with hidden layers approximates arbitrary continuous mapping from input to output. The universal approximation theorem states that a multilayer network composed of at least one hidden layer can approximate any continuous function if the number of hidden units is large enough and the parameters (weights and thresholds) are appropriately adjusted.

A future historian might call the 21st century the century of neural networks. Since the seminal work of Krizhevsky et al. outperformed conventional image classification approaches in the ImageNet classification competition [7], deep neural networks prevail in various practical applications. Despite empirical success, the deep-network approach is counterintuitive from the point of view of conventional machine learning [14]. Although deep neural networks have billions or trillions of tunable weight parameters, the networks hardly overfit to training data and can generalize well to test data not used for training. Also, we do not understand theoretically the advantages of stacking many layers, so designing a deep neural network is still an art of trial and error rather than science. The lack of theoretical understanding of deep neural networks impedes a systematic and optimal network structure design for a given application.

This survey revisits Funahashi's proof of the universal approximation theorem [3]. The theorem justified the training of neural networks using arbitrary input-output mappings and played a crucial role in developing neural networks in the 1980s. We think it is essential to reexamine Funahashi's proof for multilayer neural networks with a single hidden layer to gain insight into how we can generalize the theorem to the case of deep neural networks. The theorem is also instrumental in guiding recent physiological experiments. A single neuron is not like a perceptron of linear separation as previously hypothesized, but can operate as a multilayer neural network that takes advantage of the non-linearity of synaptic input in dendritic trees [4, 1]. By depositing Funahashi's theorem in an accessible way, this survey aims to mediate a deeper understanding of deep neural networks and the brain.

Theorem 1.1 seems to cover bounded functions. However, if we use some linear combinations, then Theorem 1.1 can cover more functions. Let $\mathrm{ReLU}(t) = \max(0, t)$ be the rectified

linear unit. Although $\mathrm{ReLU}(t)$ is not a bounded function, the function $\phi(t) = \mathrm{ReLU}(t-1) - \mathrm{ReLU}(t)$ falls within the scope of Theorem 1.1. Therefore, the conclusion of Theorem 1.1 is true even for the case of $\phi(t) = \mathrm{ReLU}(t)$. The same applies to the function $\phi(t) = \mathrm{ReLU}(t)^k$. In [5, 6], the authors replaced the $\max$-norm with Banach lattices and generalized the condition on $\phi(t)$. Going through a similar argument, one can generalize the results in [5, 6] to the n-dimensional case.

Here we collect the notation and the preliminary facts in this paper.

(1) The set $\mathbb{N}_0$ consists of all nonnegative integers.
(2) Given $x, w \in \mathbb{R}^n$, we write the Euclidean inner product by $x \cdot w$. We also write $\|x\| = \sqrt{x \cdot x}$.
(3) Given $R > 0$, we write $B(R) = \{x \in \mathbb{R}^n : \|x\| < R\}$.
(4) Let $E \subset \mathbb{R}^n$ be a measurable set. The characteristic function $\chi_E(x)$ is defined by
$$\chi_E(x) = \begin{cases} 1 & (x \in E), \\ 0 & (x \notin E). \end{cases}$$
Furthermore, $|E|$ is the Lebesgue measure of E.
(5) Let $E \subset \mathbb{R}^n$ be a measurable set that satisfies $|E| > 0$ and $1 \le p \le \infty$. The Lebesgue space $L^p(E)$ consists of all measurable functions $f(x)$ on E satisfying $\|f\|_{L^p(E)} < \infty$, where
$$\|f\|_{L^p(E)} = \begin{cases} \left(\displaystyle\int_E |f(x)|^p \, dx \right)^{1/p} & (1 \le p < \infty), \\ \mathrm{ess.sup}_{x \in E} |f(x)| & (p = \infty). \end{cases}$$
If $f(x) \in L^1(E)$, then we say that $f(x)$ is integrable over E. If $E = \mathbb{R}^n$, then we merely say that $f(x)$ is integrable.
(6) Let $f(x)$ be a function defined in $\mathbb{R}^n$. The closure of the set $\{x \in \mathbb{R}^n : f(x) \neq 0\}$ is said to be the support of $f(x)$ and denoted by $\mathrm{supp} f$.
(7) The set $C(\mathbb{R}^n)$ is the set of all continuous functions in $\mathbb{R}^n$. In addition, the set $C_{\mathrm{c}}(\mathbb{R}^n)$ is the set of all $f \in C(\mathbb{R}^n)$ satisfying that $\mathrm{supp} f$ is compact.
(8) The set $C^\infty(\mathbb{R}^n)$ is the set of all infinitely differentiable functions on $\mathbb{R}^n$. In addition, the set $C^\infty_{\mathrm{c}}(\mathbb{R}^n)$ is the set of all $f \in C^\infty(\mathbb{R}^n)$ whose support is compact.
(9) The Schwartz class $\mathcal{S}(\mathbb{R}^n)$ consists of all functions $f \in C^\infty(\mathbb{R}^n)$ satisfying
$$\sum_{\alpha \in \mathbb{N}_0^n, j \in \mathbb{N}_0, |\alpha| + j \le N} \sup_{x \in \mathbb{R}^n} (1 + |x|)^j \, |\partial^\alpha f(x)| < \infty$$
for all $N \in \mathbb{N}_0$, where we write
$$|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_n, \quad \partial^\alpha f(x) = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_n^{\alpha_n}}(x)$$
for $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n) \in \mathbb{N}_0^n$.
(10) Given a complex number z, we can uniquely write $z = x + iy$, where $x, y \in \mathbb{R}$. We write $\mathrm{Re}(z) = x$ with this in mind.
(11) Given a function $f(x)$ on $\mathbb{R}^n$, we formally define the Fourier transform by
$$\mathcal{F}[f](w) = \hat{f}(w) = \int_{\mathbb{R}^n} f(x) e^{-ix \cdot w} \, dx \quad (w \in \mathbb{R}^n).$$
Then the inverse Fourier transform is defined by
$$\mathcal{F}^{-1}[f](x) = (2\pi)^{-n} \int_{\mathbb{R}^n} f(w) e^{ix \cdot w} \, dw \quad (x \in \mathbb{R}^n).$$

Let $f(x) \in C_c^\infty(\mathbb{R}^n)$. A fundamental result on Fourier analysis is that the convergence of the limits

$$\mathcal{F}[f](w) = \lim_{R\to\infty} \int_{B(R)} f(x) e^{-ix\cdot w}\, dx, \quad \mathcal{F}^{-1}[f](x) = (2\pi)^{-n} \lim_{R\to\infty} \int_{B(R)} f(w) e^{ix\cdot w}\, dw$$

take places uniformly over $w \in \mathbb{R}^n$ and that these operators satisfies

$$\mathcal{F}^{-1}[\mathcal{F}f](x) = f(x).$$

In the rest of this section, we recall a famous theorem in general topology, which plays a vital role in proving the main theorem.

Theorem 1.2 (Tietze extension theorem)**.** *Let $f : K \to \mathbb{R}$ be a continuous function defined over a compact set $K \subset \mathbb{R}^n$. Then there exists $g(x) \in C_c(\mathbb{R}^n)$ such that $g(x) = f(x)$ on K.*

We will give a self-contained proof of Theorem 1.2 as an appendix in Section 3. See [12] for the proof of the theorem in general topological spaces.

2. Proof of the main theorem

The next lemma is used to get some information from the function $\phi(t)$.

Lemma 2.1 (Lemma 1 in Funahashi [3])**.** *Let $\phi(t)$ be the same function as Theorem 1.1. Then there exist constants δ, $\alpha > 0$ such that $\psi(t) \in L^1(\mathbb{R})$ and that $\hat{\psi}(1) \neq 0$, where*

$$\psi(t) = \phi(t/\delta + \alpha) - \phi(t/\delta - \alpha).$$

In particular, $\psi(t)$ is real-valued because $\phi(t)$ is real-valued.

Proof. Let $L, L' > 0$ be large numbers. Then

$$\begin{aligned}\int_{-L'}^{L} \psi(t)\, dt &= \delta \int_{-L'/\delta+\alpha}^{L/\delta+\alpha} \phi(s)\, ds - \delta \int_{-L'/\delta-\alpha}^{L/\delta-\alpha} \phi(s)\, ds \\ &= \delta \int_{L/\delta-\alpha}^{L/\delta+\alpha} \phi(s)\, ds - \delta \int_{-L'/\delta-\alpha}^{-L'/\delta+\alpha} \phi(s)\, ds \in [0, 4\delta\alpha \sup|\phi|].\end{aligned}$$

Thus, since $L, L' > 0$ are arbitrary, $\psi(t)$ is integrable.

It remains to show that $\hat{\psi}(1) \neq 0$ for some suitable choice of $\delta > 0$. If $\hat{\psi}(1) = 0$ for all $\delta > 0$, then we would have $\mathcal{F}[\phi(\cdot + \alpha) - \phi(\cdot - \alpha)] = 0$. Thus, $\phi(t + \alpha) = \phi(t - \alpha)$. Putting $u = t - \alpha$, we have $\phi(u) = \phi(u + 2\alpha)$. This means that $\phi(t)$ is a periodic function with period 2α. From the periodicity and the assumption that $\phi(t)$ is increasing, $\phi(t)$ is a constant on $[0, 2\alpha]$. Again, from the periodicity, $\phi(t)$ is a constant on $\mathbb{R}$. But this contradicts the assumption that $\phi(t)$ is not constant. □

Rougly speaking, the idea of Funahashi is to apply the Fourier inversion forumula to have information on $\phi(t)$. Since Theorem 1.1 is stated in discrete form, while the Fourier inversion concerns the continuous representation, the integral over the whole space $\mathbb{R}^n$. Therefore, we need a tool that transforms continuous representations into discrete representations. Lemma 2.2 below serves this purpose.

Lemma 2.2 (Lemma 2 in Funahashi [3]). *Let $A > 0$, $K \subset \mathbb{R}^n$ be a compact set, and let $h(w,x)$ be a continuous function on $[-A,A]^n \times K$. Define the functions $H(x)$ and $H_N(x)$ $(N \in \mathbb{N})$ on K by*

$$H(x) = \int_{[-A,A]^n} h(w,x)\,dw,$$
$$H_N(x) = \left(\frac{2A}{N}\right)^n \sum_{k_1,k_2,\dots,k_n=0}^{N-1} h\left(-A+\frac{2k_1A}{N}, -A+\frac{2k_2A}{N}, \dots, -A+\frac{2k_nA}{N}, x\right).$$

Then for all $\varepsilon > 0$, there exists $N_0 \in \mathbb{N}$ such that $\max_{x\in K} |H(x) - H_N(x)| < \varepsilon$ for all $N \ge N_0$.

Proof. First, we abbreviate $\mathbf{1} = (1,1,\dots,1) \in \mathbb{R}^n$ to shorten the equations under calculation. On the other hand, $\mathbf{k} \in \{1,2,\dots,N-1\}^n$ means that $\mathbf{k} = (k_1,k_2,\dots,k_n)$ with every integer $k_j \in \{0,1,\dots,N-1\}$ $(j = 1,2,\dots,n)$. Thus we write

$$\sum_{\mathbf{k}\in\{1,2,\dots,N-1\}^n} = \sum_{k_1,k_2,\dots,k_n=0}^{N-1}.$$

Then, for any $\mathbf{k} = (k_1,k_2,\dots,k_n) \in \{0,1,\dots,N-1\}^n$,

$$\left(-A+\frac{2k_1A}{N}, -A+\frac{2k_2A}{N}, \dots, -A+\frac{2k_nA}{N}\right) = -A\mathbf{1} + \frac{2A}{N}\mathbf{k}$$

and

$$H_N(x) = \left(\frac{2A}{N}\right)^n \sum_{k_1,k_2,\dots,k_n=0}^{N-1} h\left(-A+\frac{2k_1A}{N}, -A+\frac{2k_2A}{N}, \dots, -A+\frac{2k_nA}{N}, x\right)$$
$$= \left(\frac{2A}{N}\right)^n \sum_{\mathbf{k}\in\{1,2,\dots,N-1\}^n} h\left(-A\mathbf{1}+\frac{2A}{N}\mathbf{k}, x\right). \tag{2.2}$$

We estimate

$$|H(x) - H_N(x)|$$
$$= \left| \int_{[-A,A]^n} h(w,x)\,dw - \left(\frac{2A}{N}\right)^n \sum_{\mathbf{k}\in\{1,2,\dots,N-1\}^n} h\left(-A\mathbf{1}+\frac{2A}{N}\mathbf{k}, x\right)\right|.$$

By the uniform continuity of $h(w,x)$, for any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$|h(w,x) - h(w',x)| < \frac{\varepsilon}{(2A)^n}$$

for any $w, w' \in \mathbb{R}^n$ satisfying $|w - w'| < \delta$. We fix $N_0 \in \mathbb{N}$ such that $\frac{2A}{N_0} \cdot \sqrt{n} < \delta$ and let $N > N_0$. Then we have

$$\left| w - \left(-A+\frac{2k_1A}{N}, -A+\frac{2k_2A}{N}, \dots, -A+\frac{2k_nA}{N}\right)\right| < \frac{2A}{N}\cdot\sqrt{n} < \delta$$

for each $(k_1,k_2,\dots,k_n) \in \{0,1,\dots,N-1\}^n$ and

$$w \in \prod_{j=1}^n \left[-A+\frac{2k_jA}{N}, -A+\frac{2(k_j+1)A}{N}\right]$$

So, we obtain

$$\left|h(w,x)-h\left(-A\mathbf{1}+\frac{2A}{N}\mathbf{k},x\right)\right|<\frac{\varepsilon}{(2A)^n} \tag{2.3}$$

for any

$$w\in\prod_{j=1}^{n}\left[-A+\frac{2k_jA}{N},-A+\frac{2(k_j+1)A}{N}\right],$$

where $\mathbf{k}=(k_1,k_2,\cdots,k_n)$. For each $\mathbf{k}=(k_1,k_2,\dots,k_n)\in\{0,1,\dots,N-1\}^n$, we put

$$C(\mathbf{k})=\prod_{j=1}^{n}\left[-A+\frac{2k_jA}{N},-A+\frac{2(k_j+1)A}{N}\right].$$

Then, by (2.2) and (2.3), we see that

$$\begin{aligned}
&|H(x)-H_N(x)|\\
&\le\sum_{\mathbf{k}\in\{1,2,\dots,N-1\}^n}\left|\int_{C(\mathbf{k})}h(w,x)\,dw-\int_{C(\mathbf{k})}h\left(-A\mathbf{1}+\frac{2A}{N}\mathbf{k},x\right)\,dw\right|\\
&=\sum_{\mathbf{k}\in\{1,2,\dots,N-1\}^n}\left|\int_{C(\mathbf{k})}\left\{h(w,x)-h\left(-A\mathbf{1}+\frac{2A}{N}\mathbf{k},x\right)\right\}dw\right|\\
&=\sum_{k_1,k_2,\dots,k_n=0}^{N-1}\left|\int_{C(\mathbf{k})}\left\{h(w,x)-h\left(-A\mathbf{1}+\frac{2A}{N}\mathbf{k},x\right)\right\}dw\right|\\
&\le\sum_{k_1,k_2,\dots,k_n=0}^{N-1}\frac{\varepsilon}{(2A)^n}\left(\frac{2A}{N}\right)^n\\
&\le\varepsilon.
\end{aligned}$$

This completes the proof. □

Lemma 2.3. *Assume that $f(x)\in L^1(\mathbb{R}^n)$ satisfies $\mathcal{F}[f](w)\in L^1(\mathbb{R}^n)$. For all $0<A<\infty$ and all $x\in\mathbb{R}^n$, we have $I_{\infty,A}(f)(x)=J_A(x)$, where $I_{\infty,A}(f)(x)$ and $J_A(f)(x)$ are defined by (2.6) and (2.7) below, respectively. In addition, both $\{J_A(f)(x)\}_{A>0}$ and $\{I_{\infty,A}(f)(x)\}_{A>0}$ converge uniformly in $\mathbb{R}^n$.*

Proof. Let $\psi(t)$ be a function as in Lemma 2.1. By the Lebesgue dominated convergence theorem, we see that

$$\lim_{A'\to\infty}\int_{-\infty}^{\infty}\psi(t)e^{-it}\chi_{[x\cdot w-A',x\cdot w+A']}(t)\,dt=\hat{\psi}(1).$$

Thus, to prove that $I_{\infty,A}(f)(x)=J_A(f)(x)$ for all $x\in\mathbb{R}^n$, it suffices to prove that

$$\begin{aligned}
&\lim_{A'\to\infty}\int_{[-A,A]^n}\hat{f}(w)e^{ix\cdot w}\left(\int_{-\infty}^{\infty}\psi(t)e^{-it}\chi_{[x\cdot w-A',x\cdot w+A']}(t)\,dt\right)dw\\
&=\int_{[-A,A]^n}\hat{f}(w)e^{ix\cdot w}\hat{\psi}(1)\,dw.
\end{aligned}\tag{2.4}$$

Fix $A>0$ for the time being. We remark that

$$\left|\hat{f}(w)e^{ix\cdot w}\left(\int_{-\infty}^{\infty}\psi(t)e^{-it}\chi_{[x\cdot w-A',x\cdot w+A']}(t)\,dt\right)\right|\le\left|\hat{f}(w)\right|\|\psi\|_{L^1(\mathbb{R})}\tag{2.5}$$

and that $\left|\hat{f}(w)\right| \|\psi\|_{L^1(\mathbb{R})}$ is independent of A' and integrable on $[-A,A]^n$. Therefore, applying the Lebesgue dominated convergence theorem again, we obtain (2.4). Furthermore, we show that $\{J_A(f)(x)\}_{A>0}$ converges to $\mathcal{F}^{-1}\left[\hat{f}\right](x)$ uniformly in $\mathbb{R}^n$. Since $\hat{f}(w)$ is integrable, we see that

$$\begin{aligned}
&\sup_{x\in\mathbb{R}^n}\left|\mathcal{F}^{-1}\left[\hat{f}\right](x) - J_A(f)(x)\right| \\
&= (2\pi)^{-n}\sup_{x\in\mathbb{R}^n}\left|\int_{\mathbb{R}^n}\hat{f}(w)e^{ix\cdot w}\left(1-\chi_{[-A,A]^n}(w)\right)\,dw\right| \\
&\le (2\pi)^{-n}\int_{\mathbb{R}^n}\left|\hat{f}(w)\right|\left(1-\chi_{[-A,A]^n}(w)\right)\,dw \\
&\to 0 \quad (A\to\infty).
\end{aligned}$$

This finishes the proof of the lemma. □

We now refer back to the proof of Theorem 1.1.

Proof of Theorem 1.1. Take $\varepsilon>0$ arbitrarily. Let $\psi(t)$ be the function defined by Lemma 2.1.

(I) First, suppose that $f(x)\in C_c^\infty(\mathbb{R}^n)$. Here $f(x)$ need not be supported on K. Let $0<A<\infty$ and $0<A'<\infty$. We define

$$\begin{aligned}
I_{A',A}(f)(x) &= \int_{[-A,A]^n}\left(\int_{-A'}^{A'}\psi(x\cdot w-w_0)\cdot\frac{1}{(2\pi)^n\hat{\psi}(1)}\hat{f}(w)e^{iw_0}\,dw_0\right)dw \\
&= \frac{1}{(2\pi)^n\hat{\psi}(1)}\int_{[-A,A]^n}\hat{f}(w)e^{ix\cdot w}\left(\int_{-\infty}^{\infty}\psi(t)e^{-it}\chi_{[x\cdot w-A',x\cdot w+A']}(t)\,dt\right)dw,
\end{aligned}$$

$$I_{\infty,A}(f)(x) = \lim_{A'\to\infty} I_{A',A}(f)(x), \tag{2.6}$$

and

$$J_A(f)(x) = (2\pi)^{-n}\int_{[-A,A]^n}\hat{f}(w)e^{ix\cdot w}\,dw. \tag{2.7}$$

So far, we know that $I_{\infty,A}(f)(x)=J_A(f)(x)$ for all $x\in\mathbb{R}^n$ and $A>0$ due to Lemma 2.3. Because $f\in C_c^\infty(\mathbb{R}^n)\subset\mathcal{S}(\mathbb{R}^n)$, we see that

$$f(x)=\mathcal{F}^{-1}\left[\hat{f}\right](x)=\lim_{A\to\infty}J_A(f)(x)=\lim_{A\to\infty}I_{\infty,A}(f)(x), \tag{2.8}$$

where the convergence in (2.8) takes place uniformly in $\mathbb{R}^n$. Thus, there exists $A_0>0$ such that for all $A>A_0$,

$$\max_{x\in\mathbb{R}^n}|f(x)-I_{\infty,A}(f)(x)|<\frac{\varepsilon}{3}. \tag{2.9}$$

Below we take $A > A_0$ arbitrarily. Now we approximate $I_{\infty,A}(f)(x)$ on K using $I_{A',A}(f)(x)$ with $A' < \infty$. We fix $x \in K$ and $0 < A' < \infty$. Then we have

$$\begin{aligned}
&|I_{\infty,A}(f)(x) - I_{A',A}(f)(x)| \\
&\leq \frac{1}{(2\pi)^n \left|\hat{\psi}(1)\right|} \int_{[-A,A]^n} \left|\hat{f}(w)\right| \left\{ \int_{\mathbb{R}\setminus[-A',A']} |\psi(x\cdot w - w_0)|\, dw_0 \right\} dw \\
&= \frac{1}{(2\pi)^n \left|\hat{\psi}(1)\right|} \int_{[-A,A]^n} \left|\hat{f}(w)\right| \left\{ \int_{-\infty}^{\infty} |\psi(t)|\, \chi_{\mathbb{R}\setminus[x\cdot w - A', x\cdot w + A']}(t)\, dt \right\} dw.
\end{aligned}$$

Because the set K is bounded, there exists $R > 0$ such that $K \subset B(R)$. Let $w \in [-A, A]^n$. Then we have $|x \cdot w| \leq \|x\|\, \|w\| \leq R \cdot \sqrt{n} A$ and

$$\begin{aligned}
&\mathbb{R} \setminus [x\cdot w - A', x\cdot w + A'] \\
&= (-\infty, x\cdot w - A') \cup (x\cdot w + A', \infty) \\
&\subset \left(-\infty, \sqrt{n}RA - A'\right) \cup \left(-\sqrt{n}RA + A', \infty\right) \\
&=: J.
\end{aligned}$$

We remark that the set J is independent of x and w. Hence we obtain

$$\begin{aligned}
&(2\pi)^n \left|\hat{\psi}(1)\right| \max_{x\in K} |I_{\infty,A}(f)(x) - I_{A',A}(f)(x)| \\
&\leq \int_{[-A,A]^n} \left|\hat{f}(w)\right| dw \cdot \left(\max_{x\in K,\, w\in[-A,A]^n} \int_{-\infty}^{\infty} |\psi(t)|\, \chi_{\mathbb{R}\setminus[x\cdot w - A', x\cdot w + A']}(t)\, dt \right) \\
&\leq \int_{[-A,A]^n} \left|\hat{f}(w)\right| dw \cdot \int_{-\infty}^{\infty} |\psi(t)|\, \chi_J(t)\, dt.
\end{aligned}$$

We note that $\lim_{A'\to\infty} |\psi(t)|\, \chi_J(t) = 0$, $|\psi(t)| \in L^1(\mathbb{R})$ and $|\psi(t)|\, \chi_J(t) \leq |\psi(t)|$. Therefore, by virtue of the Lebesgue dominated convergence theorem, we have $\lim_{A'\to\infty} \int_{-\infty}^{\infty} |\psi(t)|\, \chi_J(t)\, dt = 0$. Namely there exists $A_0' > 0$ such that for all $A' > A_0'$,

$$\max_{x\in K} |I_{\infty,A}(f)(x) - I_{A',A}(f)(x)| < \frac{\varepsilon}{3}. \tag{2.10}$$

Combining (2.9) and (2.10) we obtain

$$\max_{x\in K} |f(x) - I_{A',A}(f)(x)| < \frac{2}{3}\varepsilon. \tag{2.11}$$

(II) Next, we consider the general case: $f(x)$ is merely a continuous function defined over K. We prove that a modified estimate of (2.11) is true. We take a real-valued extension $g(x) \in C_{\mathrm{c}}(\mathbb{R}^n)$ of $f(x)$. This is possible due to the Tietze extension theorem (Theorem 1.2). Let $\rho(x) \in C_{\mathrm{c}}^{\infty}(\mathbb{R}^n)$ be such that $0 \leq \rho(x) \leq \chi_{B(1)}(x)$ for all $x \in \mathbb{R}^n$ and $\|\rho\|_{L^1(\mathbb{R}^n)} = 1$. Write $\rho_\beta(x) = \beta^{-n}\rho(\beta^{-1}x)$. Define the convolution $\rho_\beta * g(x)$ by $\rho_\beta * g(x) = \int_{\mathbb{R}^n} \rho_\beta(x-y) g(y)\, dy$. We employ the operation $g(x) \mapsto \rho_\beta * g(x)$, which is called the mollifier. Applying the mollifier to $g(x)$, we find $\beta \in (0,1)$ such that

$$\|g - \rho_\beta * g\|_{L^\infty(\mathbb{R}^n)} < \frac{\varepsilon}{3}.$$

A geometric observation shows that $\mathrm{supp} g \subset \mathrm{supp}(\rho_\beta * g)$ and that $\mathrm{supp}(\rho_\beta * g)$ is contained in a fixed compact set L, the set of all points x whose distance from x does not exceed 1. Since

$\rho_\beta * g(x) \in C_{\rm c}^\infty(\mathbb{R}^n)$, we can apply (2.11) to the function $\rho_\beta * g(x)$. That is, there exist $0 < A_0 < \infty$ and $0 < A_0' < \infty$ such that for all $A_0 < A < \infty$ and $A_0' < A' < \infty$,

$$\max_{x \in \mathrm{supp}(\rho_\beta * g)} |\rho_\beta * g(x) - I_{A',A}(\rho_\beta * g)(x)| < \frac{2}{3}\varepsilon.$$

Recall that $g(x)$ is an extension of $f(x)$. Hence,

$$\max_{x \in K} |f(x) - I_{A',A}(\rho_\beta * g)(x)| = \max_{x \in K} |g(x) - I_{A',A}(\rho_\beta * g)(x)|.$$

Therefore, we get

$$\begin{aligned}
&\max_{x \in K} |f(x) - I_{A',A}(\rho_\beta * g)(x)| \\
&\le \max_{x \in \mathrm{supp} g} |g(x) - I_{A',A}(\rho_\beta * g)(x)| \\
&\le \max_{x \in \mathrm{supp} g} |g(x) - \rho_\beta * g(x)| + \max_{x \in \mathrm{supp} g} |\rho_\beta * g(x) - I_{A',A}(\rho_\beta * g)(x)| \\
&\le \|g - \rho_\beta * g\|_{L^\infty(\mathbb{R}^n)} + \max_{x \in \mathrm{supp}(\rho_\beta * g)} |\rho_\beta * g(x) - I_{A',A}(\rho_\beta * g)(x)| \\
&< \varepsilon. \qquad (2.12)
\end{aligned}$$

(III) Finally, we prove the conclusion of the theorem applying (2.12). We note that $f(x)$ is real-valued but that $I_{A',A}(\rho_\beta * g)(x)$ is complex-valued. This means that $H(x) = \mathrm{Re}\,(I_{A',A}(\rho_\beta * g)(x))$ is a more suitable candicate of the approximation of f:

$$\begin{aligned}
|f(x) - I_{A',A}(\rho_\beta * g)(x)| &\ge |\mathrm{Re}\,(f(x) - I_{A',A}(\rho_\beta * g)(x))| \\
&= |f(x) - H(x)|,
\end{aligned}$$

that is, $\max_{x \in K} |f(x) - H(x)| < \varepsilon$. Meanwhile, applying Lemma 2.2 to $H(x)$, there exists a natural number N_0 such that $\max_{x \in K} |H(x) - H_N(f)(x)| < \varepsilon$ holds for all $N \ge N_0$, where

$$\begin{aligned}
H_N(f)(x) &= \left(\frac{2A}{N}\right)^n \sum_{k_1,k_2,\dots,k_n=0}^{N-1} h\left(-A + \frac{2k_1A}{N}, -A + \frac{2k_2A}{N}, \dots, -A + \frac{2k_nA}{N}, x\right), \\
h(w,x) &= \int_{-A'}^{A'} \psi(x \cdot w - w_0)\gamma(w,w_0)\, dw_0, \\
\gamma(w,w_0) &= \mathrm{Re}\left(\frac{1}{(2\pi)^n \hat{\psi}(1)} \mathcal{F}[\rho_\beta * g](w) e^{iw_0}\right).
\end{aligned}$$

Hence we have

$$\max_{x \in K} |f(x) - H_N(f)(x)| < 2\varepsilon \qquad (2.13)$$

using the triangle inequality. At this moment, we could manage to find $H_N(f)(x)$ which approximates $f(x)$. However, $H_N(f)(x)$ does not satisfy the requirement of the statement. So, we apply Lemma 2.2 to $H_N(f)(x)$ once again to construct the desired function $\tilde{f}(x)$.

This can be achieved as follows: Using the same notation as in Lemma 2.2, then

$$\begin{aligned}
&\left(\frac{2A}{N}\right)^{-n} H_N(f)(x)\\
&= \sum_{k_1,k_2,\dots,k_n=0}^{N-1} h\left(-A+\frac{2k_1A}{N}, -A+\frac{2k_2A}{N}, \dots, -A+\frac{2k_nA}{N}, x\right)\\
&= \sum_{\mathbf{k}\in\{0,1,\dots,N-1\}^n} h\left(-A\mathbf{1}+\frac{2A}{N}\mathbf{k}, x\right)\\
&= \int_{-A'}^{A'} \sum_{\mathbf{k}\in\{0,1,\dots,N-1\}^n} \psi\left(x\cdot\left(-A\mathbf{1}+\frac{2A}{N}\mathbf{k}\right)-w_0\right)\gamma\left(-A\mathbf{1}+\frac{2A}{N}\mathbf{k}, w_0\right)\,dw_0.
\end{aligned}$$

To approximate $\left(\frac{2A}{N}\right)^{-n} H_N(f)(x)$ by a Riemann sum, abbreviate

$$\begin{aligned}
&\frac{2A'}{M}\sum_{m=0}^{M-1}\sum_{\mathbf{k}\in\{0,1,\dots,N-1\}^n} \psi\left(x\cdot\left(-A\mathbf{1}+\frac{2A}{N}\mathbf{k}\right)-\left(-A'+\frac{2mA'}{M}\right)\right)\\
&\quad\times\gamma\left(-A\mathbf{1}+\frac{2A}{N}\mathbf{k}, -A'+\frac{2mA'}{M}\right)
\end{aligned}$$

to $R_M(f)(x)$, where $M\in\mathbb{N}$. Using Lemma 2.2, we can find $M_0\in\mathbb{N}$ such that for any $M>M_0$,

$$\max_{x\in K}\left|\left(\frac{2A}{N}\right)^{-n} H_N(f)(x)-R_M(f)(x)\right| < \left(\frac{2A}{N}\right)^{-n}\varepsilon.$$

Estimate (2.13) and the above inequality lead the estimate

$$\max_{x\in K}\left|f(x)-\left(\frac{2A}{N}\right)^{n} R_M(f)(x)\right| < 3\varepsilon. \tag{2.14}$$

We prove that $\left(\frac{2A}{N}\right)^{n} R_M(f)(x)$ is the desired function $\tilde{f}(x)$. Note that $R_M(f)(x)$ can be expressed as

$$\begin{aligned}
&R_M(f)(x)\\
&= \frac{2A'}{M}\sum_{m=0}^{M-1}\sum_{\mathbf{k}\in\{0,1,\dots,N-1\}^n} \psi\left((x,-1)\cdot\left(-A\mathbf{1}+\frac{2A}{N}\mathbf{k}, -A'+\frac{2mA'}{M}\right)\right)\\
&\quad\times\gamma\left(-A\mathbf{1}+\frac{2A}{N}\mathbf{k}, -A'+\frac{2mA'}{M}\right).
\end{aligned}$$

To deform this expression, we put

$$\mathbf{\Omega}(m,\mathbf{k}) = \left(-A\mathbf{1}+\frac{2A}{N}\mathbf{k}, -A'+\frac{2mA'}{M}\right)\in\mathbb{R}^{n+1}$$

for every m, $\mathbf{k}$. The set $\{\mathbf{\Omega}(m,\mathbf{k}) : m=0,1,\dots,M-1, \mathbf{k}\in\{0,1,\dots,N-1\}^n\}$ consists of N^nM vectors. Thus every $\mathbf{\Omega}(m,\mathbf{k})$ can be expressed as $\mathbf{\Omega}(m,\mathbf{k})=\mathbf{\Omega}(\ell)$ $(\ell=1,2,\dots,N^nM)$. Because $\mathbf{\Omega}(\ell)\in\mathbb{R}^{n+1}$, we write

$$\mathbf{\Omega}(\ell) = (\Omega_{\ell,1}, \Omega_{\ell,2}, \dots, \Omega_{\ell,n+1}).$$

Then, by the definition of ψ, we have

$$\begin{aligned}
R_M(f)(x) &= \frac{2A'}{M}\sum_{\ell=1}^{N^nM}\psi\left((x,-1)\cdot\mathbf{\Omega}(\ell)\right)\gamma(\mathbf{\Omega}(\ell)) \\
&= \frac{2A'}{M}\sum_{\ell=1}^{N^nM}\gamma(\mathbf{\Omega}(\ell))\psi\left(\sum_{j=1}^{n}x_j\Omega_{\ell,j}-\Omega_{\ell,n+1}\right) \\
&= \frac{2A'}{M}\sum_{\ell=1}^{N^nM}\gamma(\mathbf{\Omega}(\ell))\phi\left(\sum_{j=1}^{n}\frac{x_j\Omega_{\ell,j}}{\delta}-\left(\frac{\Omega_{\ell,n+1}}{\delta}-\alpha\right)\right) \\
&\qquad -\frac{2A'}{M}\sum_{\ell=1}^{N^nM}\gamma(\mathbf{\Omega}(\ell))\phi\left(\sum_{j=1}^{n}\frac{x_j\Omega_{\ell,j}}{\delta}-\left(\frac{\Omega_{\ell,n+1}}{\delta}+\alpha\right)\right).
\end{aligned}$$

By rearranging the right-hand side, we can find real constants c_ℓ, θ_ℓ, $w_{\ell j}$, $\ell=1,2,\dots,2N^nM$, $j=1,2,\dots,n$ such that

$$\left(\frac{2A}{N}\right)^n R_M(f)(x)=\sum_{\ell=1}^{2N^nM}c_\ell\phi\left(\sum_{j=1}^{n}w_{\ell j}x_j-\theta_\ell\right)\quad(x=(x_1,x_2,\dots,x_n)\in\mathbb{R}^n).$$

Since (2.14) is nothing but (1.1) with ε replaced by 3ε, it follows that $\left(\frac{2A}{N}\right)^n R_M(f)(x)$ is the desired function $\tilde{f}(x)$. □

If a function $f(x)$ is continuous in a compact set K, then we see that

$$\|f\|_{L^2(K)}=\left(\int_K|f(x)|^2\,dx\right)^{1/2}\le|K|^{1/2}\cdot\max_{x\in K}|f(x)|.$$

Thus we easily obtain the following corollary:

Corollary 2.1. *In Theorem 1.1, one has*

$$\left\|f-\tilde{f}\right\|_{L^2(K)}<|K|^{1/2}\varepsilon.$$

3. Appendix–Proof of the Tietze extension theorem

Let $\mathrm{ReLU}(t)=\max(0,t)$. We write

$$\mu(t)=\mathrm{ReLU}(t+1)-2\mathrm{ReLU}(t)+\mathrm{ReLU}(t-1)\quad(t\in\mathbb{R}).$$

Note that $\mu(t)$ vanishes outside $(-1,1)$ and that $\mu(t)=1-|t|$ for $t\in[-1,1]$. We set

$$\nu(x)=\nu(x_1,x_2,\dots,x_n)=\prod_{j=1}^{n}\mu(x_j),$$

so that

$$\sum_{\mathbf{k}\in\mathbb{Z}^n}\nu(x-\mathbf{k})=1.$$

Lemma 3.4. *Let $K \subset \mathbb{R}^n$ be a compact set and $f(x)$ be a continuous function on K. Write $M = \max_{y\in K} |f(y)|$. There exists a continuous function $g(x)$ defined on $\mathbb{R}^n$ such that*

$$\sup_{x\in K} |f(x) - g(x)| \le \frac{2}{3}M$$

and that

$$\sup_{y\in\mathbb{R}^n} |g(y)| \le \frac{2}{3}M.$$

Proof. Since $f(x)$ is continuous in the compact set K, $f(x)$ is uniformly continuous in K. Thus, we can find $\delta > 0$ such that $|f(x) - f(y)| < \frac{1}{12}M$ for all $x, y \in K$ such that $|x - y| < \delta$. Set

$$h(x) = \min\left(\max\left(-\frac{2}{3}M, f(x)\right), \frac{2}{3}M\right) \quad (x \in K).$$

Note that

$$h(x) = \begin{cases} -\frac{2}{3}M & (f(x) \le -\frac{2}{3}M), \\ f(x) & (-\frac{2}{3}M \le f(x) \le \frac{2}{3}M), \\ \frac{2}{3}M & (\frac{2}{3}M \le f(x)). \end{cases} \tag{3.15}$$

Since $f(x)$ is continuous in K, so $h(x)$ is also continuous in K. By (3.15) and $-M \le f(x) \le M$, it is easy to see that

$$|f(x) - h(x)| \le \frac{1}{3}M.$$

Next, we prove

$$|h(x) - h(y)| < \frac{1}{3}M \tag{3.16}$$

for all $x, y \in K$ such that $|x - y| < \delta$. Note that if $h(x) = \frac{2}{3}M$, then

$$-\frac{1}{12}M < f(y) - f(x) < \frac{1}{12}M \text{ and } \frac{2}{3}M \le f(x)$$

yield

$$\frac{7}{12}M = -\frac{1}{12}M + \frac{2}{3}M \le -\frac{1}{12}M + f(x) < f(y).$$

This implies that $\frac{7}{12}M < h(y) \le \frac{2}{3}M = h(x)$. Therefore, we have

$$|h(x) - h(y)| \le \frac{1}{12}M < \frac{1}{3}M.$$

From the symmetry, we see that (3.16) holds when the case $h(x) = \frac{2}{3}M$ or $h(y) = \frac{2}{3}M$.

To complete the proof of (3.16), it remains to consider the case

$$h(x) = \max\left(-\frac{2}{3}M, f(x)\right) \text{ and } h(y) = \max\left(-\frac{2}{3}M, f(y)\right).$$

Note that

$$\max(a, b) = \frac{1}{2}\left(a + b + |a - b|\right), \quad ||a| - |b|| \le |a - b|$$

for $a, b \in \mathbb{R}$. Hence, we obtain

$$\begin{aligned}|h(x)-h(y)| &\le \frac{1}{2}\cdot|f(x)-f(y)| + \frac{1}{2}\left|\left|f(x)+\frac{2}{3}M\right| - \left|f(y)+\frac{2}{3}M\right|\right| \\ &\le \frac{1}{2}\cdot\frac{1}{12}M + \frac{1}{2}|f(x)-f(y)| \\ &\le \frac{1}{12}M < \frac{1}{3}M.\end{aligned}$$

Finally we construct $g(x)$. Choose an integer A large enough so that $2A\delta > 1$. Denote by U the set of all $\mathbf{k} \in \mathbb{Z}^n$ such that $\{x \in \mathbb{R}^n \,:\, x - A^{-1}\mathbf{k} \in [-A^{-1}, A^{-1}]^n\} \cap K \neq \emptyset$. From the definition of U, it follows that

$$\sum_{\mathbf{k}\in U} \nu(Ax-\mathbf{k}) = 1 \quad (x \in K).$$

For each $\mathbf{k} \in U$, choose $y_{\mathbf{k}} \in \{x \in \mathbb{R}^n \,:\, x - A^{-1}\mathbf{k} \in [-A^{-1}, A^{-1}]^n\} \cap K$. We put

$$g(x) = \sum_{\mathbf{k}\in U} h(y_{\mathbf{k}})\nu(Ax-\mathbf{k}) \quad (x \in \mathbb{R}^n).$$

Then $g(x)$ vanishes outside the set $\{w \in \mathbb{R}^n \,:\, w = y+z, y \in K, z \in [-A^{-1}, A^{-1}]^n\}$ and satisfies

$$g(x) - h(x) = \sum_{\mathbf{k}\in U} (h(y_{\mathbf{k}}) - h(x))\nu(Ax-\mathbf{k}) \quad (x \in K).$$

This equality implies that

$$|g(x) - h(x)| \le \frac{1}{3}M.$$

Since $|f(x) - h(x)| \le \frac{1}{3}M$ it follows that $|f(x) - g(x)| \le \frac{2}{3}M$. Furthermore, since $|h(x)| \le \frac{2}{3}M$ for all $x \in K$, it follows that $|g(x)| \le \frac{2}{3}M$ for all $x \in \mathbb{R}^n$. Thus, the proof is complete. □

With Lemma 3.4 in mind, let us prove Theorem 1.2. Let $M = \max_{x\in K} |f(x)|$. Without loss of generality, we can assume $M = 1$. We define the sequence of functions $\{g_k(x)\}_{k=1}^{\infty}$ as follows. First, we choose $g_1(x)$ as in Lemma 3.4. That is,

$$|f(x) - g_1(x)| \le \frac{2}{3} \text{ on } K$$

and $|g_1(x)| \le \frac{2}{3}$ hold.

Then define $l_1(x) = f(x) - g_1(x)$. Next apply Lemma 3.4 to the function $l_1(x)$ to have a function $g_2(x)$ satisfying

$$|l_1(x) - g_2(x)| \le \frac{2}{3}\max_{y\in K}|l_1(y)| = \left(\frac{2}{3}\right)^2 \quad (x \in K)$$

and

$$|g_2(x)| \le \frac{2}{3}\max_{y\in K}|l_1(y)| = \left(\frac{2}{3}\right)^2 \quad (x \in \mathbb{R}^n).$$

Next, define $l_2(x) = f(x) - g_1(x) - g_2(x)$ and use Lemma 3.4 for the function $l_2(x)$. We repeat this procedure to have the functions $\{g_k(x)\}_{k=1}^{\infty}$ and $\{l_k(x)\}_{k=1}^{\infty}$ satisfying

$$l_k(x) = f(x) - g_1(x) - g_2(x) - \cdots - g_k(x) = f(x) - \sum_{s=1}^{k} g_s(x) \quad (x \in K),$$

$$|l_k(x) - g_{k+1}(x)| = \left| f(x) - \sum_{s=1}^{k+1} g_s(x) \right| \leq \frac{2}{3} \max_{y \in K} |l_k(y)| \leq \left(\frac{2}{3}\right)^{k+1} \tag{3.17}$$

and

$$|g_{k+1}(x)| \leq \frac{2}{3} \max_{y \in K} |l_k(y)| \leq \left(\frac{2}{3}\right)^{k+1} \quad (x \in \mathbb{R}^n). \tag{3.18}$$

From (3.17) and (3.18), we learn that

$$g(x) = \sum_{k=1}^{\infty} g_k(x)$$

converges uniformly over $x \in \mathbb{R}^n$ and that $g(x)$ agrees with $f(x)$ over K. Thus, the proof is complete.

Ethics approval and consent to participate. Not applicable.

Consent for publication. Not applicable.

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

Author contributions. The authors contributed equally to the correctness of this paper.

Funding. This work was partially supported by Grant-in-Aid for Scientific Research (C) (23K03156), the Japan Society for the Promotion of Science (Sawano).

Acknowledgment. This work was partly supported by MEXT Promotion of Distinctive Joint Research Center Program JPMXP0723833165.

References

[1] D. Beniaguev, I. Segev, and M. London, *Single cortical neurons as deep artificial neural networks*, Neuron, **109(17)** (2021), 2727–2739. e2723.

[2] T. M. Cover, *Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition*, IEEE transactions on electronic computers(3) (1965), 326–334.

[3] K. Funahashi, *On the approximate realization of continuous mappings by neural networks*, Neural Networks **2** (1989), 183–192.

[4] A. Gidon, T. A. Zolnik, P. Fidzinski, F. Bolduan, A. Papoutsi, P. Poirazi, P. and M. E. Larkum, *Dendritic action potentials and computation in human layer 2/3 cortical neurons*, Science, **367(6473)** (2020), 83–87.

[5] N. Hatano, M. Ikeda, I. Ishikawa and Y. Sawano, *A Global Universality of Two-Layer Neural Networks with ReLU Activations*, Journal of Function Spaces, vol. 2021, Article ID 6637220, 3 pages, **2021**. https://doi.org/10.1155/2021/6637220

[6] N. Hatano, M. Ikeda, I. Ishikawa and Y. Sawano, *Global universality of the two-layer neural network with the k-rectified linear unit*, Journal of Function Spaces, vol. 2024, Article ID 3262798, 6 pages, **2024**. https://doi.org/10.1155/2024/3262798

[7] A. Krizhevsky, I. Sutskever and G. E. Hinton, *Imagenet classification with deep convolutional neural networks*, Advances in neural information processing systems, **25** (2012).

[8] W. S. McCulloch, and W. Pitts, *A logical calculus of the ideas immanent in nervous activity*. Bulletin of Mathematical Biology, **52** (1990), 99–115.

[9] M. Minsky and S. Papert, *Perceptrons: An introduction to computational geometry*, Cambridge tiass., HIT, **479**(480) (1969), 104.

[10] A. B. Novikoff, *On convergence proofs on perceptrons*, Paper presented at the Proceedings of the Symposium on the Mathematical Theory of Automata (1962).

[11] F. Rosenblatt, *The perceptron: a probabilistic model for information storage and organization in the brain*, Psychological Review, **65(6)** (1958), 386.

[12] W. Rudin, *Real and Complex Analysis* (Third Edition), McGraw-Hill, New York, 1987.

[13] D. E. Rumelhart, D. E. Hinton and G. E. R. J. Williams, *Learning representations by back-propagating errors*, Nature, **323(6088)** (1986), 533–536.
[14] T. J. Sejnowski, *The unreasonable effectiveness of deep learning in artificial intelligence*, Proceedings of the National Academy of Sciences, **117(48)** (2020), 30033–30038.

AUTHOR 1
TOKYO CITY UNIVERSITY
FACULTY OF LIBERAL ARTS AND SCIENCES
1-28-1, TAMADUTSUMI SETAGAYA-KU TOKYO 158-8557
Email address: izuki@tcu.ac.jp

AUTHOR 2
OTEMON GAKUIN UNIVERSITY
2-1-15 NISHIAI IBARAKI, OSAKA 567-8502
Email address: taka.noi.hiro@gmail.com

AUTHOR 3
CHUO UNIVERSITY
DEPARTMENT OF MATHEMATICS, GRADUATE SCHOOL OF SCIENCE AND ENGINEERING
1-13-27, KASUGA BUNKYO-KU TOKYO 112-8551
ORCID: 0000-0003-2844-8053
Email address: yoshihiro-sawano@celery.ocn.ne.jp

AUTHOR 4
TOKYO CITY UNIVERSITY
FACULTY OF INFORMATION TECHNOLOGY
1-28-1, TAMADUTSUMI SETAGAYA-KU TOKYO 158-8557
Email address: htanaka@tcu.ac.jp