A Generalized Normalized Gradient Descent Algorithm

Danilo P. Mandic

***Abstract*—A generalized normalized gradient descent (GNGD) algorithm for linear finite-impulse response (FIR) adaptive filters is introduced. The GNGD represents an extension of the normalized least mean square (NLMS) algorithm by means of an additional gradient adaptive term in the denominator of the learning rate of NLMS. This way, GNGD adapts its learning rate according to the dynamics of the input signal, with the additional adaptive term compensating for the simplifications in the derivation of NLMS. The performance of GNGD is bounded from below by the performance of the NLMS, whereas it converges in environments where NLMS diverges. The GNGD is shown to be robust to significant variations of initial values of its parameters. Simulations in the prediction setting support the analysis.**

***Index Terms*—Adaptive filtering, gradient adaptive learning rate, nonlinear prediction, normalized least mean square.**

I. Introduction

THE LEAST mean square (LMS) algorithm is a simple, yet most frequently used, algorithm for adaptive finite-impulse response (FIR) filters. It is described by the following [1]:

$$e(k) = d(k) - \mathbf{x}^T(k)\mathbf{w}(k) \tag{1}$$

$$\mathbf{w}(k+1) = \mathbf{w}(k) + \mu e(k)\mathbf{x}(k) \tag{2}$$

where $e(k)$ is the instantaneous error at the output of the filter for the time instant k, $d(k)$ is the desired signal, $\mathbf{x}(k) = [x(k-1), \ldots, x(k-N)]^T$ is the input signal vector, N is the length of the filter, $(\cdot)^T$ is the vector transpose operator, and $\mathbf{w}(k) = [w_1(k), \ldots, w_N(k)]^T$ is the filter coefficient (weight) vector. The parameter μ is the step size (learning rate) that defines how fast the algorithm is converging along the error performance surface defined by a cost function $E(k) = (1/2)e^2(k)$ and is critical to the performance of LMS. Ideally, we want an algorithm for which the speed of convergence is fast and the steady-state misadjustment is small when operating in a stationary environment, whereas in a nonstationary environment the algorithm should change the learning rate according to the dynamics of the input signal, so as to achieve as good a performance as possible.

To that cause, the normalized LMS (NLMS) algorithm has been introduced [1]. The step size of NLMS was found to be $\eta(k) = \mu/\| \mathbf{x}(k) \|_2^2$, $0 < \mu < 2$, where $\| \cdot \|_2$ denotes the Euclidean norm. The derivation and analysis of NLMS rest upon the usual independence assumptions,[1] and in theory, value $\mu = 1$ provides the fastest convergence [1], whereas in practice, the step size of the NLMS algorithm needs to be considerably smaller.[2] To preserve stability for close-to-zero input vectors, the optimal NLMS learning rate is usually modified as $\mu/\| \mathbf{x} \|_2^2 \rightarrow \mu/(\| \mathbf{x}(k) \|_2^2 + \varepsilon)$, where ε is a small positive constant.

However, input signals with unknown and possibly very large dynamical range, an ill-conditioned tap input autocorrelation matrix and coupling between different signal modes can lead to divergence of LMS and a poor performance if not divergence of NLMS. To deal with these problems, a number of gradient adaptive step size LMS algorithms have been developed in the last two decades, examples of which are algorithms by Kuzminskiy [2], Mathews [3], and Benveniste [4]; the mathematical description of the latter two is given in the Appendix. The condition for optimal adaptation in this sense is $dE/d\mu = 0$.[3] Benveniste's algorithm was derived rigorously, without taking into account the independence assumptions, resulting in computationally demanding learning rate updates, whereas Mathews' algorithm uses instantaneous gradients for learning rate adaptation. Based upon Benveniste's algorithm, to reduce its computational complexity, a class of variable step size algorithms was recently proposed in [5]. Other improvements include imposing hard constraints on the lower and upper bounds for the step size, superimposing regression on the step size sequence [6], and reducing the computational complexity by employing sign algorithms [7]. Morgan and Kratzer [8] provide a review of the existing adaptive step size NLMS-based algorithms, whereas in [9], graded updates (individual step sizes for every filter coefficient) are thoroughly analyzed in the LMS setting.

A major disadvantage of the algorithms based upon estimators of $\partial E(k)/\partial \mu$ is their sensitivity to the time correlation between input signal samples and to the value of the additional step size parameter that governs the gradient adaptation of the step size. To this cause, a generalized normalized gradient descent (GNGD) algorithm is proposed here, which is based upon the NLMS, where an additional stabilization and faster convergence are introduced by making the compensation term ε in the denominator of the NLMS step size gradient adaptive. Unlike

Manuscript received August 13, 2002; revised January 15, 2003. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Soo-Chang Pei.

The author is with the Department of Electrical and Electronic Engineering, Imperial College, London SW7 2BT, U.K. (e-mail: d.mandic@ic.ac.uk).

Digital Object Identifier 10.1109/LSP.2003.821649

[1]The independence assumptions used in the analysis of adaptive filters are: 1) sequences $\mathbf{x}(k)$ and $\mathbf{w}(k)$ are zero mean, stationary, jointly normal, and with finite moments; 2) the successive increments of tap weights are independent of one another; and 3) the error and $\mathbf{x}(k)$ sequences are statistically independent of one another.

[2]In almost all analyses of the class of LMS adaptive filters, it is assumed that the filter coefficients are statically independent of the input data currently in filter memory, an assumption that is incorrect for shift-input data.

[3]Notice that in the steady state this condition leads to $\mu(\infty) = 0$.

other variable step size algorithms, which operate satisfactorily if the initial learning rate is set close to the optimal learning rate of LMS (which is not known beforehand), the GNGD is robust to changes in the initialization of the step size adaptation parameter ρ, compensation term ε, and learning rate μ. The analysis is supported by simulations on colored, nonlinear and nonstationary signals.

II. GNGD Algorithm

Due to noise, ill-conditioned tap input correlation matrix, close-to-zero value of the input data vector, or a large learning rate μ, the NLMS algorithm (3) is not optimal for many practical settings

$$\begin{aligned}\mathbf{w}(k+1) &= \mathbf{w}(k) + \frac{\mu}{\| \mathbf{x}(k) \|_2^2 + \varepsilon} e(k)\mathbf{x}(k) \\ &= \mathbf{w}(k) + \eta(k) e(k)\mathbf{x}(k). \end{aligned} \tag{3}$$

To that cause, parameter ε in (3) is made gradient adaptive as

$$\varepsilon(k+1) = \varepsilon(k) - \rho \nabla_{\varepsilon(k-1)} E(k). \tag{4}$$

Using the chain rule, the gradient $\nabla_{\varepsilon(k-1)} E(k)$ can be evaluated as

$$\begin{aligned}\frac{\partial E(k)}{\partial \varepsilon(k-1)} &= \frac{\partial E(k)}{\partial e(k)} \frac{\partial e(k)}{\partial y(k)} \frac{\partial y(k)}{\partial \mathbf{w}(k)} \frac{\partial \mathbf{w}(k)}{\partial \eta(k-1)} \frac{\partial \eta(k-1)}{\partial \varepsilon(k-1)} \\ &= \frac{e(k)e(k-1)\mathbf{x}^T(k)\mathbf{x}(k-1)}{\left(\| \mathbf{x}(k-1) \|_2^2 + \varepsilon(k-1)\right)^2}. \end{aligned} \tag{5}$$

The proposed GNGD algorithm is therefore described by

$$\begin{aligned} y(k) &= \mathbf{x}^T(k)\mathbf{w}(k) \\ e(k) &= d(k) - y(k) \\ \mathbf{w}(k+1) &= \mathbf{w}(k) + \eta(k) e(k)\mathbf{x}(k) \\ \eta(k) &= \frac{\mu}{\| \mathbf{x}(k) \|_2^2 + \varepsilon(k)} \\ \varepsilon(k) &= \varepsilon(k-1) - \rho\mu \frac{e(k)e(k-1)\mathbf{x}^T(k)\mathbf{x}(k-1)}{\left(\| \mathbf{x}(k-1) \|_2^2 + \varepsilon(k-1)\right)^2}. \end{aligned} \tag{6}$$

Notice that there is a fundamental difference between the variable step size algorithms with a "linear" multiplicative adaptation factor (Mathews' and Benveniste's; see the Appendix) and GNGD, which employs a nonlinear update of the adaptive learing rate $\eta(k)$. The merit of the proposed algorithm is that its learning rate provides compensation for the assumptions in the derivation of NLMS, and therefore, due to its robustness and improved stability, GNGD is well suited for processing of nonlinear and nonstationary signals.

A. Stability, Robustness, and Computational Complexity of GNGD Algorithm

The classical analysis of the GNGD in terms of convergence in the mean, mean square and steady state follows the well-known analysis from the literature [3], [10]. The adaptive step size η of GNGD is essentially bounded by the stability limits of the step size of the NLMS algorithm. To find the lower bound on the compensation term ε, consider the uniform convergence condition

$$|e(k+1)| \leq |1 - \eta(k) \| \mathbf{x}(k) \|_2^2 \| |e(k)|. \tag{7}$$

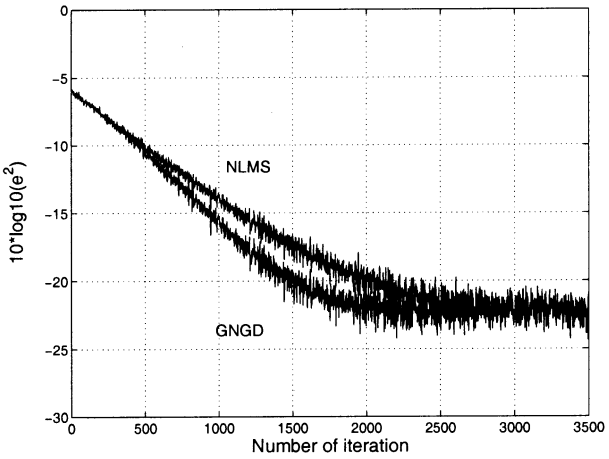

Fig. 1. Performance comparison between NLMS and GNGD on a colored signal (10) for $\mu = 0.001$.

Hence

$$|1 - \eta(k) \| \mathbf{x}(k) \|_2^2 | < 1 \tag{8}$$

which gives $0 < \mu/(\| \mathbf{x}(k) \|_2^2 + \varepsilon(k)) < 2/\| \mathbf{x}(k) \|_2^2$. For $\mu = 1$, the lower bound for stability of GNGD with respect to $\varepsilon(k)$ is

$$\varepsilon(k) > -\frac{\| \mathbf{x}(k) \|_2^2}{2}. \tag{9}$$

Computational complexity of GNGD lies in between the complexity of Mathews' and Benveniste's algorithms and is roughly twice that of NLMS. To reduce computational complexity of GNGD, and prevent distrubance in the steady state, it is possible to impose hard bounds on $\varepsilon(k)$, or to stop its adaptation after convergence. In the experiments, however, for generality, no such constraints were imposed. Due to the nonlinear nature of learning rate adaptation, GNGD is responsive and robust to the initialization of its parameters.

III. Experiments

For the experiments, the order of the FIR adaptive filter was $N = 10$, and 100 runs of independent trials were performed and averaged in the prediction setting. The performance of GNGD was first compared to that of NLMS and then to performances of other variable step size algorithms. For generality, linear, nonlinear and nonstationary signals were used in simulations. The linear signal was white noise $\{x(k)\}$ with zero mean and unit variance, passed through an AR filter (colored input) given by

$$\begin{aligned} y(k) = 1.79y(k-1) - 1.85y(k-2) \\ +1.27y(k-3) - 0.41y(k-4) + x(k). \end{aligned} \tag{10}$$

Speech was used as a nonstationary signal, whereas a nonlinear signal was [11]

$$y(k+1) = \frac{y(k)}{1 + y^2(k)} + x^3(k). \tag{11}$$

For GNGD, the initial values were $\rho = 0.15$ whereas μ was varied according to the aim of the experiment. The initial value $\varepsilon(0)$ was set to zero for most of the experiments. Fig. 1 illustrates the GNGD exhibiting faster convergence and similar

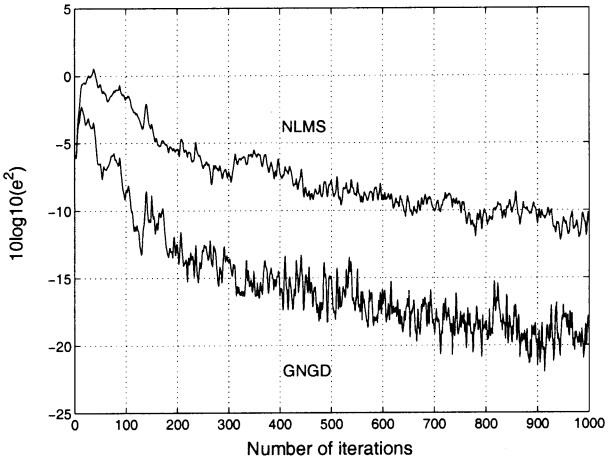

Fig. 2. Performance comparison between NLMS and GNGD for a nonlinear signal (11) for $\mu = 1.99$.

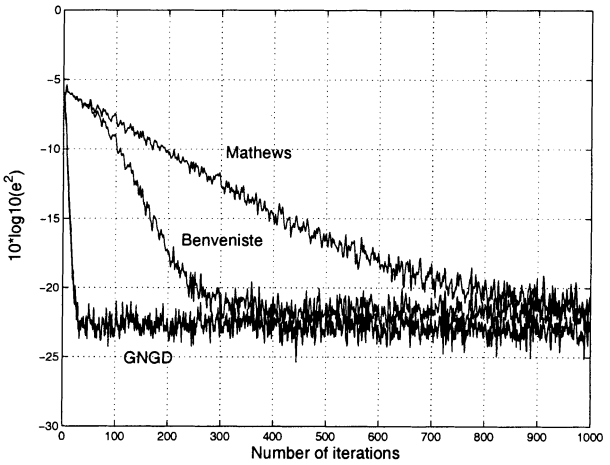

Fig. 3. Performance comparison of GNGD, Mathews', and Benveniste's algorithms for signal (10).

steady state performance to that of NLMS for a colored signal (10), for a relatively small μ. A performance comparison of GNGD and NLMS on prediction of nonlinear signal (11) is shown in Fig. 2. Learning rate $\mu = 1.99$ was chosen to be close to the stability bound of NLMS, and GNGD comprehensively outperformed NLMS. Due to its nature, the GNGD performance was similar or better than that of NLMS when NLMS was stable, whereas GNGD was convergent in cases when NLMS was not $(\mu > 2)$. Fig. 3 provides performance comparison between the GNGD algorithm and Mathews' and Benveniste's algorithms for signal (10). In this case, GNGD outperformed the other two variable step size algorithms. In a general case, however, depending on the character of a signal, GNGD exhibited better, similar, or slightly worse performance than the other two variable step size algorithms. However, its advantage over the other two considered algorithms was excellent stability and robustness over a whole spectrum of signals.

Fig. 4 illustrates sensitivity of the prediction gain $R_p = 10\log_{10}(\text{var}(y)/\text{var}(e))$ for a variation of μ and ρ for nonlinear signal (11). The GNGD was clearly able to provide stabilization of its NLMS type adaptation, for a range

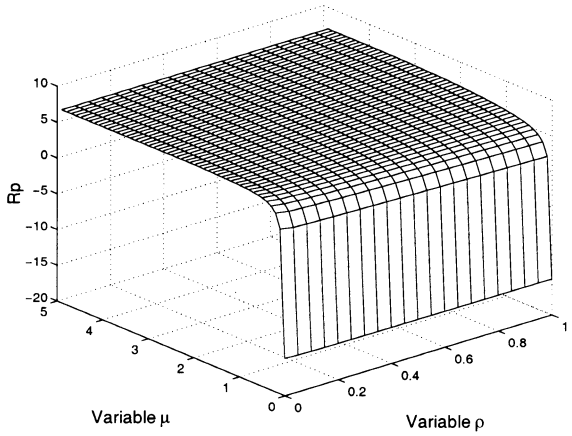

Fig. 4. Robustness of Rp to the variation of μ and ρ for nonlinear signal (11).

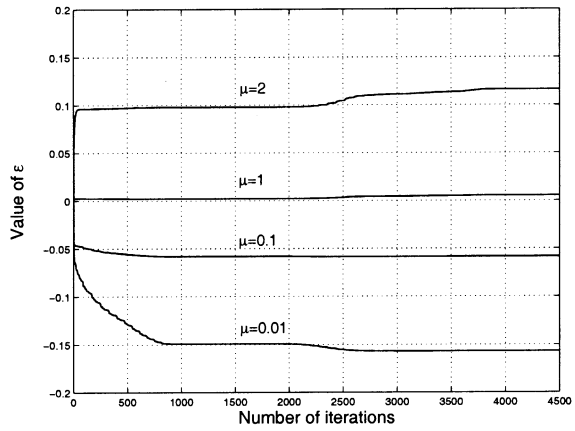

Fig. 5. Time variation of $\varepsilon(k)$ for prediction of a speech signal for different values of the learning rate μ.

of values of parameter μ, exhibiting very low sensitivity to the initialization of its parameters ρ and $\varepsilon(0)$. This is not the case with Mathews' and Benveniste's algorithms, which perform well if they are initialized with a very small learning rate and are very sensitive to the choice of parameter ρ.

Fig. 5 shows the variation of $\varepsilon(k)$ for prediction of a nonstationary speech signal. As μ increases toward suboptimal and unstable performance of NLMS, GNGD increases $\varepsilon(k)$, which in turn decreases the step size $\eta(k)$ and the algorithm is stabilized. On the other hand, when μ is very small (bottom of the diagram) $\varepsilon(k)$ goes toward negative values, increasing $\eta(k)$ and speeding up convergence of the algorithm.

IV. Conclusion

A generalized normalized gradient descent algorithm for linear adaptive filters has been proposed. It has been derived as an extension of the normalized least square algorithm where the learning rate comprises an additional adaptive factor, which stabilizes NLMS and makes GNGD suitable for filtering of nonlinear and nonstationary signals. Unlike the previously proposed gradient adaptive step size algorithms, GNGD has been shown to be robust to the initialization of its parameters. Simulations on stationary, nonstationary and nonlinear signals justify the proposed approach.

APPENDIX

The algorithm proposed in [3] is based upon a gradient adaptation of the learning rate μ of LMS from (2) by steepest descent, based upon $\partial E(k)/\partial\mu$. The step size update in Mathews' algorithm, which utilizes the independence assumptions, is given by

$$\begin{aligned}\mu(k) &= \mu(k-1) - \frac{\rho}{2}\frac{\partial}{\partial\mu(k-1)}e^2(k) \\ &= \mu(k-1) + \rho e(k)e(k-1)\mathbf{x}^T(k)\mathbf{x}(k-1) \quad (12)\end{aligned}$$

and is shown to be very sensitive to the choice of initial values of ρ and $\mu(0)$ [3], [5]. Given by

$$\begin{aligned}\mu(k) &= \mu(k-1) + \rho e(k)\mathbf{x}^T(k)\boldsymbol{\psi}(k) \\ \boldsymbol{\psi}(k) &= \left[\mathbf{I} - \mu(k-1)\mathbf{x}(k-1)\mathbf{x}^T(k-1)\right] \\ &\quad \cdot \boldsymbol{\psi}(k-1) + e(k-1)\mathbf{x}(k-1) \quad (13)\end{aligned}$$

Benveniste's algorithm is based upon the exact derivation of the adaptive learning rate and is computationaly demanding, since it requires matrix multiplications.

REFERENCES

[1] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1985.

[2] A. M. Kuzminskiy, "Self-adjustment of an adaptation coefficient of a noise compensator in a nonstationary process," *Izvestiya VUZ. Radioelektron.*, vol. 29, no. 3, pp. 103–105, 1986.

[3] V. J. Mathews and Z. Xie, "A stochastic gradient adaptive filter with gradient adaptive step size," *IEEE Trans. Signal Processing*, vol. 41, pp. 2075–2087, June 1993.

[4] A. Benveniste, M. Metivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximation*. New York: Springer-Verlag, 1990.

[5] W.-P. Ang and B. Farhang-Boroujeny, "A new class of gradient adaptive step-size LMS algorithms," *IEEE Trans. Signal Processing*, vol. 49, pp. 805–810, Apr. 2001.

[6] T. Aboulnasr and K. Mayyas, "A robust variable step-size LMS-type algorithm: Analysis and simulations," *IEEE Trans. Signal Processing*, vol. 45, pp. 631–639, Mar. 1997.

[7] J. B. Evans, P. Xue, and B. Liu, "Analysis and implementation of variable step size adaptive algorithms," *IEEE Trans. Signal Processing*, vol. 41, pp. 2517–2535, Aug. 1993.

[8] D. R. Morgan and S. G. Kratzer, "On a class of computationally efficient, rapidly converging, generalized NLMS algorithms," *IEEE Signal Processing Lett.*, vol. 3, pp. 245–247, Aug. 1996.

[9] M. Milisavljevic, "Multiple environment optimal update profiling for steepest descent algorithms," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, vol. VI, 2001, pp. 3853–3856.

[10] D. P. Mandic and J. A. Chambers, *Recurrent Neural Networks for Prediction: Architectures, Learning Algorithms and Stability*. Chichester, U.K.: Wiley, 2001.

[11] K. S. Narendra and K. Parthasarathy, "Identification and control of dynamical systems using neural networks," *IEEE Trans. Neural Networks*, vol. 1, pp. 4–27, Jan. 1990.