TR－A－0024

On the Approximate Realization of Continuous Mappings by Neural Networks

船橋　賢一

Ken-ichi Funahasi

1988. 5. 16

ＡＴＲ視聴覚機構研究所

On the Approximate Realization of Continuous Mappings by Neural Networks

Ken-ichi Funahashi

ATR Auditory and Visual Perception Research Laboratories
Twin 21 Bldg. MID Tower 2-1-61 Shiromi Higashi-ku
Osaka 540 Japan
Telephone +81-6-949-1835

On the Approximate Realization of Continuous Mappings by Neural Networks

Ken-ichi Funahashi

ATR Auditory and Visual Perception Research Laboratories

Abstract:

In this paper, we prove that any continuous mapping can be approximately realized by Rumelhart-Hinton-Williams' four-layer neural networks whose output functions for hidden units are sigmoid functions. We also show that for the approximate realization of continuous mapping, output functions need not always be sigmoid but can also be the sigmoid-like functions defined in this paper. This fact is proved by applying a lemma to the Kolmogorov-Arnol'd-Sprecher theorem.

Key words:

neural network, back propagation, output function, sigmoid function, sigmoid-like, hidden layer, unit, continuous mapping.

1. Introduction

Since McCulloch-Pitts(1943), there have been many studies of mathematical models of neural networks . Recently, Hopfields, 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 use of these models. In particular, the back propagation algorithm[1] proposed by Rumelhart-Hinton-Williams provides a learning rule for multi-layer 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 neural network.

Hecht-Nielsen[6] pointed out that Kolmogorov's theorem and Sprecher[4]'s refinement, 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. Poggio[5] has also pointed this out. That the output function of each unit of this network is not concrete monotonic increasing function like the sigmoid function is, however, a difficult point.

On the application to pattern recognition, Lippmann[2] asserts that arbitrary complex decision regions, including concave regions, can be formed using four-layer networks, but this is only an intuitive assertion.

In this paper, we apply a lemma to the Kolmogorov-Arnol'd-Sprecher theorem and show mathematically that any continuous mapping can be approximated by four-layer networks whose units have a sigmoid output function, except for those of the input and output layer. We also show that output functions need not always be sigmoid and that approximate realization of continuous mappings is possible using sigmoid-like functions defined below. McCulloch-Pitts showed that any logical circuit can be designed using their model. Correspondingly, our assertion especially shows that any continuous mapping can be approximately represented by the Rumelhart-Hinton-Williams multi-layer network.

2. Rumelhart-Hinton-Williams' Neural Network

The Rumelhart-Hinton-Williams multi-layer network that we consider here is a feed-forward type network with connections

between 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, weights w_i , threshold θ, and differentiable function ϕ as follows:

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

The learning rule of this network is known as the back propagation algorithm[1]. 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 output signal of the network is minimized. It is standard to use a monotonic increasing function such as the sigmoid function as each unit's output function.

3. Kolmogorov-Arnol'd-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, and $\mathbf{x}=(x_1,...,x_n)$ the points in Euclidian space $\mathbf{R}^n$.

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 [3] and Arnol'd refuted this conjecture and proved the following theorem.

Theorem(Kolmogorov).

Any continuous functions $f(x_1,...,x_n)$ 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[4] 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, $0 < \varepsilon < \delta$, such that every real continuous function of n variables, $f(x)$, defined on I^n, can be represented as

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

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

Hecht-Nielsen[6] pointed out that this theorem means that any continuous mapping $f : I^n \rightarrow R^m$ is represented by a form of four-layer neural network with hidden units whose output functions are ψ, $\chi_i (i=1,...,m)$.

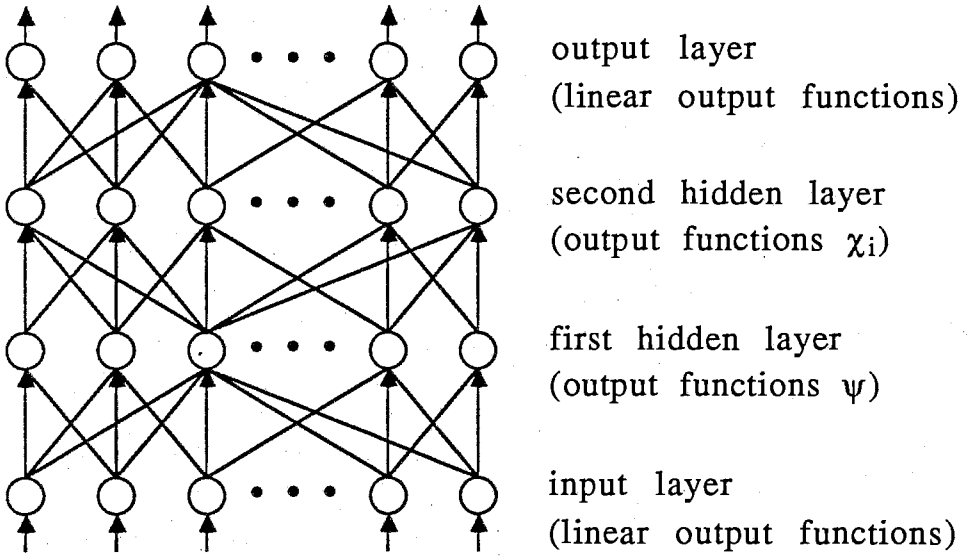

FIGURE. a four-layer network representing a continuous mapping

4. Approximate Realization of Continuous Mappings by Neural Networks

In section 3, we reviewed Kolmogorov's theorem and its refinement from the point of view of neural networks. We shall consider the possibility of representing continuous mappings by neural networks whose output functions in hidden layers are sigmoid $\phi(x) = 1/(1+\exp(-x))$, for example. 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 is used, then an input-output mapping of this network is analytic and generally cannot represent all continuous mappings.

4.1 Sigmoid-like Functions

Therefore, the possibility of an approximate realization of continuous mappings by neural networks must be discussed. We will prove that any continuous mapping can be approximated by input-output mappings of four-layer networks whose output functions for hidden layers are sigmoid, or sigmoid-like as defined below. Therefore, for the first time, we introduce the concept of sigmoid-like functions.

Definition.

A continuous function $\phi(x)$ is called *sigmoid-like* if and only if $\phi(x)$ is a bounded function where $\phi(\infty) = \lim_{x \to \infty} \phi(x)$ and $\phi(-\infty) = \lim_{x \to -\infty} \phi(x)$ exist, $\phi(\infty) - \phi(-\infty) = 1$, and its derivative $\phi'(x) = \frac{d}{dx}\phi(x)$ is summable and non-negative.

Remark. A sigmoid-like function $\phi(x)$ has the property that if we set $\phi_\varepsilon(x) = \phi(x/\varepsilon)$ $(\varepsilon > 0)$, then the derivatives $\phi_\varepsilon'(x) = (1/\varepsilon)\ \phi'(x/\varepsilon)$ converge, in the sense of the generalized function[9], to the δ function as $\varepsilon \to 0$.

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

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

Example 3. For $\phi(x)$ where $\phi(x) = 0$ $(x < 0)$, $\phi(x) = x$ $(0 \le x < 1)$ and $\phi(x) = 1$ $(x \ge 1)$, $\phi_\varepsilon'(x) = 0$ $(x < 0$ or $x \ge \varepsilon)$, $\phi_\varepsilon'(x) = 1/\varepsilon$ $(0 \le x < \varepsilon)$, and $\phi(x)$ is a sigmoid-like function.

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

Sigmoid-like functions $\phi(x)$ where $\phi(-\infty)=0$ are appropriate as output functions in the neural model because if we set $\phi_\varepsilon(x) = \phi(x/\varepsilon)$ $(\varepsilon > 0)$ then these converge to the threshold function in the McCulloch-Pitts neural model and Perceptrons as $\varepsilon \to +0$. A key to the success of the back propagation algorithm in a multi-layer network is in the use of differentiable functions as output functions. Here, the class of output functions is not limited, but this paper will be limited to the point of view that the concept of

sigmoid-like functions can be used to represent a model of the output function of a neuron.

4.2 Approximate Realization of Continuous Mappings

The following is the fundamental theorem for the approximate realization of continuous mappings.

Fundamental Theorem.

Let $\phi(x)$ be a sigmoid-like function. Any continuous mapping $f : I^n \to \mathbf{R}^m$ $(n \geq 2)$ can be approximated in the sense of uniform topology on I^n by input-output mappings of four-layer networks whose output functions for hidden layers are $\phi(x)$, and whose output functions for input and output layers are linear. That is to say, for any continuous mapping $f : I^n \to \mathbf{R}^m$ and arbitrary $\varepsilon > 0$, there is a four-layer network whose input-output mapping is given by $\tilde{f} : I^n \to \mathbf{R}^m$ and such that

$$\max_{x \in I^n} d(f(x), \tilde{f}(x)) < \varepsilon ,$$

where $d(,)$ is a metric which induces the usual topology of $\mathbf{R}^n$.

We will prove the above fundamental theorem in the next section.

McCulloch-Pitts shows that one can design any logical circuit using their model. Correspondingly, the above theorem shows that any continuous mapping can be approximately represented by multi-layer networks.

Let ψ be a strictly increasing continuous function such that $\psi((-\infty,\infty)) = (0,1)$. If the mapping f takes values in $(0,1)^m$, ψ can be taken as an output function for the output units. That is to say, the following is obtained.

Corollary.

Any continuous mapping $f : I^n \to (0,1)^m$ can be approximated by input-output mappings of four-layer neural networks whose output functions for hidden layers are a sigmoid-like function ϕ, and output functions for the output layer are ψ as stated above.

(proof)

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

q.e.d.

For the application of neural network 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. This corollary shows that if one uses four-layer networks, any decision region can be formed by a neural network. In particular, a monotonic increasing sigmoid-like function, defined above as the output function of each unit, can be chosen.

5. Proof of the Fundamental Theorem

The Kolmogorov-Arnol'd-Sprecher theorem and the following lemma are used to prove the fundamental theorem.

Lemma.

Let $g(x)$ be a continuous function on $\mathbf{R}$ and $\phi(x)$ a sigmoid-like function. For an arbitrary compact subset (bounded closed subset) K of $\mathbf{R}$ and an arbitrary $\varepsilon > 0$, there are an integer N and constant a_i, b_i, c_i $(i=1,...,N)$ such that

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

holds on K.

(*proof*)

There is a continuous function $\tilde{g}(x)$ on $\mathbf{R}$ which has a compact support such that $\tilde{g}(x) = g(x)$ on K. We may prove the lemma for $\tilde{g}(x)$ and so we may initially suppose that $g(x)$ has a compact support. For the arbitrary $\varepsilon > 0$, 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

$$| g(x) - c(x) | < \varepsilon/2 \qquad (1)$$

on $\mathbf{R}$ and whose step variances are less than $\varepsilon/4$. 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 the sigmoid-like function $\phi(x)$, set $\phi_\alpha(x) = \phi(x/\alpha)$ $(\alpha > 0)$. Then $\phi_\alpha'(x) = \frac{d}{dx}\phi_\alpha(x)$ 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 $2\varepsilon'$="minimum width of steps" and obtain

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

Divide the integrand of the right term into $(-\infty, -\varepsilon')$, $[-\varepsilon', \varepsilon']$, (ε', ∞) and estimate these using the definition of a sigmoid-like function. For example,

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

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

$$| \; c(x) - c*\phi_{\alpha'}(x) \; | < \varepsilon/4$$

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,

$$| \; c(x) - \sum_{i=1}^{N} c_i \, \phi_{\alpha}(x-x_i) \; | < \varepsilon/2 \qquad (2).$$

Using (1) and (2) we obtain

$$| \; g(x) - \sum_{i=1}^{N} c_i \, \phi_{\alpha}(x-x_i) | < \varepsilon$$

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 lemma is proved.

q.e.d.

Next we prove the fundamental theorem.

Proof of the fundamental theorem

We represent $f(\mathbf{x}) = (f_1(\mathbf{x}),...,f_m(\mathbf{x}))$. We apply Sprecher's theorem to $f_p(\mathbf{x})$ $(p=1,...,m)$ and represent $f_p(\mathbf{x})$ by the form

$$f_p(\mathbf{x}) = \sum_{j=1}^{2n+1} \chi_p \left[\sum_{i=1}^{n} \lambda^i \psi(x_i+\bar{\varepsilon}(j-1))+j-1 \right] \quad (p=1,...,m) \; ,$$

where λ and $\bar{\varepsilon}$ are constants. We apply our lemma to functions χ_p, ψ, and approximate these functions using the sigmoid-like function ϕ.

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

$$\tau_j : x \rightarrow \sum_{i=1}^{n} \lambda^i \psi(x_i+\bar{\varepsilon}(j-1))+j-1 \quad (j=1,\ldots,2n+1)$$

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,...,m) are approximated by

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

so that

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

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 $|\chi_{p,N}(x) - \chi_{p,N}(y)| < \varepsilon/(4n+2)$ (p=1,...,m).

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

$$|\tau_j(x) - \tau_{j,N'}(x)| < \min(\eta, \delta) \qquad (3),$$

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

$$\psi_{N'}(x) = \sum_{i=1}^{N'} \tilde{c}_i\, \phi(\tilde{a}_i x+\tilde{b}_i) \qquad (4)$$

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

$$\tau_{j,N'}(x) = \sum_{i=1}^{n} \lambda^i \psi_{N'}(x_i+\bar{\varepsilon}(j-1))+j-1 \qquad (5)$$

so that the above inequality(3) is satisfied. Using a transformation

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

it is seen that $f_p(x)$ (p=1,...,m) are approximated by

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

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.

6. Neural Network and Information Processing of the Brain

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

The approximate realization of continuous mappings using neural networks, which are simple models of the neural system, suggests that there are several gnostic cells in the brain and also shows the possibility of studying information processing in the brain through neural network approach.

7. Summary

As mentioned above, it is proved that any continuous mapping can be approximately represented by four-layer neural networks using the Kolmogorov-Arnol'd-Sprecher theorem and a lemma.

Presently, for application of neural networks to pattern recognition or related engineering fields, up to four-layer networks are used[7][8]. The fundamental theorem proved here provides the mathematical base and its use would be fundamental in further discussions of neural network system theory.

Acknowledgement

The author wishes to thank Drs. Y. Tohkura, T. Inui and S. Miyake for their valuable comments on the manuscript.

References

[1]Rumelhart, D.E.,and McClelland, J.L.(1986). *Parallel distributed processing*, vol.1, chap. 8, MIT Press, Cambridge MA.

[2]Lippmann, R.P.(1987). An introduction to computing with neural nets, *IEEE ASSP Magazine*, vol.4, April, 4-22.

[3]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; *Am. Math. Soc. Transl.* **28**, 55-59 (1963)

[4]Sprecher, D.A.(1965). On the structure of continuous functions of several variables, *Trans. Amer. Math. Soc..*,**115**, 340-355, March

[5]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.

[6]Hecht-Nielsen, R.(1987). Kolmogorov mapping neural network existence theorem, 1*st. International Conference on Neural Networks, IEEE*, June, 3 pp.
[7]Waibel, A., Hanazawa, T., Hinton, G. et al.(1988). Phoneme recognition: neural networks vs. hidden Markov models, *1988 International Conference on Acoustic, Speech, and Signal Processing*, 107-110
[8]Tamura, S. and Waibel, A.(1988). Noise reduction using connectionist models, *1988 International Conference on Acoustic, Speech, and Signal Processing*, 553-556
[9]Gel'fand, I.M. and Shilov, G.E.(1964). *Generalized functions*, vol.1, chap.1, Academic Press.