数理解析研究所講究録
第804巻 1992年 18-37

Neural Networks, Approximation Theory, and Dynamical Systems

Ken-ichi Funahashi and Yuichi Nakamura
船橋 賢一 中村 雄一
Department of Information and Computer Sciences
Toyohashi University of Technology

Abstract In this paper, firstly we discuss the capability problem of feedforward neural networks from the aspect of approximation theory. Secondly we prove that any finite time trajectory of a given n-dimensional dynamical system can be approximately realized by the internal state of output units of continuous time recurrent neural networks with n output units, some hidden units, and an appropriate initial conditions. The essential idea of the proof is to embed the n-dimensional dynamical system into a higher dimensional one by the approximate realization theorem of continuous mappings of three-layer neural networks. As a corollary, we also show that any continuous curve can be approximated by outputs of a recurrent neural network.

§1. Introduction

Neural networks are divided into two types namely, feedforward networks and recurrent networks from the architectural aspect. For the former networks without feedback connection, ever since the back propagation learning algorithm was proposed by Rumelhart-Hinton-Williams[19], a lot of application was made mainly to the static information processings such as pattern recognition. On the theoretical capability of this networks, Funahashi[9], Hornik-Stinchcombe-White[11] and Cybenko[8] proved mathematically that a given continuous mapping on a compact set can be realized by three-layer feedforward neural networks with any precision. In this paper we discuss the related problems from the aspect of approximation theory.

The nonlinear dynamical behavior of the latter networks is suitable for the spatio-temporal information processings. The theoretical studies for the recurrent networks

have been mainly concerned with the stability of convergence of the trajectory to the equilibrium point (e.g. Hirsch[12]). Hopfield network with symmetrical weight of connection has been applied to the content addressable memory and combimational optimization.

The learning algorithms which employed the steepest descent method for the modification of recurrent network weight have been proposed both by Williams-Zipser[25] in case of discrete time system and by Pineda[18], Perlmutter[16][17] and Sato[20] etc. for continuous time system. Sato-Murakami[21] proposed both the modified recurrent network in order to approximate the dynamical system and its learning algorithm by the use of approximate realization theorem by three-layer networks. Further they applied the algorithm to approximation of nonlinear dynamical systems. Since the network they concerned is far from the ordinary recurrent networks, the theoretical capability for the recurrent network is still opened for question. Seidl-Lorentz[22] proved the approximation theorem for the trajectory of discrete dynamical system by the use of approximate realization theorem by three-layer networks. The main goal of this study is to elucidate the theoretical capability of the continuous time recurrent networks. In this paper we will prove that the internal state of the output units of the continuous time recurrent network approximate the finite time trajectory of the dynamical system with any precision. The proof are yielded by approximate realization theorem by three-layer networks and the fundamental theorems on dynamical systems.

§ 2. What are neural networks

Neural networks we consider in this paper are not natural neural networks, and are artificial neural networks which are characterized by non-linearity and pararell prosessings for engineering systems. Neural networks are classified into two types, multilayer feedforward networks and recurrent networks in terms of architecture. There are two distinctive type of learning methods, one has supervisor node the other does not. The Hebbian learning rule is one of the unsupervised learning method whereas the back-propagation algorithm is the example of supervised one.

§ 3. Multilayer feedforward neural networks

Multilayer feedforward neural networks which are called multilayer Perceptrons consist of input , output, and some hidden layers, and have connections between layers from input to output (See Figure 2). Each layer consists of information processing unit, simply called unit which is a model of neuron. Input-output relationship of unit is given by

$$y = \phi(\Sigma w_i x_i - \theta),$$

where $\{x_i\}$ is inputs to the unit, y is output, and w_i are connection weight to the unit (See Figure 1). The function ϕ is called output function of the unit, and θ is called the threshold. Usually, a nonconstant, increasing and bounded function are used for output function ϕ , but in some case a linear function $\phi(x)=x$ is used for output layer. In the following, as output function of units, non-constant increasing bounded continuous functions are called sigmoid functions. In this case, a multilayer feedforward network with analog n-inputs and m-outputs defines a continuous mapping $f : R^n \to R^m$. The mapping f is called the input-output mapping of the network.

§ 4. Capabilities of feedforward neural networks and approximation theory

Feedforward networks which Heaviside function H(x) ($H(x)=0$, $x<0$; $H(x)=1$, $x \geqq 0$) is used as output functions of units is called Perceptron by Rosenblatt. McCulloch-Pitts showed that multilayer Perceptron with {0, 1} inputs can realize any logical function. The learning algorithm of Perceptrons without hidden layer was given by Rosenblatt. For multilayer feedforward network with differentiable output functions, back-propagation algorithm is well known as a learning algorithm of the case with hidden layers (Rumelhart-Hinton-Williams[19]).

For application to pattern recognition, feedforward networks without hidden layer can classify only linear separable categories, but computer simulation show that feedforward networks with hidden layers give good performance for several

applications. The following result gives mathematical verification for these fact (For proof, see Funahashi[9]).

Theorem 1. (Funahashi)

Let $\sigma(x)$ be a sigmoid function(i.e. a non-constant, increasing and bounded continuous function on R). Let K be a compact subset of R^n, and $f(x_1,\ldots,x_n)$ be a continuous function on K. Then, for an arbitrary $\varepsilon>0$, there exist an integer N, constants c_i, $\theta_i(i=1,\ldots,N)$ and $w_{ij}(i=1,\ldots,N;\ j=1,\ldots,n)$ such that

$$\max_{x\in K}\left| f(x_1,\cdots,x_n) - \sum_{i=1}^{N} c_i\,\sigma\left(\sum_{j=1}^{n} w_{ij}x_j - \theta_i\right)\right| < \varepsilon$$

holds.

This theorem show that three-layer feedforward neural networks whose output layer has linear units can approximate any continuous mapping $f : R^n \to R^m$ uniformly on an arbitrary compact set.

Theorem 1'

Let K be a compact subset of R^n, and $f : K \to R^m$ be a continuous mapping. Then, for an arbitrary $\varepsilon>0$, there exist an integer N, an $m\times N$ matrix A, an $N\times n$ matrix B, and an N dimensional vector θ such that

$$\max_{x\in K}\left| F(x) - A\sigma(Bx+\theta)\right| < \varepsilon$$

holds, where $\sigma : R^N \to R^N$ be a sigmoid mapping defined by

$$\sigma({}^t(u_1,\ldots,u_N)) = {}^t(\sigma(u_1),\ldots,\sigma(u_N)).$$

Similar results have been obtained by Cybenko[8] and Hornik-Stinchcombe-White[11].

From theorem 1, it is shown that three-layer networks whose output units have sigmoid output function, e.g.

$$\sigma(x)=1/(1+\exp(-x))$$

can approximate any continuous mapping $f : R^n \to (0,1)^n$ on an arbitrary compact set with any precision. Further it is shown that more layer networks have the same property.

For the proof of Theorem 1, Fourier integrals, especially Irie-Miyake's integral formula[15] and Paley-Wiener Theorem are used. Cybenko[8] showed that for output function $\sigma(x)$ which is continuous, but not increasing and

$$\lim_{x\to\infty}\sigma(x)=1 \quad , \quad \lim_{x\to-\infty}\sigma(x)=0 \ ,$$

a similar result holds, by the use of Hahn-Banach Theorem and Riesz Theorem in functional analysis (cf. Yosida[24]). Hornik-Stinchcombe-White[11] showed that, for increasing but generally discontinuous $\sigma(x)$, similar result holds by the use of a sigmoid function called cosine squashier which is defined by cosine function and Stone-Weierstrass theorem.

In the study of capability of multilayer feedforward networks stated above, feedforward networks are considered as approximator of continuous mappings, but in engineering application, networks have capability of identifying input-output relationship by learning of finite samples. This property is called the generalization of learning. Theoretical study of generalization have been begun by Baum-Haussler[2] based on the result of Vapnik-Chaevonenkis, but only Heaviside output function case are studied.

However, in the problem of function approximation, there have been remained many mathematical problems. Theorem 1 for multi-layer feedforward networks is similar to Weierstrass approximation theorem by polynomial functions, but is not similar in the following points. In Weierstrass theorem, parameters are linear, but in Theorem 1, weights w_{ij} are nonlinear parameters. From this result, in L^2-approximation by polynomials of finite degree, the best approximation exist, but in the approximation by using three-layer networks with finite hidden units the best approximation property does not hold generally. Recently, we showed that multiplier $f(x,y)=xy$ can be approximated with any precise on compact subset by three layer networks with four hidden unit whose output function is C^2(cf. Toda-Funahashi-Usui[23]). This show that when the sigmoid function $\sigma(x)=1/(1+\exp(-x))$ is used and K contains internal points, function $f(x,y)=xy$ has not best approximation in the set of input-output functions of networks with four hidden unit and liner output function, and arbitrary near functions. However it is conjectured that except particular functions, any

function have best L^2-approximation in three-layer networks with finite hidden units. And it is conjectured that four-layer networks are better than three-layer networks generally in the number of hidden units. Funahashi[9] conjectured this and Chester[5] showed a example. For solving these problem, capability of feedforward networks must be further studied from the view point of approximation theory.

§ 5. Approximate realization of identity mappings

Recently, it is studied that feedforward networks have capability of a sort of multivariate analysis by learning. One of the study is realizing identity mapping by feedforward network on the input data, and was begun from the study by Cottrell et al.[6] study of coding image by three-layer networks with few hidden units. The another is relation between pattern recognition by feedforward networks and discriminant analysis.

Theoretical study of the former problem was firstly done by Boulard-Kamp[3] in the case of three-layer networks with linear hidden units(See also Baldi-Hornik[1]).

Theorem (Boulard-Kamp)

Let $K=\{x_i\ ;\ i=1,\ldots,N\}$ be a finite set of R^n, and f be the input-output mapping of three-layer network with n inputs, n outputs and k (<n) hidden units. Then the minimum of

$$\frac{1}{N}\sum_{i=1}^{N} |x_i - f(x_i)|^2 ,$$

where $|\cdot|$ is the Euclidian norm of R^n, is equal to the mean-squared error of approximation of $\{x_i\}$ by K-L(Karhunen-Loeve) transformation with k terms (i.e. approximation by k principal components).

This theorem treats the case of three-layer networks with linear units, so the input-output mapping is composite of affine transforms. Therefore, the essential problem is to study the case of three-layer networks whose hidden units have sigmoid output functions. Funahashi[10] proved the following:

Theorem 2. (Funahashi)

Let f be the input-output mappings of three-layer networks with n inputs, n output and k(<n) hidden units such that only hidden units have nonlinear output function. Then the mean squared error between f and identity mapping id:

$$\frac{1}{N}\sum_{i=1}^{N} | x_i - f(x_i) |^2 ,$$

is greater than or equal to the error by K-L transformation with k terms. In the case of C^1-sigmoid function, the mean squared error can be approached to the k-terms approximation error.

Therefore, on the approximate realization of identity mappings by three-layer networks, the performance is lower than the K-L transfer-mation method, due to the non-linearity of hidden units. Our theorem shows that when the learning of network proceeds ideally, k-hidden units capture the k-principal component of data. This corresponds to experimental study by Cottrell-Munro[7]. To obtain better performance than K-L transformation method, we must use five-layer networks with few middle layer units. This is suggested by Theorem 1 (approximate realization theorem of continuous mappings by three-layer networks).

§ 6. Continuous time recurrent neural networks

There are two types of recurrent neural networks; discrete time neural networks and continuous time one. In this paper, we study continuous time recurrent neural networks.

The dynamics of a continuous time recurrent neural network with m units which are concerned in this paper is described by the following system of ordinary differential equations.

$$\frac{du_i(t)}{dt} = -\frac{u_i(t)}{\tau_i} + \sum_{j=1}^{m} w_{ij}\sigma(u_j(t)) + I_i(t) , \quad (i = 1, \cdots, m) \qquad (1)$$

where $u_i(t)$ is the internal state of unit i, τ_i is the time constant of unit i, w_{ij} are connection weights, $I_i(t)$ are the inputs to the system, and $\sigma(u_i(t))$ is the output of unit i. σ is called the output function and C^1-sigmoid functions (non-constant, bounded, monotone increasing functions) are used. As σ ,

$$\sigma(x)=1/(1+\exp(-x))$$

is usually used.

In the following, we deal with recurrent neural networks with the same time constant and without inputs (i.e. $I_i(t)=0$, see Figure 3). We set $u(t)={}^t(u_1(t),\ldots,u_m(t))$ and $W=(w_{ij})$ be an $m\times m$ weights matrix. Let $\sigma:R^m\to R^m$ be denoted by a sigmoid mapping

$$\sigma({}^t(u_1,\ldots,u_m))={}^t(\sigma(u_1),\ldots,\sigma(u_m)),$$

then the vector expression of (1) is

$$u'(t) = -\frac{1}{\tau}u(t) + W\sigma(u(t)) . \qquad (2)$$

§7. Approximation realization theorems of dynamical system trajectories

Let points of n-dimensional Euclidian space R^n be denoted by $x={}^t(x_1,\ldots,x_n)$ and the Euclidian norm of x defined by $|x|$.

The dynamical system on an open set of R^n means a system defied by autonomous ordinary differential equations which has solution in the open set. Let the output function σ of recurrent neural networks be a C^1-sigmoid function. As recurrent neural networks studied here have no inputs, some of units are called output units and the other are called hidden units.

In this paper, we prove the following theorems.

Theorem 3.

Let W be an open subset of R^n, $F : W\to R^n$ be a C^1-mapping, and K be a compact subset of W. There is a subset $V\subset K$ such that any solution x(t) with initial value x(0) in V of an ordinary differential equation

$$x' = F(x) \quad , \quad x(0)\in V \qquad (3)$$

is defined on $I=[0,T]$ ($T<\infty$) and x(t) is included in K for any $t\in I$. Then, for an arbitrary $\varepsilon>0$, there exist an integer N and a recurrent neural network with n output units and N hidden units such that for solution x(t) satisfying (3) and an appropriate initial state of the network,

$$\max_{t\in I} |x(t) - u(t)| < \varepsilon ,$$

where $u(t)={}^t(u_1(t),\ldots u_n(t))$ is the internal state of output units of the network.

As a corollary of the above theorem, we obtain the following:

Theorem 3'.

Let $W\subset R^n$ and $F : W\rightarrow R^n$ be same above, and suppose that $x'=F(x)$ defines a dynamical system on W. Let K be a compact subset of W and we consider trajectories of the system on interval $I=[0,T]$. Then, for an arbitrary $\varepsilon >0$, there exist an integer N and a recurrent neural network with n output units and N hidden units such that for any trajectory $\{ x(t); 0\leqq t\leqq T\}$ of the system with initial value $x(0)\in K$ and an appropriate initial state of the network,

$$\max_{t\in I} |x(t) - u(t)| < \varepsilon ,$$

where $u(t)={}^t(u_1(t),\ldots u_n(t))$ is the internal state of output units of the network.

We obtain the following:

Corollary 1.

Let σ be a strictly increasing C^1-sigmoid function such that $\sigma(R) = (0,1)$. Let W be an open subset of $(0,1)^n$, $F : W\rightarrow (0,1)^n$ be a C^1-mapping, and suppose that $x'=F(x)$ defines a dynamical system on W. Let K be a compact subset of W and we consider trajectories of the system on interval $I=[0,T]$. Then, for an arbitrary $\varepsilon >0$, there exist an integer N and a recurrent neural network with n output units and N hidden units such that for any trajectory $\{x(t); 0\leqq t\leqq T\}$ of the system with initial value $x(0)\in K$ and an appropriate initial state of the network,

$$\max_{t\in I} |x(t) - y(t)| < \varepsilon ,$$

where $y(t)={}^t(y_1(t),\ldots,y_n(t))$ is the output of the recurrent network with the sigmoid output function σ.

As a corollary of Theorem 1, we also obtain the following:

Theorem 4.

Let be $f : I=[0,T] \to R^n$ be a continuous curve, where $0<T<\infty$. Then, for an arbitrary $\varepsilon > 0$, there exist an integer N and a recurrent networks with n outputs and N hidden units such that

$$\max_{t \in I} | f(t) - u(t) | < \varepsilon ,$$

where $u(t)={}^t(u_1(t),\ldots u_n(t))$ is the internal state of output units of the network.

Samely as Corollary 1, the following corollary can be proved from Theorem 4.

Corollary 2.

Let σ be a strictly increasing C^1-sigmoid function such that $\sigma(R) = (0,1)$. Let be $f : I=[0,T] \to (0,1)^n$ be a continuous curve, where $0<T<\infty$. Then, for an arbitrary $\varepsilon > 0$, there exist an integer N and a recurrent neural network with n output units and N hidden units such that

$$\max_{t \in I} | f(t) - y(t) | < \varepsilon ,$$

where $y(t)={}^t(y_1(t),\ldots,y_n(t))$ is the output of the recurrent network with the sigmoid output function σ.

§ 8. Preliminaries

In the following, we state the basic facts of theory of dynamical systems which are used in the proofs of our theorems(See e.g. Hirsch-Smale[14]).

Let W be a open subset of R^n. A mapping $F : W \to R^n$ is said to be Lipschitz on W if there exists a constant L such that

$$| F(x) - F(y) | \leq L | x - y |$$

for all $x, y \in W$. We call L a Lipschitz constant for F. We call F locally Lipschitz if each point of W has a neighborhood W_0 in w such that the restriction $F \mid W_0$ is

Lipschitz.

Lemma 1.

Let the mapping $F : W \to R^n$ be C^1. Then F is locally Lipschitz. Moreover, if $A \subset W$ is compact, then the restriction $F \mid A$ is Lipschitz.

(For proof, see Hirsch-Smale [14], chap.8, §3.Lemma and §6. Lemma)

Lemma 2.

Let $F:W \to R^n$ be a C^1-mapping and $x_0 \in W$. Then there is some $a > 0$ and a unique solution $x:(-a,a) \to W$ of the differential equation

$$x' = F(x)$$

satisfying the initial condition $x(0)=x_0$.

(For proof, see Hirsch-Smale [14], chap 8, §2. Theorem 1)

Lemma 3.

Let W be an open subset of R^n and $F : W \to R^n$ be a C^1-mapping. Let x(t) be a solution on a maximal open interval $J=(\alpha, \beta) \subset R$ with $\beta < \infty$. Then given any compact subset $K \subset W$, there is some $t \in (\alpha, \beta)$ with $x(t) \notin K$.

(For proof, see Hirsch-Smale[14], chap. 8, §5. Theorem)

Lemma 4.

Let $F : R^n \to R^n$ be a bounded C^1-mapping. Then, the differential equation

$$x' = -\frac{1}{\tau}x + F(x)$$

where $\tau > 0$, has an unique solution on $[0, \infty)$.

(proof)

From assumption, we can take a constant $M > 0$ such that

$$|F_i(x)| \leq M \qquad (\forall i = 1,\cdots,n)$$

for all $x \in R^n$. By comparing the solution x(t) with solutions of the following equations

$$y' = -\frac{1}{\tau}y + M$$

$$y' = -\frac{1}{\tau}y - M$$

we can easily show that

$$|x_i(t)| \leq \max\{|x_i(0)|, \tau M\} = C_i .$$

We set $C = \max\{C_i\}$, then the solution $x(t)$ satisfy

$$|x(t)| \leq \sqrt{n}\, C$$

on the existing interval of the solution. From Lemma 3, $x(t)$ exists on the interval $[0, \infty)$. q.e.d.

This Lemma 4 guarantees that the equation (1) of recurrent neural network has an unique solution on $[0, \infty)$, because the output function σ is bounded and C^1.

Lemma 5.

Let $F, \tilde{F} : W \to R^n$ be Lipschitz continuous mappings and L be a Lipschitz constant of F. Suppose that for all $x \in W$,

$$|F(x) - \tilde{F}(x)| < \varepsilon.$$

If $x(t)$, $y(t)$ are solutions of

$$x' = F(x)$$

$$y' = \tilde{F}(y)$$

respectively on some interval J such that $x(t_0) = y(t_0)$, then

$$|x(t) - y(t)| \leq \frac{\varepsilon}{L}(\exp L|t - t_0| - 1)$$

for all $t \in J$. (For proof, see Hirsch-Smale[14], chap. 15, § 1. Theorem 3)

§ 9. Proof of the theorems

Under the above preliminaries, we will prove theorems stated in section 7.

Proof of Theorem 3.

Step 1.

For given $\varepsilon > 0$, we choose η so that $0 < \eta < \min\{\varepsilon, \lambda\}$, where λ is the distance between K and the boundary ∂W of W. We set

$$K_\eta = \{ x \in \mathbf{R}^n ; \exists z \in K, |x - z| \leq \eta \},$$

then K_η is a compact subset of W, because K is compact. Therefore, by Lemma 1, F is Lipschitz on K_η. We also choose $\varepsilon_1 > 0$ so that

$$\varepsilon_1 < \frac{\eta L_F}{2(\exp L_F T - 1)},$$

where L_F is a Lipschitz constant of $F \mid K_\eta$.

By the approximate realization theorem of continuous mappings by three-layer neural networks, there exist an integer N, an $n \times N$ matrix A, an $N \times n$ matrix B and an N-dimensional vector θ such that

$$\max_{x \in K_\eta} \left| F(x) - A\sigma(Bx + \theta) \right| < \frac{\varepsilon_1}{2}. \tag{4}$$

We define a C^1-mapping $\tilde{F} : R^n \to R^n$

$$\tilde{F}(x) = -\frac{1}{\tau}x + A\sigma(Bx + \theta), \tag{5}$$

where τ is choose large enough so that the following conditions are satisfied:

(a) $\forall x \in K_\eta ; \left|\frac{x}{\tau}\right| < \frac{\varepsilon_1}{2}$

(b) $\left|\frac{\theta}{\tau}\right| < \frac{\eta L_G}{2(\exp L_G T - 1)}$ and $\left|\frac{1}{\tau}\right| < \frac{L_G}{2}$,

where L_G is a constant and $L_G/2$ is a Lipschitz constant for the mapping $W\sigma : R^{n+N} \to R^{n+N}$ which will be defined later (W is defined by A and B).

Then, by (4) and (5)

$$\max_{x \in K_\eta} \left| F(x) - \tilde{F}(x) \right| < \varepsilon_1 \tag{6}$$

holds. We set x(t) and $\tilde{x}(t)$ the solutions of the following equations

$$x' = F(x),$$
$$\tilde{x}' = \tilde{F}(\tilde{x}),$$

with initial condition $x(0) = \tilde{x}(0) = x_0 \in V$, respectively. Then, by Lemma 5, for any $t \in I$,

$$\left| x(t) - \tilde{x}(t) \right| \le \frac{\varepsilon_1}{L_F}(\exp L_F t - 1) \le \frac{\varepsilon_1}{L_F}(\exp L_F T - 1).$$

Therefore, by the condition of ε

$$\max_{t \in I} \left| x(t) - \tilde{x}(t) \right| < \frac{\eta}{2} \tag{7}$$

holds.

Step 2.

We consider the following dynamical system defined by $\tilde{F}$ stated in step 1.

$$x' = -\frac{1}{\tau}x + A\sigma(Bx + \theta) . \tag{8}$$

We set $z=Bx+\theta$, then

$$z' = Bx' = -\frac{1}{\tau}z + C\sigma(z) + \frac{1}{\tau}\theta ,$$

where $C=BA$ and C is $N\times N$ matrix. We set

$$xz = {}^t(x_1,\ldots,x_n,z_1,\ldots,z_N)$$

and we define a mapping $G : R^{n+N}\to R^{n+N}$ by

$$G(xz) = -\frac{1}{\tau}xz + W\sigma(xz) + \frac{1}{\tau}\theta_1 \tag{9}$$

where W is $(n+N)\times(n+N)$ matrix and θ_1 is (n+N) matrix defined by

$$W=\begin{pmatrix} 0 & A \\ 0 & C \end{pmatrix} \quad , \quad \theta_1=\begin{pmatrix} 0 \\ \theta \end{pmatrix}$$

respectively. Then, by Lemma 2, first n components of the solution of the equation of

$$(xz)' = G(xz) \quad , \quad z(0) = Bx(0) + \theta$$

is equivalent to the solution of the system (8).

Now, we define a mapping $\tilde{G} : R^{n+N}\to R^{n+N}$ by the use of τ and W stated above, as the following:

$$\tilde{G}(xz) = -\frac{1}{\tau}xz + W\sigma(xz) . \tag{10}$$

Then the dynamical system defined by $\tilde{G}$,

$$(xz)' = -\frac{1}{\tau}xz + W\sigma(xz) \tag{11}$$

is realized by a recurrent neural network, if we set x(t) the internal state of n output unit and z(t) the internal state of N hidden unites. As G and $\tilde{G}$ are C^1-mappings, and $\sigma'(x)$ is bounded function, so the mapping $xz \to W\sigma(xz)$ is Lipschitz on R^{n+N} and let $L_G/2$ be its Lipschitz constant. Then L_G is a Lipschitz constant of G as $L_G/2$ is Lipschitz constant of $-xz/\tau$ by condition (b) of τ .

Using (9), (10) and the condition (b) of τ , we see that for any $xz\in R^{n+N}$

$$|G(xz) - \tilde{G}(xz)| = \left|\frac{\theta}{\tau}\right| < \frac{\eta L_G}{2(\exp L_G T - 1)}$$

holds. Therefore we set xz(t), uh(t) the solutions of the following equations

respectively:

$$(\tilde{x}z)' = G(\tilde{x}z)\ ,\ \begin{cases}\tilde{x}(0) = x_0 \in V\\ z(0) = Bx_0 + \theta\end{cases}$$

$$(uh)' = \tilde{G}(uh)\ ,\ \begin{cases}u(0) = x_0 \in V\\ h(0) = Bx_0 + \theta\end{cases}$$

then, by Lemma 5 we see

$$\max_{t\in I} |\tilde{x}z(t) - uh(t)| \le \frac{\eta}{2} \qquad (12)$$

holds, where $\tilde{x}(t)$ is same as $\tilde{x}(t)$ on (7).

Step3.

Using (7) and (12) stated above, for a given $\varepsilon > 0$, we can construct a recurrent neural networks with internal state uh(t) by τ and W stated above. For x(t) satisfying (3), if we set the initial state of the network by

$$u(0) = x(0) \quad \text{and}$$

$$h(0) = Bx(0) + \theta\ ,$$

we see

$$\max_{t\in I} |x(t) - u(t)| \le \frac{\eta}{2} + \frac{\eta}{2} = \eta < \varepsilon$$

q.e.d.

Remark

The recurrent network constructed in the above proof has connections between hidden units and have connections from hidden units to outputs units, but have no connection from output units to hidden units. It is obvious from the method of the proof that we can construct a recurrent network with very small connections from output units to hidden units which satisfies the condition of the Theorem.

Proof of Theorem 3'.

Because the flow $\phi_t(x)$ of the dynamical system is a continuous mapping $R \times W \to W$ ($(t,x) \to \phi_t(x)$) (see Hirsch-Smale[14]), the set of trajectories on time interval I

whose initial points are in the compact set K:

$$\tilde{K} = \{ x(t) \in \mathbf{R}^n ; x(0) \in K , 0 \leq t \leq T \}$$

is a compact subset of W. By corresponding K and $\tilde{K}$ to V and K in Theorem 3 respectively, our Theorem is proved. q.e.d.

Proof of Corollary 1.

By continuity of $\sigma^{-1}: (0,1) \to R$, $W_1 = \sigma^{-1}(W)$ is open subset of R^n, and $K_1 = \sigma^{-1}(K)$ is compact subset of W_1. For $x \in (0,1)^n$, let $u \in R^n$ be denoted by

$$^t(u_1, \ldots, u_n) = \sigma^{-1}(^t(x_1, \ldots, x_n)).$$

Then, by the sigmoid mapping σ, the given dynamical system $x' = F(x)$ on W is transformed to a dynamical system defined by

$$\frac{du_i}{dt} = \frac{1}{\sigma'(u_i)} F_i(\sigma(u_1), \cdots, \sigma(u_n)) \qquad (i = 1, \cdots, n)$$

on $W_1 \subset R^n$. From this fact, our Corollary can be easily proved by the use of Theorem 3'. q.e.d.

Proof of Theorem 4.

Using a mollifier, we can take C^∞-curve $\tilde{f} : (-\delta, T+\delta) \to R^n$ for some $\delta > 0$ such that

$$\max_{t \in I} | f(t) - \tilde{f}(t) | < \frac{\varepsilon}{2}.$$

We set $g(t) = (\tilde{f}(t), t) \in R^n \times R = R^{n+1}$ for $t \in [0,T]$, then g is an injective mapping and so there exists a one-dimensional compact C^∞-submanifold M of R^{n+1} such that $g([0,T]) \subset M$.

Taking a tubular neighborhood V of M in R^{n+1} (see M.W.Hirsh[13] Theorem 5.1), we can easily construct of a system of ordinary differential equations $x' = F(x)$ defined in V such that $F \in C^\infty$ on V and $g([0,T])$ is a part of a trajectory of the system with $x(0) = g(0)$. Using Theorem 3, there exists a recurrent networks with n+1 output units such that

$$\max_{t \in I} | g(t) - \tilde{u}(t) | < \frac{\varepsilon}{2},$$

where $\tilde{u}(t) = ^t(u_1(t), \ldots u_{n+1}(t))$ is the internal state of output units. Considering the

projection $\tilde{f}(t)$ of $g(t)$ to R^n by $\pi : R^{n+1} \to R^n$ $({}^t(x_1, \ldots x_{n+1}) \to {}^t(x_1, \ldots x_n))$, we obtain a recurrent network with n output units whose internal states $u(t) = {}^t(u_1(t), \ldots u_n(t))$ satisfy

$$\max_{t \in I} |\tilde{f}(t) - u(t)| < \frac{\varepsilon}{2}.$$

Therefore we obtain

$$\max_{t \in I} |f(t) - u(t)| < \varepsilon.$$

q.e.d.

§ 10. Summary

Firstly, we discussed the capability of multilayer feedforward networks from the viewpoint of approximation theory and discussed related problem on function approximation.

Secondary, We proved that the finite time trajectories of a given n-dimensional dynamical system are approximated by the internal states of output units of a recurrent neural networks with n output units, N hidden units and appropriate initial states. The important point of the proof is the use of the approximate realization theorem of continuous mappings by three-layer feedforward neural networks to embed the given dynamical system into a higher dimensional dynamical system which defines a recurrent neural network. We consider that one of the capability problems of continuous time recurrent neural networks is solved in a form of existence theorem of networks which approximates trajectories of a given dynamical system. Our theorem are the first step to studying the capability problems of continuous time recurrent neural networks unlimately.

Acknowledgment

We would like to thank Dr. M. Sakakibara for critical reading of the manuscript.

References

[1]Baldi, P. and Hornik, K., "Neural networks and principal component analysis:

Learning from examples without local minima", Neural Networks, vol.2, no.1, pp.53-58 (1989).

[2]Baum, E.B. and Haussler, D., "What size net gives valid generalization ?", Neural Computation, vol.1, pp.151-160 (1980).

[3]Bourlard, H. and Kamp, Y., "Auto-association by multilayer perceptrons and singular value decomposition", Biological Cybernetics, vol.59, pp.291-294 (1988).

[4]Cheney, E.W., " Introduction to Approximation Theory", McGrow-Hill, Inc. (1966).

[5]Chester, D.L., "Why two hidden layers are better than one", Proc. of IJCNN, Washington, D.C., January 15-19, vol.1, pp.265-268 (1990).

[6]Cottrell, G.W., Munro, P. and Zipser, D., "Image compression by back propagation: An example of extensional programming", N.E.Sharkey(Ed.), Advances in Cognitive Science, vol. 3, Norwood NJ, Ablex

[7]Cottrell, G.W. and Munro, P., "Principal component analysis of image via back propagation", SPIE 1001 Visual Communications and Image Processing '88, pp.1070-1076 (1988).

[8]Cybenko, G., "Approximation by superpositions of a sigmodial function", Mathematics of Control. Signals and Systems, vol.2,pp303-314 (1989).

[9]Funahashi, K., "On the approximate realization of continuous mappings by neural networks", Neural Networks, vol. 2.3, pp.183-191 (1989).

[10]Funahashi, K., "Approximate realization of identity mappings by three-layer neural networks", Electronics and Communications in Japan, Part 3, vol.73, No.11, pp.61-68, Scripta Technica, Inc. (1990).

(Translated from Denshi Joho Tsushin Gakkai Ronbunshi, vol.73-A, No.1, pp.139-145, January (1991).)

[11]Hornik, K., Stinchcombe, M. and White, H., "Multilayer feedforward networks are universal approximations", Neural Networks. vol.2, pp359-366 (1989).

[12]Hirsch, M.W., "Convergent activation dynamics continuous time networks", Neural Networks. vol.2,pp331-349 (1989).

[13]Hirsch, M.W., "Differential Topology", Springer-Verlag, (1976).

[14]Hirsch, M.W. and Smale, S., "Differential Equations, Dynamical Systems, and Liner Algebra", Academic Press, Inc., (1974).

[15]Irie, B. and Miyake, S., "Capabilities of three-layered Perceptrons", Proc. of ICNN, vol.1, pp.641-648 (1988).

[16]Pearlmutter, B.A., "Learning state space trajectories in recurrent neural networks", Proc. of IJCNN, vol.2,pp365-372 (1989).

[17]Pearlmutter, B.A., "Learning state space trajectories in recurrent neural networks", Neural Computation, vol.1, pp263-269 (1989).

[18]Pineda, F.J., "Generalization of backpropagation to recurrent neural networks", Physical Review Letters, vol.18,pp2229-2232 (1987).

[19]Rumelhart, D.E., Hinton, G.E. and Williams, R.J., "Learning internal representations by error propagation", In D.E.Rumelhart, J.L.McClelland and the PDP Research Group(Eds.), Parallel Distributed Processing, vol. 1, chap. 8, pp.318-362. Cambridge, MA:MIT Press (1986).

[20]Sato, M., "A learning algorithm to teach spatio-temporal patterns to recurrent neural networks", Biological Cybernetics, vol.1, pp256-263 (1990).

[21]Sato, M. and Murakami, Y., "Learning nonlinear dynamics by recurrent neural networks", In H.Kawakami(Ed.), Proc. of the Symposium on "Some Problems on the Theory of Dynamical Systems in Applied Sciences", Advanced Series in Dynamical Systems vol.10,World Scientific, Singapore (1991).

[22]Seidl, D.R. and Lorenz, R.D., "A structure by which a recurrent neural network can approximate a nonlinear dynamic system", Proc. of IJCNN,Seattle,WA, vol.2, pp709-714 (1991).

[23]Toda, N., Funahashi, K. and Usui, S., "Polynomial functions can be realized by finite size multilayer feedforward neural networks", Proc. of IJCNN, Singapore, pp.343-348 (1991).

[24]Yosida, K., "Functional Analysis", New York: Springer-Verlag (1968).

[25]Williams, R.J., and Zipser, D., "A learning algorithm for continually running fully recurrent neural networks", Neural Computation,vol.1,pp256-263 (1990).

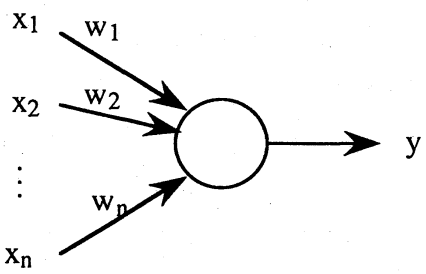

Figure 1. Unit

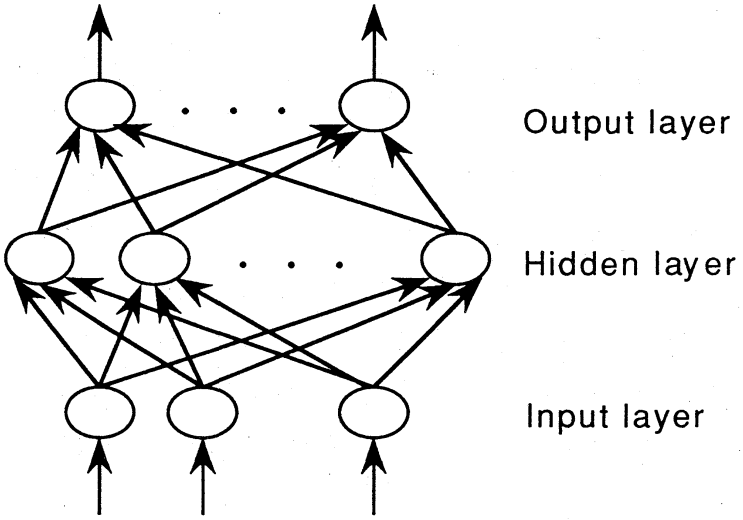

Figure 2. Three-layer feedforward network

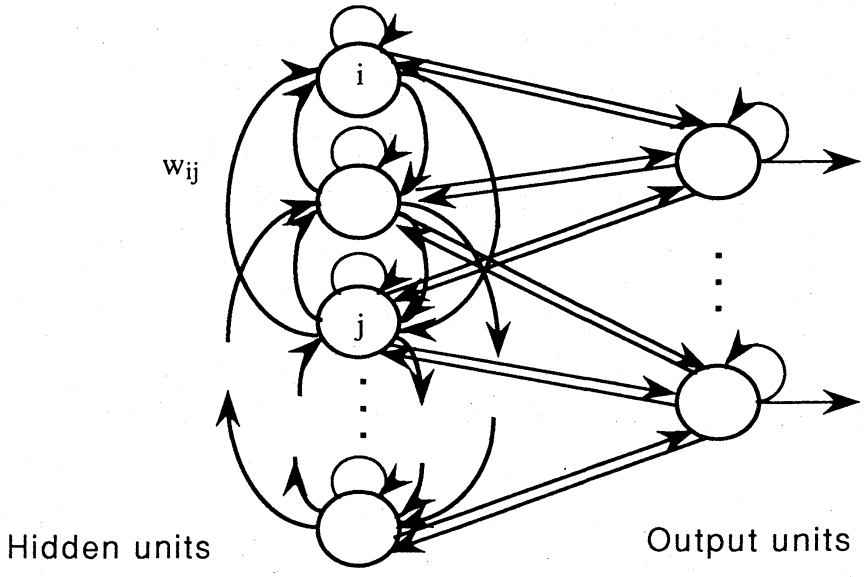

Figure 3. Recurrent neural network