
PERGAMON

AVAILABLE AT
www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

Neural Networks 16 (2003) 873–880

Neural Networks

www.elsevier.com/locate/neunet

2003 Special issue

Stochastic error whitening algorithm for linear filter estimation with noisy data

Yadunandana N. Rao*, Deniz Erdogmus, Geetha Y. Rao, Jose C. Principe

Computational NeuroEngineering Laboratory, Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611, USA

Abstract

Mean squared error (MSE) has been the most widely used tool to solve the linear filter estimation or system identification problem. However, MSE gives biased results when the input signals are noisy. This paper presents a novel stochastic gradient algorithm based on the recently proposed error whitening criterion (EWC) to tackle the problem of linear filter estimation in the presence of additive white disturbances. We will briefly motivate the theory behind the new criterion and derive an online stochastic gradient algorithm. Convergence proof of the stochastic gradient algorithm is derived making mild assumptions. Further, we will propose some extensions to the stochastic gradient algorithm to ensure faster, step-size independent convergence. We will perform extensive simulations and compare the results with MSE as well as total-least squares in a parameter estimation problem. The stochastic EWC algorithm has many potential applications. We will use this in designing robust inverse controllers with noisy data.

Keywords: Error whitening criterion; Total-least squares; LMS; Mean squared error; Error whitening

1. Introduction

It was been a widely acknowledged fact that the mean squared error (MSE) criterion is optimal for linear filter estimation when there are no noisy perturbations on the data (Haykin, 1996). In adaptive filter theory, the Wiener solution for the MSE criterion is used to derive recursive algorithms like RLS and the more popular stochastic gradient based LMS algorithm (Haykin, 1996). An important property of the Wiener solution is that if the adaptive filter is sufficiently long enough, then the prediction error signal for stationary data is white (Haykin, 1996). This very nice property is true only when the input data is noise-free. It has been long recognized that the MSE-based filter optimization approaches are unable to produce the optimal weights associated with the noise free data due to the biasing of the input covariance matrix by the additive noise (Douglas, 1996). For many real-world applications, the 'noise-free' assumption is easily violated and using MSE-based methods for parameter estimation will result in severe degradation in performance. Researchers have proposed several techniques to combat and suppress the bias in MSE-based methods. For instance, the subspace methods coupled with the Wiener solution can result in superior filter estimates. However, finding the right subspace dimension and the optimal subspace projections is a non-trivial problem. Moreover, subspace based Wiener filtering methods can only reduce the bias; they can never remove the bias completely. An important statistical tool called total least squares (TLS) (Golub & Van Loan, 1989) can be utilized to eliminate this bias completely. The major stumbling block for the TLS that severely limits its practicability is the requirement that the variances of the noisy perturbations on the input and desired signals be identical (Douglas, 1996; Rao & Principe, 2002). Recently, we proposed a novel criterion called the error whitening criterion (EWC) (Principe, Rao, & Erdogmus, 2003; Rao, Erdogmus, & Principe, 2003) that presents a different approach to whiten the error sequence at the output of an adaptive filter even in the presence of noisy inputs. This criterion enforces zero autocorrelation of the error signal beyond a certain lag; hence the name EWC.

In Section 2, we will motivate the theory of EWC; briefly state some of its interesting properties and then present an online stochastic gradient algorithm called EWC-LMS. In Section 3, we will discuss the convergence of this algorithm to the optimal EWC solution. Section 4

* Corresponding author.

0893-6080/03/$ - see front matter © 2003 Published by Elsevier Science Ltd.
doi:10.1016/S0893-6080(03)00109-6

talks about extensions to the EWC-LMS algorithm for increased robustness and faster convergence. In Section 5, we will present case studies and compare the performance of the stochastic EWC algorithm with TLS and the regular LMS algorithms. In the same section, will also demonstrate the usefulness of this algorithm in an inverse controller design application. Discussions and conclusions are in Section 6.

2. Error whitening criterion

The classical Wiener solution tries to minimize the zero-lag autocorrelation of the error, i.e. $E(e_k^2)$. In the presence of additive white noise, the zero-lag input autocorrelation is always biased by the noise power. Therefore, when we minimize the MSE, we always end up with a biased parameter estimate. This bias increases with increasing noise power. So, instead of working with zero-lag correlations, we propose to analyze the error autocorrelation at a non-zero lag. Consider the problem of identifying a linear system characterized by the parameter vector $\mathbf{w}_T \in \Re^N$ as shown in Fig. 1. Suppose noisy training data pair $(\hat{\mathbf{x}}_k, \hat{d}_k)$ is provided, where $\hat{\mathbf{x}}_k \in \Re^N = \mathbf{x}_k + \mathbf{v}_k$ and $\hat{d}_k \in \Re^1 = d_k + u_k$ with $\mathbf{x}_k$ as the noise-free input vector at discrete time index k; $\mathbf{v}_k$, the additive white noise vector entering the input; d_k being the noise-free desired signal and u_k as the additive white noise entering the desired signal. We further assume that the noises entering the system $\mathbf{v}_k$ and u_k are uncorrelated with the data pair and also uncorrelated with each other. Let the weight vector (filter) that generated the noise-free data pair $(\mathbf{x}_k, d_k)$ be $\mathbf{w}_T$, of dimension N. Without loss of generality, we will assume that the length of $\mathbf{w}$, the estimated weight vector is greater than N. Since $d_k = \mathbf{x}_k^T \mathbf{w}_T$, the error is simply given by $\hat{e}_k = \mathbf{x}_k^T(\mathbf{w}_T - \mathbf{w}) + u_k - \mathbf{v}_k^T \mathbf{w}$. The corresponding error autocorrelation at some arbitrary lag L can be determined as

$$\rho_{\hat{e}}(L) = [\mathbf{w}_T - \mathbf{w}]^T E[\mathbf{x}_k \mathbf{x}_{k-L}^T][\mathbf{w}_T - \mathbf{w}] + \mathbf{w}^T E[\mathbf{v}_k \mathbf{v}_{k-L}^T]\mathbf{w} \tag{1}$$

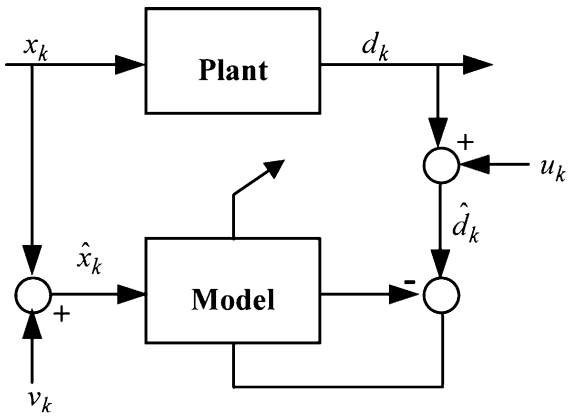

Fig. 1. System identification block diagram.

From Eq. (1), it is obvious that if $L \geq M$, where M is the length of the true filter $\mathbf{w}$, $E[\mathbf{v}_k \mathbf{v}_{k-L}^T] = \mathbf{0}$. Assuming that the matrix $E[\mathbf{x}_k \mathbf{x}_{k-L}^T]$ exists and is full rank, $\rho_{\hat{e}}(L) = 0$ only when $\mathbf{w} = \mathbf{w}_T$. Therefore, if we make the error autocorrelation at any lag $L \geq M$ zero, then the estimated weight vector will be exactly equal to the true weight vector. This is the motivation behind the EWC, which partially whitens the error signal by making $\rho_{\hat{e}}(L) = 0$ for $L \geq M$. Since the goal is to make $\rho_{\hat{e}}(L) = 0$, a suitable cost function to derive a stochastic gradient algorithm is $|\rho_{\hat{e}}(L)|$. Using Bluestein's identity (Taylor & Mellott, 1998), we can write $\hat{e}_k \hat{e}_{k-L}$ as

$$\hat{e}_k \hat{e}_{k-L} = 0.5[\hat{e}_k^2 + \hat{e}_{k-L}^2 - (\hat{e}_k - \hat{e}_{k-L})^2] \tag{2}$$

Taking the expectations on both sides and recognizing the fact that $E(\hat{e}_k^2) = E(\hat{e}_{k-L}^2)$, we get

$$E(\hat{e}_k \hat{e}_{k-L}) = E(\hat{e}_k^2) - 0.5E(\hat{e}_k - \hat{e}_{k-L})^2 \tag{3}$$

For convenience, we define $\hat{\hat{e}}_k = (\hat{e}_k - \hat{e}_{k-L})$ and use a constant β instead of -0.5. We can rewrite Eq. (3) as

$$E(\hat{e}_k \hat{e}_{k-L}) = E(\hat{e}_k^2) + \beta E(\hat{\hat{e}}_k^2) \tag{4}$$

The cost function for the EWC can now be formally stated as

$$J(\mathbf{w}) = |E(\hat{e}_k^2) + \beta E(\hat{\hat{e}}_k^2)| \tag{5}$$

The form in Eq. (5) is appealing because, it includes the MSE as a special case when $\beta = 0$. With $\beta = -0.5$, the above cost function becomes $|\rho_{\hat{e}}(L)|$ which when minimized would result in the unbiased estimate of the true weight vector. Another interesting result is that the sensitivity of $\rho_{\hat{e}}(L)$, given by, $\partial\rho_{\hat{e}}(L)/\partial\mathbf{w} = -2[\mathbf{w}_T - \mathbf{w}]E[\mathbf{x}_k \mathbf{x}_{k-L}^T]$ is zero if $(\mathbf{w}_T - \mathbf{w}) = \mathbf{0}$. Thus, if $(\mathbf{w}_T - \mathbf{w})$ is not in the null space of $E[\mathbf{x}_k \mathbf{x}_{k-L}^T]$ or if $E[\mathbf{x}_k \mathbf{x}_{k-L}^T]$ is full rank, then only $(\mathbf{w}_T - \mathbf{w}) = \mathbf{0}$ makes $\rho_{\hat{e}}(L) = 0$ and $\partial\rho_{\hat{e}}(L)/\partial\mathbf{w} = \mathbf{0}$ simultaneously. This property has a useful implication. Consider any cost function of the form $J(\mathbf{w})^p$, $p > 0$. Then the performance surface is not necessarily quadratic and the stationary points of this new cost are given by $J(\mathbf{w}) = 0$ or $\partial J(\mathbf{w})/\partial\mathbf{w} = \mathbf{0}$. Using the above property, we immediately see that both $J(\mathbf{w}) = 0$ and $\partial J(\mathbf{w})/\partial\mathbf{w} = \mathbf{0}$ yield the same solution. Optimization on Eq. (5) without the absolute value operator is impossible using a constant sign gradient algorithm, as the stationary point can then be a global maximum, minimum or a saddle point (Principe et al., 2003). The stochastic instantaneous gradient of the EWC cost function in Eq. (5) is

$$\partial J(\mathbf{w})/\partial\mathbf{w} = -2\text{sign}(\hat{e}_k^2 + \beta\hat{\hat{e}}_k^2)(\hat{e}_k\hat{\mathbf{x}}_k + \beta\hat{\hat{e}}_k\hat{\hat{\mathbf{x}}}_k) \tag{6}$$

where $\hat{\hat{e}}_k = (\hat{e}_k - \hat{e}_{k-L})$ and $\hat{\hat{\mathbf{x}}}_k = (\hat{\mathbf{x}}_k - \hat{\mathbf{x}}_{k-L})$ as defined before. The stationary point is a global minimum and using gradient descent, we can write the EWC-LMS

algorithm as

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \eta \text{sign}(\hat{e}_k^2 + \beta\hat{\dot{e}}_k^2)(\hat{e}_k\hat{\mathbf{x}}_k + \beta\hat{\dot{e}}_k\hat{\dot{\mathbf{x}}}_k) \tag{7}$$

where $\eta > 0$ is a small step-size parameter. Note that when $\beta = 0$, Eq. (7) reduces to the renowned LMS algorithm (Widrow & Stearns, 1985). We are specifically interested in using Eq. (7) with $\beta = -0.5$. In Section 3, we will present the convergence analysis of Eq. (7) and derive some useful results.

3. Convergence analysis

Theorem 1. *In the noise-free case (deterministic signals), EWC-LMS given in* (7) *converges to the stationary point* $\mathbf{w}_* = \mathbf{w}_\text{T}$ *provided that the step size satisfies the following inequality at every update*

$$0 < \eta < \frac{2|e_k^2 + \beta\dot{e}_k^2|}{\|e_k\mathbf{x}_k + \beta\dot{e}_k\dot{\mathbf{x}}_k\|^2} \forall k \tag{8}$$

Proof. From the arguments presented in Section 2 and owing to the quadratic nature of the EWC performance surface, it is clear that the EWC-LMS algorithm in Eq. (7) has a single stationary point (global minimum) $\mathbf{w}_* = \mathbf{w}_\text{T}$. The formal proof is trivial and is omitted here. Consider the weight error vector defined as $\boldsymbol{\varepsilon}_k = \mathbf{w}_* - \mathbf{w}_k$. Subtracting both sides of Eq. (7) from $\mathbf{w}_*$, we get $\boldsymbol{\varepsilon}_{k+1} = \boldsymbol{\varepsilon}_k - \eta$ sign$(e_k^2 + \beta\dot{e}_k^2)(e_k\mathbf{x}_k + \beta\dot{e}_k\dot{\mathbf{x}}_k)$. Taking the norm of this error vector we get

$$\begin{aligned}\|\boldsymbol{\varepsilon}_{k+1}\|^2 =& \|\boldsymbol{\varepsilon}_k\|^2 - 2\eta\text{sign}(e_k^2 + \beta\dot{e}_k^2)\boldsymbol{\varepsilon}_k^\text{T}(e_k\mathbf{x}_k + \beta\dot{e}_k\dot{\mathbf{x}}_k) \\ &+ \eta^2\|e_k\mathbf{x}_k + \beta\dot{e}_k\dot{\mathbf{x}}_k\|^2\end{aligned}$$

In case of noise-free data, $\boldsymbol{\varepsilon}_k^\text{T}\mathbf{x}_k = e_k$ and $\boldsymbol{\varepsilon}_k^\text{T}\dot{\mathbf{x}}_k = \dot{e}_k$. Using these two equations we get

$$\|\boldsymbol{\varepsilon}_{k+1}\|^2 = \|\boldsymbol{\varepsilon}_k\|^2 - 2\eta|e_k^2 + \beta\dot{e}_k^2| + \eta^2\|e_k\mathbf{x}_k + \beta\dot{e}_k\dot{\mathbf{x}}_k\|^2 \tag{9}$$

By allowing the error vector norm to decay asymptotically by making $\|\boldsymbol{\varepsilon}_{k+1}\|^2 < \|\boldsymbol{\varepsilon}_k\|^2$, we obtain the bound in Eq. (8). The error vector will eventually converge to zero, i.e. $\lim_{k\to\infty} \times \|\boldsymbol{\varepsilon}_k\|^2 \to 0$, which implies that $\lim_{k\to\infty} \mathbf{w}_k \to \mathbf{w}_* = \mathbf{w}_\text{T}$. □

Observe that when $\beta = 0$, the upper bound on the step-size in Eq. (8) reduces to $0 < \eta < 2/\|\mathbf{x}_k\|^2$, which is nothing but the step-size bound for LMS in the case of deterministic signals.

Theorem 2. *In the noisy data case, EWC-LMS given in* (7) *with* $\beta = -0.5$ *converges to the stationary point* $\mathbf{w}_* = \mathbf{w}_\text{T}$ *in the mean provided that the step size is bound by the inequality*

$$0 < \eta < \frac{2|E(\hat{e}_k^2 - 0.5\hat{\dot{e}}_k^2)|}{E\left[\|\hat{e}_k\hat{\mathbf{x}}_k - 0.5\hat{\dot{e}}_k\hat{\dot{\mathbf{x}}}_k\|\right]^2} \tag{10}$$

Proof. Again, it is clear that the only stationary point of Eq. (7) with $\beta = -0.5$ is $\mathbf{w}_* = \mathbf{w}_\text{T}$ even in the presence of noise where $\mathbf{w}_\text{T}$ is the true weight vector that generated the noise-free data pair $(\mathbf{x}_k, d_k)$. Following the same steps as in the proof of the previous theorem, the dynamics of the error vector norm can be determined by the difference equation

$$\begin{aligned}\|\boldsymbol{\varepsilon}_{k+1}\|^2 = \|\boldsymbol{\varepsilon}_k\|^2 &- 2\eta\text{sign}(\hat{e}_k^2 + \beta\hat{\dot{e}}_k^2)\boldsymbol{\varepsilon}_k^\text{T}(\hat{e}_k\hat{\mathbf{x}}_k + \beta\hat{\dot{e}}_k\hat{\dot{\mathbf{x}}}_k) \\ &+ \eta^2\|\hat{e}_k\hat{\mathbf{x}}_k + \beta\hat{\dot{e}}_k\hat{\dot{\mathbf{x}}}_k\|^2\end{aligned} \tag{11}$$

Using Eq. (11), a recursive expression for the error vector norm at an iteration index k in terms of the initial error vector norm $\|\boldsymbol{\varepsilon}_0\|^2$ can be written as

$$\begin{aligned}\|\boldsymbol{\varepsilon}_k\|^2 = \|\boldsymbol{\varepsilon}_0\|^2 &- 2\eta\sum_{j=0}^{k-1}\text{sign}(\hat{e}_j^2 + \beta\hat{\dot{e}}_j^2)\boldsymbol{\varepsilon}_j^\text{T}(\hat{e}_j\hat{\mathbf{x}}_j + \beta\hat{\dot{e}}_j\hat{\dot{\mathbf{x}}}_j) \\ &+ \eta^2\sum_{j=0}^{k-1}\|\hat{e}_j\hat{\mathbf{x}}_j + \hat{\dot{e}}_j\hat{\dot{\mathbf{x}}}_j\|^2\end{aligned} \tag{12}$$

With the absolute value operator replacing the sign, we get

$$\begin{aligned}&\|\boldsymbol{\varepsilon}_0\|^2 - \|\boldsymbol{\varepsilon}_k\|^2 + \eta^2\sum_{j=0}^{k-1}\|\hat{e}_j\hat{\mathbf{x}}_j + \beta\hat{\dot{e}}_j\hat{\dot{\mathbf{x}}}_j\|^2 \\ &< 2\eta\sum_{j=0}^{k-1}|\boldsymbol{\varepsilon}_j^\text{T}(\hat{e}_j\hat{\mathbf{x}}_j + \beta\hat{\dot{e}}_j\hat{\dot{\mathbf{x}}}_j)|\end{aligned} \tag{13}$$

Dividing Eq. (13) on both sides by k and letting $k \to \infty$, we have

$$\eta E\|\hat{e}_j\hat{\mathbf{x}}_j + \beta\hat{\dot{e}}_j\hat{\dot{\mathbf{x}}}_j\|^2 < \lim_{k\to\infty}\frac{1}{k}\frac{\|\boldsymbol{\varepsilon}_k\|^2}{\eta} + 2E|\boldsymbol{\varepsilon}_j^\text{T}(\hat{e}_j\hat{\mathbf{x}}_j + \beta\hat{\dot{e}}_j\hat{\dot{\mathbf{x}}}_j)|$$

and hence

$$\tfrac{\eta}{2}E\|\hat{e}_j\hat{\mathbf{x}}_j + \beta\hat{\dot{e}}_j\hat{\dot{\mathbf{x}}}_j\|^2 < E|\boldsymbol{\varepsilon}_j^\text{T}(\hat{e}_j\hat{\mathbf{x}}_j + \beta\hat{\dot{e}}_j\hat{\dot{\mathbf{x}}}_j)| \tag{14}$$

Note that we have used the facts that

$$\lim_{k\to\infty}\frac{1}{k}\sum_{j=0}^{k-1}f(\cdot) = E[f(\cdot)]$$

(ergodicity) and $\eta > 0$. If Eq. (14) holds, then it is trivial to show that $E\|\boldsymbol{\varepsilon}_{k+1}\|^2 < E\|\boldsymbol{\varepsilon}_k\|^2$. Note that the expectation operator depicts averaging over the entire input data with the weight fixed at the iteration index specified in the error vector norm. Thus, we can only state that the mean error norm decreases over time. Note that, the right-hand side of Eq. (14) is still hard to compute due to the absolute

value operator. However, using Jensen's inequality for convex functions, $E|X| \geq |E(X)|$, we can deduce a loose upper bound for the step-size as

$$0 < \eta < 2\frac{|E[\boldsymbol{\varepsilon}_j^{\mathrm{T}}(\hat{e}_j\hat{\mathbf{x}}_j + \beta\hat{\hat{e}}_j\hat{\hat{\mathbf{x}}}_j)]|}{E\|\hat{e}_j\hat{\mathbf{x}}_j + \beta\hat{\hat{e}}_j\hat{\hat{\mathbf{x}}}_j\|^2} \tag{15}$$

Now, we can simplify Eq. (15) further. The evaluation of the terms $E(\boldsymbol{\varepsilon}_k^{\mathrm{T}}\hat{e}_k\hat{\mathbf{x}}_k)$ and $E(\boldsymbol{\varepsilon}_k^{\mathrm{T}}\hat{\hat{e}}_k\hat{\hat{\mathbf{x}}}_k)$ are tedious and is omitted here. It can be shown that

$$E(\boldsymbol{\varepsilon}_k^{\mathrm{T}}\hat{e}_k\hat{\mathbf{x}}_k) = \boldsymbol{\varepsilon}_k^{\mathrm{T}}\mathbf{R}\boldsymbol{\varepsilon}_k - \boldsymbol{\varepsilon}_k^{\mathrm{T}}\mathbf{V}\mathbf{w}_k \tag{16}$$

$$E(\boldsymbol{\varepsilon}_k^{\mathrm{T}}\hat{\hat{e}}_k\hat{\hat{\mathbf{x}}}_k) = \boldsymbol{\varepsilon}_k^{\mathrm{T}}(2\mathbf{R} - \mathbf{R}_L)\boldsymbol{\varepsilon}_k - 2\boldsymbol{\varepsilon}_k^{\mathrm{T}}\mathbf{V}\mathbf{w}_k$$

where $\mathbf{R} = E[\mathbf{x}_k\mathbf{x}_k^{\mathrm{T}}]$, $\mathbf{R}_L = E[\mathbf{x}_k\mathbf{x}_{k-L}^{\mathrm{T}} + \mathbf{x}_{k-L}\mathbf{x}_k^{\mathrm{T}}]$, $\mathbf{V} = E[\mathbf{v}_k\mathbf{v}_k^{\mathrm{T}}]$. Since we assumed that the noise is white, $\mathbf{V} = \sigma_v^2 I$, where σ_v^2 represents the variance of the input noise. Now, with $\beta = -0.5$,

$$E(\boldsymbol{\varepsilon}_k^{\mathrm{T}}\hat{e}_k\hat{\mathbf{x}}_k) - 0.5(\boldsymbol{\varepsilon}_k^{\mathrm{T}}\hat{\hat{e}}_k\hat{\hat{\mathbf{x}}}_k) = 0.5\boldsymbol{\varepsilon}_k^{\mathrm{T}}\mathbf{R}_L\boldsymbol{\varepsilon}_k \tag{17}$$

Using $\boldsymbol{\varepsilon}_k = \mathbf{w}_* - \mathbf{w}_k = \mathbf{w}_{\mathrm{T}} - \mathbf{w}_k$ and $d_k = \mathbf{x}_k^{\mathrm{T}}\mathbf{w}_{\mathrm{T}}$, Eq. (17) can be further reduced to

$$E(\boldsymbol{\varepsilon}_k^{\mathrm{T}}\hat{e}_k\hat{\mathbf{x}}_k) - 0.5E(\boldsymbol{\varepsilon}_k^{\mathrm{T}}\hat{\hat{e}}_k\hat{\hat{\mathbf{x}}}_k) = E(e_k e_{k-L})$$

$$= E(\hat{e}_k^2 - 0.5\hat{\hat{e}}_k^2) \tag{18}$$

Substituting the numerator of Eq. (15) with the above result, we immediately get the upper bound in Eq. (10). If the step-size chosen satisfies this condition, then $E\|\boldsymbol{\varepsilon}_{k+1}\|^2 < E\|\boldsymbol{\varepsilon}_k\|^2$ and the error vector norm asymptotically converges to zero in the mean. Thus, $\lim_{k\to\infty} E\|\boldsymbol{\varepsilon}_k\|^2 \to 0$ and $\lim_{k\to\infty} E(\mathbf{w}_k) \to \mathbf{w}_* = \mathbf{w}_{\mathrm{T}}$. □

We would like to mention that the upper bound on step-size given by Eq. (10) is computable using only the data samples. For the LMS algorithm ($\beta = 0$), if the input and desired signals are noisy, the upper bound on the step-size is dependent on the true weight vector as well as on the variance of the noise, which makes it impractical. Since the EWC-LMS algorithm with $\beta = -0.5$ minimizes $|\rho_{\hat{e}}(L)|$, the effect of finite step-sizes on the steady state $|\rho_{\hat{e}}(L)|$ would be a good performance index. This is analogous to the excess-MSE in LMS (Haykin, 1996).

Theorem 3. *With $\beta = -0.5$, the steady state excess error autocorrelation at lag $L \geq M$, i.e. $|\rho_{\hat{e}}(L)|$ is always bound by*

$$|\rho_{\hat{e}}(L)| \leq \tfrac{\eta}{2}E(\hat{e}_\infty^2)[\mathrm{Tr}(\mathbf{R} + \mathbf{V})] + 2\eta[\sigma_u^2 + \|\mathbf{w}_\infty\|\|\mathbf{w}_*\|\mathrm{Tr}(\mathbf{V})] \tag{19}$$

where $\mathbf{R} = E[\mathbf{x}_k\mathbf{x}_k^{\mathrm{T}}]$, and $\mathbf{V} = E[\mathbf{v}_k\mathbf{v}_k^{\mathrm{T}}]$ and $\mathrm{Tr}(\cdot)$ denotes the matrix trace. The noise variances in the input and desired signals are represented by σ_v^2 and σ_u^2, respectively.

Proof. Following the footsteps of the previous proofs, we start with the error dynamics equation given by Eq. (11). Since we are interested in the dynamics near convergence (steady state) we let $k \to \infty$. Applying the expectation operator to both sides of Eq. (11) will give

$$E\|\boldsymbol{\varepsilon}_{k+1}\|^2 = E\|\boldsymbol{\varepsilon}_k\|^2 - 2\eta E[\mathrm{sign}(\hat{e}_k^2 - 0.5\hat{\hat{e}}_k^2)\boldsymbol{\varepsilon}_k^{\mathrm{T}}(\hat{e}_k\hat{\mathbf{x}}_k - 0.5\hat{\hat{e}}_k\hat{\hat{\mathbf{x}}}_k)] + \eta^2 E\|\hat{e}_k\hat{\mathbf{x}}_k - 0.5\hat{\hat{e}}_k\hat{\hat{\mathbf{x}}}_k\|^2 \tag{20}$$

Expanding the terms $\boldsymbol{\varepsilon}_k^{\mathrm{T}}\hat{e}_k\hat{\mathbf{x}}_k$, $\boldsymbol{\varepsilon}_k^{\mathrm{T}}\hat{\hat{e}}_k\hat{\hat{\mathbf{x}}}_k$ and simplifying we get

$$\begin{aligned} E\|\boldsymbol{\varepsilon}_{k+1}\|^2 =& E\|\boldsymbol{\varepsilon}_k\|^2 + \eta^2 E\|\hat{e}_k\hat{\mathbf{x}}_k - 0.5\hat{\hat{e}}_k\hat{\hat{\mathbf{x}}}_k\|^2 \\ &+ 2\eta E[\mathrm{sign}(\hat{e}_k^2 - 0.5\hat{\hat{e}}_k^2)[\mathbf{w}_*^{\mathrm{T}}(\mathbf{v}_k\mathbf{v}_k^{\mathrm{T}} - 0.5\dot{\mathbf{v}}_k\dot{\mathbf{v}}_k^{\mathrm{T}})\mathbf{w}_k]] \\ &+ 2\eta E[\mathrm{sign}(\hat{e}_k^2 - 0.5\hat{\hat{e}}_k^2)(u_k^2 - 0.5\dot{u}_k^2)] \\ &- 2\eta E|\hat{e}_k^2 - 0.5\hat{\hat{e}}_k^2| \end{aligned} \tag{21}$$

Letting $E\|\boldsymbol{\varepsilon}_{k+1}\|^2 = E\|\boldsymbol{\varepsilon}_k\|^2$ as $k \to \infty$, we see that

$$\begin{aligned} &E|(\hat{e}_k^2 - 0.5\hat{\hat{e}}_k^2)| \\ &= \tfrac{\eta}{2}E\|\hat{e}_k\hat{\mathbf{x}}_k - 0.5\hat{\hat{e}}_k\hat{\hat{\mathbf{x}}}_k\|^2 + \eta E[\mathrm{sign}(\hat{e}_k^2 - 0.5\hat{\hat{e}}_k^2) \\ &\quad \times[\mathbf{w}_*^{\mathrm{T}}(\mathbf{v}_k\mathbf{v}_k^{\mathrm{T}} - 0.5\dot{\mathbf{v}}_k\dot{\mathbf{v}}_k^{\mathrm{T}})\mathbf{w}_k + u_k^2 - 0.5\dot{u}_k^2]] \end{aligned} \tag{22}$$

By Jensen's inequality, $E|(\hat{e}_k^2 - 0.5\hat{\hat{e}}_k^2)| \geq |E(\hat{e}_k^2 - 0.5\hat{\hat{e}}_k^2)|$, and therefore we have

$$\begin{aligned} |\rho_{\hat{e}}(L)| \leq& \tfrac{\eta}{2}E\|\hat{e}_k\hat{\mathbf{x}}_k - 0.5\hat{\hat{e}}_k\hat{\hat{\mathbf{x}}}_k\|^2 + \eta E[\mathrm{sign}(\hat{e}_k^2 - 0.5\hat{\hat{e}}_k^2) \\ &\times[\mathbf{w}_*^{\mathrm{T}}(\mathbf{v}_k\mathbf{v}_k^{\mathrm{T}} - 0.5\dot{\mathbf{v}}_k\dot{\mathbf{v}}_k^{\mathrm{T}})\mathbf{w}_k + u_k^2 - 0.5\dot{u}_k^2]] \end{aligned} \tag{23}$$

Note that, we used the relation $\rho_{\hat{e}}(L) = E(\hat{e}_k^2 - 0.5\hat{\hat{e}}_k^2)$ in the above equation. The first term on the RHS of Eq. (23) can be easily evaluated by invoking the assumption that $\|\hat{\mathbf{x}}_k\|^2$ and $\hat{e}_k^2$ are uncorrelated in steady state

$$E\|\hat{e}_k\hat{\mathbf{x}}_k - 0.5\hat{\hat{e}}_k\hat{\hat{\mathbf{x}}}_k\|^2 = E(\hat{e}_k^2)[\mathrm{Tr}(\mathbf{R}) + \mathrm{Tr}(\mathbf{V})] \tag{24}$$

The above assumption is commonly used in computing the steady-state excess-MSE for stochastic LMS algorithm (Al-Naffouri & Sayed, 2001). Importantly, this assumption is less restrictive and more natural when compared to the independence theory that was frequently used in the past (Haykin, 1996). The second term in RHS of Eq. (23) is involved and has no closed form expression even with Gaussianity assumptions that are typically made in the analysis of sign-LMS algorithm (Al-Naffouri & Sayed, 2001). Even the validity of Gaussianity assumption is questionable as discussed by Eweda (Eweda, 2000) who proposed additional, reasonable constraints on the noise probability density function to overcome the Gaussianity and independence assumptions (Eweda, 2000) that lead to a more generic misadjustment upper bound for the sign-LMS algorithm. Nevertheless, the analyses of stochastic algorithms (with or without sign) in the existing literature explicitly assume that the input signal is 'noise-free' that simplifies the problem to a great extent. In this paper we particularly deal with input noise and refrain from making any assumptions in deriving an upper bound for excess error autocorrelation. We proceed by rewriting Eq. (23) using

the identity $E[\text{sign}(a)b] \le E|b|$ as

$$|\rho_{\hat{e}}(L)| \le \frac{\eta}{2}E(\hat{e}_k^2)[\text{Tr}(\mathbf{R})+\text{Tr}(\mathbf{V})] + \eta E|[\mathbf{w}_*^{\text{T}}(\mathbf{v}_k\mathbf{v}_k^{\text{T}} - 0.5\dot{\mathbf{v}}_k\dot{\mathbf{v}}_k^{\text{T}})\mathbf{w}_k + u_k^2 - 0.5\dot{u}_k^2]| \quad (25)$$

We know that $|a+b| \le |a|+|b|$ and $E(u_k u_{k-L}) = 0$. Therefore,

$$E|[u_k^2 - 0.5\dot{u}_k^2]| \le E(u_k^2) + 0.5E(\dot{u}_k^2) = 2\sigma_u^2 \quad (26)$$

Similarly,

$$E|\mathbf{w}_*^{\text{T}}(\mathbf{v}_k\mathbf{v}_k^{\text{T}} - 0.5\dot{\mathbf{v}}_k\dot{\mathbf{v}}_k^{\text{T}})\mathbf{w}_k|$$

$$\le E|\mathbf{w}_*^{\text{T}}\mathbf{v}_k\mathbf{v}_k^{\text{T}}\mathbf{w}_k| + 0.5E|\mathbf{w}_*^{\text{T}}\dot{\mathbf{v}}_k\dot{\mathbf{v}}_k^{\text{T}}\mathbf{w}_k| \quad (27)$$

Since the individual terms $\mathbf{w}_*^{\text{T}}\mathbf{v}_k\mathbf{v}_k^{\text{T}}\mathbf{w}_k$ and $\mathbf{w}_*^{\text{T}}\dot{\mathbf{v}}_k\dot{\mathbf{v}}_k^{\text{T}}\mathbf{w}_k$ are not necessarily positive we use the Cauchy–Schwartz inequality to continue further

$$\mathbf{w}_*^{\text{T}}\mathbf{v}_k\mathbf{v}_k^{\text{T}}\mathbf{w}_k \le \|\mathbf{w}_k\|\|\mathbf{w}_*\|\|\mathbf{v}_k\|^2,$$
$$\mathbf{w}_*^{\text{T}}\dot{\mathbf{v}}_k\dot{\mathbf{v}}_k^{\text{T}}\mathbf{w}_k \le \|\mathbf{w}_k\|\|\mathbf{w}_*\|\|\dot{\mathbf{v}}_k\|^2 \quad (28)$$

we know that $E[\mathbf{v}_k\mathbf{v}_{k-L}^{\text{T}}]$ is $\mathbf{0}$ for $L \ge M$. Therefore,

$$E|\mathbf{w}_*^{\text{T}}\mathbf{v}_k\mathbf{v}_k^{\text{T}}\mathbf{w}_k| + 0.5E|\mathbf{w}_*^{\text{T}}\dot{\mathbf{v}}_k\dot{\mathbf{v}}_k^{\text{T}}\mathbf{w}_k| \le 2\|\mathbf{w}_*\|\|\mathbf{w}_k\|\text{Tr}(\mathbf{V}) \quad (29)$$

Using Eqs. (26) and (29) in Eq. (25), and letting $k \to \infty$, we get the result in Eq. (19). The term $E(\hat{e}_\infty^2)$ represents the residual MSE and is given by

$$E(\hat{e}_\infty^2) = \boldsymbol{\varepsilon}_\infty^{\text{T}}\mathbf{R}\boldsymbol{\varepsilon}_\infty + \sigma_u^2 + \mathbf{w}_\infty^{\text{T}}\mathbf{V}\mathbf{w}_\infty \le \|\boldsymbol{\varepsilon}_\infty\|^2\lambda_{\max} + \sigma_u^2 + \|\mathbf{w}_\infty\|\sigma_v^2$$

where $\lambda_{\max}$ is the maximum eigenvalue of $\mathbf{R}$. □

Observe that by reducing the step-size, one can arbitrarily reduce the steady-state excess error autocorrelation at lag $L \ge M$. Extensive simulations have confirmed this fact and the results have been reported by Principe et al. (2003).

4. Algorithm extensions

EWC-LMS is a stochastic gradient algorithm and like all stochastic algorithms, it has tradeoffs between speed of convergence and accuracy of the asymptotic solution. Choosing the right step-size to ensure a desired tradeoff is a non-trivial problem. Recursive algorithms can be derived for EWC, but incur additional computational costs. In this section, we will explore some simple extensions that can improve the robustness and speed of convergence of EWC-LMS.

Normalization with eigenvalue of Hessian. It is a well-known fact that, a gradient algorithm operating around the vicinity of the stable stationary point converges if the step-size is upper bound by $2/|\lambda_{\max}|$, where $|\lambda_{\max}|$ denotes the absolute maximum eigenvalue of the Hessian matrix of the performance surface. Since, the cost function of EWC is quadratic in nature, the Hessian matrix is simply given by, $\mathbf{H} = \mathbf{R} + \beta\mathbf{S}$, where $\mathbf{R} = E[\mathbf{x}_k\mathbf{x}_k^{\text{T}}]$ and $\mathbf{S} = E[\dot{\mathbf{x}}_k\dot{\mathbf{x}}_k^{\text{T}}]$. For notational convenience, we will assume that the signals are noise-free, and this will not affect the discussions to follow. Since we use $\beta = -0.5$, the Hessian matrix $\mathbf{H}$ can have mixed eigenvalues and this complicates the algorithms for online estimation of the absolute maximum eigenvalue. From the triangle inequality (Golub & Van Loan, 1989)

$$\|\mathbf{H}\|_2 \le \|\mathbf{R}\|_2 + |\beta|\|\mathbf{S}\|_2\sqrt{\lambda_{\max}(\mathbf{R})} + |\beta|\sqrt{\lambda_{\max}(\mathbf{S})} \quad (30)$$

where $\|\cdot\|_2$ denotes the matrix norm. Since, both $\mathbf{R}$ and $\mathbf{S}$ are positive-definite matrices, we can write

$$\|\mathbf{H}\|_2 \le \sqrt{\text{Tr}(\mathbf{R})} + |\beta|\sqrt{\text{Tr}(\mathbf{S})} \le \sqrt{E\|\mathbf{x}_k\|^2} + |\beta|\sqrt{E\|\dot{\mathbf{x}}_k\|^2} \quad (31)$$

In the stochastic framework, we can include this in the update equation in Eq. (7) to give us a step-size normalized EWC-LMS update rule given by

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \frac{\eta\text{sign}(e_k^2 + \beta\dot{e}_k^2)(e_k\mathbf{x}_k + \beta\dot{e}_k\dot{\mathbf{x}}_k)}{(\|\mathbf{x}_k\| + |\beta|\|\dot{\mathbf{x}}_k\|)^2} \quad (32)$$

Note that when $\beta = 0$, Eq. (32) reduces to the well-known normalized LMS (NLMS) algorithm (Haykin, 1996).

Normalized gradient EWC-LMS. Another way of improving the convergence speed of EWC-LMS is to use the upper bound in Eq. (10) for the step-size in a stochastic framework. This is a weak upper bound for guaranteed convergence, and including this in the update equation, we get

$$\mathbf{w}_{k+1} = \mathbf{w}_k + 2\frac{(e_k^2 + \beta\dot{e}_k^2)(e_k\mathbf{x}_k + \beta\dot{e}_k\dot{\mathbf{x}}_k)}{(\|e_k\mathbf{x}_k + \beta\dot{e}_k\dot{\mathbf{x}}_k\|^2 + \delta)} \quad (33)$$

Note that in Eq. (33), the sign of the gradient is now given explicitly by the instantaneous EWC cost term itself. The term δ, a small positive constant compensates for the numerical instabilities when the signal has zero power or when the error goes to zero, which can happen in the noiseless case even with finite number of samples. Once again, we would like to state that with $\beta = 0$, Eq. (33) defaults to NLMS algorithm. The caveat is that, both Eqs. (32) and (33) do not satisfy the principle of minimum disturbance or they do not correspond to minimum norm update unlike the NLMS algorithm (Haykin, 1996). We have verified the faster convergence of the normalized EWC algorithms in Eqs. (32) and (33) with extensive simulations. The results are omitted here owing to space constraints. The drawback with the update equation in Eq. (33) is the increased misadjustment with noises injected in the training data. This is in agreement with the fact that misadjustment in the NLMS algorithm is high when compared with the standard LMS for noisy signals. This can be further controlled by inserting a small decaying step-size instead of the constant step-size of 2 in Eq. (33). It is easy to show

that $\mu = 2$ is the largest possible step-size for guaranteed convergence of Eq. (33).

EWC with multiple lags. The EWC-LMS algorithm we have proposed ensures that the error autocorrelation at any single lag $L \geq M$ is minimized. With decreasing SNR values (< -10 dB), the Hessian matrix $\mathbf{H} = \mathbf{R} + \beta\mathbf{S}$ (with $\beta = -0.5$) is mostly determined by the noise covariance matrix. This can degrade the performance of the EWC-LMS and we might be forced to use very small step-sizes (slow convergence) to achieve good results. This problem can be alleviated by incorporating multiple lags in the EWC cost function. Instead of minimizing the error autocorrelation at a single lag, we can add similar constraints at additional lags. The corresponding stochastic EWC-LMS algorithm is then given by

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \sum_{l=M}^{l=M+\Delta} \eta_l \text{sign}(\hat{e}_k^2 + \beta\hat{\hat{e}}_{kl}^2)(\hat{e}_k\hat{\mathbf{x}}_k + \beta\hat{\hat{e}}_{kl}\hat{\hat{\mathbf{x}}}_{kl}) \quad (34)$$

Note that Δ is the total number of lags (constraints) for the error autocorrelation. We have verified that Eq. (34) is more robust than the single lag EWC-LMS algorithm. However, the additional robustness comes at an increase in the computational cost. In the case when $\Delta = M$, the complexity of Eq. (34) becomes $O(M^2)$. Further analysis on Eq. (34) is beyond the scope of this paper and will be given in a later paper.

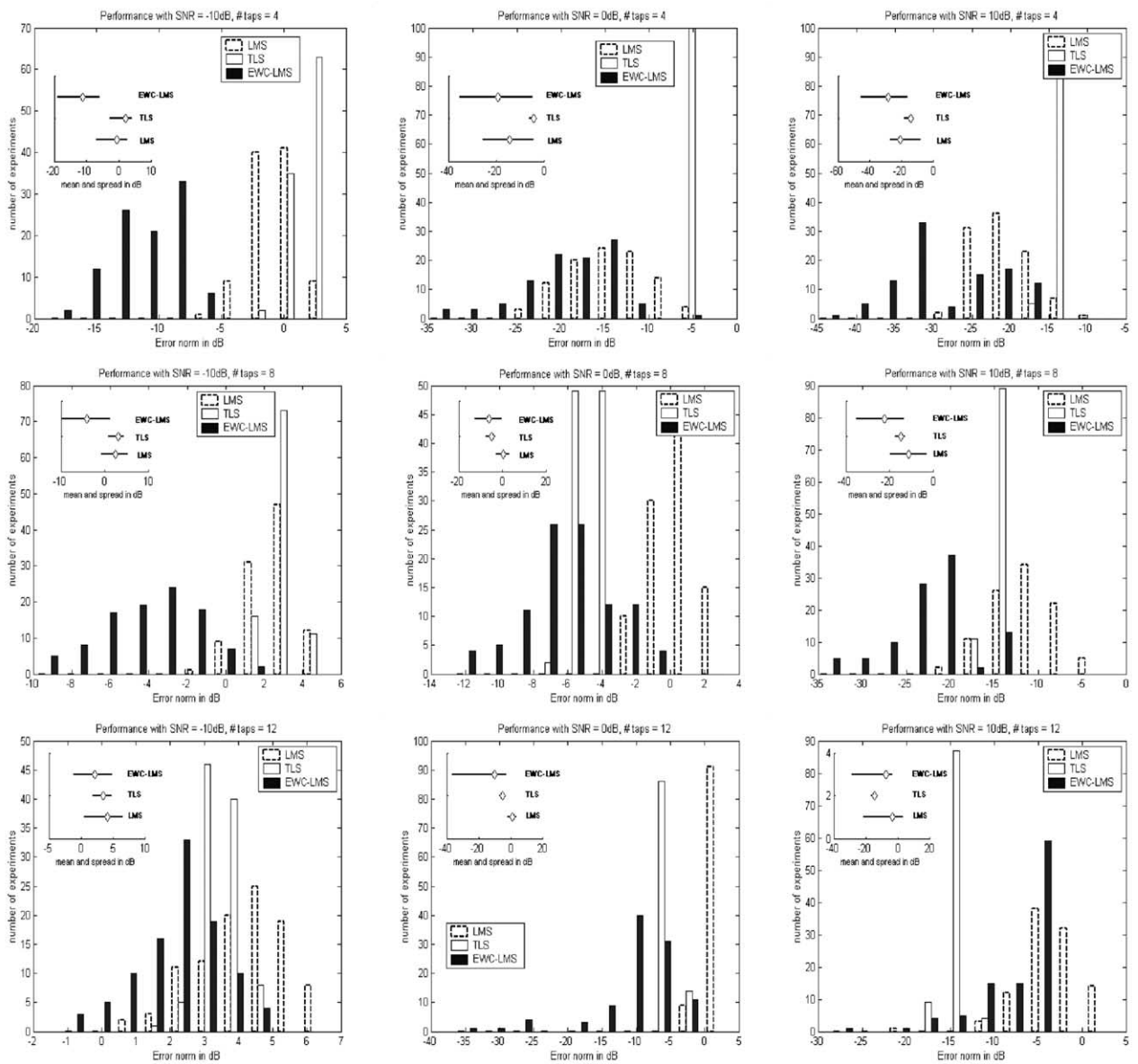

Fig. 2. Histogram plots of the error vector norm for EWC-LMS, LMS algorithms and the numerical TLS solution.

5. System identification and inverse control

System identification using EWC-LMS. We will now verify the noise rejecting capability of EWC-LMS algorithm when $\beta = -0.5$ in a system identification problem. A noise-free, sufficiently colored input signal of 50,000 samples is passed through an *unknown* system to form the noise-free desired signal. Uncorrelated white Gaussian noise is added to the input signal. Clean desired signal is used, as the noise in the desired averages out automatically in stochastic LMS type algorithms. The input SNR was set at -10, 0 and 10 dB. We chose the order of the unknown system to be 4, 8 and 12 and performed 100 Monte Carlo runs calculating the error vector norm in each case using

$$\text{error norm} = 20 \log 10[\|\mathbf{w}_T - \mathbf{w}_*\|] \quad (35)$$

where $\mathbf{w}_*$ is the solution given by EWC-LMS after one complete presentation of the training data and $\mathbf{w}_T$ represents the unknown system. We ran the regular LMS algorithm as well as the numerical TLS method (batch type). The step-sizes for both LMS and EWC-LMS algorithms were varied to get the best possible results in terms of the error vector norm given by Eq. (35). Fig. 2 shows the histograms of the error vector norms for all three methods. The inset plots in Fig. 2 show the summary of the histograms for each method. EWC-LMS performs significantly better than LMS at low SNR values (-10 and 0 dB), while performances are on par for SNR greater than 10 dB. Batch type numerical TLS method gives best results when the SNR is high. As we have stated before, TLS suffers if the noise variances in input and desired are not the same.

Inverse modeling and control. System identification is the first step in the design of an inverse controller. Specifically, we wish to design a system that controls the plant to produce a predefined output. Fig. 3 shows a block diagram of model reference inverse control (Widrow & Walach, 1995). In this case, the adaptive controller is designed so that the controller-plant pair would track the response generated by the reference model for any given input (command). Clearly, we require the plant parameters (which are typically unknown) to devise the controller. Once we have a model for the plant, the controller can be easily designed using conventional MSE minimization techniques. In this example, we will assume that the plant

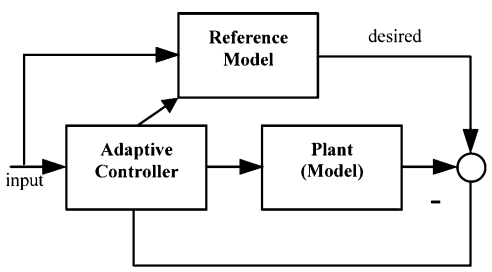

Fig. 3. Block diagram for model reference inverse control.

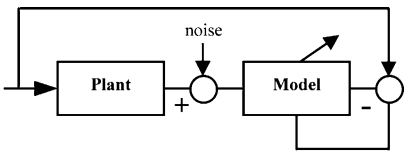

Fig. 4. Block diagram for inverse modeling.

is an all-pole system with transfer function $P(z) = 1/(1 + 0.8z^{-1} - 0.5z^{-2} - 0.3z^{-3})$. The reference model is chosen to be an FIR filter with five taps. The block diagram for the plant identification is shown in Fig. 4. Notice that the output of the plant is corrupted with additive white noise due to measurement errors. The SNR at the plant output was set to 0 dB. We then ran the EWC-LMS and LMS algorithms to estimate the model parameters given noisy input and desired signals. The model parameters thus obtained are used to derive the controller (Fig. 3) using standard backpropagation of error. We then tested the adaptive controller-plant pair for trajectory tracking by feeding the controller-plant pair with a non-linear time series and observing the responses. Ideally, the controller-plant pair must follow the trajectory generated by the reference model. Fig. 5(top) shows the tracking results for both controller-plant pairs along with the reference output. Fig. 5(bottom) shows a histogram of the tracking errors. Note that the errors with EWC-LMS controller are all concentrated around zero, which is desirable. In contrast, the errors produced with the MSE based controller are significant and this can be worse if the SNR levels drop further. Fig. 6 shows the magnitude and phase responses of the reference models along with the generated controller-model pairs. Note that, the EWC controller-model pair matches very closely with the desired transfer function, whereas MSE controller-model pair produces a significantly

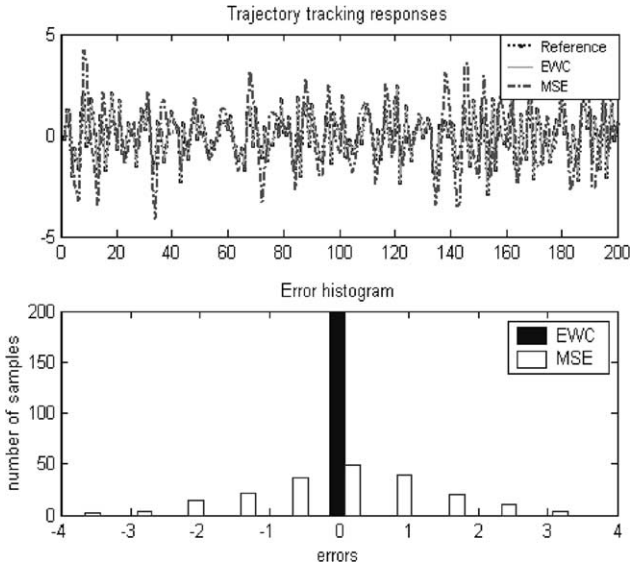

Fig. 5. Tracking results and error histograms.

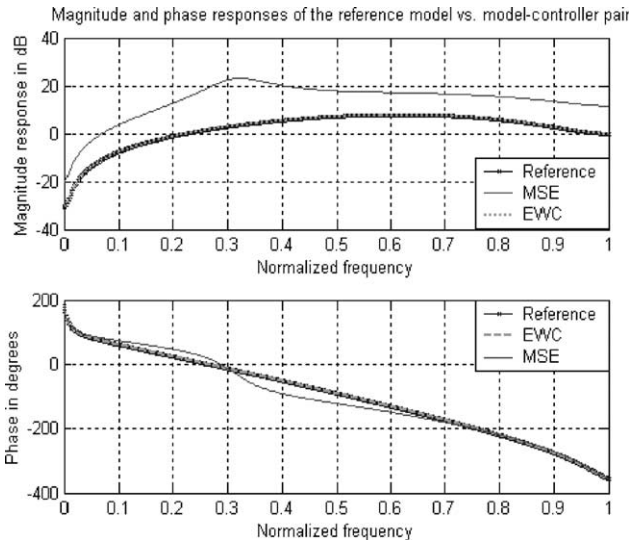

Fig. 6. Magnitude and phase responses of the reference model and designed model-controller pairs.

different transfer function. This clearly demonstrates the advantages offered by EWC.

6. Conclusions

In this paper, we presented a stochastic gradient algorithm for the recently proposed EWC, which includes MSE as a special case. MSE and TLS methods give highly biased parameter estimates if additive white noise with arbitrary variance is present in the input. However, EWC can be used to accurately estimate the underlying parameters of a linear system in the presence of additive white noise. We discussed some interesting properties of this new criterion, and then proposed an on-line, stochastic gradient algorithm with linear complexity in the number of parameters. Convergence of the stochastic gradient algorithm was derived making minimal assumptions and upper bounds on the step-size and the steady-state excess error autocorrelation were determined. We then proposed some extensions for improving the convergence speed and robustness of the EWC-LMS algorithm. Extensive Monte-Carlo simulations were carried out to show the superiority of the new criterion in a FIR system identification problem. We then successfully applied this method for designing a linear inverse controller and obtained superior results when compared with MSE based methods. Currently, further research is in progress to extend the criterion to handle colored noise and non-linear system identification.

Acknowledgements

This work was partially supported by the National Science Foundation under Grant NSF ECS-9900394.

References

Al-Naffouri, T. Y., & Sayed, A. H. (2001). Adaptive filters with error nonlinearities: mean-square analysis and optimum design. *EURASIP Journal of Applied Signal Processing*, *4*, 192–205.

Douglas, S. C. (1996). Analysis of an anti-hebbian adaptive FIR filtering algorithm. *IEEE Transactions on Circuits and Systems. II: Analog and Digital Signal Processing*, *43*(11).

Eweda, E. (2000). Convergence analysis of the sign algorithm without the independence and gaussian assumptions. *IEEE Transactions on Signal Processing*, *48*(9).

Golub, G. H., & Van Loan, C.F. (1989). *Matrix computations*. Baltimore, MD: Johns Hopkins University Press.

Haykin, S. (1996). *Adaptive filter theory*. Upper Saddle River, NJ: Prentice Hall.

Principe, J. C., Rao, Y. N., & Erdogmus, D. (2003). Error whitening wiener filters: theory and algorithms. In S. Haykin, & B. Widrow (Eds.), *Advances in LMS filters, Chap. 10*. New York, NY: Wiley.

Rao, Y. N., Erdogmus, D., & Principe, J. C. (2003). Error whitening criterion for linear filter adaptation. *Proceedings of IJCNN'03, Oregon*, in press.

Rao, Y. N., & Principe, J. C (2002). Efficient total least squares method for system modeling using minor component analysis. *Proceedings of NNSP'02* (pp. 259–268).

Taylor, F. J., & Mellott, J. (1998). *Hands-on digital signal processing*. New York, NY: McGraw Hill.

Widrow, B., & Stearns, S. D. (1985). *Adaptive signal processing*. Englewood Cliffs, NJ: Prentice Hall.

Widrow, B., & Walach, E. (1995). *Adaptive inverse control*. Englewood Cliffs, NJ: Prentice Hall.