Stability Analysis for Dynamical Neural Network Systems

S. Lam and Y. S. Hung

Department of Electrical and Electronic Engineering,
The University of Hong Kong,
Pokfulam Road,
HONG KONG.

Abstract

In this paper, the small gain theorem will be used to establish a criterion for the stability of a feedback system containing a feedforward neural network. A method for the determination of the gain of a piecewise-linear feedforward neural network is introduced and applied to the stability analysis for a control system consisting of a LTI SISO system with a dynamic ANN controller.

1. Introduction

Artificial neural networks have been used extensively in the control and identification of dynamical systems. New learning control algorithms, capable of controlling partially known, complex nonlinear systems, have been proposed [1], [2], [3]. It is noted, however, that the stability of control systems containing neural networks is rarely addressed in the literature. This may be due to the inherent complexity of multilayer neural network, especially the nonlinearity of the activation function, which does not readily lend itself to stability analysis. In order to facilitate analysis, the piecewise-linear saturation function is used as the activation function for the nodes in the hidden layer rather than the usual sigmoid function.

This paper is organized in the following way. In section 2, the small gain theorem will be used to establish a criterion for the stability of a feedback system containing a neural network. The stability result is given in terms of tha gain of a state map. In section 3, a method for the determination of the gain of a piecewise-linear feedforward neural network with one hidden layer will then be introduced. Some concluding remarks will be given in section 4.

The notation for multilayer feedforward neural network introduced in [2] will be used. A multilayer feedforward neural network denoted by $\mathfrak{N}^{N}_{n_0,n_1,\cdots,n_N}$ has n_0 input, n_N outputs and $(N-1)$ hidden layers, the ith of which contains n_i nodes.

2. Stability Theory

Consider an autonomous discrete-time dynamical system described by the state equation

$$\mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k) \tag{1}$$

where $\mathbf{x}_k \in \Re^n$ is the state vector, $k \in \mathrm{N}$ denotes the time and $\mathbf{f}: \Re^n \to \Re^n$. The system can be represented as shown in Fig. 1. We are interested in the stability of the system when the map $\mathbf{f}$ contains a neural network.

Without loss of generality, we will assume that the origin is an equilibrium state. i.e. $\mathbf{f}(\mathbf{0}) = \mathbf{0}$, as this condition can always be met by translating any equilibrium state under study to the origin through coordinate transformation.

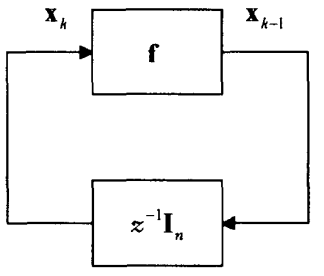

Fig. 1

Define the gain of the network as

$$g(\mathbf{f}) = \sup_{\|\mathbf{x}\| \neq 0} \frac{\|\mathbf{f}(\mathbf{x})\|}{\|\mathbf{x}\|} \tag{2}$$

The system (1) is stable if

$$\frac{\|\mathbf{x}_{k+1}\|}{\|\mathbf{x}_k\|} = \frac{\|\mathbf{f}(\mathbf{x}_k)\|}{\|\mathbf{x}_k\|} \leq 1 \tag{3}$$

Therefore, a sufficient condition for the system to be stable is that the gain of the network is less than unity.

As a motivation for our study, we will show that a dynamical system containing a neural network controller can be cast in the form of Fig. 1.

Let a n_pth order SISO LTI system be described by the state space model:

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{A}_p \mathbf{x}_k + \mathbf{B}_p u_k \\ y_k &= \mathbf{C}_p \mathbf{x}_k \end{aligned} \tag{4}$$

0-7803-1901-X/94 $4.00 ©1994 IEEE

Suppose the system is connected in a feedback configuration with a controller comprising of n_c delay elements and a neural network of the structure $\mathfrak{R}^2_{n_c+1,n_1,1}$ as shown in Fig. 2.

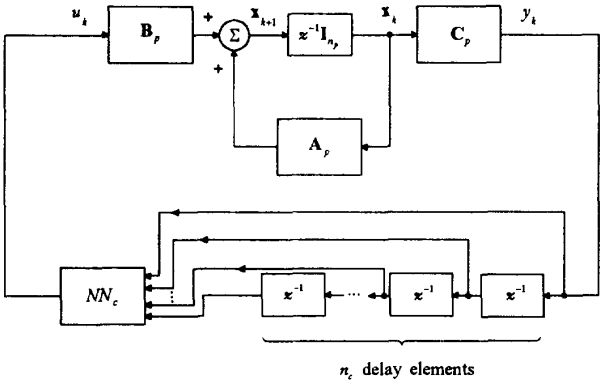

Fig. 2

The neural network controller gives

$$\begin{aligned} u_k &= NN_c(\mathbf{v}_k) \\ &= \mathbf{W}_2\,\sigma(\mathbf{W}_1\mathbf{v}_k + \mathbf{b}_1) + \mathbf{b}_2 \end{aligned} \tag{5}$$

where $\mathbf{v}_k = (y_k, y_{k-1}, \cdots, y_{k-n_c})^T$, $\mathbf{W}_1$ is the connection weights from the input to the hidden layer, $\mathbf{W}_2$ is that from the hidden layer to the output, $\mathbf{b}_1$ and $\mathbf{b}_2$ are the biases to the hidden layer and the output layer respectively, and $\sigma(\bullet)$ is the activation function of the nodes in the hidden layer. The activation function of the output layer nodes is the identify function.

The delay line can be described by the state-space model:

$$\begin{aligned} \mathbf{t}_{k+1} &= \mathbf{A}_c\,\mathbf{t}_k + \mathbf{B}_c\,y_k \\ \mathbf{v}_k &= \mathbf{C}_c\,\mathbf{t}_k + \mathbf{D}_c\,y_k \end{aligned} \tag{6}$$

where

$$\mathbf{A}_c = \begin{bmatrix} 0 & \mathbf{I}_{n_c-1} \\ \vdots & \\ 0 & \cdots\ 0 \end{bmatrix},\quad \mathbf{B}_c = \begin{bmatrix} 0 \\ \vdots \\ 1 \end{bmatrix},\quad \mathbf{C}_c = \begin{bmatrix} \mathbf{I}_{n_c} \\ 0 \end{bmatrix},\quad \mathbf{D}_c = \begin{bmatrix} 0 \\ \vdots \\ 1 \end{bmatrix} \tag{7}$$

As a result, the closed-loop system shown in Fig. 2 is of order $n_p + n_c$. Now, define a cascaded state vector as $\mathbf{z}_k = (\mathbf{x}_k^T, \mathbf{t}_k^T)^T$

Combining (4) and (6) and making use of (5), we have

$$\begin{aligned} \mathbf{z}_{k+1} &= \begin{bmatrix} \mathbf{x}_{k+1} \\ \mathbf{t}_{k+1} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{A}_p\,\mathbf{x}_k + \mathbf{B}_p\,u_k \\ \mathbf{A}_c\,\mathbf{t}_k + \mathbf{B}_c\,y_k \end{bmatrix} \\ &= \begin{bmatrix} \begin{bmatrix} \mathbf{A}_p & \mathbf{0}_{n_c} \end{bmatrix}\mathbf{z}_k + \mathbf{B}_p\mathbf{W}_2\,\sigma\left(\begin{bmatrix} \mathbf{W}_1\mathbf{D}_c\mathbf{C}_p & \mathbf{W}_1\mathbf{C}_c \end{bmatrix}\mathbf{z}_k + \mathbf{b}_1\right) + \mathbf{B}_p\mathbf{b}_2 \\ \begin{bmatrix} \mathbf{B}_c\mathbf{C}_p & \mathbf{A}_c \end{bmatrix}\mathbf{z}_k \end{bmatrix} \end{aligned} \tag{8}$$

This shows that the system depicted in Fig. 2 can be put into the form of (1) and therefore we can apply the condition (3) to assess the stability of the system. The next question that arises is how one may check condition (3) if the function $\mathbf{f}$ contains a neural network.

3. Gain Estimation for a Neural Network

The sigmoid function (hyperbolic tangent) commonly used in multilayer feedforward networks makes the estimation of gain analytically untractable. As an approximation, a piecewise-linear saturation function will be used instead in our analysis.

The saturation activation function is defined by

$$\begin{aligned} \sigma(x) &= \mathrm{sat}(x) \\ &= \begin{cases} x, & |x| \le 1 \\ \mathrm{sgn}(x), & |x| > 1 \end{cases} \end{aligned} \tag{9}$$

Consider a feedforward network belonging to $\mathfrak{R}^2_{n_0,n_1,n_2}$. Such a network has n_0 inputs, n_1 nodes in the single hidden layer and n_2 outputs. The outputs are related to the inputs through equation (5).

Alternatively, we can write

$$\begin{aligned} \mathbf{y} &= \mathbf{f}(\mathbf{x}) \\ &= \mathbf{b}_2 + \sum_{j=1}^{n_1} \mathbf{w}_{2j}\,\sigma(\mathbf{w}_{1j}\mathbf{x} + b_{1j}) \end{aligned} \tag{10}$$

where $\mathbf{x} \in \Re^{n_0}$, $\mathbf{y} \in \Re^{n_2}$, $\mathbf{f}: \Re^{n_0} \to \Re^{n_2}$, $\mathbf{w}_{1j}$ is the jth row of $\mathbf{W}_1$, $\mathbf{w}_{2j}$ is the jth column of $\mathbf{W}_2$, and b_{1j} is the jth element of $\mathbf{b}_1$.

In order for the gain of $\mathbf{f}$ to be finite, we require that $\mathbf{y} = \mathbf{0}$ when $\mathbf{x} = \mathbf{0}$. This implies that

$$\mathbf{b}_2 = -\sum_{j=1}^{n_1} \mathbf{w}_{2j}\,\mathrm{sat}(b_{1j}) \tag{11}$$

As a result, (10) becomes

$$\mathbf{y}(\mathbf{x}) = \sum_{j=1}^{n_1} \mathbf{w}_{2j}\left[\mathrm{sat}(\mathbf{w}_{1j}\mathbf{x} + b_{1j}) - \mathrm{sat}(b_{1j})\right] \tag{12}$$

We will next exploit the piecewise-linear nature of (12) by partitioning the input space.

3.1 Geometry of the input space

Each processing element (neuron), has two decision boundaries given by the $(n_0 - 1)$-dimensional hyperplanes $\mathbf{w}_{1j}\mathbf{x} = -1 \pm b_{1j}$ $(j = 1 \cdots n_1)$. These two hyperplanes partition the n_0-dimensional input space into three regions, namely one linear and two saturation regions. As a result, the input space is partitioned into a cell complex (termed arrangement in [4])

induced by the $2n_1$ (n_0-1)-dimensional hyperplanes. For a 2-dimensional input space, the arrangement consists of vertices (intersections of lines), edges (maximal connected components of the lines containing no vertex), and regions (maximal connected components of $\Re^2$ containing no edge or vertex). The notion of 2-dimensional arrangement is easily generalized to three and higher dimensions. In general, the arrangement consists of open convex n_0-dimensional polytopes and various open convex k-dimensional polytopes (k-faces) bounding them, for $0 \le k \le n_0 - 1$. In each of these faces, $\mathbf{y}$ is affine in $\mathbf{x}$. An algorithm to construct a representation for the cell complex defined by n hyperplanes in d-dimensions in optimal $O(n_d)$ time was proposed in [4]. After the arrangement is determined, we can find the maximum gain over each region.

3.2 Gain Evaluation in Cell Complex

In a n_0-dimensional arrangement, we will show that the gain over the n_0-dimensional polytopes is bounded by that of the bordering k-faces ($0 \le k \le n_0 - 1$). First, let us start from single input single output case.

Case 1 : Single Input Single Output (SISO)

Consider $x, y \in \Re$, $f : \Re \to \Re$ with

$$y = f(x)$$
$$= \sum_{j=1}^{n_1} w_j^2 \left[\mathrm{sat}(w_j^1 x + b_j^1) - \mathrm{sat}(b_j^1) \right]$$

In this case, the arrangement consists of open line segments bounded by points. Consider one of these segments, say (a,b). Over this interval, x can be expressed as an affine combination of the end-points.

$$\forall x \in (a,b), \quad x = \alpha a + (1-\alpha) b, \quad \text{for some } 0 < \alpha < 1,$$

As y is an affine function of x, so it can be expressed as

$$y = f(x) = \alpha f(a) + (1-\alpha) f(b) \quad 0 < \alpha < 1, \tag{13}$$

Let $\quad G(x) = \dfrac{|f(x)|}{|x|}$

Now, we have to consider two cases: 1). If y does not change sign in the interval (a,b), $G(x)$ is monotonic in x. Therefore, maximum of G occurs at a or b. 2). If y changes sign in the interval (a,b) at a point d, i.e. $f(d) = 0, d \in (a,b)$, the interval (a,b) can be subdivided into two sub-intervals $(a,d]$ and $[d,b)$. In each of these two intervals, $G(x)$ is monotonic. Therefore, G is bounded at endpoints a, d, or b. Since $G(d) = 0$, maximum G occurs at either of the end-points (i.e. a or b). Hence

$$g(f) = \sup_{|x| \ne 0} G(x)$$
$$= \max\{G(a), G(b)\}$$

Case 2 : Single Input Multi-Output (SIMO)

Consider $x \in \Re, \mathbf{y} \in \Re^{n_2}, \mathbf{f} : \Re \to \Re^{n_2}$, with

$$\mathbf{y} = \mathbf{f}(x)$$
$$= \sum_{j=1}^{n_1} \mathbf{w}_{2j} \left[\mathrm{sat}(w_{1j} x + b_{1j}) - \mathrm{sat}(b_{1j}) \right] \tag{14}$$

The input space is partitioned into intervals over which the output $\mathbf{y}$ is affine in x. By a similar argument as above, in any one of the intervals, (a,b), we have

$$\forall x \in (a,b), \quad x = \alpha a + (1-\alpha) b, \quad \text{for some } 0 < \alpha < 1,$$
$$\mathbf{f}(x) = \alpha \mathbf{f}(a) + (1-\alpha) \mathbf{f}(b)$$

As the norm operator is convex by the triangular inequality, it follows that

$$\|\mathbf{f}(x)\|_2 \le \alpha \|\mathbf{f}(a)\|_2 + (1-\alpha) \|\mathbf{f}(b)\|_2 \tag{15}$$

Comparing (15) with (13), we can apply the result of the SISO case to show that $G(x)$ is maximum at the end-points.

Case 3 : Multi-Inputs Multi-Output (SIMO)

Consider $\mathbf{x} \in \Re^{n_0}, \mathbf{y} \in \Re^{n_2}, \mathbf{f} : \Re^{n_0} \to \Re^{n_2}$, with

$$\mathbf{y} = \mathbf{f}(\mathbf{x})$$
$$= \sum_{j=1}^{n_1} \mathbf{w}_{2j} \left[\mathrm{sat}(\mathbf{w}_{1j}^T \mathbf{x} + b_{1j}) - \mathrm{sat}(b_{1j}) \right] \tag{16}$$

In any one of the open n_0-dimensional polytope $F_j^{n_0}$, $\mathbf{y}$ can be written as

$$\mathbf{y} = \mathbf{J}_j \mathbf{x} + \mathbf{c}_j \quad \mathbf{x} \in F_j^{n_0} \tag{17}$$

where $\mathbf{J}_i$ is the Jacobian of $\mathbf{f}$ in the region $F_j^{n_0}$.

If we fix all the elements of $\mathbf{x}$ except x_i, we can apply the result of SIMO case to show that the maximum $G(\mathbf{x})$ cannot occur in the interior of the n_0-dimensional polytope. Hence, we only need to consider the k-faces bounding the n_0-dimensional polytopes ($0 \le k \le n_0 - 1$).

Let the k-face (denoted by F_i^k) be a k-simplex. i.e. the vectors $\mathbf{p}_i - \mathbf{p}_0$, $i = 1, \cdots, k$ are linearly independent where $\mathbf{p}_0, \mathbf{p}_1, \cdots, \mathbf{p}_k$ are its vertices. Using the barycentric coordinates, any point $\mathbf{x}$ in F_i^k can be written as

$$\mathbf{x} = \alpha_1 \mathbf{p}_1 + \alpha_2 \mathbf{p}_2 + \cdots + \alpha_k \mathbf{p}_k + (1 - \alpha_1 - \cdots - \alpha_k) \mathbf{p}_0$$
$$= \mathbf{X}\,\mathbf{k} \tag{18}$$

where

$$\mathbf{X} = [\mathbf{p}_1 - \mathbf{p}_0 \quad \mathbf{p}_2 - \mathbf{p}_0 \quad \cdots \quad \mathbf{p}_k - \mathbf{p}_0 \quad \mathbf{p}_0]$$

$$\mathbf{k} = \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_k & 1 \end{bmatrix}^T \tag{19}$$

subjected to the constraint that $\mathbf{x}$ lies in the interior of the k-face, i.e.

$$\alpha_i > 0, \quad i = 1 \cdots k, \qquad \sum_{i=1}^{k} \alpha_i < 1 \tag{20}$$

(Note: If F_i^k is not a k-simplex, i.e. it has more than $k+1$ vertices, we can choose $k+1$ vertices $\mathbf{p}_0, \mathbf{p}_1, \cdots, \mathbf{p}_k$ s.t. $\mathbf{p}_i - \mathbf{p}_0, i = 1, \cdots, k$ are linearly independent. In the case where F_i^k is unbounded, we can choose $k+1$ affine independent points $\mathbf{p}_0, \mathbf{p}_1, \cdots, \mathbf{p}_k$ on it. In both cases, the constraint (20) will need to be modified.)

Using equation (18), $\mathbf{y}$ can be written as

$$\begin{aligned} \mathbf{y} &= \mathbf{J}\mathbf{x} + \mathbf{c} \\ &= \mathbf{J}\mathbf{X}\mathbf{k} + \mathbf{c} \\ &= \mathbf{Y}\mathbf{k} \end{aligned} \tag{21}$$

where

$$\mathbf{Y} = \begin{bmatrix} \mathbf{J}(\mathbf{p}_1 - \mathbf{p}_0) & \cdots & \mathbf{J}(\mathbf{p}_k - \mathbf{p}_0) & \mathbf{J}\mathbf{p}_0 + \mathbf{c} \end{bmatrix} \tag{22}$$

The gain is

$$\begin{aligned} G(\mathbf{x}) &= \frac{\| \mathbf{y} \|_2}{\| \mathbf{x} \|_2} \\ &= \left(\frac{\mathbf{k}^T \mathbf{Y}^T \mathbf{Y} \mathbf{k}}{\mathbf{k}^T \mathbf{X}^T \mathbf{X} \mathbf{k}} \right)^{1/2} \end{aligned} \tag{23}$$

The stationary points of G can be found by setting the gradient $\nabla_{\mathbf{k}} G^2 = \mathbf{0}$. After some simplifications, we have

$$\mathbf{Y}^T \mathbf{Y} \mathbf{k} - G^2 \mathbf{X}^T \mathbf{X} \mathbf{k} = \mathbf{0} \tag{24}$$

which is a generalized singular value problem. By using the generalized singular value decomposition (GSVD) [5] (which is essentially a simultaneous diagonalization of both the matrices $\mathbf{Y}^T\mathbf{Y}$ and $\mathbf{X}^T\mathbf{X}$), given $\mathbf{Y} \in \Re^{n_2 \times (k+1)}$ and $\mathbf{X} \in \Re^{n_0 \times (k+1)}$, we can find orthogonal matrices $\mathbf{U} \in \Re^{n_2 \times n_2}$ and $\mathbf{V} \in \Re^{n_0 \times n_0}$ and an invertible matrix $\mathbf{W} \in \Re^{(k+1) \times (k+1)}$ s.t.

$$\begin{aligned} \mathbf{U}^T \mathbf{X} \mathbf{W} &= \mathbf{C} \\ &= \mathrm{diag}(c_1, \cdots, c_q) \quad c_i \geq 0, \quad q = \min(n_0, (k+1)) \end{aligned} \tag{25}$$

and

$$\begin{aligned} \mathbf{V}^T \mathbf{Y} \mathbf{W} &= \mathbf{S} \\ &= \mathrm{diag}(s_1, \cdots, s_r) \quad s_i \geq 0, \quad r = \min(n_2, (k+1)) \end{aligned} \tag{26}$$

The columns of $\mathbf{W} = \begin{bmatrix} \mathbf{w}_1 & \cdots & \mathbf{w}_{k+1} \end{bmatrix}$ are the generalized singular vectors of the pair $(\mathbf{Y}, \mathbf{X})$, satisfying

$$(s_i^2 \mathbf{Y}^T \mathbf{Y} - c_i^2 \mathbf{X}^T \mathbf{X}) \mathbf{w}_i = \mathbf{0} \qquad i = 1 \cdots k+1$$

It follows that if $s_i \neq 0$, then $(\mathbf{Y}^T\mathbf{Y} - \sigma_i^2 \mathbf{X}^T \mathbf{X})\mathbf{w}_i = \mathbf{0}$ where $\sigma_i = c_i / s_i$ is the gain at one of the stationary points of G. The vector $\mathbf{k}$ corresponding to σ_i can then be found by scaling the vector $\mathbf{w}_i$ so that the last element is unity. If it satisfies (20), it means that there is a stationary point of G in the interior of the k-face F_i^k. The maximum σ_i is the maximum gain of G among all the stationary points (including the local maximum, if any) in the interior of F_i^k. If there is no local maximum in F_i^k, the gain over the k-face is bounded by the gain over the bordering j-faces, $0 \leq j < k$, and we will apply the procedure to these j-faces.

The maximum gain of the network is then the maximum of the maximum gain over all the k-faces, $0 \leq k \leq n_0 - 1$.

4. Conclusion

We have shown how the small gain theorem can be used to assess the stability of a dynamical system containing a multilayer feedforward neural network and delay elements. In order to apply the small gain theorem, it becomes necessary to estimate the gain of a feedforward neural network. A method to estimate the gain of a feedforward neural network with one hidden layer of nodes with piecewise-linear activation function is proposed. This provides a means for applying the stability result developed here to the kind of systems shown in Fig. 2.

5. References

[1] Miller W. T., Sutton R. S. and Werbos P. J., *Neural Networks for Control.* Cambridge, MA: MIT Press, 1990.

[2] Narendra, K. S., and Parthasarathy, P., "Identification and Control of Dynamical Systems Using Neural Networks," *IEEE Trans. Neural Networks*, vol. 1, pp. 4-27, Mar. 1990.

[3] Narendra , K. S., and Parthasarathy, P., "Gradient Methods for the Optimization of Dynamical Systems Containing Neural Networks," *IEEE Trans. Neural Networks*, vol. 2, pp. 257-262, Mar. 1991.

[4] Edelsbrunner, H., O'Rourke, J., and, Seidel, R., "Constructing arrangements of lines and hyperplanes with Applications," *SIAM. J. Comput.*, pp.341-363, Vol. 15, No.2, May, 1986.

[5] Golub, G. H., and Van Loan, C. F., *Matrix Computations*, pp.471-472, 1989, John Hopkins University Press, Baltimore.