A CLASS OF GRADIENT-ADAPTIVE STEP SIZE ALGORITHMS FOR COMPLEX-VALUED NONLINEAR NEURAL ADAPTIVE FILTERS

Su Lee Goh and Danilo P. Mandic

Imperial College London
Exhibition Road, London, SW7 2BT, UK
E-mail: {su.goh,d.mandic}@imperial.ac.uk

ABSTRACT

A class of variable step-size algorithms for complex-valued nonlinear neural adaptive finite impulse response (FIR) filters realised as a dynamical perceptron is proposed. The adaptive step-size is updated using gradient descent to give variable step-size complex-valued nonlinear gradient descent (VSCNGD) algorithms. These algorithms are shown to be capable of tracking signals with rich and unknown dynamics, and exhibit faster convergence and smaller steady state error than the standard algorithms. Further, the analysis of stability and computational complexity is provided. Simulations in the prediction setting support the approach.

1. INTRODUCTION

Real-valued adaptive filters have been used for processing signals in various disiplines such as acoustics, communications and seismology. The least mean square (LMS) algorithm [1] is one of the most common approaches to train linear adaptive filters. Despite its roboustness, this algorithm is relatively slow at converging to the optimal least squares solution. A number of variable step-size least mean square (VSLMS) algorithms have been developed to speed up convergence of linear adaptive filters [2, 3, 4]. Generally, the idea behind variable step-size is to have large step-sizes when the estimated errors are large at the early stages of adaptation, and smaller step-sizes when approaching steady-state convergence. This helps to avoid the trade-off between fast convergence and misadjustment, commonly experienced with the fixed step-size LMS.

To that end, Benveniste *et al.* [2] propose and analyse an adaptive step-size algorithm based on the gradient of the instantaneous squared error with respect to the step-size. Benveniste's algorithm, in fact, performs time-varying low pass filtering of the noisy instantaneous gradients in the update of the step-size. This algorithm was derived rigorously without making the usual independence assumptions, which results in better performance but increased computational complexity as compared to standard LMS. Attempts to reduce the computational complexity of this algorithm include the Mathews and Xie [3] and Ang and Farhang-Boroujeny [4] algorithm. In [4], a fixed parameter low pass filter replaces time-varying filtering of the instantaneous gradients in the step-size update from [2], whereas in Mathews and Xie's algorithm, only raw instantaneous gradients are used, which makes this algorithm sensitive to initial conditions and noise. One advantage of algorithms from [3] and [4] over Benveniste's algorithm is their relative simplicity, at the cost of possible performance degradation.

Recently, there has been much research directed towards development and analysis of complex-valued adaptive filters, especially nonlinear ones. The applications of these filters are emerging and the theory of complex-valued nonlinear adaptive filters is following this development [5]. Our aim is to extend the class of gradient adaptive step-size algorithms to the case of complex-valued nonlinear adaptive filters. The derivation of the proposed class of variable step-size complex-valued nonlinear gradient descent (VSCNGD) algorithms follows the approaches from [2, 3, 4]. Notice, however, that extensions to the nonlinear case in the complex domain are non-trivial. For generality, we focus on filters with a 'fully' complex nonlinear activation function (AF)[1] of a neuron, where the nonlinearity within the complex AF must be analytic and bounded almost everywhere in the complex domain, $\mathbb{C}$ [6]. This way, the Cauchy-Riemann[2] equations are satisfied which makes it possible to use gradient descent. The analysis is supported by simulations on colored, nonlinear and real-world signals.

The work of D. Mandic is partially supported by EPSRC grant GR/S71149/01.

[1] In a previously frequently used split-complex AF, the real and imaginary components of the input signal x are separated and fed through the real-valued AF $f_R(x) = f_I(x)$, $x \in \mathbb{R}$. A split-complex AF is therefore given as $f(x) = f_R(Re(x)) + jf_I(Im(x))$, hence these functions are not analytic.

[2] Cauchy-Riemann equations state that the partial derivatives of a function $f(z) = u(x,y) + jv(x,y)$ along the real and imaginary axes should be equal: $f'(z) = \frac{\partial u}{\partial x} + j\frac{\partial v}{\partial x} = \frac{\partial v}{\partial y} - j\frac{\partial u}{\partial y}$. This way $\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}$, $\frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y}$.

2. CLASS OF VARIABLE STEP-SIZE CNGD ALGORITHMS

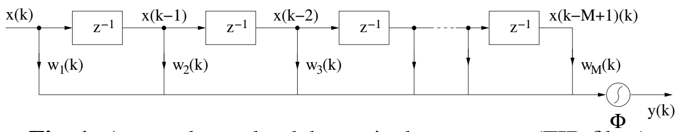


Fig. 1. A complex-valued dynamical perceptron (FIR filter)

The weight update of the standard CNGD algorithm shown for nonlinear FIR adaptive filter (Fig. 1) is expressed as [7]:

$$\mathbf{w}(k+1) = \mathbf{w}(k) + \eta e(k)\left[\Phi'(\mathbf{x}^T(k)\mathbf{w}(k))\right]^* \mathbf{x}^*(k) \quad (1)$$

where $e(k) = d(k) - \Phi(\mathbf{x}^T(k)\mathbf{w}(k))$ denotes the instantaneous error at the output of the filter at the time instant k, $d(k)$ is the desired signal, $\mathbf{x}(k) = [x(k-1), \ldots, x(k-M)]^T$ is the input signal, M is the length of the filter, $(\cdot)^T$ is the vector transpose operator, $(\cdot)^*$ is the complex conjugate operator, and $\mathbf{w}(k) = [w_1(k), \ldots, w_M(k)]^T$ is the filter coefficient vector. The parameter η is the step-size and is critical to the convergence of the algorithms, whereas Φ denotes the complex activation function.

2.1. Variable Step-Size CNGD (VSCNGD1) Algorithm

To cater for the unknown dynamics of the inputs and their possible nonstationary nature, we propose to make the step-size η in (1) gradient adaptive, as

$$\eta(k) = \eta(k-1) - \rho \nabla_{\eta} E(k)_{|\eta=\eta(k-1)} \quad (2)$$

where $E(k) = \frac{1}{2}e(k)e^*(k) = \frac{1}{2}|e(k)|^2$ is the cost function. The gradient $\nabla_{\eta} E(k)$ can be evaluated as

$$\nabla_{\eta} E(k) = \frac{1}{2}\left[e(k)\frac{\partial e^*(k)}{\partial \eta(k-1)} + e^*(k)\frac{\partial e(k)}{\partial \eta(k-1)}\right] \quad (3)$$

To calculate the two partial derivatives from (3), it is necessary to use the Cauchy-Riemann equations to obtain[3]

$$\frac{\partial e^*(k)}{\partial \eta(k-1)} = -\mathbf{x}^H(k)\left\{\Phi'(\mathbf{x}^T(k)\mathbf{w}(k))\right\}^* \frac{\partial \mathbf{w}^*(k)}{\partial \eta(k-1)} \quad (4)$$

$$\frac{\partial e(k)}{\partial \eta(k-1)} = -\mathbf{x}^T(k)\Phi'(\mathbf{x}^T(k)\mathbf{w}(k))\frac{\partial \mathbf{w}(k)}{\partial \eta(k-1)} \quad (5)$$

For simplicity, we denote $\Phi(\mathbf{x}^T(k)\mathbf{w}(k)) = \Phi(k)$ and $\frac{\partial \mathbf{w}(k)}{\partial \eta(k-1)} = \boldsymbol{\psi}(k)$. Thus, from (1) we have

$$\begin{aligned}
\boldsymbol{\psi}(k) &= \boldsymbol{\psi}(k-1) \\
&+ \frac{\partial \eta(k-1)}{\partial \eta(k-1)} e(k-1)\left\{\Phi'(k-1)\right\}^* \mathbf{x}^*(k-1) \\
&+ \eta(k-1)\frac{\partial e(k-1)}{\partial \eta(k-1)}\left\{\Phi'(k-1)\right\}^* \mathbf{x}^*(k-1) \\
&+ \eta(k-1)e(k-1)\frac{\partial\left\{\Phi'(k-1)\right\}^*}{\partial \eta(k-1)}\mathbf{x}^*(k-1) \\
&+ \eta(k-1)e(k-1)\left\{\Phi'(k-1)\right\}^* \frac{\partial \mathbf{x}^*(k-1)}{\partial \eta(k-1)} \quad (6)
\end{aligned}$$

[3]Full derivation will be omitted due to space limitation.

From (6), we now arrive at $\boldsymbol{\psi}(k)$, which gives VSCNGD1

$$\begin{aligned}
&\boldsymbol{\psi}(k) = \boldsymbol{\psi}(k-1)\,[\mathbf{I} \\
&- \eta(k-1)\left|\Phi'(k-1)\right|^2 \mathbf{x}^*(k-1)\mathbf{x}^T(k-1) \\
&+ \eta(k-1)e(k-1)\left\{\Phi''(k-1)\right\}^* \mathbf{x}^*(k-1)\mathbf{x}^H(k-1)\big] \\
&+ e(k-1)\left\{\Phi'(k-1)\right\}^* \mathbf{x}^*(k-1) \quad (7)
\end{aligned}$$

Finally we obtain the gradient update as $\nabla_{\eta} E(k) = -\frac{1}{2}\big[e(k) \left\{\Phi'(k)\right\}^* \mathbf{x}^H(k)\boldsymbol{\psi}^*(k) + e^*(k)\Phi'(k)\mathbf{x}^T(k)\boldsymbol{\psi}(k)\big]$. Notice that Ψ consists of the instantaneous gradient term $e(k-1)\left\{\Phi'(k-1)\right\}^* \mathbf{x}^*(k-1)$ and a filtered version of Ψ (the term in the square brackets).

2.2. Variable Step-Size CNGD (VSCNGD2) Algorithm

The adaptive step-size in (7) is rigorously derived by extending the approach from Benveniste to the nonlinear complex case. Next, for simplicity, following [4], we replace the square bracket term in equation (7) by a constant $0 < \alpha < 1$. This leads to the VSCNGD2 algorithm given by

$$\boldsymbol{\psi}(k) = \alpha\boldsymbol{\psi}(k-1) + e(k-1)\left\{\Phi'(k-1)\right\}^* \mathbf{x}^*(k-1) \quad (8)$$

For $\alpha < 1$, equation (7) represents a low pass filter that takes the weighted average of the present and past observations of the instantaneous gradients $e(k-1)\left\{\Phi'(k-1)\right\}^* \mathbf{x}^*(k-1)$.

2.3. Variable Step-Size CNGD (VSCNGD3) Algorithm

When $\alpha = 0$, (which analogous to [3]), the VSCNGD3 algorithm, which is the simplest of the three proposed algorithms is obtained. Table 1 summarizes the three proposed algorithms, selection of initial parameters, and the number of complex multiplications involved.

3. STABILITY ANALYSIS

We next employ the contraction mapping principle to illustrate the mean square convergence of the proposed algorithms and set bounds on the values of the step-size. By the Contraction Mapping Theorem (CMT), function $F : \mathbf{Z} \rightarrow \mathbf{Z}$ is a contraction if [8]

$$|F(x) - F(y)| \leq \gamma\,|x - y| \quad \forall x, y \in \mathbf{Z} \quad (9)$$

where $0 \leq \gamma < 1$. According to Banach's fixed point theorem, contractive functions have at most one fixed point, where for every $x \in \mathbf{Z}$, $\left|F^n(x) - F^{n+1}(x)\right| \rightarrow 0$ as $n \rightarrow \infty$. To make use of CMT, let $\mathbf{v}(k) = \mathbf{w}(k) - \mathbf{w}_{opt}$, and subtract (1) $\mathbf{w}_{opt}$ from both side of (1), to give

$$\begin{aligned}
\mathbf{v}(k+1) &= \mathbf{v}(k) + \eta(k)e(k)\left[\Phi'(k)\right]^* \mathbf{x}^*(k) \\
&= \mathbf{v}(k) - \eta(k)\left[\Phi'(k)\right]^* \mathbf{x}^*(k)\big[\Phi(\mathbf{x}^T(k)\mathbf{w}(k)) \\
&- \Phi(\mathbf{x}^T(k)\mathbf{w}_{opt})\big] + \eta(k)e_{opt}\left[\Phi'(k)\right]^* \mathbf{x}^*(k) \quad (10)
\end{aligned}$$

Table 1. Classification of VSCNGD algorithms and complexity

Algorithms	Variable step-size update	Parameters	No. of. mul.
VSCNGD1	$\boldsymbol{\psi}(k) = \boldsymbol{\psi}(k-1)\left[\mathbf{I} - \eta(k-1)\left\|\Phi'(k-1)\right\|^2 \mathbf{x}^*(k-1)\mathbf{x}^T(k-1)\right.$ $+\eta(k-1)e(k-1)\left\{\Phi''(k-1)\right\}^* \mathbf{x}^*(k-1)\mathbf{x}^H(k-1)\big]$ $+e(k-1)\left\{\Phi'(k-1)\right\}^* \mathbf{x}^*(k-1)$	$\eta(0)$=0.05, ρ=0.0002	16M
VSCNGD2	$\boldsymbol{\psi}(k) = \alpha\boldsymbol{\psi}(k-1) + e(k-1)\left\{\Phi'(k-1)\right\}^* \mathbf{x}^*(k-1)$	α=0.95 $\eta(0)$=0.05, ρ=0.0002	9M
VSCNGD3	$\boldsymbol{\psi}(k) = e(k-1)\left\{\Phi'(k-1)\right\}^* x^*(k-1)$	$\eta(0)$=0.05, ρ=0.0002	8M

The activation function of a neuron used here is a complex hyperbolic tangent, $\Phi(x) = \frac{e^{\beta x} - e^{-\beta x}}{e^{\beta x} + e^{-\beta x}}$. Using the condition from (9), $\forall x, y \in [a,b]$, $\exists \xi \in (a,b)$, such that [9]

$$|tanh(x) - tanh(y)| \leq |tanh'(\xi)|\,|x-y| \qquad (11)$$

Combining (11) and (10), we have

$$\mathbf{v}(k+1) \leq \left[\mathbf{I} - \eta(k)\left[\Phi'(k)\right]^* \Phi'(\xi)\mathbf{x}^*(k)\mathbf{x}^T(k)\right]\mathbf{v}(k) \quad (12)$$

where the inhomogeneous part of (10) $\eta(k)e_{\,opt}\left[\Phi'(k)\right]^* \mathbf{x}^*(k)$ can be ignored in the convergence analysis. By taking the expectation of squared l_2-norm in (12), we obtain

$$E\left\|\mathbf{v}(k+1)\right\|_2^2 \leq \Lambda[E[B^2]]E\left\|\mathbf{v}(k)\right\|_2^2 \qquad (13)$$

where $\Lambda[E[B^2]]$ is the maximum eigenvalue of $E[B^2] = E\left[\mathbf{I} - \eta(k)\left[\Phi'(k)\right]^* \Phi'(\xi)\mathbf{x}^*(k)\mathbf{x}^T(k)\right]^2$. Combining the condition from (13) with CMT principle (9), we have $\Lambda[E[B^2]] \leq \gamma^2 \leq 1$. Thus, the upper bound for the step-size $\eta(k)$ is given as

$$0 < \eta(k) < \frac{2}{\left[2\lambda + tr(\mathbf{R}^*_{\mathbf{xx}})\right]\left|\left[\Phi'(k)\right]^* \Phi'(\xi)\right|} \qquad (14)$$

where λ is the maximum eigenvalue of the autocorrelation matrix $\mathbf{R}^*_{\mathbf{xx}}$.

4. SIMULATIONS

In all the experiments, the order of the nonlinear adaptive filter was chosen to be $M = 4$, with $\beta = 1$. Simulations were undertaken by averaging 200 iterations of independent trials on prediction of complex-valued benchmark colored and nonlinear signals as well as on real-life signals. The colored signal was a complex linear $AR(4)$ process given by $r(k) = 1.79r(k-1) - 1.85r(k-2) + 1.27r(k-3) + 0.41r(k-4) + n(k)$ with complex white Gaussian noise (CWGN) $n(k) \sim \mathcal{N}(0,1)$ as the driving input. The CWGN can be expressed as $n(k) = n^r(k) + jn^i(k)$. The real and imaginary components of CWGN were mutually statistically independent sequences having equal variances so that $\sigma_n^2 = \sigma_{n^r}^2 + \sigma_{n^i}^2$. The nonlinear input signals were NL1 and NL2 [10], given respectively in (15) and (16)

$$z(k) = \frac{z(k-1)z(k-2)\left[z(k-1)+2.5\right]}{1 + z^2(k-1) + z^2(k-2)} + r(k-1) \quad (15)$$

$$\begin{aligned} z(k) &= \frac{5z(k-1)z(k-2)}{1 + z^2(k-1) + z^2(k-2) + z^2(k-3)} \\ &+ \; r(k-1) + 0.8r(k-2) \end{aligned} \qquad (16)$$

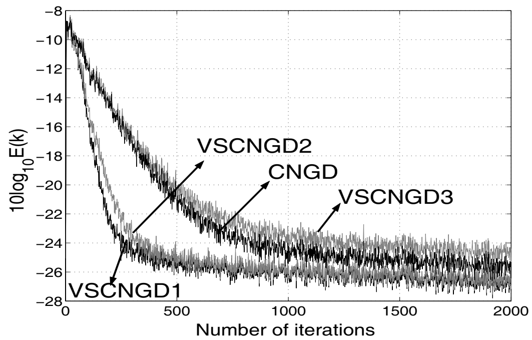
To further verify the approach, we also tested the proposed algorithms on real-world signals, including the radar and wind signals. The measurement used to assess the performance was the prediction gain $R_p = 10log_{10}\left(\frac{\sigma_x^2}{\hat{\sigma}_e^2}\right)[dB]$ [1], where σ_x^2 denotes the variance of the input signal $x(k)$, and $\hat{\sigma}_e^2$ denotes the estimated variance of the forward prediction error $e(k)$. Figures 2a and 3a show the performance of CNGD, VSCNGD1, VSCNGD2 and VSCNGD3 algorithms on colored ($AR(4)$) and nonlinear (15) inputs. Observe that VSCNGD1 has the fastest convergence compared to the rest of the algorithms for both inputs. In a general case, depending on the signal, VSCNGD3 exhibited similar or slightly worst performance than the CNGD algorithm. Figures 2b and 3b illustrate a time variation of the step-size $\eta(k)$. Table 2 shows the comparison of prediction gains between the CNGD and the proposed algorithms for both benchmark and real-life radar and wind complex signals. In all the cases, VSCNGD1 gave best performance, followed by VSCNGD2 and VSCNGD3.

Table 2. Prediction gain R_p for CNGD and proposed algorithms on various signals

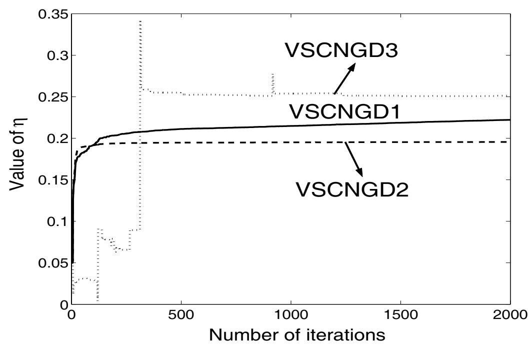
R_p [dB]	AR(4)	NL2	Wind	Radar
CNGD	5.010	1.710	14.121	10.888
VSCNGD1	6.606	4.774	16.7957	14.060
VSCNGD2	6.271	4.493	16.093	15.384
VSCNGD3	4.099	1.877	14.297	13.162

5. CONCLUSIONS

The step-size in the CNGD algorithm for simple complex-valued nonlinear neural adaptive filters has been made adaptive using a gradient descent based approach to give a class of variable step-size complex-valued nonlinear gradient descent (VSCNGD) algorithms. These algorithms have been developed for a general complex nonlinear activation function of a neuron. The convergence analysis has been performed and the proposed algorithms have been shown to

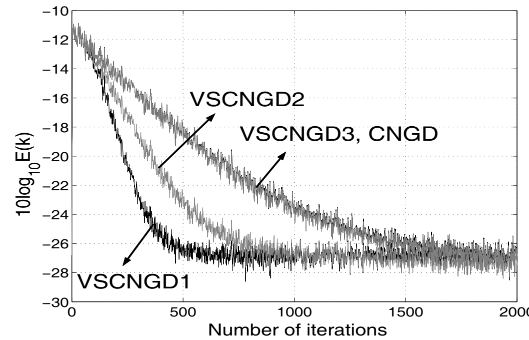


(a) Performance comparison between algorithms

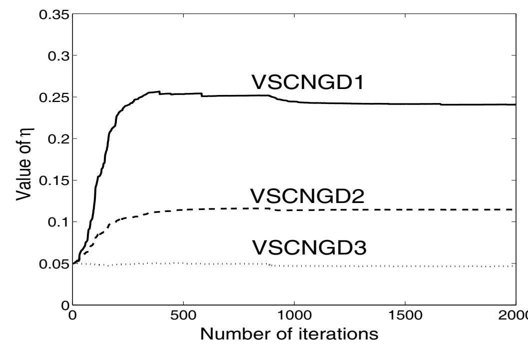


(b) Time variation of $\eta(k)$

Fig. 2. Performance of CNGD, VSCNGD1, VSCNGD2 and VSCNGD3 on prediction of colored ($AR(4)$) input



(a) Performance comparison between algorithms



(b) Time variation of $\eta(k)$

Fig. 3. Performance of CNGD, VSCNGD1, VSCNGD2 and VSCNGD3 on prediction of NL1 input

converge faster than the standard CNGD algorithm. Experiments on both (linear and nonlinear) benchmark and real-life complex signals support the analysis.

6. REFERENCES

[1] S. Haykin and Liang Li, "Nonlinear adaptive prediction of nonstationary signals," *IEEE Transactions on Signal Processing*, vol. 43, no. 2, pp. 526–535, 1995.

[2] A. Benveniste, M. Metivier, and P. Priouret, *Adaptive algorithms and stochastic approximation*, New York: Springer-Verlag, 1990.

[3] V. J Mathews and Z. Xie, "A stochastic gradient adaptive filter with gradient adaptive step size," *IEEE Transactions on Signal Processing*, vol. 41, no. 6, pp. 2075 – 2087, 1993.

[4] W. P. Ang and B. Farhang-Boroujeny, "A new class of gradient adaptive step-size LMS algorithms," *IEEE Transactions on Signal Processing*, vol. 49, no. 4, pp. 805 – 810, 2001.

[5] A. Hirose, *Complex-valued Neural Networks Theories and Applications, Series on Innovative Intelligence - Vol. 5*, World Scientific, 2003.

[6] T. Kim and T. Adali, "Approximation by fully complex MLP using elementary transcendental activation functions," *In Proceedings of XI IEEE Workshop on Neural Networks for Signal Processing*, pp. 203–212, 2001.

[7] A. I. Hanna and D. P. Mandic, "A complex-valued nonlinear neural adaptive filter with a gradient adaptive amplitude of the activation function," *Neural Networks*, vol. 16, no. 2, pp. 155–159, 2003.

[8] L. Gorniewicz, S. A. Marano, and M. Slosarski, "Fixed points of contractive multivalued maps," *Proceedings of the American Mathematical Society*, vol. 124, no. 9, pp. 2675–2683, 1996.

[9] I. R. Krcmar, M. M. Bozic, and D. P. Mandic, "Global asymptotic stability for RNNs with a bipolar activation function," *Neural Network Applications in Electrical Engineering, NEUREL 2000*, pp. 33 – 36, 2000.

[10] K. S. Narendra and K. Parthasarathy, "Identification and control of dynamical systems using neural networks," *IEEE Transactions on Neural Networks*, vol. 1, no. 1, pp. 4–27, 1990.