Forthcoming in: *Neural Computation*, 2002.

Universal Approximation of Multiple Nonlinear Operators by Neural Networks

Andrew D. Back
Windale Technologies
Brisbane, QLD 4075, Australia

Tianping Chen
Department of Mathematics
Fudan University
Shanghai, 200433, China

Recently, there has been interest in the observed capabilities of some classes of neural networks with fixed weights to model multiple nonlinear dynamical systems. While this property has been observed in simulations, open questions exist as to how this property can arise. In this paper, we propose a theory which provides a possible mechanism by which this multiple modeling phenomena can occur.

1 Introduction

Understanding the mechanisms by which neural networks can approximate functions, functionals and operators has generated significant interest in recent years. The property of universal function approximation has been proven for various feedforward neural network models (see for example Hornik, Stinchcombe & White 1989). In dynamic[1] networks with embedded time delays, these results have been extended to the case of universal functional approximation (Chen & Chen 1993) and universal operator approximation (Sandberg 1991, Chen & Chen 1995). Universal approximation properties for recurrent neural networks over a finite time interval have been proven by Sontag (1992).

Related work in continuously or discretely time-varying dynamical systems exists in the literature under various names, eg. multiple controllers (Narendra, Balakrishnan & Ciliz 1995), *modular neural networks* (Jacobs, Jordan, Nowlan & Hinton 1991, Schmidhuber 1992), *hybrid systems* (Branicky, Borkar & Mitter 1994, Brockett 1993, Branicky 1996, Lemmon & Antsaklis 1995). Recently, several researchers have shown independently that some types of neural network are capable of learning to model several different dynamical systems within one structure (Feldkamp, Puskorius & Moore 1997, Younger, Conwell & Cotter 1999, Hochreiter & Schmidhuber 1997, Hochreiter, Younger & Conwell 2001).

In this paper we provide a theoretical explanation of a mechanism by which neural network models can approximate dynamical systems that change continuously or switch between some form of ‘characteristic’ behaviours. We refer to such systems as *multiple nonlinear operator* (MNO) models. In the next section we present results on the universal approximation of multiple nonlinear operators by neural networks followed by a brief example.

[1] We refer to neural networks with time-delay or recurrent connections as *dynamic* neural network models.

2 Universal Approximation of Multiple Nonlinear Operators

2.1 Functional and Operator Approximation

It is well known that nonlinear dynamical systems can be treated as functionals and operators (Chen & Chen 1995, Sandberg 1991, Stiles, Sandberg & Ghosh 1997). The time delay neural network (TDNN) functional (Chen & Chen 1995) is defined on a compact set in $C_{[a,b]}$, (the space of all continuous functions), and is given by

$$G(u) = \sum_{i=1}^{N} c_i g \left(\sum_{j=1}^{m} \xi_{ij} u\left(t_j\right) + \theta_i \right) \tag{1}$$

where $u(t)$ is a continuous real-valued output at time t.

A more general class of model which has significant implications for modeling dynamic systems, are those capable of approximating arbitrary nonlinear operators. For dynamic models, this means mapping from one time-varying sequence (a function of time) to another. Recently, Chen & Chen (1995) proved universal approximation for the following operator model

$$G(u)(y) = \sum_{k=1}^{N} \sum_{i=1}^{M} c_i^k g \left(\sum_{j=1}^{m} \xi_{ij}^k u\left(x_j\right) + \theta_i^k \right) g\left(\omega_k \cdot y + \varsigma_k\right) \tag{2}$$

where $G(u)(y)$ is an arbitrary nonlinear operator with inputs u and y.

The Chen network can be regarded as a feedforward network with two weight layers, in which the output weights have each been replaced by a two layer functional approximating network. The idea of these 'parameter replacement' networks, has also been proposed independently elsewhere (Priestley 1980, Back & Chen 1998, Schmidhuber 1992). In this previous work universal operator approximation properties were not proven however.

2.2 Multiple Nonlinear Operator Approximation

Consider a multiple nonlinear operator H, defined broadly as a set of mappings of the form $H : F_i \rightarrow F_{i+1}$, where $i = 1, 2, ...$ is the index of the discrete functionals in the discrete multiple model case, or $H : F_a \rightarrow F_b$ corresponding to the continuous multiple model case. Such a model encapsulates the basic framework of multiple model capabilities observed recently in various neural networks (Feldkamp et al. 1997).

Following a similar approach to (Chen & Chen 1995), it is possible to construct a network which performs the task of universally approximating a multiple nonlinear operator. We obtain the following theorem:

Theorem 1 *Suppose that g is a nonpolynomial, continuous, bounded function, X is a Banach Space, $K_1 \subseteq X$ is a compact set in X and $K_2 \subseteq \mathcal{R}^n, K_3 \subseteq \mathcal{R}^n, K_4 \subseteq \mathcal{R}^n$ are compact sets in $\mathcal{R}^n$, V is a compact set in $C\left(K_1\right)$, Z is a compact set in $C\left(K_3\right)$, G is a nonlinear continuous operator, which maps V into $C\left(K_2\right) \times C\left(K_3\right)$, then for any $\epsilon > 0$, there are positive integers M, N, P, m, p, constants $c_i^{kh}, \varphi_{il}^{kh}, \rho_i^{kh}, \varsigma_k, \xi_{ij}^k, \theta_i^k \in \mathcal{R}$, points $\omega_k \in \mathcal{R}^n$ $x_j \in K_1, x_l \in K_3$ $i = 1, \cdots, M$,*

$k = 1, \cdots, N$, $h = 1, \cdots, P$, $j = 1, \cdots, m$, $l = 1, \cdots, p$ *such that*

$$\left| \begin{array}{c} G(u)(y)(z) - \\ \sum_{k=1}^{N} \sum_{i=1}^{M} \sum_{h=1}^{P} c_i^{kh} g\left(\sum_{l=1}^{p} \varphi_{ij}^{kh} z_l + \rho_i^{kh}\right) \\ g\left(\sum_{j=1}^{m} \xi_{ij}^{k} u(x_j) + \theta_i^k\right) g(\omega_k \cdot y + \varsigma_k) \end{array} \right| < \epsilon \tag{3}$$

holds for all $u \in V$, $y \in K_2$ *and* $z \in Z$.

Proof. From Theorems 3 and 5 in (Chen & Chen 1995), and following the same approach as in Theorem 5 (Chen & Chen 1995), we assume that G is a continuous operator which maps a compact set Z in $C(K_3)$ into $C(K_4)$. Hence $G(Z) = \{G(z) : z \in Z\}$ is also a compact set in $C(K_4)$. From Theorem 3 (Chen & Chen 1995) we have for any $\epsilon > 0$, a positive integer N, real numbers $c_i^k(G(z)), \theta_i^k, \varsigma_k, \xi_{ij}^k \in \mathcal{R}$, vectors $\omega_k \in \mathcal{R}^n$, $i = 1, \cdots, N$, $j = 1, \cdots, m$, $k = 1, \cdots, N$ such that

$$\left| \begin{array}{c} G(u)(y)(z) - \\ \sum_{k=1}^{N} \sum_{i=1}^{M} c_i^k(G(z))\, g\left(\sum_{j=1}^{m} \xi_{ij}^{k} u(x_j) + \theta_i^k\right) \\ g(\omega_k y + \varsigma_k) \end{array} \right| < \frac{\epsilon}{2} \tag{4}$$

holds for all $y \in K_2$, $v \in V$, and $z \in Z$.

As before, we note that since G is a continuous operator, for each $k = 1, \cdots, N$, $c_i^k(G(z))$ is a continuous functional defined on Z. Hence, by applying Theorem 4 (Chen & Chen 1995) for each $i = 1, \cdots, N$, $k = 1, \cdots, N$, it is possible to determine positive integers N_k, m_k, constants $c_i^{kh}, \varphi_{ij}^{kh}, \rho_i^{kh}, \in \mathcal{R}$, and $x_j \in K_1$, $h = 1, \cdots, N_k$, $j = 1, \cdots, m_k$,such that

$$\left| c_i^k(G(z)) - \sum_{h=1}^{N_k} c_i^{kh} g\left(\sum_{j=1}^{m_k} \varphi_{ij}^{kh} z_j + \rho_i^{kh}\right) \right| < \frac{\epsilon}{2L} \tag{5}$$

holds for all $i = 1, \cdots, N$, $k = 1, \cdots, N$ and $z \in Z$ where

$$L = \sum_{k=1}^{N} \sup_{y \in K_2} |g(\omega_k \cdot y + \varsigma_k)| \tag{6}$$

Therefore, substituting (5) into (4), we have

$$\left| \begin{array}{c} G(u)(y)(z) - \\ \sum_{k=1}^{N} \sum_{i=1}^{M} \sum_{h=1}^{P} c_i^{kh} g\left(\sum_{j=1}^{m_k} \varphi_{ij}^{kh} z_j + \rho_i^{kh}\right) \\ g\left(\sum_{j=1}^{m} \xi_{ij}^{k} u(x_j) + \theta_i^k\right) g(\omega_k \cdot y + \varsigma_k) \end{array} \right| < \epsilon \tag{7}$$

Let $P = \max_k \{N_k\}$, $p = \max_k \{m_k\}$, $c_i^{kh} = 0 \,\forall\, N_k < h \leq P$, and $\varphi_{il}^{kh} = 0 \,\forall\, m_k < l \leq p$. Hence we have

$$\left| \begin{array}{c} G(u)(y)(z) - \\ \sum_{k=1}^{N} \sum_{i=1}^{M} \sum_{h=1}^{P} c_i^{kh} g\left(\sum_{l=1}^{p} \varphi_{il}^{kh} z_l + \rho_i^{kh}\right) \\ g\left(\sum_{j=1}^{m} \xi_{ij}^{k} u(x_j) + \theta_i^k\right) g(\omega_k \cdot y + \varsigma_k) \end{array} \right| < \epsilon \tag{8}$$

□

This model can be interpreted as the z input acting as a dynamic 'gating' mechanism which by virtue of its universal functional approximation capabilities on a compact set, allows the selection of a particular operator, defined by the encompassed Chen network. The mapping of the z input allows the effective output weights c_i^{kh} to be varied such that the desired operators can be chosen. Thus, the operator map which is given by the mapping on $u(x_j)$ can now be adjusted arbitrarily for every z input. The implication is that $G(u)(y)(z) \subseteq K_o$ and we obtain a network capable of universally approximating any multiple nonlinear operator on a compact set by allowing the input z to select any unique operator mapping. Hence, a fixed weight network of this type is capable of approximating multiple dynamic systems, as observed in some classes of neural networks.

Example

An example multiple nonlinear operator, demonstrating the model, can be synthesized as follows. Let there exist an MNO defined by $H_m(u)(x)(t) = H_m(u)x(t)$, where

$$H_m(u) = \frac{1}{1 - \zeta_1(u,t)\, q^{-1} + \zeta_2(u,t)\, q^{-2}} \tag{9}$$

We introduce coefficient functions $\zeta_i(t)\,,\, i = 1, 2$ given by

$$\begin{aligned} \zeta_i(t) &= c_{i0} + c_{i1}\zeta_i^{'}(t) \\ \zeta_i^{'}(t) &= \frac{b_{i0}u(t) + b_{i1}u(t-1)}{1 + a_{i1}q^{-1} + a_{i2}q^{-2}} \end{aligned} \tag{10}$$

The characteristics of the model H_m, varies as a function of the ancillary input signal u. It is not difficult to select model parameters to observe widely varying model characteristics.

The resulting MNO model encompasses multiple operators of the general form given in (9). The model can also be viewed as an extension of the usual bilinear structure, described in terms of the difference equation form of nonlinear pole-zero model given by

$$\begin{aligned} y(t) &= x(t) + \zeta_1(t)\, y(t-1) + \zeta_2(t)\, y(t-2) \\ \zeta_i(t) &= c_{i1} + c_{i2}\left(b_0 x(t) + b_1 x(t) - a_{i1}\zeta_i(t-1) + a_{i2}\zeta_i(t-2)\right) \end{aligned} \tag{11}$$

where $x(t)$ is the input and $y(t)$ is the output.

3 Conclusions

In this paper, we have proposed a new theorem of multiple nonlinear operator approximation. This theorem provides an explanation of how the phenomena of multiple models can occur in dynamic neural networks. The proposed theorem is an extension of Chen and Chen's earlier results on universal operator approximation. It is hoped that the work presented here will serve to stimulate additional research into this potentially powerful capability of neural networks.

Acknowledgments

This work was performed while the first author was at the RIKEN Brain Science Institute, The Institute of Physical and Chemical Research, Japan. The first author acknowledges fruitful discussions with S. Amari and L. Feldkamp. The second author is supported by NSF of China.

References

Back, A. & Chen, T. (1998), Approximation of hybrid systems by neural networks, *in* N. Kasabov, R. Kozma, K. Ko, R. O'Shea, G. Coghill & T. Gedeon, eds, 'Proc of 1997 Int Conf on Neural Information Processing ICONIP'97', Vol. 1, Springer-Verlag, Singapore, pp. 326–329.

Branicky, M. (1996), Studies in Hybrid Systems, PhD thesis, MIT.

Branicky, M. S., Borkar, V. S. & Mitter, S. K. (1994), A unified framework for hybrid control: background, model, and theory, Technical Report LIDS-P-2239, Laboratory for Information and Decision Systems, Dept of Electrical Engineering and Computer Science, MIT.

Brockett, R. (1993), Hybrid models for motion control systems, *in* H. Trentelman and J. C. Willems, Eds.,'Perspectives in Control', Boston: Birkhauser, pp. 29–54.

Chen, T. & Chen, H. (1993), 'Approximations of continuous functionals by neural networks with application to dynamic systems', *IEEE Trans. Neural Networks* **4**(6), 910–918.

Chen, T. & Chen, H. (1995), 'Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems', *IEEE Trans. Neural Networks* **6**(4), 911–917.

Feldkamp, L., Puskorius, G. & Moore, P. (1997), 'Adaptive behaviour from fixed weight dynamic networks', *Information Sciences* **98**, 217–235.

Hochreiter, S. & Schmidhuber, J. (1997), 'Long short-term memory', *Neural Computation* **9**(8), 1735–1780.

Hochreiter, S., Younger, A. S. & Conwell, P. R. (2001), Learning to learn using gradient descent, *in* 'Lecture Notes on Comp. Sci. 2130, Proc. Intl. Conf. on Artificial Neural Networks (ICANN-2001)', Springer: Berlin, Heidelberg, pp. 87–94.

Hornik, K., Stinchcombe, M. & White, H. (1989), 'Multilayer feedforward networks are universal approximators', *Neural Networks* **2**(5), 359–366.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J. & Hinton, G. E. (1991), 'Adaptive mixtures of local experts', *Neural Computation* **3**(1), 79–87.

Lemmon, M. & Antsaklis, P. (1995), 'Inductively inferring valid logical models for continuous-state dynamical systems', *Theoretical Computer Science* **138**, 201–220.

Narendra, K., Balakrishnan, J. & Ciliz, M. (1995), 'Adaptation and learning using multiple models switching and tuning', *IEEE Control Systems Magazine* **15**(3), 37–51.

Priestley, M. (1980), 'State-dependent models: a general approach to non-linear time series analysis', *J. Time Series Analysis* **1**(1), 47–71.

Sandberg, I. (1991), 'Approximation theorems for discrete-time systems', *IEEE Trans. Circuits, Syst.* **38**(5), 564–566.

Schmidhuber, J. (1992), 'Learning to control fast-weight memories: An alternative to dynamic recurrent networks', *Neural Computation* **4**(1), 131–139.

Sontag, E. (1992), Neural nets as systems models and controllers, *in* 'Proc. Seventh Yale Workshop on Adaptive and Learning Systems', Yale University, pp. 73–79.

Stiles, B. W., Sandberg, I. W. & Ghosh, J. (1997), 'Complete memory structures for approximating nonlinear discrete-time mappings', *IEEE Transactions on Neural Networks* **8**(6), 1397–1409.

Younger, A., Conwell, P. & Cotter, N. (1999), 'Fixed-weight on-line learning', *IEEE Trans. on Neural Networks* **10**(2), 272–283.