Neural Networks, Vol. 2, pp. 183–192, 1989
Printed in the USA. All rights reserved.

0893-6080/89 $3.00 + .00

ORIGINAL CONTRIBUTION

On the Approximate Realization of Continuous Mappings by Neural Networks

KEN-ICHI FUNAHASHI

ATR Auditory and Visual Perception Research Laboratories

(*Received* 6 *May* 1988; *revised and accepted* 14 *September* 1988)

Abstract—*In this paper, we prove that any continuous mapping can be approximately realized by Rumelhart-Hinton-Williams' multilayer neural networks with at least one hidden layer whose output functions are sigmoid functions. The starting point of the proof for the one hidden layer case is an integral formula recently proposed by Irie–Miyake and from this, the general case (for any number of hidden layers) can be proved by induction. The two hidden layers case is proved also by using the Kolmogorov-Arnold-Sprecher theorem and this proof also gives non-trivial realizations.*

Keywords—Neural network, Back propagation, Output function, Sigmoid function, Hidden layer, Unit, Realization, Continuous mapping.

1. INTRODUCTION

Since McCulloch-Pitts (1943), there have been many studies of mathematical models of neural networks. Recently, Hopfield, Hinton, Rumelhart, Sejnowski and others have tried many concrete applications such as pattern recognition, and have shown that it is possible to clarify the mechanism of human information processing by the use of these models. In particular, the back propagation algorithm (generalized delta rule) proposed by Rumelhart, Hinton, and Williams (1986) provides a learning rule for multilayer networks. Many applications of this algorithm have been shown recently. However, there has been little theoretical research on the capability of the Rumelhart–Hinton–Williams multilayer network.

On the application to pattern recognition, Lippmann (1987) asserts that arbitrary complex decision regions, including concave regions, can be formed using four-layer networks, but this is only an intuitive assertion. Wieland and Leighton (1987) showed an example of a three-layer network with thresholding units which partitions a space into concave subspaces. Huang and Lippmann (1987) demonstrated by simulations that three-layer networks can form several complex decision regions in pattern recognition application. However, it has been known that any piecewise-linear decision region (which is not necessarily convex) can be realized by a multilayer network (Duda & Fossum, 1966). It's learning algorithm was also proposed (Amari, 1967) based on the same principle as the generalized delta rule. There are also other applications of multilayer networks for forming mappings, such as NETtalk by Sejnowski and Rosenberg (1987).

Hecht-Nielsen (1987) pointed out that Kolmogorov's theorem (Kolmogorov, 1957) and Sprecher's refinement (Sprecher, 1965), which are both known as negative solutions of Hilbert's thirteenth problem, show that any continuous mapping can be represented by a form of four-layer neural network. Uesaka (1971) and Poggio (1983) have also pointed this out. However, the assertion has a problem in that the output function of each unit of this network is not a given sigmoid function.

Irie and Miyake (1988) obtained an integral formula which suggests the realization of functions of several variables by three-layer networks by analogy with the principle of the computerized tomography (CT). But in this integral formula, the output function $\psi(x)$ must satisfy the condition of absolute integrability, so that it cannot be a sigmoid function. Moreover, the function to be realized is given by an integral representation and the formula does not di-

The author wishes to thank Drs. Y. Tohkura, T. Inui and S. Miyake, and Mr. T. Okamoto for their valuable comments on the manuscript. The author also would like to thank anonymous reviewers whose constructive suggestions have improved the quality of this paper.

Requests for reprints should be sent to Ken-ichi Funahashi, ATR Auditory and Visual Perception Research Laboratories, Twin 21 Building, MID Tower 2-1-61 Shiromi, Higashi-ku, Osaka 540, Japan.

rectly give the realization theorem of functions by networks with finite units.

In neural networks of the feed-forward type by Rumelhart–Hinton–Williams, bounded and monotone increasing differentiable functions such as the sigmoid function $\phi(x) = 1/(1 + e^{-x})$ are used as output functions of units. This is a different point from the McCulloch-Pitts model and perceptron which use heaviside function as output functions of units and is the reason why it is possible to derive a learning algorithm for multilayer networks.

In a feed-forward type network, it's input–output relationship defines a mapping which is called an input–output mapping of the network. We studied the problem of network capabilities from the point of view of input–output mappings.

In this paper, we started from an integral formula recently proposed by Irie and Miyake (1988) and proved the theorem which guarantees the approximate realization of continuous mappings by three-layer (one hidden layer) networks whose output functions for hidden layer are sigmoid, and whose output functions for input and output layers are linear in the sense of uniform topology. It is easy to prove the theorem for k (≥ 3)-layer networks by using the theorem for a three-layer case. But the proof of the theorem for the case $k > 3$ gives only trivial approximate realization of given mappings. Therefore we show another proof for the four-layer case by using the Kolmogorov–Arnold-Sprecher theorem (Kolmogorov, 1957; Sprecher, 1965).

McCulloch-Pitts showed that any logical circuit can be designed using their model. Correspondingly, our assertion shows that any continuous mapping can be approximately represented by the Rumelhart–Hinton–Williams multilayer network.

2. MULTILAYER NEURAL NETWORKS

The Rumelhart–Hinton–Williams multilayer network that we consider here is a feed-forward type network with connections between adjoining layers only. Networks generally have hidden layers between the input and output layers. Each layer consists of computational units. The input–output relationship of each unit is represented by inputs x_i, output y, connection weights w_i, threshold θ, and differentiable function ϕ as follows:

$$y = \phi\left(\sum_{i=1}^{k} w_i x_i - \theta\right).$$

The learning rule of this network is known as the back propagation algorithm (Rumelhart, Hinton, & Williams, 1986). The back propagation algorithm is an algorithm that uses a gradient descent method to modify weights and thresholds so that the error between the desired output and the output signal of the network is minimized. We generally use a bounded and monotonic increasing differentiable function which is called sigmoid function for each unit's output function.

If a multilayer network has n input units and m output units, then the input–output relationship defines a continuous mapping from n-dimensional Euclidean space to m-dimensional Euclidean space. We call this mapping the input–output mapping of the network. We study the problem of network capabilities from the point of view of input–output mappings. It is observed that for the study of mappings defined by multilayer networks it is sufficient to consider networks whose output functions for hidden layers are the above $\phi(x)$ and whose output functions for input and output layers are linear.

3. APPROXIMATE REALIZATION OF CONTINUOUS MAPPINGS BY NEURAL NETWORKS

We shall consider the possibility of representing continuous mappings by neural networks whose output functions in hidden layers are sigmoid, for example, $\phi(x) = 1/(1 + e^{-x})$. It is simply noted here that general continuous mappings cannot be exactly represented by Rumelhart–Hinton–Williams' networks. For example, if a real analytic output function such as the sigmoid function $\phi(x) = 1/(1 + e^{-x})$ is used, then an input–output mapping of this network is analytic and generally cannot represent all continuous mappings.

Let points of n-dimensional Euclidean space $\mathbf{R}^n$ be denoted by $\mathbf{x} = (x_1, \ldots, x_n)$ and the norm of $\mathbf{x}$ defined by $|\mathbf{x}| = (\sum_{i=0}^{n} x_i^2)^{1/2}$.

We prove the following theorems and corollaries in this paper.

Theorem 1.

Let $\phi(x)$ be a nonconstant, bounded and monotone increasing continuous function. Let K be a compact subset (bounded closed subset) of $\mathbf{R}^n$ and $f(x_1, \ldots, x_n)$ be a real valued continuous function on K. Then for an arbitrary $\epsilon > 0$, there exists an integer N and real constants $c_i, \theta_i (i = 1, \ldots, N)$, $w_{ij} (i = 1, \ldots, N, j = 1, \ldots, n)$ such that

$$\tilde{f}(x_1, \ldots, x_n) = \sum_{i=1}^{N} c_i \phi\left(\sum_{j=1}^{n} w_{ij} x_j - \theta_i\right)$$

satisfies $\max_{\mathbf{x} \in K} |f(x_1, \ldots, x_n) - \tilde{f}(x_1, \ldots, x_n)| < \epsilon$. In other words, for an arbitrary $\epsilon > 0$, there exists a three-layer network whose output functions for the hidden layer are $\phi(x)$, whose output functions for input and output layers are linear and which has an input–output function $\tilde{f}(x_1, \ldots, x_n)$ such that

$\max_{\mathbf{x}\in K} |f(x_1, \ldots, x_n) - \tilde{f}(x_1, \ldots, x_n)| < \epsilon$.

The above theorem easily leads to the following general theorem.

Theorem 2.

Let $\phi(x)$ be a nonconstant, bounded and monotone increasing continuous function. Let K be a compact subset (bounded closed subset) of $\mathbf{R}^n$ and fix an integer $k \geq 3$. Then any continuous mapping $f: K \to \mathbf{R}^m$ defined by $\mathbf{x} = (x_1, \ldots, x_n) \to (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))$ can be approximated in the sense of uniform topology on K by input–output mappings of k-layer (k-2 hidden layers) networks whose output functions for hidden layers are $\phi(x)$, and whose output functions for input and output layers are linear. In other words, for any continuous mapping $f: K \to \mathbf{R}^m$ and an arbitrary $\epsilon > 0$, there exists a k-layer network whose input–output mapping is given by $\tilde{f}: K \to \mathbf{R}^m$ such that $\max_{\mathbf{x}\in K} d(f(\mathbf{x}), \tilde{f}(\mathbf{x})) < \epsilon$, where $d(,)$ is a metric which induces the usual topology of $\mathbf{R}^m$.

Corollary 1.

Let $\phi(x)$, K be as above and fix an integer $k \geq 3$. Then any mapping $f: \mathbf{x} \in K \to (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x})) \in \mathbf{R}^m$ where $f_i(\mathbf{x})$ $(i = 1, \ldots, m)$ are summable on K, can be approximated in the sense of L^2-topology on K by input–output mappings of k-layer (k-2 hidden layers) networks whose output functions for hidden layers are $\phi(x)$ and whose output functions for input and output layers are linear. In other words, for an arbitrary $\epsilon > 0$, there exists a k-layer network whose input–output mapping is given by $\tilde{f}: \mathbf{x} \in K \to (\tilde{f}_1(\mathbf{x}), \ldots, \tilde{f}_m(\mathbf{x})) \in \mathbf{R}^m$ such that

$$d_{L^2(K)}(f, \tilde{f}) = \left(\sum_{i=1}^{m} \int_K |f_i(x_1, \ldots, x_n) - \tilde{f}_i(x_1, \ldots, x_n)|^2 \, d\mathbf{x} \right)^{1/2} < \epsilon.$$

Corollary 2.

Let K be as above and fix an integer $k \geq 3$. Let $\phi(x)$ be a strictly increasing continuous function such that $\phi((-\infty, \infty)) = (0, 1)$. Then any continuous mapping $f: K \to (0, 1)^m$ can be approximated in the sense of uniform topology on K by input–output mappings of $k(\geq 3)$-layer neural networks whose output functions for hidden and output layers are $\phi(x)$.

Proof. Set $f(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))$. As $\phi^{-1}: (0, 1) \to (-\infty, \infty)$ is continuous, the theorem 2 is applied to the mapping $\mathbf{x} \to \phi^{-1}f(\mathbf{x}) = (\phi^{-1}f(\mathbf{x}), \ldots, \phi^{-1}f_m(\mathbf{x}))$ and the corollary is obtained easily.

q.e.d.

Remark 1.

Usual output functions such as the sigmoid function $1/(1 + e^{-x})$ used for back-propagation neural networks satisfy the condition of $\phi(x)$ that $\phi(x)$ is a nonconstant, bounded and monotone increasing continuous function.

Remark 2.

Any mapping is approximately realized by a three-layer (one hidden layer) network. However, it should be theoretically studied in the future that the possibility of $k > 3$-layer networks can realize a given mapping with less costs (number of units or connections) than three-layer networks, within error ϵ.

For the application of neural networks to pattern recognition, if m is the number of recognized categories, usually m output units corresponding to these categories are used, and the system is allowed to learn to take values near 1 only for units corresponding to the input categories. Corollaries show that if one uses multilayer networks with hidden layers, any decision region can be formed by a neural network. In particular, a strictly increasing continuous function, as the output function of each unit, can be chosen.

In this paper, we call bounded and monotone increasing continuous functions, sigmoid functions. In particular, a sigmoid function $\phi(x)$ having a weak derivative which is summable has the property that if we set $\phi_\epsilon(x) = \phi(x/\epsilon)(\epsilon > 0)$, then the derivatives $\phi'_\epsilon(x) = (1/\epsilon)\phi'(x/\epsilon)$ converge, in the sense of the generalized function (see, e.g., Gel'fand & Shilov, 1964), to the δ function as $\epsilon \to 0$. That is to say, if $\phi(\infty) - \phi(-\infty) = 1$, then for any smooth function $g(x)$ with compact support,

$$\lim_{\epsilon \to +0} \int_{-\infty}^{\infty} \phi'_\epsilon(x) \cdot g(x) \, dx = g(0).$$

The following examples are included in the class of sigmoid functions considered here.

Example 1. For $\phi(x) = 1/(1 + \exp(-x))$, $\phi'_\epsilon(x) = 1/\epsilon \exp(-x/\epsilon)/(1 + \exp(-x/\epsilon))^2$ and $\phi(x)$ is a sigmoid function.

Example 2. For $\Phi(x) = 1/\sqrt{2\pi} \int_{-\infty}^{x} \exp(-t^2/2) \, dt$, $\Phi'_\epsilon(x) = \epsilon/\sqrt{2\pi} \exp(-x^2/2\epsilon)$ and $\Phi(x)$ is a sigmoid function.

Example 3. For $\phi(x)$ where $\phi(x) = 0(x < 0)$, $\phi(x) = x(0 < x < 1)$ and $\phi(x) = 1(x \geq 1)$, $\phi'_\epsilon(x) = 0(x < 0 \text{ or } x \geq \epsilon)$, $\phi'_\epsilon(x) = 1/\epsilon(0 \leq x < \epsilon)$, and $\phi(x)$ is a sigmoid function.

In the McCulloch-Pitts neural model and perceptron, a threshold function $\phi(x) = 1(x \geq 0)$, $= 0(x < 0)$ is used as the output function.

Sigmoid functions $\phi(x)$ where $\phi(-\infty) = 0$ and $\phi(\infty) = 1$ are appropriate as output functions in the neural model because if we set $\phi_\epsilon(x) = \phi(x/\epsilon)(\epsilon > 0)$ then these converge to the threshold function in the McCulloch–Pitts neural model and perceptron as $\epsilon \to +0$.

McCulloch–Pitts shows that one can design any logical circuit using their model. Correspondingly, theorem 2 above shows that any continuous mapping can be approximately represented by multilayer networks with sigmoid output functions.

4. PRELIMINARY 1 (MOLLIFIERS, FOURIER TRANSFORMS)

Fundamental matters used in this paper are reviewed here.

Let $L^p(\mathbf{R}^n)(p \geq 1)$ denote the space of all measurable functions $f(\mathbf{x})$ on $\mathbf{R}^n$ which satisfy

$$\int_{\mathbf{R}^n} |f(\mathbf{x})|^p \, d\mathbf{x} < \infty.$$

The norm of $f \in L^p(\mathbf{R}^n)$ is defined by

$$\|f(\mathbf{x})\|_{L^p} = \left(\int_{\mathbf{R}^n} |f(\mathbf{x})|^p \, d\mathbf{x} \right)^{1/p},$$

and the convergence $f_n(\mathbf{x}) \to f(\mathbf{x})$ in $L^p(\mathbf{R}^n)$ is defined by

$$\lim_{n \to \infty} \|f_n(\mathbf{x}) - f(\mathbf{x})\|_{L^p} = 0.$$

Generally, for any measurable set K, $L^p(K)(p \geq 1)$ is defined similarly.

Let $\rho(\mathbf{x})$ be a function on $\mathbf{R}^n$ which satisfies the following conditions:

(i) $\rho(\mathbf{x}) \geq 0$, $\rho(\mathbf{x})$ has continuous partial derivatives of all orders and the support is contained in the unit sphere $|\mathbf{x}| \leq 1$.
(ii) $\int_{\mathbf{R}^n} \rho(\mathbf{x}) \, d\mathbf{x} = 1$

Then, for $\epsilon > 0$, set $\rho_\epsilon(\mathbf{x}) = (1/\epsilon)^n \rho(\mathbf{x}/\epsilon)$.

If $u(\mathbf{x}) \in L^1_{\text{loc}}$, that is, $u(\mathbf{x})$ is locally summable, consider

$$\rho_\epsilon * u(\mathbf{x}) = \int_{\mathbf{R}^n} \rho_\epsilon(\mathbf{x} - \mathbf{y}) u(\mathbf{y}) \, d\mathbf{y},$$

then the following assertions hold: (a) $\rho_\epsilon * u(\mathbf{x}) \in C^\infty$, that is, $\rho_\epsilon * u(\mathbf{x})$ has continuous partial derivatives of all orders, and the support of $\rho_\epsilon * u(\mathbf{x})$ is contained in the ϵ neighborhood of support of $u(\mathbf{x})$; (b) if $u(\mathbf{x})$ is a continuous function with compact support, then $\rho_\epsilon * u(\mathbf{x}) \to u(\mathbf{x})$ uniformly on $\mathbf{R}^n$ as $\epsilon \to +0$; and (c) if $u(\mathbf{x}) \in L^p(\mathbf{R}^n)(p \geq 1)$ then $\rho_\epsilon * u(\mathbf{x}) \to u(\mathbf{x})$ in L^p as $\epsilon \to +0$. The operator $\rho_\epsilon *$ is called a mollifier.

For $f(\mathbf{x}) \in L^1(\mathbf{R}^n)$, Fourier transform

$$\hat{f}(\boldsymbol{\xi}) = \int_{\mathbf{R}^n} e^{-i\langle \mathbf{x}, \boldsymbol{\xi} \rangle} f(\mathbf{x}) \, d\mathbf{x}, \tag{1}$$

where $\langle \mathbf{x}, \boldsymbol{\xi} \rangle = \sum_{i=1}^n x_i \xi_i$, can be defined and set $\hat{f}(\boldsymbol{\xi}) = \mathcal{F} f(\boldsymbol{\xi})$.

If $f(\mathbf{x})$ satisfies an additional condition that $f(\mathbf{x})$ has continuous partial derivatives of order up to n, then $f(\mathbf{x})$ can be represented at each point by inverse Fourier transform of $\hat{f}(\boldsymbol{\xi})$ as follows.

$$f(\mathbf{x}) = (2\pi)^{-n} \int_{\mathbf{R}^n} e^{i\langle \mathbf{x}, \boldsymbol{\xi} \rangle} \hat{f}(\boldsymbol{\xi}) \, d\boldsymbol{\xi}.$$

The Plancherel theorem especially asserts that $\mathcal{F}$ can be extended one to one onto mapping $\mathcal{F}: L^2(\mathbf{R}^n) \to L^2(\mathbf{R}^n)$ and for $f(\mathbf{x}) \in L^1 \cap L^2(\mathbf{R}^n)$, $\mathcal{F} f(\boldsymbol{\xi})$ is equal to the one defined by (1). Furthermore, for $f(\mathbf{x}) \in L^2(\mathbf{R}^n)$,

$$\left\| \mathcal{F} f(\boldsymbol{\xi}) - \int_{|\mathbf{x}| \leq A} e^{-i\langle \mathbf{x}, \boldsymbol{\xi} \rangle} f(\mathbf{x}) \, d\mathbf{x} \right\|_{L^2} \longrightarrow 0 (A \longrightarrow +\infty)$$

(see e.g., Yosida, 1968).

5. PRELIMINARY 2 (IRIE–MIYAKE'S INTEGRAL FORMULA)

The following theorem is a starting point for proof of Theorem 1.

Theorem (Irie-Miyake)

Let $\psi(x) \in L^1(\mathbf{R})$, that is, let $\psi(x)$ be absolutely integrable and $f(x_1, \ldots, x_n) \in L^2(\mathbf{R}^n)$. Let $\Psi(\boldsymbol{\xi})$ and $F(w_1, \ldots, w_n)$ be Fourier transforms of $\psi(x)$ and $f(x_1, \ldots, x_n)$ respectively.

If $\Psi(1) \neq 0$, then

$$f(x_1, \ldots, x_n) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \psi\left(\sum_{i=1}^n x_i w_i - w_0 \right) \frac{1}{(2\pi)^n \Psi(1)} F(w_1, \ldots, w_n) \times \exp(i w_0) \, dw_0 \, dw_1 \cdots dw_n.$$

Remark

This formula precisely asserts that if we set

$$I_{\infty,A}(x_1, \ldots, x_n) = \int_{-A}^{A} \cdots \int_{-A}^{A} \left[\int_{-\infty}^{\infty} \psi\left(\sum_{i=1}^n x_i w_i - w_0 \right) \times \frac{1}{(2\pi)^n \Psi(1)} F(w_1, \ldots, w_n) \times \exp(i w_0) \, dw_0 \right] dw_1 \cdots dw_n$$

then

$$\lim_{A\to\infty} \|I_{\infty,A}(x_1, \ldots, x_n) - f(x_1, \ldots, x_n)\|_{L^2} = 0.$$

Connecting this formula with three-layer networks, Irie and Miyake (1988) assert that arbitrary functions can be represented by a three-layer network with an infinite number of computational units. In this formula, w_0 corresponds to threshold, w_i corresponds to connection weight and $\psi(x)$ corresponds to the output function of the units. However, the sigmoid function $1/(1 + e^{-x})$ does not satisfy the condition of this formula that $\psi(x)$ be absolutely integrable and so the formula does not directly give the realization theorem of functions by networks.

6. PRELIMINARY 3 (SEVERAL LEMMAS)

We prepare several Lemmas for proof of our theorem 1.

Lemma 1.

Let $\phi(x)$ be a nonconstant, bounded and monotone increasing continuous function. For $\alpha > 0$, if we set

$$g(x) = \phi(x + \alpha) - \phi(x - \alpha),$$

then $g(x) \in L^1(\mathbf{R})$, that is,

$$\int_{-\infty}^{\infty} |g(x)|\, dx < \infty.$$

Furthermore, for some $\delta > 0$, if we set

$$g_\delta(x) = \phi(x/\delta + \alpha) - \phi(x/\delta - \alpha),$$

then the value of Fourier transform $G_\delta(\xi)$ of $g_\delta(x)$ at $\xi = 1$ is non-zero.

Proof. Let $|g(x)| \le M$. For $L > M$,

$$\begin{aligned}\int_{-L}^{L} |g(x)|\, dx &= \int_{-L}^{L} g(x)\, dx = \int_{-L+\alpha}^{L+\alpha} \phi(x)\, dx \\ &\quad - \int_{-L-\alpha}^{L-\alpha} \phi(x)\, dx \\ &= \int_{L-\alpha}^{L+\alpha} \phi(x)\, dx \\ &\quad - \int_{-L-\alpha}^{-L+\alpha} \phi(x)\, dx \le 4\alpha M.\end{aligned}$$

Therefore,

$$\lim_{L\to\infty} \int_{-L}^{L} |g(x)|\, dx < \infty.$$

We show that for some $\delta > 0$, $G_\delta(1) \neq 0$. If the assertion does not hold, then for any $\delta > 0$,

$$\int_{-\infty}^{\infty} (\phi(x/\delta + \alpha) - \phi(x/\delta - \alpha))e^{-ix}\, dx = 0.$$

By the change of the variable,

$$\int_{-\infty}^{\infty} (\phi(x + \alpha) - \phi(x - \alpha))e^{-ix\delta}\, dx = 0 \quad \text{(for any } \delta > 0\text{).} \quad (1)$$

Taking the complex conjugate of the above equation (1),

$$\int_{-\infty}^{\infty} (\phi(x + \alpha) - \phi(x - \alpha))e^{ix\delta}\, dx = 0 \quad \text{(for any } \delta > 0\text{).} \quad (2)$$

Since the Fourier transform $G_1(\xi)$ of $g_1(x) = \phi(x + \alpha) - \phi(x - \alpha) \in L^1(\mathbf{R})$ is continuous, so, from (1) and (2), $G_1(\xi)$ is identically zero. Therefore,

$$\phi(x + \alpha) - \phi(x - \alpha) \equiv 0.$$

This is a contradiction, because $\phi(x)$ is not constant. q.e.d.

Remark

Lemma 1 holds for $\phi(x)$ which is locally summable.

Lemma 2

Let $A_i > 0$ $(i = 1, \ldots, m)$, K be a compact subset (bounded closed subset) of $\mathbf{R}^n$ and $h(x_1, \ldots, x_m, t_1, \ldots, t_n)$ be a continuous function on $[-A_1, A_1] \times \cdots \times [-A_m, A_m] \times K$.

Then the function defined by the integral

$$H(\mathbf{t}) = \int_{-A_1}^{A_1} \cdots \int_{-A_m}^{A_m} h(x_1, \ldots, x_m, t_1, \ldots, t_n)\, dx_1 \ldots dx_m$$

can be approximated uniformly on K by the Riemann sum

$$\begin{aligned}H_N(\mathbf{t}) &= \frac{2A_1 \cdots 2A_m}{N^m} \\ &\quad \times \sum_{k_1 \cdots k_m = 0}^{N-1} h\left(-A_1 + \frac{k_1 \cdot 2A_1}{N}, \ldots,\right. \\ &\quad \left. - A_m + \frac{k_m \cdot 2A_m}{N}, t_1, \ldots, t_n\right).\end{aligned}$$

In other words, for an arbitrary $\epsilon > 0$, there exists a natural number N_0 such that for $N \ge N_0$,

$$\max_{\mathbf{t}\in K} |H(\mathbf{t}) - H_N(\mathbf{t})| < \epsilon.$$

Proof. The function $h(\mathbf{x}, \mathbf{t})$ is continuous on the compact set $[-A_1, A_1] \times \cdots \times [-A_m, A_m] \times K$, so $h(\mathbf{x},$

t) is uniformly continuous. Therefore for any $\epsilon > 0$, we can take the integer N_0 such that if $N \geq N_0$ and

$$\left| x_i - \left(A_i + \frac{k_i \cdot 2A_i}{N} \right) \right| < \frac{2A_i}{N} \ (i = 1, \ldots, m) \text{ then}$$

$$\left| h(x_1, \ldots, x_m, t_1, \ldots, t_n) - h\left(-A_1 + \frac{k_1 \cdot 2A_1}{N}, \ldots, -A_m + \frac{k_m \cdot 2A_m}{N}, t_1, \ldots, t_n \right) \right| < \frac{\epsilon}{2A_1 \cdots 2A_m}.$$

Assertion of the Lemma is obvious from this inequality. q.e.d.

7. PROOF OF THEOREMS

We will prove our theorems in Section 3 under the above preliminaries.

Proof of Theorem 1

Step 1. Because $f(\mathbf{x})(\mathbf{x} = (x_1, \ldots, x_n))$ is a continuous function on a compact subset K of $\mathbf{R}^n$, $f(\mathbf{x})$ can be extended to be a continuous function on $\mathbf{R}^n$ with compact support. We also denote this by $f(\mathbf{x})$.

If we operate the mollifier $\rho_\alpha *$ on $f(\mathbf{x})$, $\rho_\alpha * f(\mathbf{x})$ is C^∞-function with compact support. Furthermore, $\rho_\alpha * f(\mathbf{x}) \to f(\mathbf{x})(\alpha \to +0)$ uniformly on $\mathbf{R}^n$. Therefore we may suppose $f(\mathbf{x})$ is a C^∞-function with compact support for proving Theorem 1. By the Paley–Wiener theorem (see, e.g., Yosida, 1968), the Fourier transform $F(\mathbf{w})(\mathbf{w} = (w_1, \ldots, w_n))$ of $f(\mathbf{x})$ is real analytic and, for any integer N, there exists a constant C_N such that

$$|F(\mathbf{w})| \leq C_N(1 + |\mathbf{w}|)^{-N}. \tag{3}$$

In particular, $F(\mathbf{w}) \in L^1 \cap L^2(\mathbf{R}^n)$.

We define $I_A(x_1, \ldots, x_n)$, $I_{\infty,A}(x_1, \ldots, x_n)$ and $J_A(x_1, \ldots, x_n)$ as follows:

$$I_A(x_1, \ldots, x_n) = \int_{-A}^{A} \cdots \int_{-A}^{A} \psi\left(\sum_{i=1}^{n} x_i w_i - w_0 \right) \cdot \frac{1}{(2\pi)^n \Psi(1)} F(w_1, \ldots, w_n) \times \exp(iw_0)\, dw_0\, dw_1 \cdots dw_n, \tag{4}$$

$$I_{\infty,A}(x_1, \ldots, x_n) = \int_{-A}^{A} \cdots \int_{-A}^{A} \left[\int_{-\infty}^{\infty} \psi\left(\sum_{i=1}^{n} x_i w_i - w_0 \right) \frac{1}{(2\pi)^n \Psi(1)} F(w_1, \ldots, w_n) \times \exp(iw_0)\, dw_0 \right] dw_1 \cdots dw_n, \tag{5}$$

$$J_A(x_1, \ldots, x_n) = \frac{1}{(2\pi)^n} \int_{-A}^{A} \int_{-A}^{A} F(w_1, \ldots, w_n) \exp\left(i \sum_{i=1}^{n} x_i w_i \right) dw_1 \cdots dw_n, \tag{6}$$

where $\psi(x) \in L^1$ is defined by

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

for some α and δ so that $\psi(x)$ satisfies Lemma 1 in Section 6.

The essential part of the proof of Irie–Miyake's integral formula is the equality

$$I_{\infty,A}(x_1, \ldots, x_n) = J_A(x_1, \ldots, x_n) \tag{7}$$

and this is derived from

$$\int_{-\infty}^{\infty} \psi\left(\sum_{i=1}^{n} x_i w_i - w_0 \right) \exp(iw_0)\, dw_0 = \exp\left(i \sum_{i=1}^{n} x_i w_i \right) \cdot \Psi(1). \tag{8}$$

In our discussion, using the estimate of $F(\mathbf{w})$, we can prove

$$\lim_{A \to \infty} J_A(x_1, \ldots, x_n) = f(x_1, \ldots, x_n)$$

uniformly on $\mathbf{R}^n$. Therefore

$$\lim_{A \to \infty} I_{\infty,A}(x_1, \ldots, x_n) = f(x_1, \ldots, x_n)$$

uniformly on $\mathbf{R}^n$. That is to say, we can state that for any $\epsilon > 0$ there exists $A > 0$ such that

$$\max_{\mathbf{x} \in \mathbf{R}^n} |I_{\infty,A}(x_1, \ldots, x_n) - f(x_1, \ldots, x_n)| < \epsilon/2. \tag{i}$$

Step 2. We will approximate $I_{\infty,A}$ by finite integrals on K. For $\epsilon > 0$, fix A which satisfies (i).

For $A' > 0$, set

$$I_{A',A}(x_1, \ldots, x_n) = \int_{-A}^{A} \cdots \int_{-A}^{A} \left[\int_{-A'}^{A'} \psi\left(\sum_{i=1}^{n} x_i w_i - w_0 \right) \times \frac{1}{(2\pi)^n \Psi(1)} F(w_1, \ldots, w_n) \times \exp(iw_0)\, dw_0 \right] dw_1 \ldots dw_n.$$

We will show that, for $\epsilon > 0$, we can take $A' > 0$ so that

$$\max_{\mathbf{x} \in K} |I_{A',A}(x_1, \ldots, x_n) - I_{\infty,A}(x_1, \ldots, x_n)| < \epsilon/2. \tag{ii}$$

Using the following equation

$$\int_{-A'}^{A'} \psi\left(\sum_{i=1}^{n} x_i w_i - w_0\right) \exp(iw_0)\, dw_0 = \int_{\sum_{i=1}^{n} x_i w_i - A'}^{\sum_{i=1}^{n} x_i w_i + A'} \psi(t) \exp(-it)\, dt \cdot \exp\left(i \sum_{i=1}^{n} x_i w_i\right),$$

the fact $F(\mathbf{x}) \in L^1$ and compactness of $[-A, A]^n \times K$, we can take A' so that

$$\left| \int_{-A'}^{A'} \psi\left(\sum_{i=1}^{n} x_i w_i - w_0\right) \exp(iw_0)\, dw_0 - \int_{-\infty}^{\infty} \psi\left(\sum_{i=1}^{n} x_i w_i - w_0\right) \exp(iw_0)\, dw_0 \right| \leq \frac{\epsilon (2\pi)^n |\Psi(1)|}{\left(2 \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} |F(x)|\, dx + 1\right)} \times \int_{-A}^{A} \cdots \int_{-A}^{A} |F(x)|\, dx \text{ on } K.$$

Therefore,

$$\max_{\mathbf{x} \in K} |I_{A',A}(x_1, \ldots, x_n) - I_{\infty,A}(x_1, \ldots, x_n)| \leq \frac{\epsilon}{\left(2 \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} |F(x)|\, dx + 1\right)} \times \int_{-A}^{A} \cdots \int_{-A}^{A} |F(x)|\, dx < \epsilon/2.$$

Step 3. From (i) and (ii), we can say that for any $\epsilon > 0$, there exist $A, A' > 0$ such that

$$\max_{\mathbf{x} \in K} |f(x_1, \ldots, x_n) - I_{A',A}(x_1, \ldots, x_n)| < \epsilon. \tag{iii}$$

That is to say, $f(\mathbf{x})$ can be approximated by the finite integral $I_{A',A}(\mathbf{x})$ uniformly on K. The integrand of $I_{A',A}(\mathbf{x})$ can be replaced by the real part and is continuous on $[-A', A'] \times \cdots \times [-A, A] \times K$, so by Lemma 2, $I_{A',A}(\mathbf{x})$ can be approximated by the Riemann sum uniformly on K.

Since

$$\psi\left(\sum_{i=1}^{n} x_i w_i - w_0\right) = \phi\left(\sum_{i=1}^{n} w_i x_i/\delta - w_0 + \alpha\right) - \phi\left(\sum_{i=1}^{n} w_i x_i/\delta - w_0 - \alpha\right),$$

the Riemann sum can be represented by a three-layer network. Therefore $f(\mathbf{x})$ can be represented approximately by the three-layer networks. q.e.d.

Proof of Theorem 2

If $k = 3$, set $f: \mathbf{x} = (x_1, \ldots, x_n) \to (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))$ and apply Theorem 1 to each $f_i(\mathbf{x})$.

For the general case, we first remark that a $k(>3)$-layer network can be represented by the composition of k-2 three-layer networks and using the realization of identity mapping by three-layer network,

Proof of Corollary 1

In the expression $f: \mathbf{x} \to (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))$, we extend $f_i(\mathbf{x})$ to functions which take value zero on $\mathbf{R}^n - K$. We also denote these by $f_i(\mathbf{x})$ $(i = 1, \ldots, m)$. We can approximate $f_i(\mathbf{x})(i = 1, \ldots, m)$ by C^∞-functions with compact support by operating mollifier $\rho_\alpha *$ on f_i and apply theorem 2 to $\rho_\alpha * f_i$. q.e.d.

The above proof of the theorem 2 for the case $k > 3$ gives only trivial approximate realizations of given mappings by k-layer networks. Therefore, we shall give a different proof for the case $k = 4$, by using the Kolmogorov–Arnold–Sprecher theorem, which gives nontrivial realizations of continuous mappings.

8. KOLMOGOROV–ARNOLD–SPRECHER'S THEOREM

Let $I = [0, 1]$ denote the closed unit interval, $I^n = [0, 1]^n (n \geq 2)$ the Cartesian product of I.

In his famous thirteenth problem, Hilbert conjectured that there are analytic functions of three variables which cannot be represented as a finite superposition of continuous functions of only two arguments. Kolmogorov (1957) and Arnold refuted this conjecture and proved the following theorem.

Theorem (Kolmogorov)

Any continuous functions $f(x_1, \ldots, x_n)$ of several variables defined on $I^n (n \geq 2)$ can be represented in the form

$$f(\mathbf{x}) = \sum_{j=1}^{2n+1} \chi_j \left(\sum_{i=1}^{n} \psi_{ij}(x_i)\right),$$

where χ_j, ψ_{ij} are continuous functions of one variable and ψ_{ij} are monotone functions which are not dependent on f.

Sprecher (1965) refined the above theorem and obtained the following:

Theorem (Sprecher)

For each integer $n \geq 2$, there exists a real, monotone increasing function $\psi(x)$, $\psi([0, 1]) = [0, 1]$, dependent on n and having the following property:
For each preassigned number $\delta > 0$ there is a rational number ϵ, $0 < \epsilon < \delta$, such that every real continuous

function of n variables, $f(\mathbf{x})$, defined on I^n, can be represented as

$$f(\mathbf{x}) = \sum_{j=1}^{2n+1} \chi\left[\sum_{i=1}^{n} \lambda^i \psi(x_i + \epsilon(j-1)) + j - 1\right],$$

where the function χ is real and continuous and λ is an independent constant of f.

Hecht-Nielsen (1987) pointed out that this theorem means that any continuous mapping $f:\mathbf{x} \in I^n \to (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x})) \in \mathbf{R}^m$ is represented by a form of four-layer neural network with hidden units whose output functions are ψ, $\chi_i (i = 1, \ldots, m)$, where ψ is used for the first hidden layer, χ_i is given by Sprecher's theorem for $f_i(\mathbf{x})$ and $\chi_i (i = 1, \ldots, m)$ are used for the second hidden layer.

9. ALTERNATIVE PROOF OF THEOREM 2 FOR THE CASE $k = 4$

In section 8, we reviewed Kolmogorov's theorem and its refinement from the point of view of neural networks. The Kolmogorov–Arnold–Sprecher theorem and the following proposition are used to prove our theorem 2 for the case $k = 4$. This proposition is a special case (one variable case) of theorem 1 in Section 3.

Proposition

Let $g(x)$ be a continuous function on $\mathbf{R}$ and $\phi(x)$ a bounded and monotone increasing continuous function. For an arbitrary compact subset (bounded closed subset) K of $\mathbf{R}$ and an arbitrary $\epsilon > 0$, there are an integer N and real constants a_i, b_i, $c_i (i = 1, \ldots, N)$ such that

$$\left| g(x) - \sum_{i=1}^{N} c_i \phi(a_i x + b_i) \right| < \epsilon$$

holds on K.

In the appendix, we shall state the direct proof of the above proposition by a different method without using Fourier transforms under the additional condition that $\phi(x)$ has a weak derivative which is summable.

Next we prove theorem 2 for the case $k = 4$ by using the Kolmogorov–Arnold–Sprecher theorem and the above proposition.

Proof. We may suppose that $K = [0, 1]^n$, because $f_p(\mathbf{x})(p = 1, \ldots, m)$ can be extended continuous functions with compact supports. We apply Sprecher's theorem to $f_p(\mathbf{x})(p = 1, \ldots, m)$ and represent $f_p(\mathbf{x})$ by the form

$$f_p(\mathbf{x}) = \sum_{j=1}^{2n+1} \chi_p\left[\sum_{i=1}^{n} \lambda^i \psi(x_i + \bar{\epsilon}(j-1)) + j - 1\right]$$

$(p = 1, \ldots, m)$, where λ and $\bar{\epsilon}$ are constants. We apply our proposition to functions χ_p, ψ, and approximate these functions using a sigmoid function ϕ.

Let $K_j (j = 1, \ldots, 2n + 1)$ be the images of $[0, 1]^n$ by mappings

$$\tau_j : \mathbf{x} \longrightarrow \sum_{i=1}^{n} \lambda^i \psi(x_i + \bar{\epsilon}(j-1)) + j - 1 \qquad (j = 1, \ldots, 2n+1)$$

and set $K = \cup K_j$. Take $\delta > 0$ and the closure K_δ of δ neighborhood of K. Continuous functions χ_p $(p = 1, \ldots, m)$ are approximated by

$$\chi_{p,N}(x) = \sum_{i=1}^{N} c_{i\cdot N} \phi(a_{i,N} x + b_{i,N}) \tag{9}$$

so that

$$|\chi_p(x) - \chi_{p,N}(x)| < \epsilon/(4n+2)(p = 1, \ldots, m) \tag{10}$$

on K_δ. As $\chi_{p,N}(x)$ are uniformly continuous on K_δ, sufficiently small η can be taken so that if $|x - y| < \eta (x, y \in K_\delta)$ then $|\chi_{p,N}(x) - \chi_{p,N}(y)| < \epsilon/(4n + 2)(p = 1, \ldots, m)$.

We apply our lemma to τ_j and approximate τ_j on $[0, 1]^n$ by $\tau_{j,N'}$ so that

$$|\tau_j(\mathbf{x}) - \tau_{j,N'}(\mathbf{x})| < \min(\eta, \delta), \tag{11}$$

where $\tau_{j,N'}(x)(j = 1, \ldots, m)$ are defined as follows: We approximate $\psi(x)$ by

$$\psi_{N'}(x) = \sum_{i=1}^{N'} \tilde{c}_i \phi(\tilde{a}_i x + \tilde{b}_i) \tag{12}$$

on $2n\bar{\epsilon}$ neighborhood of $[0, 1]$ and set

$$\tau_{j,N'}(\mathbf{x}) = \sum_{i=1}^{n} \lambda^i \psi_{N'}(x_i + \bar{\epsilon}(j-1)) + j - 1 \tag{13}$$

so that the above inequality (11) is satisfied. Using a transformation

$$\begin{aligned} &\sum_{j=1}^{2n+1} \chi_p[\tau_j(\mathbf{x})] - \sum_{j=1}^{2n+1} \chi_{p,N}[\tau_{j,N'}(\mathbf{x})] \\ &= \sum_{j=1}^{2n+1} \chi_p[\tau_j(\mathbf{x})] - \sum_{j=1}^{2n+1} \chi_{p,N}[\tau_j(\mathbf{x})] \\ &+ \sum_{j=1}^{2n+1} \chi_{p,N}[\tau_j(\mathbf{x})] - \sum_{j=1}^{2n+1} \chi_{p,N}[\tau_{j,N'}(\mathbf{x})], \end{aligned}$$

it is seen that $f_p(\mathbf{x})(p = 1, \ldots, m)$ are approximated by

$$\sum_{j=1}^{2n+1} \chi_{p,N}[\tau_{j,N'}(\mathbf{x})] \quad (p = 1, \ldots, m)$$

on $[0, 1]^n$ so that the errors are less than ϵ. Looking at the form of this approximation, the theorem is obtained. q.e.d.

10. NEURAL NETWORK AND INFORMATION PROCESSING IN THE BRAIN

In the Rumelhart–Hinton–Williams multilayer neural network, input and output values of each unit correspond to pulse-frequencies in a neuron and thus each unit, disregarding time characteristics, is a very simple model of the neuron. When a neural network is implemented for pattern recognition in engineering fields, output units correspond to gnostic cells in the brain.

The approximate realization of continuous mappings using neural networks, which are simple models of the neural system, suggest that there are several gnostic cells in the brain. It also shows the possibility of revealing information processing in the brain through neural network approaches.

11. SUMMARY

We proved the approximate realization theorem of continuous functions by three-layer networks. This theorem leads to the approximate realization theorem of continuous mappings by $k(\geq 3)$-layer networks and we showed that any mapping whose components are summable on compact subset, can be approximately represented by $k(\geq 3)$-layer networks in the sense of L^2-norm. We also showed an alternative proof of the theorem for the case $k = 4$ by using the Kolmogorov–Arnold–Sprecher theorem and a proposition which is a special case of the three-layer case. We consider that one of the problems of analyzing neural network capabilities is solved in the form of the existence theorem of networks which are approximately capable of representing any mapping given.

Presently, for application of neural networks to pattern recognition or related engineering fields, up to four-layer networks are used (Waibel, Hanazawa, Hinton, Shikano, & Lang, 1988; Tamura & Waibel, 1988). The theorems proved here provide that the mathematical base and their use would be fundamental in further discussions of neural network system theory.

REFERENCES

Amari, S. (1967). A theory of adaptive pattern classifiers, *IEEE Transactions on Electronic Computers*, **EC-16**, 299–307.

Duda, R. O., & Fossum, H. (1966). Pattern classification by iteratively determined linear and piecewise linear discriminant functions. *IEEE Transactions on Electronic Computers*, **EC-15**, 220–232.

Gel'fand, I. M., & Shilov, G. E. (1964). *Generalized functions*, (Vol. 1, Chap. 1). New York: Academic Press.

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

Huang, W. Y., & Lippmann, R. P. (1987). Neural net and traditional classifiers. In D. Z. Anderson (Ed.), *Neural information processing Systems, Denver, Colorado*, 1987 (pp. 387–396). New York: American Institute of Physics.

Irie, B., & Miyake, S. (1988). Capabilities 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. *Doklady Akademii Nauk SSSR*, **144**, 679–681; *American Mathematical Society Translation*, **28**, 55–59 [1963].

Lippmann, R. P. (1987, April). An introduction to computing with neural nets. *IEEE ASSP Magazine*, **4**, pp. 4–22.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the idea immanent in nervous activity. *Bulletin of Mathematical Biophysics*, **5**, 115–133.

Poggio, T. (1983). Visual algorithms. In O. J. Braddick & A. C. Sleigh (Eds.), *Physical and biological processing of images* (pp. 128–135). New York: Springer-Verlag.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by error propagation. In D. E. Rumelhart, J. L. McClelland and the PDP Research Group (Eds.), *Parallel distributed processing* (Vol. 1, pp. 318–362). Cambridge, MA: MIT Press.

Sejnowski, T. J., & Rosenberg, C. R. (1987). Parallel networks that learn to pronounce English text. *Complex Systems*, **1**, 145–168.

Sprecher, D. A. (1965). On the structure of continuous functions of several variables. *Transactions of the American Mathematical Society*, **115**, 340–355.

Tamura, S. and Waibel, A. (1988). Noise reduction using connectionist models. 1988 *International Conference on Acoustic, Speech, and Signal Processing*, pp. 553–556.

Uesaka, Y. (1971). Analog perceptrons: On additive representation of functions. *Information and Control*, **19**, 41–65.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. (1988). Phoneme recognition: neural networks vs. hidden Markov models. 1988 *International Conference on Acoustic, Speech, and Signal Processing*, pp. 107–110.

Wieland, A., & Leighton, R. (1987). Geometric analysis of neural network capabilities, *IEEE First International Conference on Neural Networks*, **3**, 385–392.

Yosida, K. (1968). *Functional analysis*. New York: Springer-Verlag.

APPENDIX (DIRECT PROOF OF THE PROPOSITION IN SECTION 9 BY A DIFFERENT METHOD)

Proof. There is a continuous function $\tilde{g}(x)$ on $\mathbf{R}$ which has a compact support such that $g(x) = \tilde{g}(x)$ on K. We may prove the proposition for $\tilde{g}(x)$ and so we may initially suppose that $g(x)$ has a compact support. We may also suppose that $\phi(\infty) - \phi(-\infty) = 1$. For the arbitrary $\epsilon > 0$, we will approximate $g(x)$ on K by a summation of sigmoid functions whose variables are shifted and scaled. Initially, we can approximate $g(x)$ by a simple function (step function) $c(x)$ with compact support so that

$$|g(x) - c(x)| < \epsilon/2 \tag{A.1}$$

on $\mathbf{R}$ and whose step variances are less than $\epsilon/4$. Here $c(x)$ is represented using the Heaviside function $H(x)$ as follows:

$$c(x) = \sum_{i=1}^{N} c_i H(x - x_i).$$

For a sigmoid function $\phi(x)$, set $\phi_\alpha(x) = \phi(x/\alpha)(\alpha > 0)$. Then $\phi'_\alpha(x) = \dfrac{d}{dx}\phi_\alpha(x)$ converge to the delta function as $\alpha \to 0$. We consider the convolution $c*\phi'_\alpha(x)$ of $c(x)$ and $\phi'_\alpha(x)$. We set $2\epsilon'$ = "minimum width of steps" and obtain

$$c(x) - c*\phi'_\alpha(x) = \int_{-\infty}^{\infty} \phi'_\alpha(y)[c(x) - c(x - y)]\, dy.$$

Divide the integrand of the right term into $(-\infty, -\epsilon')$, $[-\epsilon', \epsilon']$, (ϵ', ∞) and estimate these using the properties of sigmoid functions. For example,

$$\left| \int_{-\epsilon'}^{\epsilon'} \phi_\alpha'(y)[c(x) - c(x - y)]\, dy \right| < \epsilon/4 \int_{-\infty}^{\infty} \phi_\alpha'(y)\, dy = \epsilon/4$$

and other terms will be arbitrarily small for a sufficiently small α. Therefore we obtain

$$|c(x) - c*\phi_\alpha'(x)| < \epsilon/4.$$

As $c*\phi_\alpha'(x) = c'*\phi_\alpha(x)$ and $c'(x)$ is given by

$$c'(x) = \sum_{i=1}^{N} c_i \delta(x - x_i)$$

and so, $c*\phi_\alpha'(x)$ is represented as follows:

$$c*\phi_\alpha'(x) = \sum_{i=1}^{N} c_i \phi_\alpha(x - x_i).$$

That is to say,

$$\left| c(x) - \sum_{i=1}^{N} c_i \phi_\alpha(x - x_i) \right| < \epsilon/2. \tag{A.2}$$

Using (A.1) and (A.2) we obtain

$$\left| g(x) - \sum_{i=1}^{N} c_i \phi_\alpha(x - x_i) \right| < \epsilon.$$

Here $\phi_\alpha(x - x_i) = \phi(x/\alpha - x_i/\alpha)$, so we set $a_i = 1/\alpha$, $b_i = -x_i/\alpha$ and the proposition is proved. q.e.d.