Neural Networks, Vol. 4, pp. 385–394, 1991
Printed in the USA. All rights reserved.

0893-6080/91 $3.00 + .00

ORIGINAL CONTRIBUTION

Representation of Functions by Superpositions of a Step or Sigmoid Function and Their Applications to Neural Network Theory

YOSHIFUSA ITO

Nagoya University College of Medical Technology, Nagoya, Japan

(Received 16 November 1989; revised and accepted 26 October 1990)

Abstract—*The starting point of this article is the inversion formula of the Radon transform; the article aims to contribute to the theory of three-layered neural networks. Let H be the Heaviside function. Then, for any function* $f \in \mathcal{S}(\mathbf{R}^n)$, *there is a function* g_f *such that f can be represented on* $\mathbf{R}^n$ *by an integral* $\int H(x \cdot \omega - t) g_f(t, \omega)\, dt\, d\mu(\omega)$, *where* μ *is the uniform measure on the unit sphere* $\mathbf{S}^{n-1}$, $t \in \mathbf{R}$ *and* $\omega \in \mathbf{S}^{n-1}$. *Furthermore, f can be approximated uniformly arbitrarily well on the whole space* $\mathbf{R}^n$ *by a finite sum of the form* $\sum_k a_k H(x \cdot \omega^{(k)} - t^{(k)})$. *Let* H_σ *be a sigmoid function on* $\mathbf{R}$ *defined by* $H_\sigma(t) = \int H(t - x \cdot \omega)\, d\sigma(x)$, *where* σ *is a spherically symmetric probability measure. Suppose that* σ *satisfies a few further conditions. Then, for any* $f \in \mathcal{S}(\mathbf{R}^n)$, *there is a function* $g_{f,\sigma}$ *such that f can be written* $\int H_\sigma(x \cdot \omega - t) g_{f,\sigma}(t, \omega)\, dt\, d\mu(\omega)$ *with the unscaled sigmoid function* H_σ *fixed beforehand. This expression can also be approximated uniformly arbitrarily well on* $\mathbf{R}^n$ *by a finite sum.*

Keywords—Three-layered neural network, Heaviside function, Sigmoid function, Radon transform, Inverse radon transform, Integral representation, Finite sum approximation.

1. INTRODUCTION

The problem of representation of a function in several variables by a neural network has been studied by many authors. This article is concerned with the theory of three-layered neural networks. Hecht-Nielsen (1987) pointed out a possible connection between Kolmogorov's celebrated result (1957) and the present problem. Wieland and Leighton (1987) have dealt with networks consisting of one or two hidden layers, where the capability of networks of both threshold and sigmoid units are analysed. Irie and Miyake (1988) have obtained an integral representation formula with an integrable kernel fixed beforehand. This representation formula is the kind which would be realized by a three-layered neural network if infinitely many units could be used. In 1989, several papers related to the present topics appeared; Carroll and Dickinson; Cybenko; Funahashi; and Hornik, Stinchcombe, and White (1989). They all have claimed that a three-layered neural network with sigmoid units on the hidden layer can approximate continuous or other kinds of functions defined on compact sets. However, their methods are different. Carroll and Dickinson used the inverse Radon transform, which consists of two successive operations. They first approximated the integration over the surface of the unit sphere (the second operation in the inversion transform; see Section 2.2 in this article) by a finite sum. Next, they approximated the respective terms in the sum by linear combinations of the scalings of a sigmoid function. Cybenko's method is a handsome combination of the Hahn–Banach theorem and the Riesz representation theorem. His proof is existential. Funahashi approximated Irie and Miyake's integral representation by a finite sum, using a kernel which can be expressed as a difference of two sigmoid functions. Hornik et al. applied the Stone–Weierstrass theorem using trigonometric functions, where their approximations were not only in the uniform topology on a compact set but also in the ρ_μ-topology. However, the latter can be attained if the uniform approximation can be

Dedicated to Professor Takeyuki Hida in honour of his 60th birthday.

The author wishes to express his deep gratitude to Professors I. Kubo, J. Gani, and the referees for their helpful discussions, careful reading of the manuscript, and useful suggestions for revising the text.

Requests for reprints should be addressed to Yoshifusa Ito, Nagoya University College of Medical Technology, 1-1-20 Daiko Minami, Higashi-ku, Nagoya 461, Japan.

achieved on an arbitrary compact set. Though their methods are thus different mutually, common features are observed among them: (a) they all have proved that their approximation formulae hold on compact sets (though the ρ_μ-topology is slightly different); and (b) scaled a sigmoid (or other) function.

We carry two policies throughout (except in the Section 5): (a) uniform approximation on the whole space $\mathbf{R}^n$; and (b) the use of a step or sigmoid function without scaling. It is obvious that every continuous function defined on $\mathbf{R}^n$ cannot be approximated uniformly on the whole space by a finite sum of sigmoid functions. Hence, a certain restriction on a function to be approximated is unavoidable. We treat rapidly decreasing continuous functions in this article. However, this restriction can be weakened to some extent as is remarked and illustrated by examples. If a sigmoid function can be used without scaling, the optimal neural connection weights as a vector is on the surface of the unit sphere. This might be occasionally convenient, though it could generally require more units. If a sigmoid function satisfies a few conditions, we can express any rapidly decreasing function as an integral over the sigmoid function without scaling.

Three exact integral representations of rapidly decreasing C^∞-functions are obtained in Section 3. They are respectively integrations over an unscaled step or sigmoid function. In Section 4, each of these integral representations are approximated by a finite sum. Thus, three linear combinations of unscaled shifted rotations of a step or sigmoid function are obtained. Two of them can approximate uniformly any rapidly decreasing continuous function on the whole space $\mathbf{R}^n$. Because a step or sigmoid function is used in this article, small deviations of the value caused by approximation could spread out far away and pile up at a distance. It is essential in Section 4 to prove that this pile-up can be avoided. When our policies are weakened or removed, many other forms of approximation formulae can be obtained as corollaries to our results. We have avoided mentioning all of them, but two typical examples are illustrated in Section 5 in order to demonstrate how to derive other approximation formulae from ours. The proofs described in this article are constructive. Hence, most of the results can be simulated by computer.

The choice of the uniform approximation, equivalently the supremum norm, is a necessity in this article. A norm defined by integration over $\mathbf{R}^n$ such as $L^p(\mathbf{R}^n)$-norm is difficult to be managed in our theory because a small change of the value caused by approximation can spread out widely. As a result, norms such as $L^p(\mathbf{R}^n)$-norm become divergent. Conversely, the supremum norm is satisfactorily useful in our theory as will be observed. Furthermore, it is convenient because it is stronger on a compact set than some other norms (see Example 5.2). Hence, the uniform topology is appropriate in this article.

Two theorems are used without proofs, for they are available in Helgason (1980) and Gel'fand, Graev, and Vilenkin (1966). Accordingly, rigorously treated in Section 3 is the space $\mathcal{S}(\mathbf{R}^n)$ of rapidly decreasing C^∞-functions defined on $\mathbf{R}^n$ as in the case of the corresponding lemmas in both monographs, though this space can be extended to that of less regular and less rapidly decreasing functions as is described above.

2. PRELIMINARIES

While summarizing the theory of the Radon transform that is the starting point of this article, it is also necessary to prepare several lemmas to be used in this article. Most of the notations are introduced in this section.

We denote by H the Heaviside function. Any shift of H is called a step function. A sigmoid function on $\mathbf{R}$ is that which is not a step function but monotonic increasing, and converges to 1 as $t \to +\infty$ and to 0 as $t \to -\infty$. We denote by $\mathbf{S}^{n-1}$ the unit sphere in the n-dimensional Euclidean space $\mathbf{R}^n$. Any unit vector ω can be regarded as an element of $\mathbf{S}^{n-1}$. A step function can be extended to $\mathbf{R}^n$ by $H(x \cdot \omega - t)$, where $\cdot$ stands for the inner product. A sigmoid function can also be extended to $\mathbf{R}^n$ in the same way. We call these extended functions by their original names, respectively. Denote by $\mathbf{P}_{t,\omega}$ the hyperplane which is perpendicular to ω and has the point $t\omega \in \mathbf{R}^n$ on it. We then have that $\mathbf{P}_{t,\omega} = \{x \in \mathbf{R}^n | x \cdot \omega = t\}$. Also denote by $m_{t,\omega}$ the uniform measure on $\mathbf{P}_{t,\omega}$ with density 1. The Radon transform of a function f is defined by

$$\check{f}(t, \omega) = \int f(x)\, dm_{t,\omega}(x). \tag{2.1}$$

If f belongs to the space $L^1(\mathbf{R}^n)$ of integrable functions with respect to the Lebesgue measure, its Radon transform is defined for almost every $t \in \mathbf{R}$ for any $\omega \in \mathbf{S}^{n-1}$. If f is integrable with respect to $m_{t,\omega}$ of each hyperplane, $\check{f}$ is defined everywhere on $\mathbf{R} \times \mathbf{S}^{n-1}$.

Let us call a function u on $\mathbf{R} \times \mathbf{S}^{n-1}$ symmetric if $u(t, \omega) = u(-t, -\omega)$, and antisymmetric if $u(t, \omega) = -u(-t, -\omega)$. Then, the Radon transform $\check{f}$ is symmetric because $\mathbf{P}_{t,\omega} = \mathbf{P}_{-t,-\omega}$. We call a function f on $\mathbf{R} \times \mathbf{S}^{n-1}$ homogeneous if

$$\int \varphi(t, \omega) t^k\, dt$$

is a homogeneous polynomial of the kth degree in the components $\omega_1, \ldots, \omega_n$ of ω.

We call a function f defined on $\mathbf{R}^n$ rapidly decreasing if

$$\lim_{|x| \to \infty} |x_1^{k_1} \cdots x_n^{k_n} f(x)| = 0$$

for any nonnegative k_i's. Similarly, a function φ defined on $\mathbf{R} \times \mathbf{S}^{n-1}$ is rapidly decreasing if

$$\lim_{t \to \infty} |t^k \varphi(t, \omega)| = 0$$

for any $k \geq 0$. Note that rapidly decreasing functions are neither differentiable nor continuous in this article unless otherwise stated. We use a differential operator ∂_ω in the direction ω for a function on $\mathbf{R}^n$: that is, $\partial_\omega = \omega_1 \partial_{x_1} + \cdots + \omega_n \partial_{x_n}$, where $\partial_{x_k} = \partial / \partial_{x_k}$. For a function φ defined on $\mathbf{R} \times \mathbf{S}^{n-1}$, we use differential operators along the great circles on $\mathbf{S}^{n-1}$, besides $\partial_t = \partial / \partial t$:

$$\partial_\alpha \varphi(t, \omega) = \lim_{s \to 0} \frac{\varphi(t, \omega + s\alpha) - \varphi(t, \omega)}{s},$$

where α is a directed great circle starting from ω and $\omega + s\alpha$ is the endpoint of the initial section of α with length s. When the differentiation is in the radial direction we use a character such as s or t for the suffix, and when it is along a great circle we use the first couple of the Greek characters.

An infinitely continuously differentiable function is called a C^∞-function. The spaces $\mathcal{S}(\mathbf{R}^n)$ and $\mathcal{S}(\mathbf{R} \times \mathbf{S}^{n-1})$ are defined in Helgason (1980). They are the Schwartz spaces which consist of rapidly decreasing C^∞-functions on the respective spaces. We denote by $\mathcal{S}_{SH}(\mathbf{R} \times \mathbf{S}^{n-1})$ a subspace consisting of symmetric homogeneous elements of $\mathcal{S}(\mathbf{R} \times \mathbf{S}^{n-1})$. In Helgason (1980) the theorem below is proved.

THEOREM A. The Radon transform is a linear one-to-one mapping of $\mathcal{S}(\mathbf{R}^n)$ onto $\mathcal{S}_{SH}(\mathbf{R} \times \mathbf{S}^{n-1})$.

We denote by μ the uniform measure with density 1 on the unit sphere $\mathbf{S}^{n-1}$ (i.e. $d\mu$ is the surface element). The operator $\square_t$ is defined by

$$\square_t \varphi(t) = \partial_t^2 \varphi(t).$$

The definition of the fractional power of $-\square_t$ and that of the principal value (v.p.) are described in Appendix. The theorem below is described in both Gel'fand et al. (1966) and Helgason (1980).

THEOREM B. Suppose that $f \in \mathcal{S}(\mathbf{R}^n)$. The inversion formula for the Radon transform of f is

$$f(x) = \frac{1}{2^n \pi^{n-1}} \int [(-\square_t)^{(n-1)/2} \check{f}(t, \omega)]_{t = x \cdot \omega} \, d\mu(\omega). \quad (2.2)$$

This inversion formula can be written concretely

$$f(x) = \begin{cases} c_n \int [\partial_t^{n-1} \check{f}(t, \omega)]_{t = x \cdot \omega} \, d\mu(\omega) & \text{for odd } n, \\ d_n \text{p.v.} \int \left[\int_{-\infty}^{\infty} \check{f}(t, \omega)(t - x \cdot \omega)^{-n} \, dt \right] d\mu(\omega) & \text{for even } n, \end{cases} \quad (2.3)$$

where

$$c_n = \frac{(-1)^{(n-1)/2}}{2^n \pi^{n-1}} \quad \text{and} \quad d_n = \frac{(-1)^{n/2}(n-1)!}{2^n \pi^n}. \quad (2.4)$$

For convenience, we introduce an operator L with respect to the variable t defined by

$$L\varphi(t) = \frac{1}{2^n \pi^{n-1}} (-\square_t)^{(n-1)/2} \varphi(t). \quad (2.5)$$

Then,

$$f(x) = \int [L\check{f}(t, \omega)]_{t = x \cdot \omega} \, d\mu(\omega). \quad (2.6)$$

Now let us prepare several lemmas. We denote by $C_0^\infty(\mathbf{R} \times \mathbf{S}^{n-1})$ a space of functions defined on $\mathbf{R} \times \mathbf{S}^{n-1}$ which are infinitely continuously differentiable with derivatives converging to 0 as $t \to \pm\infty$. Lemmas 2.1 and 2.2 are proven in Appendix.

LEMMA 2.1. Let n be even and $\varphi \in \mathcal{S}(\mathbf{R} \times \mathbf{S}^{n-1})$. Then:

1. The principal value

$$I\varphi(t, \omega) = \text{p.v.} \int_{-\infty}^{\infty} p^{-n} \varphi(t + p, \omega) \, dp \quad (2.7)$$

is well defined for each (t, ω).
2. $I\varphi(t, \omega) \in C_0^\infty(\mathbf{R} \times \mathbf{S}^{n-1})$.
3. The operations I, ∂_t and ∂_α are mutually commutative.
4. For all the differential operators of the form $\partial_t^k \partial_\alpha \cdots \partial_\beta$, there exists a positive constant $M_{k,\alpha,\dots,\beta}$ such that

$$|\partial_t^k \partial_\alpha \cdots \partial_\beta (I\varphi)(t, \omega)| < \frac{M_{k,\alpha,\dots,\beta}}{|t|^{n+k} + 1}. \quad (2.8)$$

We introduce an operator L_ω on $\mathbf{R}^n$ defined by

$$L_\omega \psi(x) = [L\psi(x + t\omega)]_{t=0}. \quad (2.9)$$

More concretely, we have that, for odd n,

$$L_\omega \psi(x) = c_n \partial_\omega^{n-1} \psi(x)$$

and, for even n,

$$L_\omega \psi(x) = d_n \text{p.v.} \int p^{-n} \psi(x + p\omega) \, dp.$$

LEMMA 2.2. Let $\psi \in \mathcal{S}(\mathbf{R}^n)$. Then, we have that

$$\partial_t \check{\psi}(t, \omega) = (\partial_\omega \psi)\check{}(t, \omega), \quad (2.10)$$

$$L\check{\psi}(t, \omega) = (L_\omega \psi)\check{}(t, \omega). \quad (2.11)$$

The lemma above will be used in Proposition 3.2 to derive the function g_f defined by eqn (3.2) by the alternative way.

LEMMA 2.3. Suppose that a function f is defined on $\mathbf{R}^n$ and continuously differentiable. If f and its derivatives $\partial_{x_k} f$, $k = 1, \cdots, n$, are all integrable with respect to the Lebesgue measure, then we have that

$$\int f(x)\, dm_{t,\omega}(x) = \int H(t - x \cdot \omega)\partial_\omega f(x)\, dx \quad (2.12)$$

for all (t, ω).

Proof. Without loss of generality, we may suppose that $\omega = (1, 0, \cdots, 0)$. Since f and $\partial_{x_1} f$ are integrable,

$$\lim_{x_1 \to -\infty} f(x_1, \cdots, x_n) = 0$$

for almost all $(x_2, \cdots, x_n)$. Hence, we have that

$$\int \cdots \int f(t, x_2, \cdots, x_n)\, dx_2 \cdots dx_n = \int \cdots \int dx_2 \cdots dx_n \int_{-\infty}^{t} \partial_{x_1} f(x_1, \cdots, x_n)\, dx_1. \quad (2.13)$$

By Fubini's theorem, the right hand side of eqn (2.11) is equal to

$$\int \cdots \int H(t - x_1)\partial_{x_1} f(x_1, \cdots, x_n)\, dx_1\, dx_2 \cdots dx_n.$$

Hence, we obtain the lemma. □

Let us denote by Σ a set of measures defined on $\mathbf{R}^n$ which are

$\Sigma 0$. Nonnegative, spherically symmetric and with total mass 1.

Then, if $\sigma \in \Sigma$ is not the delta function δ_0 at the origin, a convolution

$$H_\sigma(t) = \int H(t - x \cdot \omega)\, d\sigma(x) \quad (2.14)$$

is a sigmoid function. Obviously, H_σ does not depend on ω. If the Fourier transform $\mathfrak{F}\sigma$ of $\sigma \in \Sigma$ satisfies the following conditions, we call $\sigma \in \Sigma_1$:

$\Sigma 1$. $\mathfrak{F}\sigma \in C^\infty(\mathbf{R}^n)$ (infinitely continuously differentiable on $\mathbf{R}^n$).
$\Sigma 2$. There are a positive integer N and a positive constant a for which

$$\mathfrak{F}\sigma(y) > \frac{a}{|y|^N + 1} \quad \text{for all} \quad y \in \mathbf{R}^n.$$

LEMMA 2.4. Suppose that $\sigma \in \Sigma_1$. Then, for any arbitrary $f \in \mathcal{S}(\mathbf{R}^n)$, there is a function $v \in \mathcal{S}(\mathbf{R}^n)$ such that $f = v * \sigma$.

Proof. Since $\mathfrak{F}\sigma \in C^\infty(\mathbf{R}^n)$, any moment of σ

$$\int |x_1|^{\alpha_1} \cdots |x_n|^{\alpha_n}\, d\sigma(x), \quad \alpha_1 \geq 0, \cdots, \alpha_n \geq 0,$$

is finite; namely, any partial derivative of $\mathfrak{F}\sigma$ is bounded. Hence, $\mathfrak{F}f/\mathfrak{F}\sigma \in \mathcal{S}(\mathbf{R}^n)$. Set $v = \mathfrak{F}^{-1}(\mathfrak{F}f/\mathfrak{F}\sigma)$. Then, v is the function we look for. □

The delta function at the origin δ_0 belongs to Σ_1. If $\varphi \in \mathcal{S}(\mathbf{R}^n)$ is spherically symmetric, nonnegative and $\int \varphi(x)\, dx = 1$, then $\alpha\delta_0 + \beta\varphi\, dx (\alpha + \beta = 1, \alpha > \beta \geq 0)$ belongs to Σ_1. The sigmoid function defined by eqn (2.14), using this measure, is not continuous. However, there are many absolutely continuous measures in Σ_1. Set $\psi(x) = Ae^{-a|x|}$, $a > 0$, $A^{-1} = \int e^{-a|x|}\, dx$. Then, the Fourier transform $\mathfrak{F}\psi(y)$ is $(1 + |y|^2/a^2)^{-(n+1)/2}$. Furthermore, there are many such measures with compact support. Let $n = 1$ and set $\psi_1(x) = 3/2(1 - |x|)^2$ for $|x| \leq 1$ and $\psi_1(x) = 0$ otherwise. Then, $\mathfrak{F}\psi_1(y) = 6y^{-2}(1 - 1/y \sin y)$. For $n > 1$, $\psi_n(x)\, dx = \Pi_{k=1}^n \psi_1(x_k)\, dx$ satisfies the conditions $\Sigma 1$ and $\Sigma 2$. Hence, if we take the spherical symmetrization of $\psi_n\, dx$, then it is an element of Σ_1. The sigmoid function defined by an absolutely continuous measure σ is continuous.

3. THE INTEGRAL REPRESENTATION

Using the lemmas so far prepared, Theorem 3.1 (the first main theorem), Proposition 3.2, and Corollaries 3.3 and 3.4 are proven. The statement of the main theorem is that an integral over the step function with a weight g_f exactly represents a function $f \in \mathcal{S}(\mathbf{R}^n)$ on the whole space $\mathbf{R}^n$. The properties of the function g_f is important for obtaining an approximate representation formula in Section 4. It is shown in Proposition 3.2 that g_f can be obtained by the alternative way. As corollaries to Theorem 3.1, it will be proven that a sigmoid function can be used instead of the step function in the integral representation. In Corollary 3.3, the function to be approximated must be a convolution. We can easily remove this restriction by applying Lemma 2.4. Thus, obtained is Corollary 3.4.

THEOREM 3.1. Suppose that $f \in \mathcal{S}(\mathbf{R}^n)$. Then, f is represented as

$$f(x) = \iint H(x \cdot \omega - t) g_f(t, \omega)\, dt\, d\mu(\omega), \quad (3.1)$$

where g_f is a function defined by

$$g_f(t, \omega) = \partial_t Lf(t, \omega), \qquad (t, \omega) \in \mathbf{R} \times \mathbf{S}^{n-1}. \quad (3.2)$$

The function g_f satisfies the conditions below:

1. $g_f \in C_0^\infty(\mathbf{R} \times \mathbf{S}^{n-1})$.
2. For any $k \geq 0$ and any directed arcs $\alpha, \cdots, \beta$, there is a positive constant $M_{k,\alpha,\cdots,\beta}$ such that

$$|\partial_t^k \partial_\alpha \cdots \partial_\beta g_f(t, \omega)| < \frac{M_{k,\alpha,\cdots,\beta}}{|t|^{n+k+1} + 1} \quad \text{for all } (t, \omega).$$

3. $\int_{-\infty}^{\infty} g_f(t, \omega)\, dt = 0$ for all ω.
4. g_f is antisymmetric.

Proof. By Theorem A, $\check{f} \in \mathcal{S}(\mathbf{R} \times \mathbf{S}^{n-1})$. For odd n, $L\check{f} = c_n \partial_t^{n-1} \check{f}$ obviously belongs to $\mathcal{S}(\mathbf{R} \times \mathbf{S}^{n-1})$.

For even n, $L\check{f}(t, \omega) = d_n I\check{f}(t, \omega)$ is differentiable and both $L\check{f}(t, \omega)$ and $\partial_t L\check{f}(t, \omega)$ are integrable by Lemma 2.1. Hence, for both odd and even n,

$$[L\check{f}(t, \omega)]_{t=x\cdot\omega} = \int H(x\cdot \omega - t)\partial_t L\check{f}(t, \omega)\, dt \quad (3.3)$$

by Lemma 2.3. From one of the forms of the inversion formula (2.6) of the Radon transform and (3.3) we have that

$$f(x) = \iint H(x \cdot \omega - t)\partial_t L\check{f}(t, \omega)\, dt\, d\mu(\omega). \quad (3.4)$$

Hence, we obtain (3.1) with (3.2).

Next, let us confirm that items 1, 2, 3, and 4 hold. For odd n, $g_f(t, \omega) = c_n\partial_t^n \check{f}(t, \omega)$. Hence, 1, 2, and 3 are obvious. Since $\check{f}$ is symmetric, g_f is antisymmetric for odd n. For even n, $g_f(t, \omega) = d_n\partial_t I\check{f}(t, \omega)$. Hence, 1, 2, and 3 are obvious by Lemma 2.1. Since $I\check{f}$ (t, ω) is symmetric, g_f is antisymmetric for even n, too. Hence, the theorem is proven. □

Since there is the inversion formula (3.1), the mapping $f \in \mathcal{S}(\mathbf{R}^n) \to g_f$ is one-to-one.

PROPOSITION 3.2. The function g_f defined in Theorem 3.1 can also be obtained by

$$g_f(t, \omega) = \int L_\omega \partial_\omega f(x)\, dm_{t,\omega}(x). \quad (3.5)$$

Proof. By Lemma 2.2, we can easily prove that the right-hand side of (3.2) is equal to that of (3.5). □

Let us obtain an integral representation formula for a convolution $f_\sigma(x) = f*\sigma(x)$.

COROLLARY 3.3. Suppose that $\sigma \in \Sigma$. Then, for $f \in \mathcal{S}(\mathbf{R}^n)$, the convolution f_σ is represented as

$$f_\sigma(x) = \iint H_\sigma(x \cdot \omega - t)g_f(t, \omega)\, dt\, d\mu(\omega). \quad (3.6)$$

Proof. This equation is straightforward:

$$\begin{aligned} f_\sigma(x) &= \int f(y)\, d\sigma(x - y) \\ &= \int \left(\int H(y \cdot \omega - t)\, d\sigma(x - y)\right) g_f(t, \omega)\, dt\, d\mu(\omega) \\ &= \iint H_\sigma(x \cdot \omega - t)g_f(t, \omega)\, dt\, d\mu(\omega). \end{aligned}$$

Hence, we obtain the corollary. □

In other words, σ can be released from f_σ and synthesized with H. In Corollary 3.3, the sigmoid function H_σ is not scaled, but the function to be approximated must be a convolution. This restriction looks strong. Nevertheless, this corollary is useful. The counterpart of this corollary in Section 4 is closely related to the well-known approximate representation formula with a scalable sigmoid function (see Example 5.1). Moreover, the corollary below follows immediately from this corollary.

COROLLARY 3.4. Let $\sigma \in \Sigma_1$. Then, any function $f \in \mathcal{S}(\mathbf{R}^n)$ is represented as

$$f(x) = \iint H_\sigma(x \cdot \omega - t)g_{f,\sigma}(t, \omega)\, dt\, d\mu(\omega), \quad (3.7)$$

where $g_{f,\sigma}$ is the function which can be obtained when f is replaced by a function $\mathcal{F}^{-1}(\mathcal{F}f/\mathcal{F}\sigma)$ in eqn (3.2).

Proof. Note that Lemma 2.4 guarantees the existence of a function $v \in \mathcal{S}(\mathbf{R}^n)$ such that $f = v * \sigma$. Applying Corollary 3.3, we obtain

$$v * \sigma(x) = \iint H_\sigma(x \cdot \omega - t)g_v(t, \omega)\, dt\, d\mu(\omega). \quad (3.8)$$

Since $v \equiv \mathcal{F}^{-1}(\mathcal{F}f/\mathcal{F}\sigma)$, we obtain the corollary. □

Thus, it is proven that any function $f \in \mathcal{S}(\mathbf{R}^n)$ can be expressed as an integral of a sigmoid function which is fixed beforehand and cannot be scaled.

Remark 3.1. By checking the proofs of Theorem 3.1, Proposition 3.2, Corollaries 3.3 and 3.4, it is almost clear that they can be extended to less regular functions. If the function f is n (resp. $n + 1$) times continuously differentiable for odd (resp. even) n and the derivatives decrease sufficiently rapidly, then the assertions of the theorem, proposition, and corollaries are all true. Furthermore, they hold in a sense even for discontinuous functions (see Examples). It is also clear that, in any case where the function g_f is obtained, Theorem 3.1 and others hold. Hence, the function f can be less rapidly decreasing. In order to describe this fact rigorously, we need start with rewriting Theorems A, B and Lemmas giving long proofs. Though we avoid the details, examples below illustrate that less regular and less rapidly decreasing functions can be expressed as the integrations of the form (3.1).

The method for obtaining the function g_f is explicit. This fact may be of significance in designing an actual neural networks. Several simple examples of g_f are illustrated below. In Example 3.1, a less regular function f defined on $\mathbf{R}$ is treated. When f is discontinuous, g_f involves the delta function. In Example 3.2, it is shown that the function g_f can be obtained by applying either Theorem 3.1 or Proposition 3.2. Example 3.3 is similar to Example 3.2. In Example 3.4, a less regular function on $\mathbf{R}^3$ is treated, where both the delta function and its derivative are involved.

Example 3.1. Suppose φ to be a function defined on $\mathbf{R}$ and differentiable. Set $f(x) \equiv \chi_{[a,b]}(x)\varphi(x)$, where $\chi_{[a,b]}$ is the indicator function of the interval. Then,

$$\check{f}(t, \omega) = \chi_{[a,b]}(t\omega)\varphi(t\omega),$$

where $\omega = \pm 1$. Hence,

$$\begin{aligned} g_f(t, \omega) = \omega\{&\chi_{[a,b]}(t\omega)\varphi'(t\omega) \\ &+ \varphi(a)\delta(t\omega - a) - \varphi(b)\delta(t\omega - b)\}. \quad (3.9) \end{aligned}$$

It can be easily confirmed that the equations below hold.

$$\begin{aligned}\int H(x\cdot\omega - t)g_f(t,\omega)\,dt\,d\mu(\omega) &= \frac{1}{2}\left\{\int H(x-t)g(t,1)\,dt + \int H(-x-t)g(t,-1)\,dt\right\}\\ &= \frac{1}{2}\chi_{[a,b]}(x)\left\{\int_a^x \varphi'(t)\,dt + \varphi(a) - \int_b^{-x}\varphi'(-t)\,dt + \varphi(b)\right\}\\ &= \chi_{[a,b]}(x)\varphi(x) - \frac{1}{2}\{\varphi(a)\delta(x-a) + \varphi(b)\delta(x-b)\}.\end{aligned} \tag{3.10}$$

The value of eqn (3.10) does not coincide with that of f at $x = a, b$. Generally, this sort of exceptional points appear if a discontinuous function is treated. If the support of φ is contained in $[a, b]$, $g_f(t, \omega) = f'(t\omega)$ and the right-hand side of eqn (3.10) is exactly equal to $\varphi(x)$. In this case, if f is $n + 1$ times continuously differentiable, then g_f is n times continuously differentiable in $t \in (a, b)$. If φ is not truncated, it must be integrable but does not need to be rapidly decreasing. Thus, this simple example illustrates that a function to be approximated need neither be infinitely differentiable nor rapidly decreasing.

Example 3.2. Set

$$f(x) \equiv \exp\left[-\frac{a^2}{2}|x|^2\right], \qquad x \in \mathbf{R}^3. \tag{3.11}$$

First, let us apply Theorem 3.1. Then,

$$\check{f}(t) = \frac{2\pi}{a^2}\exp\left[-\frac{a^2}{2}t^2\right]. \tag{3.12}$$

Hence, we obtain

$$\begin{aligned} g_f(t,\omega) &= \partial_t L\check{f}(t) = -\frac{1}{8\pi^2}\partial_t^3\check{f}(t)\\ &= \frac{1}{4\pi}\{a^4t^3 - 3a^2t\}\exp\left[-\frac{a^2}{2}t^2\right].\end{aligned} \tag{3.13}$$

Next, let us apply Proposition 3.2. Then

$$\partial_\omega^3 f(x) = \{-a^6(x\cdot\omega)^3 + 3a^4(x\cdot\omega)\} \times \exp\left[-\frac{a^2}{2}|x|^2\right]. \tag{3.14}$$

From this equation, we have

$$\begin{aligned} g_f(t,\omega) &= \int (L_\omega\partial_\omega f)(x)\,dm_{t,\omega}(x)\\ &= \frac{1}{4\pi}\{a^4t^3 - 3a^2t\}\exp\left[-\frac{a^2}{2}t^2\right].\end{aligned} \tag{3.15}$$

Thus, we have obtained the same g_f in both cases. We can confirm that the function f can be recovered from the function g_f:

$$\begin{aligned}\int H(x\cdot\omega - t)g_f(t,\omega)\,dt\,d\mu(\omega) &= \frac{1}{4\pi}\int d\mu(\omega)\int_{-\infty}^{x\cdot\omega}\{a^4t^3 - 3a^2t\} \times \exp\left[-\frac{a^2}{2}t^2\right]dt\\ &= -\frac{1}{4\pi}\int\{(ax\cdot\omega)^2 - 1\} \times \exp\left[-\frac{a^2}{2}(x\cdot\omega)^2\right]d\mu(\omega)\\ &= \exp\left[-\frac{a^2}{2}|x|^2\right].\end{aligned} \tag{3.16}$$

Example 3.3. Set

$$f(x) \equiv |x|^2\exp\left[-\frac{a^2}{2}|x|^2\right], \qquad x \in \mathbf{R}^3. \tag{3.17}$$

Then, by either eqn (3.2) or eqn (3.5),

$$g_f(t,\omega) = \frac{1}{4\pi}\{a^4t^5 + 7a^2t^3 - 6t\}\exp\left[-\frac{a^2}{2}t^2\right]. \tag{3.18}$$

Hence,

$$f(x) = \frac{1}{4\pi}\iint\{a^4t^5 + 7a^2t^3 - 6t\} \times \exp\left[-\frac{a^2}{2}t^2\right]H(x\cdot\omega - t)\,dt\,d\mu(\omega). \tag{3.19}$$

Example 3.4. Suppose that f is a characteristic function of a sphere in $\mathbf{R}^3$ with centre at the origin and radius a:

$$f(x) = \begin{cases}1, & |x| \le a,\\ 0, & |x| > a.\end{cases} \tag{3.20}$$

Then

$$\check{f}(t) = \chi_{[-a,a]}(t)\pi(a^2 - t^2). \tag{3.21}$$

Hence,

$$\begin{aligned} g_f(t,\omega) &= c_3\partial_t^3\check{f}(t)\\ &= \frac{1}{4\pi}\{\delta(t+a) - \delta(t-a) - a\delta'(t-a) - a\delta'(t+a)\}.\end{aligned} \tag{3.22}$$

Hence,

$$\begin{aligned}\iint H(x\cdot\omega - t)g_f(t,\omega)\,dt\,d\mu(\omega) &= \frac{1}{4\pi}\int\{\chi_{[-a,a]}(x\cdot\omega) - a\delta(x\cdot\omega - a) - a\delta(x\cdot\omega + a)\}\,d\mu(\omega).\end{aligned} \tag{3.23}$$

For $|x| \le a$, this quantity is equal to 1. For $|x| > a$, let us divide the integral on the right-hand side of eqn (3.23):

$$\int_{x\cdot\omega<-a} + \int_{|x\cdot\omega|\le a} + \int_{x\cdot\omega>a}.$$

The first and third integrals are equal to $4\pi a/|x|$ and 0, respectively. The second integral is equal to

$$-4\pi \int_0^{\pi/2} \sin\theta\,\delta(|x|\cos\theta - a)\,d\theta = -\frac{4\pi a}{|x|}.$$

Hence, eqn (3.23) is equal to 0 for $|x| > 0$. There is a reason that we may define the value of

$$2\pi a \int_0^{\pi/2} \delta(x \cdot \omega - a)\,d\mu(\omega)$$

to be 1/2 for $|x| = a$. If we use this definition, the value of eqn (3.23) is 1/2 for $|x| = a$.

If the dimension is higher than 1 and the function $f(x)$ is not spherically symmetric, then the calculation for obtaining g_f is too lengthy to be given in full in this article, which is intended to present the theory.

4. APPROXIMATION OF THE REPRESENTATIONS

Let us approximate the integral representations in Theorem 3.1 and Corollaries 3.3 and 3.4 by finite sums of a step or sigmoid function, respectively. In the present theory, replacement of the respective integrals in eqns (3.1), (3.6), and (3.7) by finite sums must be performed very carefully. A deviation of the value of the representation caused by such replacement can spread out widely and accumulate somewhere, because a step or sigmoid functions is used. Hence, even if a good approximation is attained on a compact set, it could be violated at a distance because of accumulation of the deviations. In order to guarantee that our approximation holds on the whole space $\mathbf{R}^n$, we have to prove that such accumulation does not take place anywhere. This is an essential difference which exists between an approximation on $\mathbf{R}^n$ and that on a compact set. The proof of the theorem below includes the method of avoiding accumulation of the deviations caused by approximation.

Let $\mathbf{Q}_x$ be a sphere having the line segment Ox between the origin O and a point $x \in \mathbf{R}^n$ as one of its diameters. Note that the point $(x \cdot \omega)\omega$ is on the sphere $\mathbf{Q}_x$. However, since $(x \cdot \omega)\omega = (x \cdot (-\omega))(-\omega)$, the set $\{(x \cdot \omega)\omega | \omega \in \mathbf{S}^{n-1}\}$ covers the sphere $\mathbf{Q}_x$ twice. We denote by μ_{Q_x} the uniform measures with density 1 on $\mathbf{Q}_x$. Let u be a symmetric function defined on $\mathbf{R} \times \mathbf{S}^{n-1}$. Then, the integration of $u(x \cdot \omega, \omega)$ over the unit sphere $\mathbf{S}^{n-1}$ is converted into that over $\mathbf{Q}_x$ by a correspondence $\omega \rightarrow (x \cdot \omega)\omega$.

Let us approximate a rapidly decreasing continuous function (not necessarily differentiable) defined on $\mathbf{R}^n$ by a finite sum of a step or sigmoid function.

THEOREM 4.1. Let f be a rapidly decreasing continuous function defined on $\mathbf{R}^n$. Then, for an arbitrary positive number ε, there are finite sets of numbers $\{a_k\}_{k=1}^N$ and $\{t^{(k)}\}_{k=1}^N$, and a finite set of unit vectors $\{\omega^{(k)}\}_{k=1}^N$ such that a finite sum

$$f_\varepsilon(x) = \sum_{k=1}^{N} a_k H(x \cdot \omega^{(k)} - t^{(k)}) \tag{4.1}$$

satisfies

$$|f(x) - f_\varepsilon(x)| < \varepsilon \qquad \text{for all} \quad x \in \mathbf{R}^n. \tag{4.2}$$

Proof. For an arbitrary $\varepsilon > 0$, there is a function $\bar{f} \in \mathcal{S}(\mathbf{R}^n)$ such that $|f(x) - \bar{f}(x)| < \varepsilon/3$. In order to approximate $\bar{f}$, let us introduce partitions of $\mathbf{R}$ and $\mathbf{S}^{n-1}$. We denote by $\Delta = \{\Delta_i\}_{i=1}^I$ a partition of $\mathbf{R}$, where $\Delta_i = [\tau_{i-1}, \tau_i]$, $\tau_{i-1} < \tau_i$, and by $\Theta = \{\Theta_j\}_{j=1}^J$ a partition of $\mathbf{S}^{n-1}$. In these partitions, I and J are finite positive integers. Set $E_{ij} = \{t\omega \in \mathbf{R}^n \mid t \in \Delta_i, \omega \in \Theta_j\}$ and define a set Λ_x by

$$\Lambda_x = \{(i, j) | E_{ij} \cap \mathbf{Q}_x \neq \emptyset\}. \tag{4.3}$$

Since $g_{\bar{f}}$ satisfies condition 2 in Theorem 3.1, $g_{\bar{f}}(t, \omega) \in L^1(dt\,d\mu)$. Hence, there are finite partitions Δ and Θ for which

$$\sum_{(i,j) \in \Lambda_x} \iint_{\Delta_i \times \Theta_j} |g_{\bar{f}}(t, \omega)|\,dt\,d\mu(\omega) < \frac{\varepsilon}{3} \qquad \text{for all} \quad x. \tag{4.4}$$

Let $t^{(i)}$ be an arbitrary point of the interval Δ_i and $\omega^{(j)}$ be an arbitrary unit vector in Θ_j. Set

$$a_{ij} = \iint_{\Delta_i \times \Theta_j} g_{\bar{f}}(t, \omega)\,dt\,d\mu(\omega). \tag{4.5}$$

Thus, we have obtained three sets $\{a_{ij}\}$, $\{t^{(i)}\}$ and $\{\omega^{(j)}\}$. We can prove that, for these sets,

$$\bar{f}_1(x) = \sum_{i,j} a_{ij} H(x \cdot \omega^{(j)} - t^{(i)}) \tag{4.6}$$

satisfies

$$|\bar{f}(x) - \bar{f}_1(x)| < \frac{2}{3}\varepsilon \qquad \text{for all} \quad x \in \mathbf{R}^n.$$

In fact, suppose that $E_{ij} \cap \mathbf{Q}_x = \emptyset$. Then,

$$\iint_{\Delta_i \times \Theta_j} H(x \cdot \omega - t) g_{\bar{f}}(t, \omega)\,dt\,d\mu(\omega) = a_{ij} H(x \cdot \omega_j - t_i), \tag{4.7}$$

because $H(x \cdot \omega - t) = H(x \cdot \omega_j - t_i) = 0$ or 1 on such a set $\Delta_i \times \Theta_j$ and a_{ij} is defined by eqn (4.5). Hence, we have that

$$\begin{aligned} &\left| \iint H(x \cdot \omega - t) g_{\bar{f}}(t, \omega)\,dt\,d\mu(\omega) - \sum_{i,j} a_{ij} H(x \cdot \omega^{(j)} - t^{(i)}) \right| \\ &\le \sum_{(i,j) \in \Lambda_x} \left| \iint_{\Delta_i \times \Theta_j} H(x \cdot \omega - t) g_{\bar{f}}(t, \omega)\,dt\,d\mu(\omega) - a_{ij} H(x \cdot \omega^{(j)} - t^{(i)}) \right| \\ &\le 2 \sum_{(i,j) \in \Lambda_x} \iint_{\Delta_i \times \Theta_j} |g_{\bar{f}}(t, \omega)|\,dt\,d\mu(\omega) < \frac{2}{3}\varepsilon \end{aligned} \tag{4.8}$$

Hence, eqn (4.6) is smaller than ε. By renumbering the suffixes and the superfixes in eqn (4.6), we obtain eqn (4.1). This concludes the proof. □

Remark 4.1. It is obvious that we can take the partitions in a way that the equalities $\Delta_{-i} = -\Delta_i$, $-I < i < I$ and $\Theta_{-j} = -\Theta_j$, $-J < j < J$ hold; that is, the partitions can be symmetric. Such partitions must be convenient because the function $g_{\bar{\gamma}}$ is asymmetric.

The corollary below is straightforward from Theorem 4.1. We use a sigmoid function H_σ without scaling.

COROLLARY 4.2. Let $\sigma \in \Sigma$ and suppose that f is a rapidly decreasing continuous function defined on $\mathbf{R}^n$. Then, for an arbitrary positive number ε, there are finite sets of numbers $\{a_k\}_{k=1}^N$ and $\{t^{(k)}\}_{k=1}^N$, and a finite set of unit vectors $\{\omega^{(k)}\}_{k=1}^N$, such that a finite sum

$$f_{\sigma,\varepsilon}(x) = \sum_{k=1}^{N} a_k H_\sigma(x \cdot \omega^{(k)} - t^{(k)}) \tag{4.9}$$

satisfies

$$|f_\sigma(x) - f_{\sigma,\varepsilon}(x)| < \varepsilon \quad \text{for all} \quad x \in \mathbf{R}^n. \tag{4.10}$$

Proof. Let f_ε be the approximation obtained in Theorem 4.1. Then, from eqn (4.2),

$$|f_\sigma(x) - f_\varepsilon * \sigma(x)| < \varepsilon \quad \text{for all} \quad x \in \mathbf{R}^n, \tag{4.11}$$

Rewrite the convolution as

$$\begin{aligned} f_\varepsilon * \sigma(x) &= \sum_{k=1}^{N} a_k \int H(y \cdot \omega^{(k)} - t^{(k)})\, d\sigma(x - y) \\ &= \sum_{k=1}^{N} a_k H_\sigma(x \cdot \omega^{(k)} - t^{(k)}). \end{aligned} \tag{4.12}$$

Thence, by setting $f_{\sigma,\varepsilon} \equiv f_\varepsilon * \sigma$, we obtain the corollary. □

It might be advantageous that the sets $\{a_k\}$, $\{t^{(k)}\}$, and $\{\omega^{(k)}\}$ in Theorem 4.1 and Corollary 4.2 do not depend on σ. We can choose a convenient σ even after ε and these three sets are decided.

Because a continuous function defined on a compact set can be uniformly approximated by a rapidly decreasing function, this result includes the well-known approximation theory with a scalable sigmoid function. Details will be described in Section 5.

The corollary below is immediate from Corollary 4.2.

COROLLARY 4.3. Let $\sigma \in \Sigma_1$ and suppose that f is a rapidly decreasing continuous function defined on $\mathbf{R}^n$. Then, for an arbitrary positive number ε, there are finite sets of numbers $\{a_k\}_{k=1}^N$ and $\{t^{(k)}\}_{k=1}^N$, and a finite set of unit fectors $\{\omega^{(k)}\}_{k=1}^N$ such that a finite sum

$$f_\varepsilon(x) = \sum_{k=1}^{N} a_k H_\sigma(x \cdot \omega^{(k)} - t^{(k)}) \tag{4.13}$$

satisfies

$$|f(x) - f_\varepsilon(x)| < \varepsilon \quad \text{for all} \quad x \in \mathbf{R}^n. \tag{4.14}$$

Proof. Similarly to Corollary 3.4, we can obtain this result. □

Thus, it is proven that a finite sum of unscaled shifted rotations of a sigmoid function can approximate any rapidly decreasing continuous (not necessarily differentiable) function on the whole space $\mathbf{R}^n$, if the sigmoid function satisfies a few conditions.

5. OTHER EXAMPLES

This section is annexed to show that we can derive several other results if our policies are weakened. Avoiding to mention each of them tediously, we describe here two typical examples. In Example 5.1, the policy of use of a sigmoid function without scaling is abandoned, but the uniform approximation holds on the whole space $\mathbf{R}^n$. In Example 5.2, the policy of approximation on the whole space is abandoned and a norm weaker on a compact set is adopted, but the sigmoid function is not scaled.

Let $\sigma \in \Sigma$. For $h > 0$, we define a scaling of σ by

$$\int f(x)\, d\sigma_h(x) = \int f(hx)\, d\sigma(x). \tag{5.1}$$

This scaling may be written symbolically $h^{-n}\sigma(x/h)$. The sigmoid function H_σ is scaled if the measure σ is scaled. The result below is derived from Corollary 4.2 under the condition that H_σ can be scaled.

Example 5.1. Suppose that a measure $\sigma \in \Sigma$ is absolutely continuous and the density of σ is an element of $\mathcal{S}(\mathbf{R}^n)$. Then, any rapidly decreasing continuous function f can be approximated uniformly arbitrarily well on $\mathbf{R}^n$ by a finite sum

$$f_\varepsilon(x) = \sum_{k=1}^{N} a_k H_{\sigma_h}(x \cdot \omega^{(k)} - t^{(k)}), \qquad h > 0. \tag{5.2}$$

In fact, for any $\varepsilon > 0$, there is a scaling of σ such that

$$|f(x) - f_{\sigma_h}(x)| < \frac{\varepsilon}{2} \quad \text{for all} \quad x \in \mathbf{R}^n.$$

Furthermore, by Corollary 4.2, there is a finite sum of the form (5.2) which satisfies

$$|f_{\sigma_h}(x) - f_\varepsilon(x)| < \frac{\varepsilon}{2} \quad \text{for all} \quad x \in \mathbf{R}^n.$$

Hence, we obtain the result. □

Set $\mathbf{W}^{(k)} = \omega^{(k)}/h$ and $T^{(k)} = t^{(k)}/h$, $k = 1, \ldots, n$. Then,

$$H_{\sigma_h}(x \cdot \omega^{(k)} - t^{(k)}) = H_\sigma(x \cdot \mathbf{W}^{(k)} - T^{(k)}).$$

Hence, the right-hand side of eqn (5.2) is written

$$f_\varepsilon(x) = \sum_{k=1}^{N} a_k H_\sigma(x \cdot \mathbf{W}^{(k)} - T^{(k)}), \quad h > 0.$$

This is the well-known form of the approximation formula.

The supremum norm is stronger on a compact set, say K, than some other norms such as $L^p(K)$-norm. Let X be a function space defined on K and endowed with a norm $\|\cdot\|$. Suppose that a set of infinitely continuously differentiable functions is dense in X and the supremum norm is stronger than $\|\cdot\|$. Then, applying the standard functional–analytical discussion, we can prove that any function of X can be approximated arbitrarily well in $\|\cdot\|$-norm by a finite sum of the form (4.1), (4.9), (4.13), or (5.2). The following is an example, where the standard discussion is applied to eqn (4.13):

Example 5.2. Let $\sigma \in \Sigma_1$. Then, any function $f \in L^p(K)$, $p \geq 1$, can be approximated arbitrarily well in $L^p(K)$-norm by a finite sum

$$\Sigma_k a_k H_\sigma(x \cdot \omega^{(k)} - t^{(k)}). \tag{5.3}$$

To confirm this, note that the supremum norm is stronger than $L^p(K)$-norm and a set of infinitely continuously differentiable functions is dense in $L^p(K)$. Then, apply the standard discussion to Corollary 4.3 and we obtain this result. □

Except that the sigmoid function H_σ is not scaled in eqn (5.3), this result is similar to Theorem 3 of Carroll and Dickinson (1989).

6. SUMMARY

1. Two policies are carried throughout except the annexed part: (a) uniform approximation on the whole space $\mathbf{R}^n$; and (b) the use of a step or sigmoid function without scaling.
2. The main tool in this article is the inverse Radon transform.
3. First, it is proven that an integration over a step function with weight g_f can exactly represent $f \in \mathcal{S}(\mathbf{R}^n)$ on the whole space $\mathbf{R}^n$ (Theorem 3.1).
4. Then, two other exact integral representations on $\mathbf{R}^n$ were derived, in each of which a nonstep sigmoid function was used instead of the step function without scaling. In the former the function to be represented must be a convolution $f * \sigma$ (Corollary 3.3).
5. This restriction is removed in the latter; it was proven that any $f \in \mathcal{S}(\mathbf{R}^n)$ can be exactly represented on the whole space as an integral of a sigmoid function without scaling (Corollary 3.4).
6. Examples of the weight g_f are illustrated (Examples 3.1 ~ 4).
7. These three integral representations are approximated by finite sums respectively (Theorem 4.1, Corollaries 4.2 and 4.3). The approximations hold uniformly on the whole space $\mathbf{R}^n$.
8. In the annexed part (Section 5), two typical examples are illustrated in order to demonstrate how other results are obtained when our policies are weakened (Examples 5.1 and 5.2).

REFERENCES

Carroll, S. M., & Dickinson, B. W. (1989). Construction of neural nets using the Radon transform. *'89IJCNN Proceedings*, **1**, 607–611.

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

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

Gel'fand, I. M., Graev, M. I., & Vilenkin, N. Ya. (1966). *Generalized Function* Vol. 5 (English translation) New York: Academic Press.

Hecht-Nielsen, R. (1987). Kolmogorov's mapping neural network existence theorem. *IEEE First International Conference on Neural Networks*, **3**, 11–13.

Helgason, S. (1980). *The Radon Transform*. Basel, Switzerland: Birkhauser.

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

Irie, B., & Miyake, S. (1988). Capability of three-layered perceptrons. *IEEE International Conference on Neural Networks*, **1**, 641–648.

Kolmogorov, A. N. (1957). On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. *Dokl. Akad. Nauk. SSSR*, **144**, 679–681. English Translation. (1963). *Math. Soc. Transl.*, **28**, 55–59.

Wieland, A., & Leighten, R. (1987). Geometric analysis of neural network capability. *IEEE First International Conference on Neural Networks*, **1**, 385–392.

APPENDIX

In order to avoid theoretical complications in the text, the definition of the fractional power of the operator $\Box_t$ and related matters and the proofs of Lemmas 2.1 and 2.2 are described here.

In order to treat the inverse Radon transform, we need the fractional power of the operator $-\Box_t$ which is defined by $\Box_t\varphi(t) = (\partial^2/\partial t^2)\varphi(t)$. This notation is used in Helgason (1980). The fractional power is defined by

$$(-\Box_t)^\alpha \sigma(t) = 4^\alpha \pi^{-1/2} \frac{\Gamma(\alpha + 1/2)}{\Gamma(-\alpha)} \text{v.p.} \int |p - t|^{-2\alpha - 1} \varphi(p)\, dp, \tag{A.1}$$

where $\alpha > 0$ is not an integer and v.p. stands for the principal value. For even n, we have

$$(-\Box_t)^{(n-1)/2}\varphi(t) = (-1)^{n/2} \frac{1}{\pi}(n-1)!\text{v.p.} \int p^{-n}\varphi(p + t)\, dp. \tag{A.2}$$

The principal value at $t = 0$ on the right-hand side of eqn (A.2) is written

$$\text{p.v.} \int_{-\infty}^{\infty} p^{-n}\varphi(p)\, dp = \int_0^{\infty} p^{-n} \Bigg\{ \varphi(p) + \varphi(-p) - 2\left[\varphi(0) + \frac{p^2}{2!}\varphi''(0) + \cdots + \frac{p^{n-2}}{(n-2)!}\varphi^{(n-2)}(0)\right]\Bigg\}\, dp \tag{A.3}$$

(Gelfand et al., 1966). If $\varphi \in \mathcal{S}(\mathbf{R})$, the principal value is well-defined.

Proof of Lemma 2.1. Though item 1 can be proved straightforwardly, we briefly illustrate the proof of this part of the lemma because the equations (A.4a), (A.4b) and (A.4c) below are useful. Writing $\varphi^{(m)} = \partial_t^m \varphi$, let us divide the principal value into three parts:

$$I\varphi(t,\omega) = \sum_{i=1}^{3} I_i\varphi(t,\omega), \quad (A.4)$$

where

$$I_1\varphi(t,\omega) = \int_0^1 p^{-n}\left\{\varphi(t+p,\omega) + \varphi(t-p,\omega) - 2\left[\varphi(t,\omega) + \frac{p^2}{2!}\varphi''(t,\omega) + \cdots + \frac{p^{n-2}}{(n-2)!}\varphi^{(n-2)}(t,\omega)\right]\right\} dp, \quad (A.4a)$$

$$I_2\varphi(t,\omega) = \int_1^\infty p^{-n}\{\varphi(t+p,\omega) + \varphi(t-p,\omega)\}\, dp, \quad (A.4b)$$

$$I_3\varphi(t,\omega) = -2\int_1^\infty p^{-n}\left\{\varphi(t,\omega) + \frac{p^2}{2!}\varphi''(t,\omega) + \cdots + \frac{p^{n-2}}{(n-2)!}\varphi^{(n-2)}(t,\omega)\right\} dp. \quad (A.4c)$$

Put

$$R_n(\varphi,t,p,\omega) \equiv p^{-n}\left\{\varphi(t+p,\omega) + \varphi(t-p,\omega) - \left[\varphi(t,\omega) + \frac{p^2}{2!}\varphi''(t,\omega) + \cdots + \frac{p^{n-2}}{(n-2)!}\varphi^{(n-2)}(t,\omega)\right]\right\}.$$

By Taylor's theorem, there exists θ, $|\theta| < 1$, such that

$$R_n(\varphi,t,p,\omega) = \frac{2}{n!}\varphi^{(n)}(t+\theta p,\omega). \quad (A.5)$$

Therefore, we have an expression

$$I_1\varphi(t,\omega) = \int_0^1 R_n(\varphi,t,p,\omega)\,dp = \int_0^1 \frac{2}{n!}\varphi^{(n)}(t+\theta p,\omega)\,dp.$$

Since $\varphi \in \mathcal{S}(\mathbf{R}\times\mathbf{S}^{n-1})$, this integral is integrable and rapidly decreasing in t. Since

$$|R_n(\varphi,t,p,\omega)| \le \sup\left\{\left|\frac{2}{n!}\varphi^{(n)}(s,\omega)\right|;\ s\in\mathbf{R},\ \omega\in\mathbf{S}^{n-1}\right\} < \infty \quad (A.6)$$

by eqn (A.5) and since $R_n(\varphi,t,p,\omega)$ is continuous in (t,ω), $I_1\varphi(t,\omega)$ is continuous in (t,ω). The second integral $I_2\varphi(t,\omega,a)$ is integrable, continuous in (t,ω) and bounded by $M(|t|^n+1)^{-1}$ with $M > 0$. The third one $I_3\varphi(t,\omega)$ is also integrable and rapidly decreasing in t. Hence, the principal value (2.7) is well-defined and continuous in (t,ω), and (2.8) holds for $k = 0$.

Set

$$\Delta_h\varphi(t,\omega) \equiv \varphi(t+h,\omega) - \varphi(t,\omega).$$

Then,

$$\frac{1}{h}\{(I_1\varphi)(t+h,\omega) - (I_1\varphi)(t,\omega)\} = \frac{1}{h} I_1(\Delta_h\varphi)(t,\omega) = \int_0^1 R_n\left(\frac{1}{h}\Delta_h\varphi,t,p,\omega\right) dp.$$

Obviously, we have that, by eqn (A.5) and the mean value theorem,

$$\lim_{h\to 0} R_n\left(\frac{1}{h}\Delta_h\varphi,t,p,\omega\right) = R_n(\varphi',t,p,\omega)$$

and

$$\left|R_n\left(\frac{1}{h}\Delta_h\varphi,t,p,\omega\right)\right| = \frac{2}{n!}\left|\frac{1}{h}(\Delta_h\varphi)^{(n)}(t+\theta p,\omega)\right| = \frac{2}{n!}|\varphi^{(n+1)}(t+\theta p+\theta_1 h,\omega)| \le \sup_{s\in\mathbf{R}}\frac{2}{n!}|\varphi^{(n+1)}(s,\omega)|,$$

where $|\theta| < 1$ and $0 < \theta_1 < 1$. Hence, we can apply Lebesgue's
where $|\theta| < 1$ and $0 < \theta_1 < 1$. Hence, we can apply Lebesgue's
dominated convergence theorem and obtain

$$\partial_t I_1\varphi(t,\omega) = \lim_{h\to 0}\frac{1}{h}\{I_1\varphi(t+h,\omega) - I_1\varphi(t,\omega)\} = I_1\varphi'(t,\omega). \quad (A.7)$$

for $i = 1$. We can further show that eqn (A.7) holds for $i = 2$, 3. Thus, we obtain

$$\partial_t I\varphi(t,\omega) = I\varphi'(t,\omega), \quad (A.8)$$

which means the commutativity between the operation I and ∂_t. Similarly we can prove the commutativity among I, ∂_t and ∂_ω's. Since $\partial_t^k\partial_{\omega_i}\cdots\partial_{\omega_j}\varphi$ is again in $\mathcal{S}(\mathbf{R}\times\mathbf{S}^{n-1})$, the principal value $I(\varphi, t,\omega)$ is infinitely differentiable and each derivative is continuous in (t,ω). By the commutativity (A.7), we obtain the expression below:

$$\begin{aligned}\partial_t^k I_2\varphi(t,\omega) &= \int_1^\infty p^{-n}\{\varphi^{(k)}(t+p,\omega) + \varphi^{(k)}(t-p,\omega)\}\,dp \\ &= -\sum_{j=1}^{k}\frac{(n+j-2)!}{(n-1)!}\{\varphi^{(k-j)}(t+1,\omega) \\ &\quad + (-1)^j\varphi^{(k-j)}(t-1,\varphi)\} \\ &\quad + \frac{(n+k-2)!}{(n-1)!}\int_1^\infty p^{-n-k} \\ &\quad \times\{\varphi(t+p,\omega) + (-1)^k\varphi(t-p,\omega)\}\,dp.\end{aligned} \quad (A.9)$$

In the last member, the integral is bounded by $M_1(|t|+1)^{-n-k}$, $M_1 > 0$, and other terms are rapidly decreasing. Since $\partial_t I_i\varphi(t,\omega)$, $i = 1, 3$, are rapidly decreasing, eqn (2.8) holds. This concludes the proof. □

Proof of Lemma 2.2: The left-hand side of the first equation is equal to

$$\lim_{h\to 0}\frac{1}{h}\left\{\int \psi(x)\,dm_{t+h,\omega}(x) - \int\psi(x)\,dm_{t,\omega}(x)\right\} = \lim_{h\to 0}\int \partial_\omega\psi(x+\theta h\omega)\,dm_{t,\omega}(x) \quad (A.10)$$

with $0 < \theta < 1$. Hence, we obtain eqn (2.10) by the dominated convergence theorem. For odd n, eqn (2.11) can be easily obtained by applying item 1 repeatedly. Now suppose that n is even. Then, the right-hand side of the second equation in this lemma is written

$$(L_\omega\psi)^\sim(t,\omega) = \frac{(-1)^{n/2}(n-1)!}{2^n\pi^n}\int\left[\int_0^1 J_1\psi(x,p,\omega)\,dp + \int_1^\infty J_2\psi(x,p,\omega)\,dp + \int_1^\infty J_3\psi(x,p,\omega)\,dp\right] dm_{t,\omega}(x), \quad (A.11)$$

where

$$J_1\psi(x,p,\omega) = p^{-n}\left\{\psi(x+p\omega) + \psi(x-p\omega) - 2\left[\psi(x) + \frac{p^2}{2!}\partial_\omega^2\psi(x) + \cdots + \frac{p^{n-2}}{(n-2)!}\partial_\omega^{n-2}\psi(x)\right]\right\}, \quad (A.12a)$$

$$J_2\psi(x,p,\omega) = p^{-n}\{\psi(x+p\omega) + \psi(x-p\omega)\}, \quad (A.12b)$$

$$J_3\psi(x,p,\omega) = -2p^{-n}\left\{\psi(x) + \frac{p^2}{2!}\partial_\omega^2\psi(x) + \cdots + \frac{p^{n-2}}{(n-2)!}\partial_\omega^{n-2}\psi(x)\right\}. \quad (A.12c)$$

By Taylor's theorem, there exists θ, $|\theta| < 1$, such that

$$J_1\psi(x,p,\omega) = \frac{2}{n!}\partial_\omega^n\psi(x+\theta p\omega).$$

Hence, the right-hand side of (A.12a) is integrable with respect to $dp\,dm_{t,\omega}$ on $[0,1]\times\mathbf{P}_{t,\omega}$. Both $J_2\psi(x,p,\omega)$ and $J_3\psi(x,p,\omega)$ are also integrable with respect to $dp\,dm_{t,\omega}$ on $[1,\infty]\times\mathbf{P}_{t,\omega}$. Hence, by Fubini's theorem, we can change the order of the integrations on the right-hand side of (A.11) and obtain

$$(L_\omega\psi)^\sim(t,\omega) = \frac{(-1)^{n/2}(n-1)!}{2^n\pi^n}\int_0^\infty p^{-n}\left[\int\left\{\psi(x+p\omega) + \psi(x-p\omega) - 2\left[\psi(x) + \frac{p^2}{2!}\partial_\omega^2\psi(x) + \cdots + \frac{p^{n-2}}{(n-2)!}\partial_\omega^{n-2}\psi(x)\right]\right\} dm_{t,\omega}(x)\right] dp.$$

Therefore, the second equation in this lemma also holds. □