Neural Processing Letters **11:** 1–5, 2000.

Towards the Optimal Learning Rate for Backpropagation

DANILO P. MANDIC[1] and JONATHON A. CHAMBERS[2]
[1] *School of Information Systems, University of East Anglia, Norwich, NR4 7TJ, UK;* [2] *Dept. of Electrical and Electronic Engineering, Imperial College of Science, Technology and Medicine, Exhibition Road, SW7 2BT London, UK, E-mail: d.mandic@uea.ac.uk*

Abstract. A backpropagation learning algorithm for feedforward neural networks with an adaptive learning rate is derived. The algorithm is based upon minimising the instantaneous output error and does not include any simplifications encountered in the corresponding Least Mean Square (LMS) algorithms for linear adaptive filters. The backpropagation algorithm with an adaptive learning rate, which is derived based upon the Taylor series expansion of the instantaneous output error, is shown to exhibit behaviour similar to that of the Normalised LMS (NLMS) algorithm. Indeed, the derived optimal adaptive learning rate of a neural network trained by backpropagation degenerates to the learning rate of the NLMS for a linear activation function of a neuron. By continuity, the optimal adaptive learning rate for neural networks imposes additional stabilisation effects to the traditional backpropagation learning algorithm.

Key words: adaptive learning rate, backpropagation, feedforward neural networks, optimal gradient learning

1. Introduction

Algorithms based upon the recursive solution of the Wiener filter, have been heavily used in both linear and nonlinear real-time applications. In the area of linear adaptive filters, the most popular algorithm is the Least Mean Square (LMS) algorithm. However, its inherent limitations have forced researchers to try to improve its performance, through, for instance, the Normalised LMS (NLMS) [1, 2], a posteriori LMS [2], or through an adaptive learning rate. The idea behind the variable learning-rate LMS is that the algorithm runs with a large learning-rate (step-size), when the algorithm is far from the optimal solution, thus having a large convergence rate, whereas the algorithm runs with a small step-size when near the optimal solution, so as to achieve a low level of misadjustment. The criteria which have been proposed for the step-size adaptation are: squared instantaneous error [3]; sign changes of successive samples of the gradient [4]; reducing the squared error at each instant [5]; cross correlation of input and error [6]; square of a time-averaging estimate of the autocorrelation of two consecutive error terms [7]. However, the use of the above algorithms is rather application oriented, and generally limited to

only time-invariant, stationary systems. In addition, these algorithms are sensitive to noise [7]. Stabilisation for LMS has been achieved through the NLMS algorithm, but its derivation is rather involved, and requires the use of the Lagrange Multipliers method [8].

In the area of nonlinear adaptive processors, backpropagation is the most widely used gradient based algorithm. However, the analysis of adaptive learning rate in backpropagation has received little attention. The most popular algorithm for backpropagation with an adaptive learning rate is the delta-bar-delta rule [9], which also suffers from sensitivity to noise and relative instability [10, 11]. A further attempt to improve backpropagation-based algorithms was via *a posteriori* gradient algorithms for neural networks [12].

Here, we derive the optimal time-varying learning rate for a nonlinear adaptive neural network based upon backpropagation, which rests upon the value of instantaneous error, rather than on some statistical, or some empirical approach, as in the linear adaptive case.

2. The Optimal Step Size for a Single Neuron Neural Network

The equations that define the adaptation in backpropagation neural networks with one neuron are

$$e(k) = d(k) - \Phi\left(\mathbf{w}^T(k)\mathbf{X}(k)\right) \tag{1}$$

$$\mathbf{w}(k+1) = \mathbf{w}(k) - \eta\nabla_{\mathbf{w}(k)}e^2(k) \tag{2}$$

where $e(k)$ is the instantaneous error at the output neuron, $d(k)$ is some teaching (desired) signal, $\mathbf{w}(k) = [w_1(k), \ldots, w_N(k)]^T$ is the weight vector, $\mathbf{X}(k) = [x_1(k), \ldots, x_N(k)]^T$ is the input vector, and $(\cdot)^T$ denotes the vector transpose. The learning rate η is supposed to be a small positive real number. The nonlinear activation function of a neuron is denoted by Φ.

Equation (2) can be rewritten as

$$\mathbf{w}(k+1) = \mathbf{w}(k) + 2\eta\Phi'\left(\mathbf{w}^T(k)\mathbf{X}(k)\right)e(k)\mathbf{X}(k) \tag{3}$$

By expanding the error term (1) with a Taylor series, we obtain

$$\begin{aligned} e(k+1) = \; & e(k) + \sum_{i=1}^{N}\frac{\partial e(k)}{\partial w_i(k)}\Delta w_i(k) \\ & + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\partial^2 e(k)}{\partial w_i(k)\partial w_j(k)}\Delta w_i(k)\Delta w_j(k) + \cdots \end{aligned} \tag{4}$$

where only the first two terms will be considered.

From (1), the first partial derivatives can be obtained as

$$\frac{\partial e(k)}{\partial w_i(k)} = -\Phi'\left(\mathbf{w}^T(k)\mathbf{X}(k)\right)x_i(k) \quad i = 1, 2, \ldots, N \tag{5}$$

and from (3), the weight correction is obtained by

$$\begin{aligned} \Delta w_i(k) &= w_i(k+1) - w_i(k) \\ &= 2\eta\Phi'\left(\mathbf{w}^T(k)\mathbf{X}(k)\right)e(k)x_i(k) \quad i = 1, 2, \ldots, N \end{aligned} \tag{6}$$

Now, combining (4), (5), and (6), we obtain

$$e(k+1) = e(k) - 2\eta\left[\Phi'\left(\mathbf{w}^T(k)\mathbf{X}(k)\right)\right]^2 e(k)\sum_{i=1}^{N} x_i^2(k) \tag{7}$$

The instantaneous squared error is therefore given by

$$e^2(k+1) = e^2(k)\left[1 - 2\eta\left[\Phi'\left(\mathbf{w}^T(k)\mathbf{X}(k)\right)\right]^2 \sum_{i=1}^{N} x_i^2(k)\right]^2 \tag{8}$$

In order to obtain the minimum of (8), we differentiate with respect to η, and obtain the optimal value of learning rate $\eta_{OPT}(k)$ for a backpropagation trained perceptron as

$$\eta_{OPT}(k) = \frac{1}{2\left[\Phi'\left(\mathbf{w}^T(k)\mathbf{X}(k)\right)\right]^2 \sum_{i=1}^{N} x_i^2(k)} \tag{9}$$

Denoting the term $\mathbf{w}^T(k)\mathbf{X}(k)$ by $net(k)$, and recognising that $\sum_{i=1}^{N} x_i^2(k) = \| \mathbf{X}(k) \|^2$, we obtain the following final expression for $\eta_{OPT}(k)$

$$\eta_{OPT}(k) = \frac{1}{2\left[\Phi'\left(net(k)\right)\right]^2 \| \mathbf{X}(k) \|_2^2} \tag{10}$$

Notice that this relationship is closely related to the learning rate in the NLMS algorithm for linear adaptive filters. Indeed, for a linear activation function of a neuron, the adaptive learning rate from (9) becomes exactly the learning rate in the NLMS algorithm. For a nonlinear activation function of a neuron, the learning rate becomes normalised by the tap input power of the input signal to a perceptron, multiplied by the squared derivative of the nonlinear activation function at the current point $net(k)$. Hence, we will refer to the result from (9) and (10) as the Normalised Backpropagation (NBP) algorithm for neural networks consisting of a single perceptron.

3. The Adaptive Step Size Algorithm for a Multilayer Backpropagation Network

The NBP algorithm for a general feedforward neural network trained by backpropagation can be derived from the corresponding algorithm for a single perceptron. For simplicity, we consider a general feedforward network with an arbitrary number of hidden layers, and corresponding neurons, and with only one

output neuron. The notion of local gradient δ is introduced [10] such that the local gradient $\delta_1^{(M)}(k)$ for the neuron in the Mth (output) layer $\mathcal{L}_M$ is

$$\delta_1^{(M)}(k) = \Phi'\left(net_1^{(M)}(k)\right) e(k) \tag{11}$$

whereas for the ith neuron in the $(l-1)$th hidden layer $\mathcal{L}_{l-1}$, the local gradient is

$$\delta_i^{(l-1)}(k) = \Phi'\left(net_i^{(l-1)}(k)\right) \sum_{j \in \mathcal{L}_l} w_{j,i}(k)\delta_j^{(l)}(k) \;\; 1 \leq l < M, \;\; i \in \mathcal{L}_{l-1} \tag{12}$$

Here, the activation of a neuron $y_i^{(l-1)}(k)$ given by $\sum_j w_{i,j}^{(l-1)}(k) y_j^{(l-2)}(k)$ is denoted by $net_i^{(l-1)}(k)$. Then, at the time instant k, the correction to the weight connecting the ith neuron in the $(l-1)$th layer and the jth neuron in the $(l-2)$nd layer $\Delta w_{i,j}^{(l-1)}(k)$ becomes (9)

$$\Delta w_{i,j}^{(l-1)}(k) = 2\eta z_j(k)\delta_i^{(l-1)}(k) \tag{13}$$

where the input signal to the ith neuron in the $(l-1)$th layer is

$$z_j(k) = \begin{cases} x_j(k), & \text{neuron } i \text{ in the first layer} \\ y_j(k), & \text{neuron } i \text{ in layer } l, \;\; 1 < l \leq M \end{cases} \tag{14}$$

Notice that due to (11)–(13), all the local gradients in a general network, and therefore all the weight corrections in the network at the time instant k, become multiplied by the instantaneous output error $e(k)$. Now, recognising the connection between the terms $\Delta w_{i,j}^{(l-1)}(k)$ and $\frac{\partial e(k)}{\partial w_{i,j}^{(l-1)}(k)}$, and undertaking the same procedure as for the case of a single perceptron, we obtain

$$\eta_{OPT}(k) = \frac{1}{2\left[\Phi'\left(net_1^{(M)}(k)\right)^2 \sum_{i \in \mathcal{L}_{M-1}} z_i^2(k) + \cdots + \sum_{m \in \mathcal{L}_1} \delta_m^2(k) \sum_{n \in \mathcal{L}_0} x_n^2(k)\right]} \tag{15}$$

Although straightforward, the backpropagation algorithm with the optimal adaptive learning rate for a general multilayer feedforward network depends on the size and topology of a particular network, and its mathematical expression can be rather clumsy. However, the learning rate η is again normalised by the sums of total tap input power to every neuron in the network, multiplied again by the square of appropriate local gradients δ.

4. Conclusions

We have derived the Normalised Backpropagation algorithm (NBP) for a class of feedforward neural networks. The algorithm runs with an adaptive learning rate which is based upon the $\mathcal{L}_2$ norm of the appropriate input vector and local

gradient at the neuron. The algorithm is derived from the instantaneous output error of the network, based upon the Taylor series expansion of the output error. No ad hoc rules, such as in the frequently considered Least Mean Square (LMS) algorithm with an adaptive step size, are included. The algorithm shows behaviour correspondent to that of the Normalised Least Mean Square (NLMS) algorithm, and by continuity, imposes additional stability on the backpropagation algorithm. This makes the NBP algorithm likely to be very useful in real time applications, such as nonlinear prediction of statistically nonstationary signals.

References

1. Ljung, L. and Soderstrom, T.: *Theory and Practice of Recursive Identification*, MIT Press, Cambridge, MA.,1983.
2. Treichler, J. R., Johnson, Jr., C. R. and Larimore, M. G.: *Theory and Design of Adaptive Filters*, John Wiley & Sons, New York, 1987.
3. Kwong, R. H. and Johnston, E. W.: A variable step size LMS algorithm, *IEEE Transactions on Signal Processing* **40**(7) (1992), 1633–1641.
4. Evans, J. B., Xue, P. and Liu, B.: Analysis and implementation of variable step size adaptive algorithms, *IEEE Transactions on Signal Processing* **41**(8) (1993), 2517–2535.
5. Mathews, V. J. and Xie, Z.: A stochastic gradient adaptive filter with gradient adaptive step size, *IEEE Transactions on Signal Processing* **41**(6) (1993), 2075–2087.
6. Shan, T. J. and Kailaith, T.: Adaptive algorithms with an automatic gain control feature, *IEEE Transactions on Acoustics, Speech and Signal Processing* **35**(1) (1988), 122–127.
7. Aboulnasr, T. and Mayyas, K.: A robust variable step-size LMS-type algorithm: Analysis and simulations, *IEEE Transactions on Signal Processing* **45**(3) (1997), 631–639.
8. Haykin, S.: *Adaptive Filter Theory*, Prentice-Hall, 3d ed., Englewood Cliffs, NJ, 1996.
9. Jacobs, R. A.: Increased rates of convergence through learning rate adaptation, *Neural Networks* **1** (1988), 295–307.
10. Haykin, S.: *Neural Networks – A Comprehensive Foundation*, Prentice Hall, Englewood Cliffs, NJ, 1994.
11. Douglas, S. C. and Cichocki, A.: On-line step-size selection for training of adaptive systems, *IEEE Signal Processing Magazine* **14**(6) (1997), 45–46.
12. Mandic, D. P. and Chambers, J. A.: 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) (1998), 365–370.