NOTE Communicated by Simon Haykin

Relationships Between the *A Priori* and *A Posteriori* Errors in Nonlinear Adaptive Neural Filters

Danilo P. Mandic
School of Information Systems, University of East Anglia, Norwich, U.K.

Jonathon A. Chambers
Signal Processing Section, Department of Electrical and Electronic Engineering, Imperial College of Science, Technology and Medicine, London, U.K.

The lower bounds for the a posteriori prediction error of a nonlinear predictor realized as a neural network are provided. These are obtained for a priori adaptation and a posteriori error networks with sigmoid nonlinearities trained by gradient-descent learning algorithms. A contractivity condition is imposed on a nonlinear activation function of a neuron so that the a posteriori prediction error is smaller in magnitude than the corresponding a priori one. Furthermore, an upper bound is imposed on the learning rate η so that the approach is feasible. The analysis is undertaken for both feedforward and recurrent nonlinear predictors realized as neural networks.

1 Introduction

A posteriori techniques have been considered in the area of linear adaptive filters (Treichler, Johnson, & Larimore, 1987; Ljung & Soderstrom, 1983; Douglas & Rupp, 1997). However, in the area of neural networks, the use of a posteriori techniques is still in its infancy. Recently it has been shown that an a posteriori approach in the neural networks framework exhibits behavior correspondent to the normalised least mean square (NLMS) algorithm in the linear adaptive filters case (Mandic & Chambers, 1998). Consequently, it is expected that the instantaneous a posteriori output error $\tilde{e}(k)$ is smaller in magnitude than the corresponding a priori error $e(k)$ (Treichler et al., 1987; Mandic & Chambers, 1998). However, little is known about the relationships between the a posteriori and a priori error, learning rate, slope in the nonlinear activation function of a neuron, and feasibility of such a neural predictor.

In the case of a single-node neural network, with a nonlinear activation function of a neuron Φ, the a priori output of a network y is given by

$$y(k) = \Phi\left(\mathbf{x}^T(k)\mathbf{w}(k)\right), \tag{1.1}$$

where $\mathbf{x}(k)$, $\mathbf{w}(k)$, and $(\cdot)^T$ denote, respectively, the input vector to a network,

the weight vector, and vector transpose operator. Function Φ is assumed to belong to the class of sigmoid functions. The updated weight vector $\mathbf{w}(k+1)$ is available from the learning algorithm before the next, updated, input vector $\mathbf{x}(k+1)$, therefore an a posteriori estimate $\bar{y}$ can be calculated as

$$\bar{y}(k) = \Phi\left(\mathbf{x}^T(k)\mathbf{w}(k+1)\right). \tag{1.2}$$

The corresponding instantaneous a priori and a posteriori errors at the output neuron of a neural network are given respectively as $e(k) = d(k) - y(k)$ and $\bar{e}(k) = d(k) - \bar{y}(k)$, where $d(k)$ is some teaching signal.

Our aim is to preserve

$$|\bar{e}(k)| \leq \gamma\,|e(k)|, \quad 0 < \gamma < 1 \tag{1.3}$$

at each iteration, for both feedforward and recurrent neural networks acting as a nonlinear predictor. This a priori learning a posteriori error algorithm corresponds to the normalized gradient-descent algorithm for neural networks (Haykin, 1996; Mandic & Chambers, 1998). In this work we seek to guarantee that the a posteriori error $\bar{e}$ is uniformly smaller in magnitude than the corresponding a priori error e.

The problem can be represented in the gradient-descent setting as

$$\begin{aligned} \mathbf{w}(k+1) &= \mathbf{w}(k) - \nabla_{\mathbf{w}} E\left(e(k)\right) \\ \bar{e}(k) &= d(k) - \Phi\left(\mathbf{x}^T(k)\mathbf{w}(k+1)\right) \end{aligned}$$

subject to

$$|\bar{e}(k)| \leq \gamma\,|e(k)|, \ 0 < \gamma < 1, \tag{1.4}$$

where the cost function $E(e(k))$ is some nonlinear function of the instantaneous a priori output error $e(k)$, typically a quadratic function of $e(k)$.

Here, we provide relationships between the a priori prediction error $e(k)$, a posteriori prediction error $\bar{e}(k)$, learning rate of a gradient-descent learning algorithm $\eta(k)$, and the slope β of a nonlinear activation function of a neuron Φ, for both the feedforward and recurrent case, with respect to objective 1.4. Constraints are imposed on the nonlinear activation function Φ, so that equation 1.3 holds. Moreover, the conditions for the learning rate η are given so that approach 1.4 is feasible. In that case, as a matter of example, we derive the relationship between the learning rate η and the slope β for the logistic nonlinear activation function of a neuron.

2 Contraction Mapping and Nonlinear Activation Functions

By the contraction mapping theorem (CMT), function K is a contraction on $[a, b] \in \mathbb{R}$ if (Gill, Murray, & Wright, 1981; Zeidler, 1986):

i. $x \in [a, b] \Rightarrow K(x) \in [a, b]$

Figure 1: Contraction mapping.

ii. $\exists \gamma \ < 1 \in \mathbb{R}^+ \ s.t. \ |K(x) - K(y)| \leq \gamma \, |x - y|, \forall x, y \in [a, b]$

as shown in Figure 1. Applying the mean value theorem (MVT) (Luenberger, 1969) to the definition of CMT, for $\forall x, y \in [a, b]$, $\exists \xi \in (a, b)$ such that

$$|\Phi(x) - \Phi(y)| = |\Phi'(\xi)(x - y)| = |\Phi'(\xi)||x - y|. \tag{2.1}$$

Now, clause $\gamma < 1$ in part ii of the CMT becomes $\gamma \geq |\Phi'(\xi)|, \xi \in (a, b)$. For the example of the logistic nonlinear activation function of a neuron

$$\Phi(x) = \frac{1}{1 + e^{-\beta x}} \tag{2.2}$$

whose first derivative is

$$\Phi'(x) = \frac{\beta e^{-\beta x}}{\left(1 + e^{-\beta x}\right)^2}, \tag{2.3}$$

$\gamma < 1 \Leftrightarrow \beta < 4$ is the condition for function Φ to be a contraction on $\forall [a, b] \in \mathbb{R}$ (Mandic & Chambers, 1999a).

3 The Case of a Feedforward Neural Filter

The gradient-descent algorithm for single-node a priori adaptation a posteriori error networks, with the cost function in the form of $E(k) = \frac{1}{2}e^2(k)$, is given by Narendra and Parthasarathy (1990, 1991):

$$\begin{aligned} \mathbf{w}(k+1) &= \mathbf{w}(k) + \eta(k)e(k)\Phi'\left(\mathbf{x}^T(k)\mathbf{w}(k)\right)\mathbf{x}(k) \\ \tilde{e}(k) &= d(k) - \Phi\left(\mathbf{x}^T(k)\mathbf{w}(k+1)\right). \end{aligned} \tag{3.1}$$

This case represents a generalization of finite impulse response (FIR) linear adaptive filters.

Multiplying the first equation in equation 3.1 from the left side by $\mathbf{x}^T(k)$ and applying the nonlinear activation function Φ on either side, we obtain

$$\begin{aligned} &\Phi\left(\mathbf{x}^T(k)\mathbf{w}(k+1)\right) \\ &\quad = \Phi\left(\mathbf{x}^T(k)\mathbf{w}(k) + \eta(k)e(k)\Phi'\left(\mathbf{x}^T(k)\mathbf{w}(k)\right)\|\mathbf{x}(k)\|_2^2\right). \end{aligned} \tag{3.2}$$

Further analysis depends on the function Φ, which can exhibit either contractive or expansive behavior.

3.1 Contractive Activation Function. If function Φ is a contraction, then

$$\Phi(a+b) \leq \Phi(a) + \Phi(b). \tag{3.3}$$

Theorem 1. *The lower bound for the a posteriori error obtained by the algorithm 3.1 with constraint 1.3 and a contractive nonlinear activation function Φ, is*

$$\bar{e}(k) > \left[1 - \eta(k)\Phi'\left(\mathbf{x}^T(k)\mathbf{w}(k)\right)\|\mathbf{x}(k)\|_2^2\right] e(k). \tag{3.4}$$

Proof. With $a = \mathbf{x}^T(k)\mathbf{w}(k)$ and $b = \eta(k)e(k)\Phi'\left(\mathbf{x}^T(k)\mathbf{w}(k)\right)\|\mathbf{x}(k)\|_2^2$, applying inequality 3.3 to 3.2 and subtracting $d(k)$ from both sides of the resulting equation, due to contractivity of Φ, we obtain

$$\bar{e}(k) \geq \left[1 - \Phi\left(\eta(k)\Phi'\left(\mathbf{x}^T(k)\mathbf{w}(k)\right)\|\mathbf{x}(k)\|_2^2\right)\right] e(k). \tag{3.5}$$

For Φ a contraction, $|\Phi(\xi)| < |\xi|$, $\forall \xi \in \mathbb{R}$, and equation 3.5 finally becomes

$$\bar{e}(k) > \left[1 - \eta(k)\Phi'\left(\mathbf{x}^T(k)\mathbf{w}(k)\right)\|\mathbf{x}(k)\|_2^2\right] e(k), \tag{3.6}$$

which is the lower bound for the a posteriori error for a contractive nonlinear activation function.

Corollary 1. *The range allowed for the learning rate $\eta(k)$ in an a priori adaptation a posteriori error algorithm, 3.1, with constraint 1.3, for the conditions given in theorem 1, is*

$$0 < \eta(k) < \frac{1}{\Phi'\left(\mathbf{x}^T(k)\mathbf{w}(k)\right)\|\mathbf{x}(k)\|_2^2}. \tag{3.7}$$

3.1.1 Some Simplifications. For Φ a contraction, $|\Phi'(\xi)| < 1$, $\forall \xi \in \mathbb{R}$. Therefore, even stricter conditions than those given in theorem 1 and corollary 1 are as follows.

Theorem 2. *The lower bound for the a posteriori error obtained by algorithm 3.1 with constraint 1.3 and a contractive nonlinear activation function Φ, is*

$$\bar{e}(k) > \left[1 - \eta(k)\|\mathbf{x}(k)\|_2^2\right] e(k). \tag{3.8}$$

Corollary 2. *The range allowed for the learning rate $\eta(k)$, for the conditions given in theorem 2, is*

$$0 < \eta(k) < \frac{1}{\|\mathbf{x}(k)\|_2^2}. \tag{3.9}$$

If, for convenience, the input data are normalized within a unit norm ($\|\mathbf{x}\|_2^2 < 1$), condition 3.9 becomes further simplified to $0 < \eta(k) < 1$.

3.2 Expansive Activation Function. If function Φ is an expansion, then $\Phi(a+b) \geq \Phi(a) + \Phi(b)$, $|\Phi(\xi)| > |\xi|$, $\forall \xi \in \mathbb{R}$, and $|\Phi'(\xi)| \geq 1$. In this case, equation 3.2 becomes

$$\bar{e}(k) \leq \left[1 - \Phi\left(\eta(k)\Phi'\left(\mathbf{x}^T(k)\mathbf{w}(k)\right)\|\mathbf{x}(k)\|_2^2\right)\right] e(k). \tag{3.10}$$

For Φ an expansion, $|\Phi(\xi)| > |\xi|$, and equation 3.4 finally becomes

$$\bar{e}(k) < \left[1 - \eta(k)\Phi'\left(\mathbf{x}^T(k)\mathbf{w}(k)\right)\|\mathbf{x}(k)\|_2^2\right] e(k), \tag{3.11}$$

which holds for negative η, which is not feasible.

4 The Case of a Recurrent Neural Filter

In this case, the gradient-descent updating equation regarding the recurrent neuron can be symbolically expressed as (Haykin, 1994)

$$\frac{\partial y(k)}{\partial \mathbf{w}(k)} = \Pi(k+1) = \Phi'\left(\mathbf{x}^T(k)\mathbf{w}(k)\right)[\mathbf{x}(k) + \mathbf{w}(k)\Pi(k)], \tag{4.1}$$

where vector Π denotes the set of corresponding gradients of the output neuron, and vector $\mathbf{x}(k)$ encompasses both the external and feedback inputs to the recurrent neuron. This case is a generalization of linear adaptive infinite impulse response (IIR) filters.

The correction to the weight vector at the time instant k becomes

$$\Delta\mathbf{w}(k) = \eta(k)e(k)\Pi(k). \tag{4.2}$$

The real time recurrent learning (RTRL) (Williams & Zipser, 1989) based learning algorithm for single-node a priori adaptation a posteriori error networks is now given by

$$\begin{aligned} \mathbf{w}(k+1) &= \mathbf{w}(k) + \eta(k)e(k)\Pi(k) \\ \bar{e}(k) &= d(k) - \Phi\left(\mathbf{x}^T(k)\mathbf{w}(k+1)\right). \end{aligned} \tag{4.3}$$

In the spirit of algorithm 1.4, following the same principle as for feedforward networks, we obtain the lower error bound for the a priori adaptation a posteriori error algorithm in single-node recurrent neural networks acting as nonlinear adaptive filters.

Theorem 3. *The lower bound for the a posteriori error obtained by algorithm 4.3 with constraint 1.3, and a contractive nonlinear activation function Φ, is*

$$\bar{e}(k) > \left[1 - \eta(k)\mathbf{x}^T(k)\Pi(k)\right]e(k), \tag{4.4}$$

whereas the range allowed for the learning rate $\eta(k)$ is given in the following corollary.

Corollary 3. *The range allowed for the learning rate $\eta(k)$ in an a priori adaptation a posteriori error algorithm, 4.3, with constraint 1.3, and the conditions given in theorem 3, is*

$$0 < \eta(k) < \frac{1}{\mathbf{x}^T(k)\Pi(k)}. \tag{4.5}$$

4.1 The Case of a General Recurrent Neural Network. For recurrent neural networks of the Williams-Zipser type (Williams & Zipser, 1989), with N neurons and one output neuron, the weight matrix update for an RTRL training algorithm can be expressed as

$$\Delta\mathbf{W}(k) = \eta(k)e(k)\frac{\partial y_1(k)}{\partial\mathbf{W}(k)} = \eta(k)e(k)\Pi_1(k), \tag{4.6}$$

where $\mathbf{W}(k)$ represents the weight matrix and $\Pi_1(k) = \frac{\partial y_1(k)}{\partial\mathbf{W}(k)}$ is the matrix of gradients at the output neuron $\pi^1_{n,l}(k) = \frac{\partial y_1(k)}{\partial w_{n,l}}$, where index n runs along the N neurons in the network, and index l runs along the inputs to the network. This equation is similar to equation 4.2, with the only difference being that weight matrix $\mathbf{W}$ replaces weight vector $\mathbf{w}$ and gradient matrix $\Pi = [\Pi_1, \ldots, \Pi_N]$ replaces gradient vector Π. Notice that in order to update matrix Π_1, a modified version of equation 4.1 has to update gradient matrices $\Pi_i, i = 1, \ldots, N$. More details about this procedure can be found in Williams and Zipser (1989) and Haykin (1994). Undertaking the analysis in the same manner as for a recurrent perceptron, we obtain the following conditions imposed on the learning rate η and the a posteriori error $\bar{e}$ for a priori learning a posteriori error recurrent neural predictor with an arbitrary size.

Corollary 4. *The lower bound for the a posteriori error obtained by an a priori learning a posteriori error RTRL algorithm, 4.6, with constraint 1.3, and a contractive nonlinear activation function Φ, is*

$$\bar{e}(k) > \left[1 - \eta(k)\mathbf{x}^T(k)\Pi_1(k)\right]e(k), \tag{4.7}$$

whereas the range of allowable learning rates $\eta(k)$ is

$$0 < \eta(k) < \frac{1}{\mathbf{x}^T(k)\Pi_1(k)}. \tag{4.8}$$

4.2 The Case of a Linear Activation Function. In the case of a linear activation function, which is neither contractive nor expansive, the nonlinear networks with a single neuron, for both the feedforward and recurrent case, degenerate respectively into linear adaptive finite impulse response (FIR) and IIR filters. A comprehensive analysis of such cases is given in Treichler et al. (1987) and Ljung and Soderstrom (1983).

5 The Case of the Logistic Activation Function

We provide a simple example of our analysis for the case of a logistic nonlinear activation function of a neuron. In section 2, we showed that the condition for the logistic activation function to be a contraction is $\beta < 4$. As such a function is monotone and increasing, the bound on its first derivative, 2.3, is $\Phi'(\xi) \leq \frac{\beta}{4}, \forall \xi \in \mathbb{R}$. That being the case, the conditions from theorem 1 and corollary 1 become, respectively,

$$\bar{e}(k) > \frac{1}{4}\left[4 - \eta(k)\beta\|\mathbf{x}(k)\|_2^2\right] e(k) \tag{5.1}$$

and

$$0 < \eta(k) < \frac{4}{\beta\|\mathbf{x}(k)\|_2^2}. \tag{5.2}$$

Based on theorem 3 and corollary 3, similar conditions can be derived for the recurrent case. These relationships considerably extend and shed additional light on the recently derived relations between the learning rate η and the slope in the nonlinear activation function β for a general a priori neural network (Thimm, Moerland, & Fiesler, 1996; Mandic & Chambers, 1999b).

6 Conclusions

We have provided relationships between the a priori and a posteriori prediction error, learning rate, and slope of the nonlinear activation function of a nonlinear adaptive filter realized by a neural network. This leads to the lower bound on the a posteriori error in a priori learning a posteriori error neural predictors, whose a posteriori output error is uniformly smaller in magnitude than the a priori one. The lower bound is derived based on the learning rate η, the first derivative of a general nonlinear activation function of a neuron around the current point on the error performance surface, and the $\mathcal{L}_2$ norm of the input vector. This has been achieved for learning algorithms based on gradient descent for both the feedforward and recurrent cases. In both cases, further conditions on the learning rate η are imposed so that the approach is feasible. In that case, a general nonlinear activation function of a neuron has to exhibit contractive behavior. The relationship

between the learning rate η and the slope β is further evaluated for the example of the logistic activation function.

References

Douglas, S. C., & Rupp, M. (1997). A posteriori updates for adaptive filters. In *Conference Record of the Thirty-First Asilomar Conference on Signals, Systems and Computers* (Vol. 2, pp. 1641–1645).

Gill, P. E., Murray, W., & Wright, M. H. (1981). *Practical optimization.* London: Academic Press.

Haykin, S. (1994). *Neural networks—A comprehensive foundation.* Englewood Cliffs, NJ: Prentice Hall.

Haykin, S. (1996). *Adaptive filter theory* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Ljung, L., & Soderstrom, T. (1983). *Theory and practice of recursive identification.* Cambridge, MA: MIT Press.

Luenberger, D. G. (1969). *Optimization by vector space methods*. New York: Wiley.

Mandic, D. P., & Chambers, J. A. (1998). A posteriori real time recurrent learning schemes for a recurrent neural network based non-linear predictor. *IEE Proceedings—Vision, Image and Signal Processing, 145*(6), 365–370.

Mandic, D. P., & Chambers, J. A. (1999a). Global asymptotic stability of nonlinear relaxation equations realised through a recurrent perceptron. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP–99)* (Vol. 2, pp. 1037–1040).

Mandic, D. P., & Chambers, J. A. (1999b). Relationship between the slope of the activation function and the learning rate for the RNN. *Neural Computation, 11*(5), 1069–1077.

Narendra, K. S., & Parthasarathy, K. (1990). Identification and control of dynamical systems using neural networks. *IEEE Transactions on Neural Networks, 1*(1), 4–27.

Narendra, K. S., & Parthasarathy, K. (1991). Gradient methods for the optimization of dynamical systems containing neural networks. *IEEE Transactions on Neural Networks, 2*(2), 252–262.

Thimm, G., Moerland, P., & Fiesler, E. (1996). The interchangeability of learning rate and gain in backpropagation neural networks. *Neural Computation, 8*, 451–460.

Treichler, J. R., Johnson, Jr., C. R., & Larimore, M. G. (1987). *Theory and design of adaptive filters*. New York: Wiley.

Williams, R., & Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural Computation, 1*, 270–280.

Zeidler, E. (1986). *Nonlinear functional analysis and its applications* (Vol. 1). Berlin: Springer-Verlag.

Received October 9, 1998; accepted May 26, 1999.