PERIODICA POLYTECHNICA SER. EL. ENG. VOL. 37, NO. 2, PP. 119-129 (1993)

AN ADAPTIVE LEAST SQUARES ALGORITHM FOR THE EFFICIENT TRAINING OF ARTIFICIAL NEURAL NETWORKS

Annamária R. VÁRKONYI-KÓCZY

Department of Measurement and Instrument Engineering
Technical University of Budapest
H-1521 Budapest, Hungary
email: koczy@mmt.bme.hu
Phone and Fax: + 36 1 166-4938

Received: September 20, 1993

Abstract

Recently a number of publications have proposed alternative methods to apply in least mean square (LMS) algorithms in order to improve convergence rate. It has been also shown that variable step size methods can provide better convergence speed than the fixed step size ones. This paper introduces a new algorithm for the on-going calculation of the step size, and investigates its applicability in the training of multilayer neural networks. The proposed method seems to be efficient at least in the case of lower level additive input noise.

Keywords: adaptive step size, training of neural networks.

1. Introduction

A number of recent publications on learning systems show that adaptation rate issues are still very important research topics since practical problems always require higher and higher rates of convergence. This paper concentrates on variable step size methods related to the LMS algorithm. The LMS algorithm is one of the possible alternatives for solving adaptation problems. As it is explained in (BELLANGER, 1992), the current state of adaptation algorithms can be characterized by the competition of the LMS algorithms with the recursive least squares (RLS) type of algorithms. The former one is computationally simple while this latter is more efficient at the price of higher computational burden. As it is shown in (WIDROW et al., 1984), the LMS algorithm seems to be more appropriate in the case of nonstationary inputs, and can be considered as a starting point for many algorithms related to nonlinear applications (see e.g. (WIDROW et al., 1990)).

Many quite recent results show that the application of variable step size can considerably improve convergence at a relatively low price. For the RLS algorithm see e.g. (DINIZ et al., 1992) where the computation of an optimal convergence factor is proposed for conventional transversal

FIR filter realization, and for the subband adaptive filtering method. A similar example is the work of Kollias (KOLLIAS et al., 1989) where a higher order approach together with variable step size has been suggested for the training of artificial neural networks.

In Section 2 of this paper it is shown that for the LMS type algorithms, based on the adaptive linear combine approach of Widrow (see. e.g. (WIDROW et al., (1985)), the coefficient error can be reduced in every step optimally if the input is not noisy, and the convergence factor μ is variable. For the variable μ a simple explicit formula is given.

In Section 3 the application of this optimal convergence factor is extended to such 'nonlinear combine as multilayer neural networks. This extension is an approximation, however, according to the simulations this approximation can improve the efficiency of the training in comparison to the conventional backpropagation algorithms even if it is extended with the momentum technique (see e.g. WIDROW et al., (1990)).

2. Optimum Convergence Rate for the LMS Algorithm

In the LMS-type adaptation schemes the parameters are updated in the following form (see e.g. (WIDROW et al., (1985)):

$$\mathbf{W}_{k+1} = \mathbf{W}_k + \mu \varepsilon_k \mathbf{X}_k \tag{1}$$

where

$\mathbf{W}_k$: is the parameter vector of the linear combiner,
$\mathbf{X}_k$: is the input vector of the linear combiner, the so-called regression vector,
ε_k : stands for the output error: $\varepsilon_k = \mathbf{d}_k - \mathbf{y}_k$, where
$\mathbf{d}_k$: is the desired sample,
$\mathbf{y}_k$: is the output sample: $\mathbf{y} = \mathbf{W}_k^T \mathbf{X}_k$, and
μ : is the (possible variable) step size.

The classical solutions for the selection of the step size suggest two alternatives (see e.g. (WIDROW et al., 1990)). The first one is the so-called μ-LMS algorithm which operates with a small constant step size. The second one is the so-called α-LMS algorithm which applies a time-varying step size of the form of

$$\mu = \frac{\alpha}{\mathbf{X}_k^T \mathbf{X}_k}; \qquad 0 < \alpha < 2. \tag{2}$$

For practical applications

$$0 < \alpha < 1$$

settings are suggested. If we investigate the step size problem through the convergence of parameters, we can utilize the stability theory approach used in the general convergence analysis of time varying and adaptive systems (see e.g. JOHNSON, 1984). This relates the convergence problem to the behaviour of such a homogeneous system where the state variables correspond to the coefficient errors. If $\mathbf{W}^*$ stands for the ideal parameter vector to be approximated, and $\Delta\mathbf{W}_k = \mathbf{W}^* - \mathbf{W}_k$ denotes the parameter error at the kth iteration, then the next step results in a parameter error of

$$\Delta\mathbf{W}_{k+1} = (\mathbf{I} - \mu\mathbf{X}_k\mathbf{X}_k^T)\Delta\mathbf{W}_k. \tag{3}$$

At this point let us introduce a generalization of the step size μ, and replace it by a full step size matrix $\mathbf{M}$. This means that the state transition matrix in (3) has a form of $\mathbf{I} - \mathbf{A}_k\mathbf{X}_k^T$ where $\mathbf{A}_k = \mathbf{M}\mathbf{X}_k$ stands for a vector. The investigation of this state transition matrix shows that the maximum reduction in the parameter error in (3) can be achieved if $\mathbf{M} = \mathbf{diag} < \mu, \mu, \ldots \mu >$, and

$$\mu = \frac{1}{\mathbf{X}_k^T\mathbf{X}_k}. \tag{4}$$

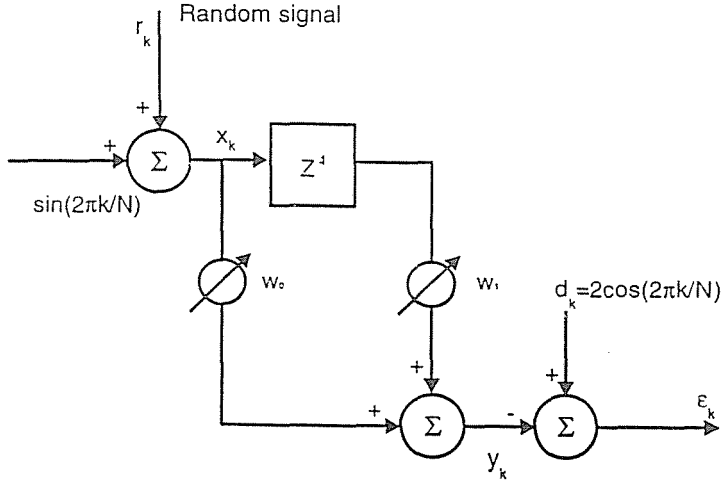

Fig. 1. The adaptive linear combiner example with a random signal added at the input

It is interesting to note that this result corresponds to the α-LMS algorithm with $\alpha = 1$. Obviously, (4) is a result where the inputs are not considered to be noisy observations. For the case of noisy inputs, depending on the

noise level, the application of $\alpha < 1$ is advisable. As a simple example, and for comparison we consider the example of Widrow (see WIDROW et al., 1985, p. 103). The adaptive system can be seen in *Fig. 1.* The input is a sine wave consisting of $N = 16$ samples for every period with a random signal added as input noise. The desired signal is the cosine waveform. For the practical case use of no input noise (the amplitude range of the noise is 0.001) the RMS error can be followed in *Fig. 2.* Two runs were accomplished for every step size setting: the name 'upper track' stands for the case with initial condition $(w_0, w_1) = (0, 0)$, while the name 'lower track' stands for the $(w_0, w_1) = (4, -10)$ initial condition case. *Fig. 3* shows the convergence on the parameter plane.

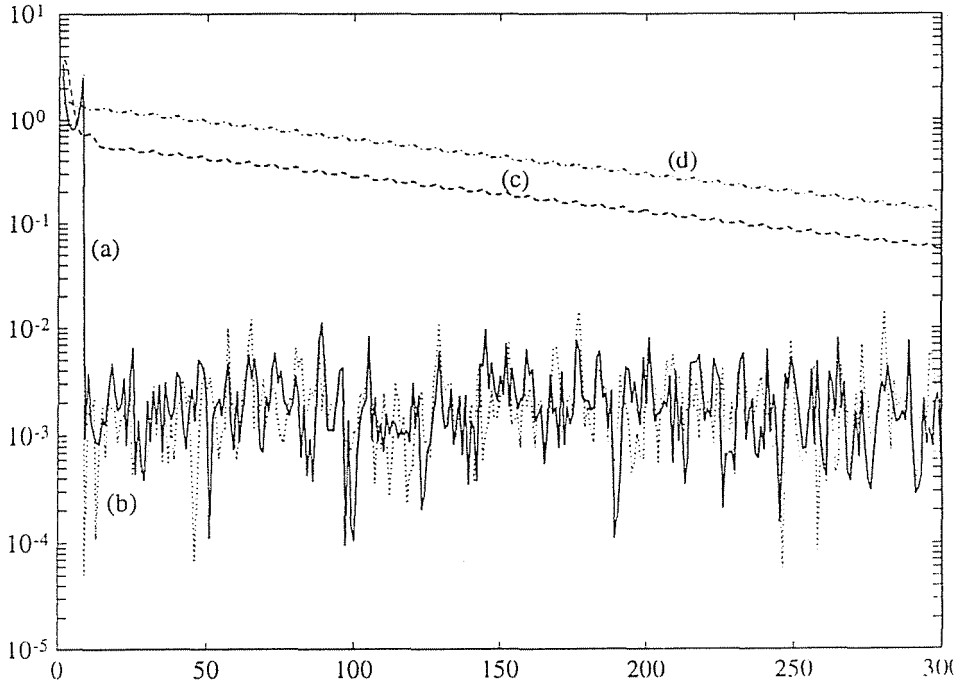

Fig. 2. RMS error versus iteration number for the example of *Fig. 1* with $N = 16$, and noise amplitude = 0.001
(a) adaptive μ 'lower track' b) adaptive μ 'upper track'
(c) $\mu = 0.1$ 'lower track' (d) $\mu = 0.1$ 'upper track'

In the case of noisy inputs, as it is known from the literature of the LMS algorithm (see e.g. WIDROW et al., 1985), as the algorithm itself does not provide proper filtering for the noise, the convergence is more problematic, and the application of (4) results in noisy parameter estimates: the

noise reduction effect of the μ-LMS algorithm with small μ over a large number of iterations is not present. However, the 'aggressivity' of the relatively large step size may improve convergence possibly with a combination of the constant and small μ-based methods. *Figs. 4–6* show the behaviour of the two algorithms in the noisy input case. It can be observed that the tracks for the variable step size case heavily oscillate, but can reach smaller RMS error sooner than the fixed μ alternatives. If the algorithm is followed by an on-line RMS calculation, then it can be easily combined with the μ-LMS algorithm.

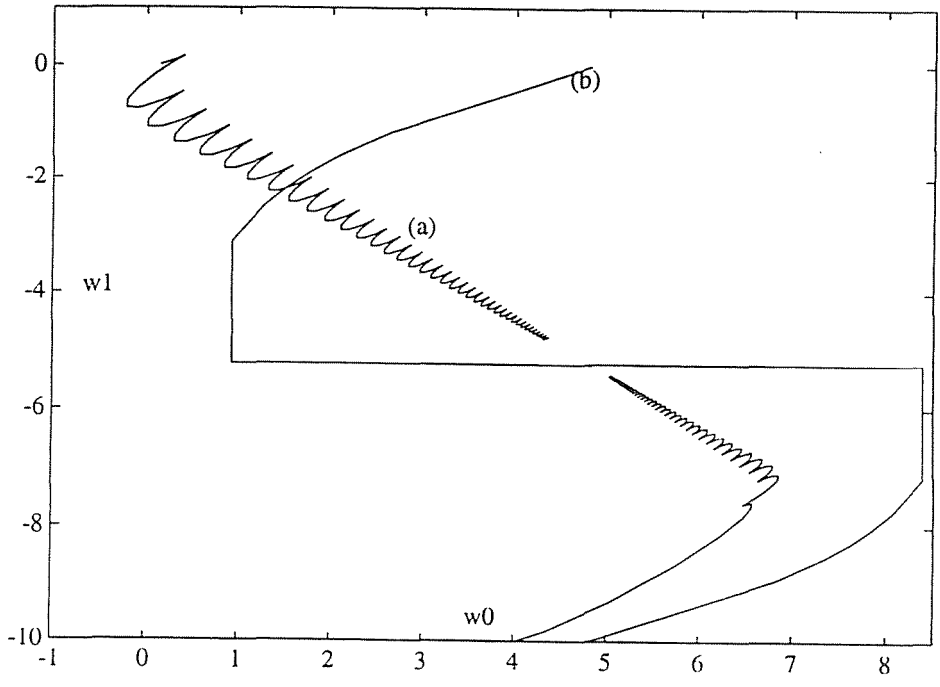

Fig. 3. Weight value tracks for the example of *Fig. 1* with $N = 16$, noise amplitude = 0.001
(a) $\mu = 0.1$ (b) adaptive μ

3. Pseudolinear Regressions with Variable Step Size

In adaptive IIR filtering (see e.g. WIDROW et al., 1985) the adaptation mechanism of the parameters is very similar to the case of (1), however, there is a considerable difference in the content. In IIR adaptive filtering

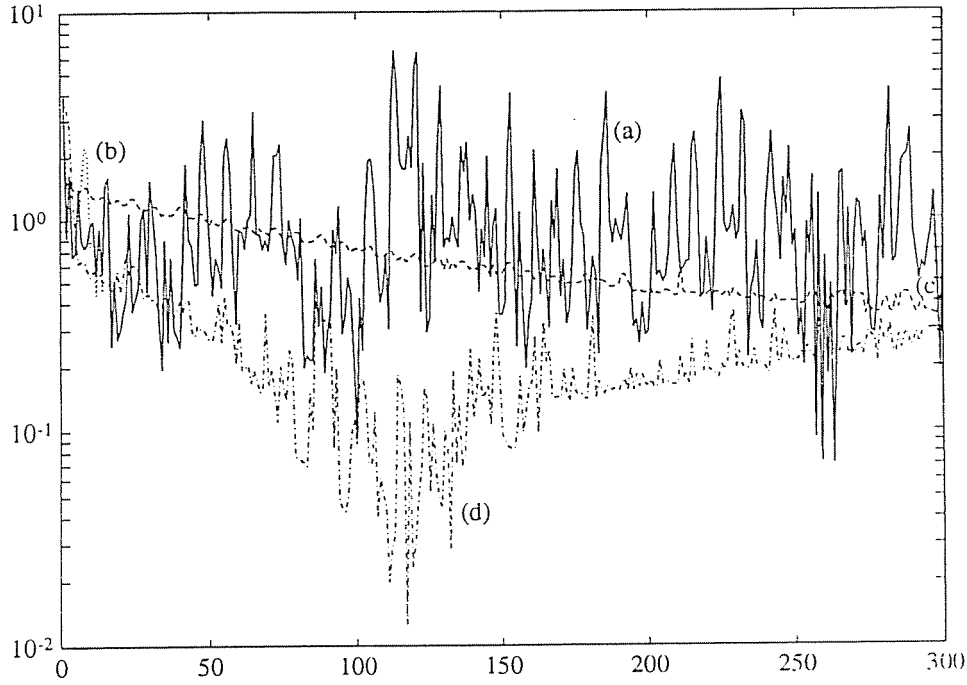

Fig. 4. RMS versus iteration number for the example of *Fig. 1* with $N = 16$, amplitude of noise = 0.34
(a) adaptive μ 'lower track' (b) adaptive μ 'upper track'
(c) $\mu = 0.1$ 'upper track' (d) $\mu = 0.1$ 'lower track'

problems the outputs cannot be regarded as linear regressions, since the regression vector has implicit parameter dependencies. For this very reason the above development is not directly applicable. However, to improve the convergence, the variable step size is a possible alternative even for these filters. In the literature several propositions can be found for time varying step size matrices (see e.g. MOHAMMAD, 1993). The relation of these propositions to the stability theory approach is still an open problem.

Another direction of generalization of the LMS-type algorithms is the training of artificial neural networks. The famous backpropagation algorithm (see e.g. WIDROW et al., 1990) is a typical example: it follows the μ-LMS rule. The second proposition of this paper is the application of result (4) to the backpropagation algorithm, even if due to the nonlinearity the parameter error 'propagation' cannot be expressed in the form of (3). However, if we apply a step size calculated according to (4), then we accept an approximation of the error $\varepsilon_k = \Delta \mathbf{W}_k \mathbf{X}_k$, where $\mathbf{X}_k$ stands for the

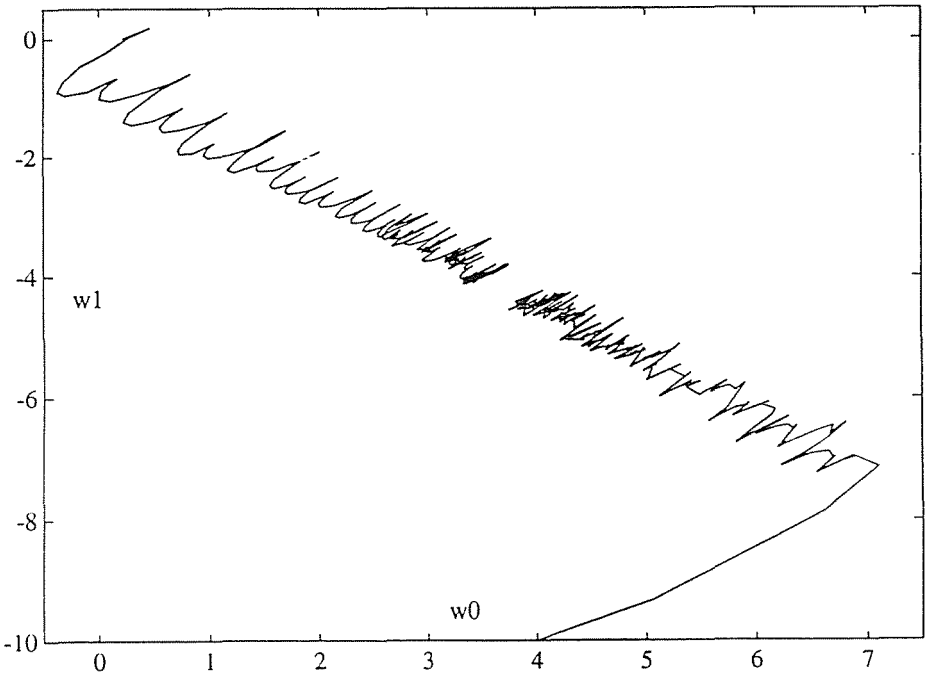

Fig. 5. Weight-value tracks for the LMS algorithm operating as in *Fig.* 1 with $N = 16$, noise amplitude 0.34, and $\mu = 0.1$

derivative of $\mathbf{y}_k$ in the usual forms of the backpropagation. The properties of this approximation are still under investigation, however, the first simulations show that this approximation results in better convergence, and again with the combination of the original algorithm, can provide faster convergence with negligible increase of computational complexity. As an illustration, first a very simple example is considered (see *Fig. 7*). The function to be approximated with a two layered three neuron network is the $\mathbf{y} = \mathbf{th}\ (\mathbf{x})$. During the iteration the error y' has been approximated. *Fig. 8* shows the result where the superiority of the varying step size can be observed. As a more complex, but still simple example is the approximation of the function $\mathbf{y} = 4\ \mathbf{x}/(\mathbf{1} + 4\ \mathbf{x}^2)$. The network is shown in *Fig. 9*, and the RMS error can be seen in *Fig. 10*. Simulations based on standard backpropagation networks with momentum technique show worse behaviour than the varying step size technique does.

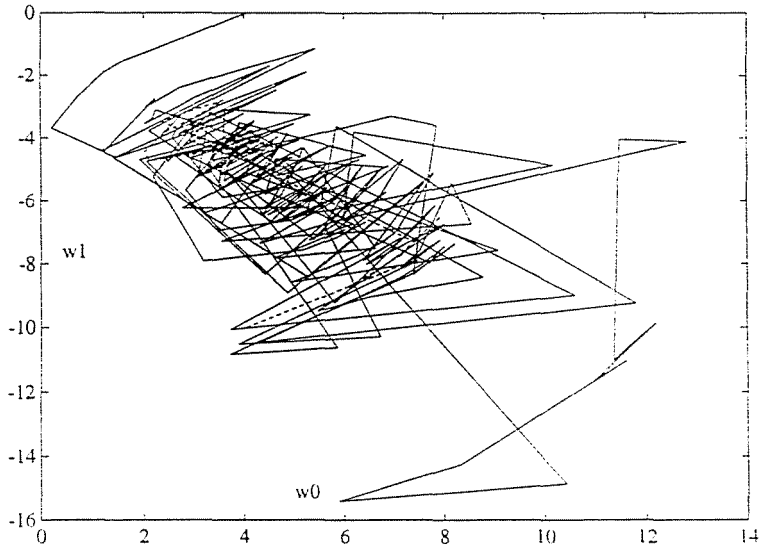

Fig. 6. Weight-value tracks for the LMS algorithm operating as in *Fig. 1* with adaptive μ, $N = 16$, and noise amplitude 0.34

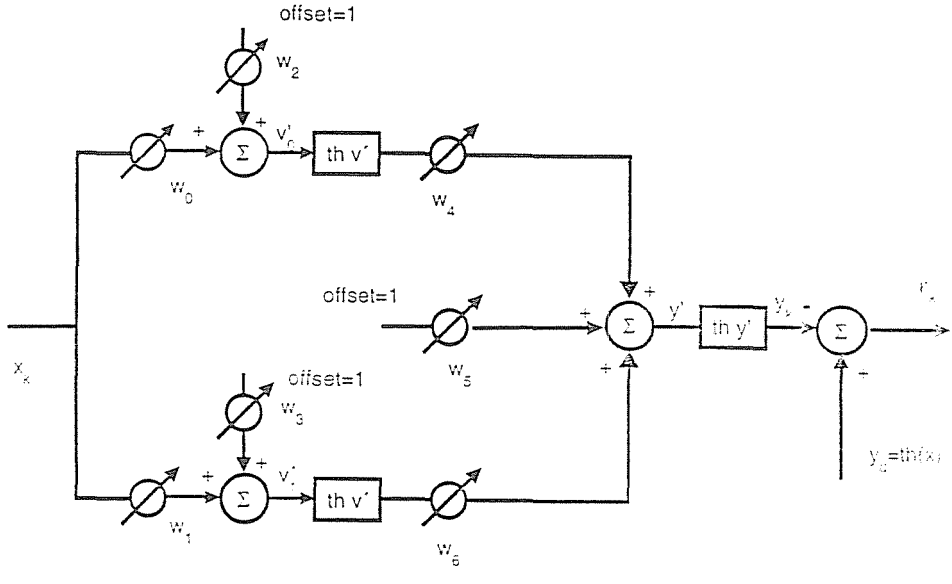

Fig. 7. The adaptive neural network for $y = th(x)$

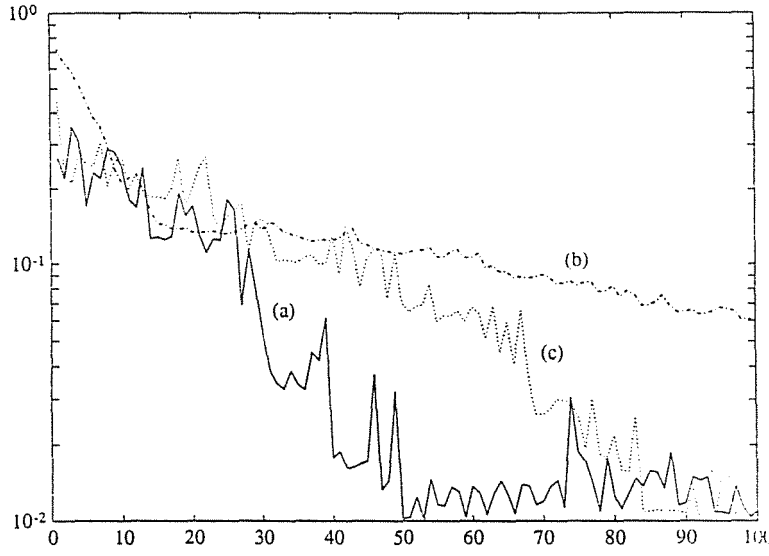

Fig. 8. RMS error of the example of *Fig. 7.*
(a) adaptive μ (b) $\mu = 0.1$
(c) $\mu = 0.4$

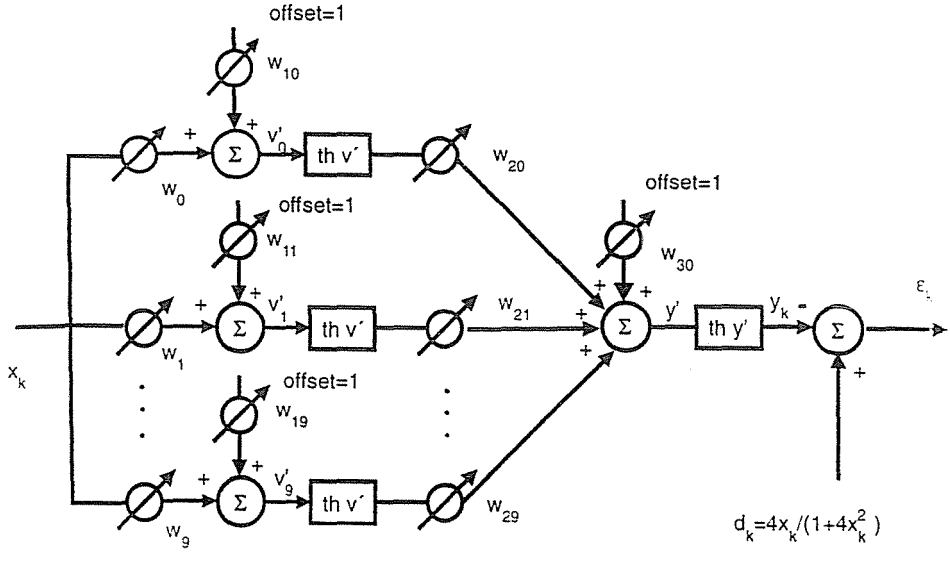

Fig. 9. The adaptive neural network for $y = 4x/(1 + 4x^2)$

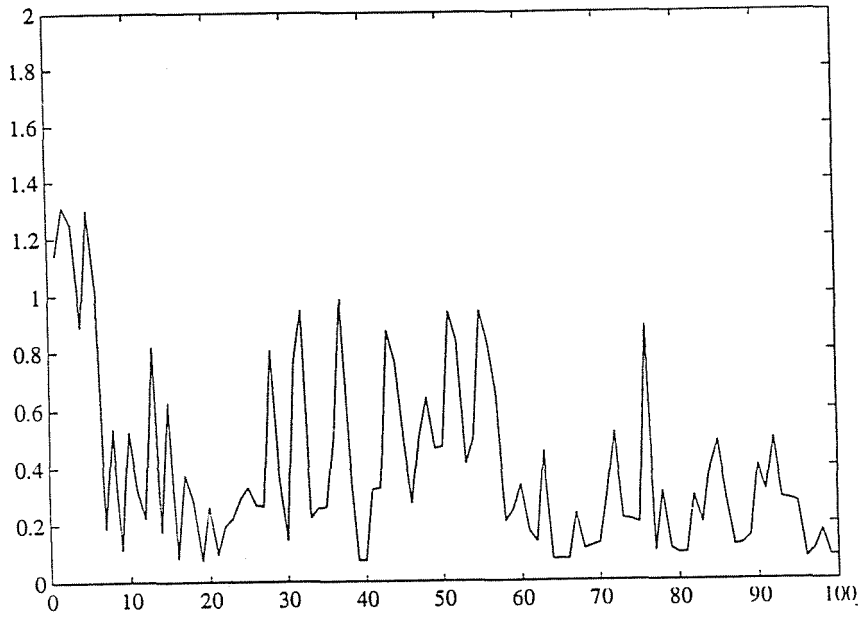

Fig. 10. The RMS error of the example of *Fig. 9*

4. Conclusions

In this paper a variable step size method has been reported for LMS-type learning schemes. The approach seems to be 'aggressive', and efficient especially for cases of low input noise. The variable step size suggested here for adaptive linear combine-type adaptive systems provides maximum parameter error reduction in every iteration step even for that case where initially a 'full' step size matrix is considered. In the second part this result has been extended to the 'pseudolinear' case, and especially to the training of artificial neural networks. The properties of the suggested algorithm in the case of larger input noise, and in the 'pseudolinear' case require further investigations.

References

BELLANGER, M. (1992): Trends in Adaptive Filtering. In SIGNAL PROCESSING VI: Theories and Applications. Vanderwalle, J., Boite, R., Moonen, M., Oosterlinck, A. (eds.). Elsevier Science Publishers B. V. pp. 47–54.

DINIZ, P. S. R. – BISCAINHO, L. W. P. (1992): Optimal Variable Step Size for the LMS/Newton Algorithm with Application to Subband Adaptive Filtering, *IEEE Transactions on Signal Processing*, Vol. 40, No. 11, pp. 2825–2829.

JOHNSON, C. R. (1984): Adaptive IIR Filtering: Current Results and Open Issues, *IEEE Transactions on Information Theory*, Vol. IT–30, No. 2, pp. 237–250.

KOLLIAS, S. – ANASTASSIOU, D. (1989): An Adaptive Least Squares Algorithm for the Efficient Training of Artificial Neural Networks, *IEEE Transactions on Circuits and Systems*, Vol. 36, No. 8, pp. 1092–1101.

MOHAMMAD, TH. N. (1993): Improved Adaptive Signal Processing Algorithms. Candidate of Technical Science Thesis. Hungarian Academy of Sciences, Budapest.

WIDROW, B. – WALACH, E. (1984): On the Statistical Efficiency of the LMS Algorithm with Nonstationary Inputs. *IEEE Transactions on Information Theory*, Vol. IT–30, No. 2, pp. 211–221.

WIDROW, B. – STEARNS, S. D. (1985): Adaptive Signal Processing. Englewood Cliffs, NJ: Prentice Hall.

WIDROW, B. – LEHR, M. A. (1990): 30 Years of Adaptive Neural Networks: Perceptron, Madaline, and Backpropagation. *Proceedings of the IEEE*, Vol. 78, No. 9, pp. 1415–1442.